

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

The Johns Hopkins University

3400 N. Charles St., Baltimore, MD 21218, USA

{stavros,vlasios,byrne}@jhu.edu

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.

Index Terms

Discriminative training, correlation modeling, adaptive training.

I. 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) [1], [2] 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) [3]. The goal of SAT is to reduce inter-speaker variability within the training set. SAT is an iterative procedure that generates a set of speaker independent state observation distributions along with matched speaker dependent transforms for all the speakers in the training set.

The transforms used in MLLT and SAT are estimated under the ML criterion [4], [1], [3]. Discriminative training under the Maximum Mutual Information (MMI) criterion [5] has recently been shown to be useful in large vocabulary conversational speech recognition (LVCSR) tasks [6]. 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) [7].

One approach to the use of DLTs is Maximum Mutual Information Linear Regression (MMILR) which was introduced by Uebel and Woodland [7], [8], who showed that it can be used for supervised speaker adaptation. Gunawardana and Byrne [9] 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. [10] 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 [11] 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) [12], where the transforms are estimated by a class separability criterion. Linear MMI Analysis (LMA) [13], on the other hand, replaces the class separability criterion of LDA with a MMI criterion. As observed by Schlüter [13], 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, [14]). 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

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

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

II. 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 [4]. 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 constrained 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.

A. 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' [4] treatment of MLLT and Gunawardana's CMLLR derivation [9].

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 S_r . Therefore we can write $\log q(\hat{\zeta}_1^i | s_1^i; \bar{\theta})$ as $\sum_s \sum_{\tau=1}^i \log q(\hat{\zeta}_\tau | s; \bar{T}_r) 1_s(s_\tau) 1_r(\mathcal{R}(s))$, where $1_s(s_\tau) = 1$ if $s_\tau = s$, 0 otherwise and similarly, $1_r(\mathcal{R}(s)) = 1$ if $r = \mathcal{R}(s)$, 0 otherwise. We can then express equation (1) as:

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

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 $D_s = \sum_{s_1^i: s_\tau = s} d'(s_1^i)$.

In the Appendix, we show an iterative procedure in which each row of T_r is optimized given the current value of all the other rows. Each row of the transform is updated by

$$[\bar{T}_r]_i = (\alpha p_i + k_i) G_i^{-1} \quad (3)$$

where α satisfies a quadratic expression (equation B1.8, [4]), p_i is the extended cofactor row vector $[0 \ c_{i1} \ \dots \ c_{im}]$, ($c_{ij} = \text{cof}(A_{ij})$) and

$$G_i = \sum_{s \in S_r} \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^i \gamma'_s(\tau; \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T + D_s J_s \right)$$

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

where J_s is defined as the matrix

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

B. Gaussian Parameter Estimation

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

$$\bar{\theta} : \sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \cdot \nabla_{\tilde{\theta}} \log q(\hat{\zeta}_\tau | s; \tilde{\theta}) + D_s \int q(\zeta; \tilde{\theta}) \nabla_{\tilde{\theta}} \log q(\zeta; \tilde{\theta}) d\zeta = 0 \quad (4)$$

where we define $\gamma'_s(\tau; \tilde{\theta}) = \gamma_s(\tau; \tilde{\theta}) - \gamma_s^g(\tau; \tilde{\theta})$. Here, the posteriors $\gamma_s(\tau; \tilde{\theta})$ and $\gamma_s^g(\tau; \tilde{\theta})$ are estimated for each state using the new transform estimates and the old Gaussian model parameters. To simultaneously update the Gaussian means and variances we take the derivative of the state dependent emission density with respect to μ_s and Σ_s^{-1} , substitute the result in equation (4) and solve for μ_s and Σ_s . The entire derivation is presented in the Appendix, where the estimate for the mean is shown to be

$$\bar{\mu}_s = \frac{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau + D_s \mu_s}{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) + D_s} \quad (5)$$

and the estimate for the variance is

$$\bar{\Sigma}_s = \frac{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau \hat{\zeta}_\tau^T \bar{T}_r^T + D_s (\Sigma_s + \mu_s \mu_s^T)}{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) + D_s} - \bar{\mu}_s \bar{\mu}_s^T. \quad (6)$$

In the implementation of the procedures reported here, the D_s values for equations (5) and (6) are calculated as described by Woodland and Povey [6] (Sec. 3, schemes *i* & *ii* with $E = 2$). 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 (3) which is iterated until the $[T_r]_i$ parameters converge. After this MMI Gaussian parameter estimates are found via equations (5) and (6).

