

DISCRIMINATIVE LINEAR TRANSFORMS FOR FEATURE NORMALIZATION AND SPEAKER ADAPTATION IN HMM ESTIMATION

Stavros Tsakalidis, Vlasios Doumptiotis, William Byrne

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218, USA
{stavros,vlasios,byrne}@clsp.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.

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 known that explicit modeling of correlations between spectral parameters in speech recognition should result in increased classification accuracy and improved descriptive power. However, computational, storage and robust estimation considerations make the use of full covariance matrices in HMM observation distributions impractical. A technique widely used to capture the correlations in the feature vector is Maximum Likelihood Linear Transformations (MLLT) [5, 4]. The transform is applied to the data and as a result each state has a full covariance Gaussian distribution. These covariance matrices are constrained by transformation matrices estimated over 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. During training, transforms are applied to the speaker independent Gaussian means of the acoustic models to derive speaker dependent models. SAT is an iterative procedure that produces a set of speaker independent state observation distributions along with a set of 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]. Recently, discriminative training under the Maximum Mutual Information (MMI) criterion [10] has been shown to be useful in large vocabulary conversational speech recognition (LVCSR) tasks [14]. Its success has triggered an interest in the use of linear transformation methods estimated under MMI criterion, rather than under ML. These are called Discriminative Linear Transforms (DLT) [12].

MMILR was introduced by Uebel and Woodland [12, 13] as a DLT estimation procedure. They showed that it can be used for supervised adaptation. Gunawardana and Byrne [7] introduced CMLLR that estimates DLT under the MMI criterion and showed that CMLLR can be used for unsupervised adaptation.

Maximum likelihood linear transforms have also been incorporated with MMI training. McDonough [9] 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 [8] combined MLLT with MMI estimation of the HMM Gaussian parameters. These transforms were found using ML estimation and then fixed throughout the subsequent MMI model estimation iterations.

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. The derivation of these procedures proceeds 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:

$$\bar{\theta} : \sum_{s:\mathcal{R}(s)=r} \sum_{\tau} \gamma_{s,\tau} \cdot \nabla_{\theta} \log q(o_{\tau}|s;\bar{\theta}) + \sum_{s:\mathcal{R}(s)=r} D_s \int q(o|s;\theta) \nabla_{\theta} \log q(o|s;\bar{\theta}) do = 0 \quad (1)$$

θ is the parameter we wish to estimate under the CML criterion; it is tied over a set of states, defined by the regression class $\{s|\mathcal{R}(s)=r\}$. We will show, in the subsequent sections, how this estimation criterion can be used for feature normalization and speaker adaptation in HMM training.

This work was partly supported by the National Science Foundation under Grant No. IIS-9982329.

2. DISCRIMINATIVE LIKELIHOOD LINEAR TRANSFORMS

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 such that $Ao + b$, 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 is the extended transformation matrix, $[b_r \ A_r]$ associated with the r th regression class; 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 reparametrization of the emission density augments the usual set of HMM parameters with the parameters of the transform. This new parameter set is defined as $\theta = (T_r, \mu_s, \Sigma_s)$.

Our goal is to estimate the transforms and HMM parameters under the CML criterion. We call this modeling approach Discriminative Likelihood Linear Transform (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.

We now outline the main points of the estimation procedure. A detailed presentation is available [11]. The derivation incorporates Gales' treatment of MLLT [3] and Gunawardana's CMLLR derivation [7]. With the HMM means and variances fixed, the transform estimate is found by differentiating the logarithm of the emission density q with respect to $[\bar{T}_r]_i$ (the i th row vector of \bar{T}_r) and substituting the result in equation (1) :

$$\beta \frac{p_i}{p_i [\bar{T}_r^T]_i} = [\bar{T}_r]_i \mathbf{G}_i - \mathbf{k}_i \quad (2)$$

where \mathbf{p}_i is the extended cofactor row vector $[0 \ c_{i1} \ \dots \ c_{im}]$, $(c_{ij} = \text{cof}(\bar{A}_{ij}))$, and

$$\mathbf{G}_i = \sum_{s: \mathcal{R}(s)=r} \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau} \gamma_{s,\tau} \zeta_{\tau} \zeta_{\tau}^T + D_s \mathbf{J}_s \right)$$

$$\mathbf{k}_i = \sum_{s: \mathcal{R}(s)=r} \frac{\mu_{s,i}}{\sigma_{s,i}^2} \left(\sum_{\tau} \gamma_{s,\tau} \zeta_{\tau}^T + D_s [\mathbf{J}_s]_1 \right)$$

$$\beta = \sum_{s: \mathcal{R}(s)=r} \left(\sum_{\tau} \gamma_{s,\tau} + D_s \right)$$

Here, $\gamma_{s,\tau}$ is estimated for each state under the initial parameters (T_r, μ_s, Σ_s) . \mathbf{J}_s is defined as the matrix

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

An iterative solution to the optimization of equation (2) is described by Gales [3]. It can be shown that the i^{th} row of the transformation