C. 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 matrix 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.

III. DISCRIMINATIVE SPEAKER ADAPTIVE TRAINING

Speaker Adaptive Training (SAT) [3] 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 distributions 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 and variances fixed to their current values. Subsequently, we compute the speaker independent means and variances 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:

$$\begin{aligned} \bar{\theta} : \sum_{s_1^i} \left[q(s_1^i | \hat{w}_1^{\hat{n}}, \hat{o}_1^i, \hat{k}_1^i; \theta) - q(s_1^i | \hat{o}_1^i, \hat{k}_1^i; \theta) \right] \cdot \nabla_{\theta} \log q(\hat{o}_1^i, \hat{k}_1^i | s_1^i; \bar{\theta}) \\ + \sum_{s_1^i} d'(s_1^i) \int q(o_1^i, \hat{k}_1^i | s_1^i; \theta) \cdot \nabla_{\theta} \log q(o_1^i, \hat{k}_1^i | s_1^i; \bar{\theta}) do_1^i = 0. \quad (7) \end{aligned}$$

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

$$\sum_{k,r,s} \sum_{\tau=1}^i \log q(\hat{o}_{\tau} | s, k; \bar{\theta}) 1_k(\hat{k}_{\tau}) 1_s(s_{\tau}) 1_r(\mathcal{R}(s)), \text{ where } 1_k(\hat{k}_{\tau}) = 1 \text{ if } k = \hat{k}_{\tau}, 0 \text{ otherwise.}$$

Equation (7) then becomes:

$$\begin{aligned} \bar{\theta} : \sum_{k,r} \sum_{s \in S_r} \sum_{\tau: \hat{k}_{\tau}=k} \gamma'_s(\tau; \theta) \nabla_{\theta} \log q(\hat{o}_{\tau} | s, k; \bar{\theta}) \\ + \sum_{k,r} \sum_{s \in S_r} D_s^{(k)} \int q(o | s, k; \theta) \nabla_{\theta} \log q(o | s, k; \bar{\theta}) do = 0. \quad (8) \end{aligned}$$

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 $D_s^{(k)} = \sum_{\tau: \hat{k}_{\tau}=k} \sum_{s_1^i: s_{\tau}=s} d'(s_1^i)$.

A. Estimation of DSAT Transforms

With the HMM parameters fixed, the parameter update relationship of equation (8) can be expressed as:

$$\begin{aligned} \bar{T}_r^{(k)} : \sum_{s \in S_r} \sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \theta) \cdot \nabla_{T_r^{(k)}} \log q(\hat{o}_\tau | s, k; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) \\ + \sum_{s \in S_r} D_s^{(k)} \int q(o | s, k; T_r^{(k)}, \mu_s, \Sigma_s) \nabla_{T_r^{(k)}} \log q(o | s, k; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) do = 0. \end{aligned} \quad (9)$$

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 (9) gives

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

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

$$\sum_{s \in S_r} \Sigma_s^{-1} \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \theta) \hat{o}_\tau + D_s^{(k)} T_r^{(k)} \xi_s \right) \xi_s^T = \sum_{s \in S_r} \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \theta) + D_s^{(k)} \right) \Sigma_s^{-1} \bar{T}_r^{(k)} \xi_s \xi_s^T. \quad (10)$$

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

B. Gaussian Parameter Estimation

We now describe the estimation scheme for the state independent Gaussian means and variances under the CML criterion. With the transforms estimated as in Section III-A, we denote the parameter set as $\tilde{\theta} = (\bar{T}_r^{(k)}, \mu_s, \Sigma_s)$. To simultaneously update the Gaussian means and variances we differentiate the state dependent emission density with respect to μ_s and Σ_s^{-1} , substitute the result in equation (8), and solve for μ_s and Σ_s . The entire derivation may be found in the Appendix, where the estimate for the mean is shown to be

$$\bar{\mu}_s = \left(\sum_k \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \tilde{\theta}) + D_s^{(k)} \right) \bar{A}_r^{(k)T} \Sigma_s^{-1} \bar{A}_r^{(k)} \right)^{-1} \times \sum_k \bar{A}_r^{(k)T} \Sigma_s^{-1} \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \tilde{\theta}) (\hat{o}_\tau - b_r^{(k)}) + D_s^{(k)} \bar{A}_r^{(k)} \mu_s \right) \quad (11)$$

and the estimate for the variance is

$$\bar{\Sigma}_s = \frac{\sum_k \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \tilde{\theta}) (\hat{o}_\tau - \bar{\mu}_s^{(k)})^2 + D_s^{(k)} \left(\Sigma_s + (\bar{A}_r^{(k)} \mu_s - \bar{A}_r^{(k)} \bar{\mu}_s)^2 \right) \right)}{\sum_k \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \tilde{\theta}) + D_s^{(k)} \right)} \quad (12)$$

where $\bar{\mu}_s^{(k)} = \bar{A}_r^{(k)} \bar{\mu}_s + \bar{b}_r^{(k)}$, are the new speaker dependent means.

The state occupancies are found via counts accumulated for each speaker using the new speaker dependent transform estimates. The $D_s^{(k)}$ values for equations (10), (11) and (12) are set on a per speaker basis. They are calculated as described by Woodland and Povey [6] (Sec. 3, scheme *ii* with $E = 2$) and guarantee that the term to be inverted in equation (11) 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 (10), after which speaker independent MMI Gaussian parameters are found via equation (11) and equation (12).

IV. EXPERIMENTAL RESULTS

A. 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 experiments, were built using HTK [15] 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 [16]. The speech was parameterized into 39-dimensional PLP cepstral coefficients with delta and acceleration components [17]. Cepstral mean and variance normalization was performed over each conversation side. The acoustic models used cross-word triphones

with decision tree clustered states [15], 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 Large Vocabulary Decoder [18], with a 33k-word trigram language model provided by SRI [19].

The recognition tests were carried out on a subset of the 2000 Hub-5 Switchboard-1 evaluation set (SWBD1) [20] and the 1998 Hub-5 Switchboard-2 evaluation set (SWBD2) [21]. 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 [15]. All states of every context-dependent phone 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 [6]. 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.

B. DLLT Results

We conducted a series of experiments to compare DLLT to MLLT. Throughout the MLLT experiments [4], we perform one update of the transforms followed by one update of the Gaussians using statistics, obtained from the Viterbi alignment. These alignments were obtained from the ML baseline and were kept fixed throughout the MLLT experiments. Similarly, during DLLT iterations, we perform one update of the transforms followed by one update of the Gaussians, as described in Section II-A. The triphone posterior probabilities used during MMI based training were calculated from the ML baseline and were kept fixed throughout the DLLT updates.

#classes			0	1	2	3	4	5	6	7	8
48	MLLT	SWBD1	41.1	39.4	39.2	39.2	39.0	38.9	38.8	38.7*	38.8
		SWBD2	51.1	50.0	50.3	50.4	50.3	50.2	50.2	50.2*	50.1
	DLLT	SWBD1	*	37.6	37.4						
		SWBD2	*	49.5	48.8						
223	MLLT	SWBD1	41.1	38.9	38.7	38.7	38.2	37.9	37.8*	38.0	37.9
		SWBD2	51.1	49.8	49.8	49.6	49.5	49.5	49.3*	49.1	49.1
	DLLT	SWBD1	*	37.0	37.0						
		SWBD2	*	48.6	48.6						
467	MLLT	SWBD1	41.1	38.4	38.2	38.2	37.7	37.9	37.8*	37.9	37.8
		SWBD2	51.1	49.6	49.5	49.3	49.2	49.0	49.0*	49.0	49.1
	DLLT	SWBD1	*	37.1	36.9						
		SWBD2	*	48.4	48.8						

TABLE I

WORD ERROR RATE (%) OF SYSTEMS TRAINED WITH MLLT AND DLLT AND TESTED ON THE SWBD1 AND SWBD2 TEST SETS FOR DIFFERENT NUMBER OF CLASSES. DLLT SYSTEMS ARE SEEDED FROM WELL TRAINED MLLT SYSTEMS, INDICATED BY ASTERISKS

Our first experiment kept the parameters of the HMM observation distributions fixed at their ML values. Throughout these experiments we used a fixed set of 467 transform classes generated by the above described clustering algorithm. The SWBD1 ML baseline Word Error Rate is 41.1%. The first and second iteration of MLLT yield Word Error Rate of 39.1% and 39.4%, showing overtraining at the second iteration. DLLT yields Word Error Rate 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.

Our next objective was to see the improvements obtained by varying the number of transformation classes. For this purpose, we trained three MLLT systems with 48, 223 and 467 transformation classes, all initialized from the ML baseline. The 48 class MLLT system has transforms tied to each state of every context-dependent phone associated with the same mono-

	MLLT		DLLT		DLLT	
	SWBD1	SWBD2	SWBD1	SWBD2	SWBD1	SWBD2
0	41.1	51.1	†	†	*	*
1†	38.4	49.6	38.2	49.2	37.1	48.4
2	38.2	49.5	37.3	48.9	36.9	48.8
3	38.2	49.3	37.8	48.8	-	-
4	37.7	49.2				
5	37.9	49.0				
6*	37.8	49.0				
7	37.9	49.0				
8	37.8	49.1				

TABLE II

WORD ERROR RATE (%) OF SYSTEMS TRAINED WITH DLLT AND TESTED ON THE SWBD1 AND SWBD2 TEST SETS FOR TWO DIFFERENT INITIALIZATION POINTS.

phone. For the 223 and 467 class MLLT systems we allow a maximum of 6 and 10 divisions of the original classes, respectively. Results are reported in Table I. In these MLLT experiments we observe significant performance gains in using the larger number of transformation classes.

We then performed DLLT, for these same three collections of transformation classes. In each case, the DLLT system is seeded from a well-trained MLLT system (indicated by *). We note that in all cases the DLLT systems outperform the MLLT systems. Best performance is obtained with the larger number of transformation classes, although the advantages are not as great as with MLLT. This suggests that DLLT can be effective with fewer transforms than MLLT. For subsequent experiments we use the 467 transformation class set.

To validate our approach we calculated the value of $\log P(W|O; \theta)$ as a function of the iteration number for the 467 class DLLT system. In iteration 0 (the MLLT baseline) is $-2.15E05$, in DLLT iteration 1 is $-1.80E05$ and in DLLT iteration 2 is $-1.53E05$. These results confirm that the MMI objective function is increasing under the estimation procedure for DLLT.

In the next set of experiments we explored the sensitivity of DLLT to different initialization points. These experiments are shown in Table II. The DLLT system is seeded after 1 iteration of

	MLLT+MMIE		DLLT+MMIE	
	SWBD1	SWBD2	SWBD1	SWBD2
0	*	*	*	*
1	37.6	48.7	36.7	48.6
2	37.0	48.4	36.8	48.9
3	36.9	48.7	-	-

TABLE III

WORD ERROR RATE (%) OF SYSTEMS TRAINED WITH MLLT+MMIE AND DLLT+MMIE IS SEEDED FROM MODELS FOUND AFTER 6 MLLT ITERATIONS.

MLLT experiments (indicated by †). After two iterations, DLLT performance is (37.3%/48.9%). For convenience we repeat the 467 class MLLT experiments of Table I. We also repeat the 467 class DLLT experiments of Table I in which estimation was initialized after 6 iterations of MLLT experiments. After two iterations, DLLT performance is (36.9%/48.8%). We find that the latter is superior to performing DLLT after only 1 iteration of MLLT. This shows the importance of a proper initialization of the DLLT procedure. We also observe, that DLLT converges in fewer iterations 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.

We next study the application of MMIE estimation of the Gaussian means and variances in a system with fixed transforms estimated via MLLT. We call this technique MLLT+MMIE following Ljolje [11], in Table III. We compare this to a similar approach in which the transforms are estimated via 1 iteration of DLLT and are then fixed prior to further MMIE Gaussian iterations. We call this technique DLLT+MMIE in Table III. Both procedures were initialized with the well trained MLLT system of Table II found at iteration 6. The similar performance found in these scenarios is not surprising, in that enough in domain data is available to allow discriminative estimation of the unconstrained observation distributions. In this case, the discriminative estimation of the unconstrained Gaussian parameters clearly dominates the initial calculations of the underlying transforms.

C. DSAT Results

We conducted a series of experiments to compare DSAT to ML-SAT estimation. In speaker adaptive training the characteristics of a speaker are estimated from the test data itself and not from some transcribed enrollment data. Although we can estimate a large number of transforms for any of the training speakers since in training we have the correct transcription and adequate amount of data, this is not the case for test speakers (unsupervised adaptation with few data). Therefore the number of transforms that can be reliably estimated is limited. Throughout these experiments we used a fixed set of 2 regression classes corresponding to speech and non-speech states. As the number of transforms increases the WER increases, since the amount of test data is small enough that more than 2 transformation matrices can not be reliably estimated.

Table IV shows the performance of the ML-SAT and DSAT estimation procedures. ML-SAT was performed with a MMIE trained model indicated at iteration 0. In this implementation of ML-SAT, the transformation parameters and the Gaussian mean and variance parameters, are estimated at each iteration via Baum-Welsh procedure, over the transcribed training data. In the DSAT experiments only the mean and the transformation parameters are reestimated under the CML criterion. The variance is not updated between iterations; we keep the variance value estimated at ML-SAT iteration 5. Furthermore the lattice link posteriors used in DSAT are those obtained using the ML baseline model (41.1%/51.1%). Our goal is to show that DSAT can improve over ML-SAT through improved estimation of the speaker dependent models. We expect that further gains could be obtained by optimizing variances as well.

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. A comparison between DSAT, as described in Section III, and ML-SAT is presented in the columns DSAT-2 and ML-SAT of Table IV. The DSAT-2 mean and transformation parameters were reestimated at each iteration under the CML criterion. The best DSAT-2 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. Since we start from a well trained MMIE system, the gains obtained from DSAT-2 are due to the fact that we incorporate speaker adaptive training into MMIE parameter estimation.

	ML-SAT		DSAT-1		DSAT-2		MMI-SAT	
	SWBD1	SWBD2	SWBD1	SWBD2	SWBD1	SWBD2	SWBD1	SWBD2
0	35.9	47.0	35.9	47.0	*	*	*	*
1	35.4	46.2	36.1	46.5	34.1	44.7	34.1	44.8
2	35.2	45.6	36.5	46.5	33.8	44.6	33.8	44.6
3	34.8	45.1	36.5	46.7	33.6	44.5	33.7	44.4
4	34.7	45.2	-	-	33.5	44.4	33.5	44.4
5*	34.5	44.8	-	-	33.4	44.2	33.6	44.6
6	34.6	45.0						
7	34.3	45.0						
8	34.3	44.7						

TABLE IV

WORD ERROR RATE (%) OF SYSTEMS TRAINED WITH ML-SAT, MMI-SAT AND DSAT ESTIMATION AND EVALUATED ON SWBD1 AND SWBD2 TEST SETS. THE ML-SAT AND DSAT-1 MODELS WERE INITIALIZED BY MMI TRAINED MODELS. THE MMI-SAT AND DSAT-2 MODELS WERE SEEDED FROM MODELS FOUND AFTER 5 ML-SAT ITERATIONS. RESULTS INCLUDE UNSUPERVISED MLLR SPEAKER ADAPTATION.

While DSAT-2 was found superior to ML-SAT, performing ML-SAT subsequent to MMI is needed for the best initialization of DSAT. In the DSAT-1 column of Table IV the performance of DSAT initialized with MMIE is presented for a fair comparison with ML-SAT. Experimental results suggest that DSAT should be applied following several iterations of ML-SAT.

Finally, we compare DSAT with MMI-SAT. The previously developed MMI-SAT procedure [22] proceeds by fixing the ML-SAT transforms prior to subsequent iterations of MMIE estimation. A comparison between DSAT and MMI-SAT is presented in the columns DSAT-2 and MMI-SAT of Table IV. The experimental results show significant improvement over ML-SAT. Also DSAT gives slightly better results after 5 iterations, an absolute difference of 0.2%/0.4%, which is attributed to the discriminative calculation of the transformation matrices.

V. CONCLUSIONS

This paper describes the integration of Discriminative Linear Transforms into MMI estimation for Large Vocabulary Speech Recognition. We have developed estimation procedures that find DLTs jointly 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 estimation procedures. 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.

APPENDIX A

DISCRIMINATIVE LIKELIHOOD LINEAR TRANSFORMS

A. DLLT Estimation

The reestimation formula for the transforms are derived from the update relationship of equation (2). This derivation involves the differentiation of the reparametrized emission density q with respect to $[T_r]_i$ and the calculation of the integral. After these steps, we are able to express (2) in such form where an iterative solution is available.

With the HMM means and variances fixed, 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

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

where p_i is the extended cofactor row vector $[0 \ c_{i1} \ \dots \ c_{im}]$, ($c_{ij} = \text{cof}(A_{ij})$).

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

$$\begin{aligned} \sum_{s \in S_r} \sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \theta) \left(\frac{p_i}{p_i [\bar{T}_r^T]_i} - [\bar{T}_r]_i \hat{Z}_{s,i} + \hat{w}_{s,i} \right) \\ + \sum_{s \in S_r} D_s \int q(\zeta; T_r) \left(\frac{p_i}{p_i [\bar{T}_r^T]_i} - [\bar{T}_r]_i Z_{s,i} + w_{s,i} \right) d\zeta = 0. \end{aligned} \quad (13)$$

The calculation of the integral in equation (13) 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

$$\begin{bmatrix} 1 & [A_r^{-1}(\mu_s - b_r)]^T \\ A_r^{-1}(\mu_s - b_r) & A_r^{-1}[\Sigma_s + (\mu_s - b_r)(\mu_s - b_r)^T]A_r^{-1T} \end{bmatrix}. \quad (14)$$

Equation (13) can then be written as

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

Rearranging yields

$$\beta \frac{p_i}{p_i [\bar{T}_r^T]_i} = [\bar{T}_r]_i G_i - k_i \quad (15)$$

where

$$G_i = \sum_{s \in S_r} \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T + D_s J_s \right)$$

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

$$\beta = \sum_{s \in S_r} \left(\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \theta) + D_s \right)$$

An iterative solution to the optimization of equation (15) is described by Gales [4], 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

$$[\bar{T}_r]_i = (\alpha p_i + k_i) G_i^{-1}$$

where α satisfies a quadratic expression (equation B1.8, [4]).

B. Gaussian Parameter Estimation

The state dependent Gaussian mean and variance parameters are estimated under the CML criterion with the use of the simplified parameter update relationship of equation (4). The derivation of the update formulas involves the gradient of the reparametrized emission density with respect to μ_s and Σ_s^{-1} and the calculation of the integral. After these steps, we solve for μ_s and Σ_s . With the transforms estimated as described, we denote the entire parameter set as $\tilde{\theta} = (\bar{T}_r, \mu_s, \Sigma_s)$.

1) *Mean estimation:* The gradient of $\log q(\zeta|s; \tilde{\theta})$ with respect to the parameter component μ_s is given by

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

Substituting into equation (4) and rearranging gives

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

Calculating the integral yields

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

Finally the update equation for μ_s is given by

$$\bar{\mu}_s = \frac{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau + D_s \mu_s}{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) + D_s} \quad (16)$$

2) *Variance estimation:* The gradient of $\log q(\zeta|s; \tilde{\theta})$ with respect to Σ_s^{-1} is given by

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

Substituting into equation (4) gives

$$\begin{aligned}\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s) (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s)^T \right) \\ + D_s \int q(\zeta|s; \tilde{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \zeta - \bar{\mu}_s) (\bar{T}_r \zeta - \bar{\mu}_s)^T \right) d\zeta = 0. \quad (17)\end{aligned}$$

Calculating the integral in the previous equation gives

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

Substituting the integral into equation (17) yields

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

Using the fact that $\bar{\mu}_s$ is given by equation (16) we can obtain the reestimation formula for the new estimate of Σ_s as

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

APPENDIX B

DISCRIMINATIVE SPEAKER ADAPTIVE TRAINING

A. Gaussian Parameter Estimation

The state independent Gaussian mean and variance parameters for DSAT are estimated under the CML criterion with the use of the parameter update relationship of equation (8). The derivation of the update formulas involves the gradient of the reparametrized emission density with respect to μ_s and Σ_s^{-1} and the calculation of the integral. After these steps, we solve for μ_s and Σ_s . The parameter set is $\tilde{\theta} = (\bar{T}_r^{(k)}, \mu_s, \Sigma_s)$.

1) *Mean estimation:* From equation (8), the Gaussian means are found as:

$$\begin{aligned} \bar{\mu}_s : \sum_k \sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \tilde{\theta}) \cdot \nabla_{\mu_s} \log q(\hat{o}_\tau | s, k; \bar{T}_r^{(k)}, \bar{\mu}_s, \Sigma_s) \\ + \sum_k D_s^{(k)} \int q(o|s, k; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) \cdot \nabla_{\mu_s} \log q(o|s, k; \bar{T}_r^{(k)}, \bar{\mu}_s, \Sigma_s) do = 0. \quad (18) \end{aligned}$$

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 (18) 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 $\bar{\mu}_s^{(k)} = \bar{A}_r^{(k)} \bar{\mu}_s + \bar{b}_r^{(k)}$, is defined as the new 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

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

2) *Variance estimation:* From equation (8), the Gaussian variance is found as:

$$\bar{\Sigma}_s^{-1} : \sum_k \sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \tilde{\theta}) \cdot \nabla_{\Sigma_s^{-1}} \log q(\hat{o}_\tau | s; \bar{T}_r^{(k)}, \bar{\mu}_s, \bar{\Sigma}_s) + \sum_k D_s^{(k)} \int q(o|s; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) \nabla_{\Sigma_s^{-1}} \log q(o|s; \bar{T}_r^{(k)}, \bar{\mu}_s, \bar{\Sigma}_s) do = 0. \quad (19)$$

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 (19) 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

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

ACKNOWLEDGMENTS

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

REFERENCES

- [1] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1998.
- [2] M. J. F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, 1999.
- [3] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *International Conference on Spoken Language Processing*, 1996, pp. 1137–1140.
- [4] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, 1998.
- [5] Y. Normandin, "Maximum mutual information estimation of hidden Markov models," in *Automatic Speech and Speaker Recognition: Advanced Topics*, Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, Eds., chapter 3, pp. 57–81. Kluwer, 1996.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *European Conference on Speech Communication and Technology*, 2001.
- [10] 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.
- [11] A. Ljolje, "The AT&T LVCSR-2001 system," in *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.
- [12] 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.
- [13] R. Schlüter, *Investigations on Discriminative Training Criteria*, Ph.D. thesis, RWTH Aachen - University of Technology, 2000.
- [14] A. Gunawardana, "Maximum mutual information estimation of acoustic hmm emission densities," Tech. Rep. CLSP Research Note No. 40, CLSP, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA, 2001.
- [15] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, Version 3.0*, July 2000.
- [16] W. Byrne, "The JHU March 2001 Hub-5 Conversational Speech Transcription System," in *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.
- [17] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.
- [18] M. Mohri and M. Riley, "Integrated context-dependent networks in very large vocabulary speech recognition," in *European Conference on Speech Communication and Technology*, 1999.
- [19] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. S onmez, F. Weng, and J. Zheng, "The SRI march 200 Hub-5 conversational speech transcription system," in *Proceeding of the Speech Transcription Workshop*. NIST, 2000.
- [20] A. Martin, M. Przybocki, J. Fiscus, and D. Pallett, "The 2000 NIST evaluation for recognition of conversational speech over the telephone," in *Proceeding of the Speech Transcription Workshop*. NIST, 2000.

- [21] A. Martin, J. Fiscus, M. Przybocki, and 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.
- [22] John McDonough, Thomas Schaaf, and Alex Waibel, “On maximum mutual information speaker-adapted training,” in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002.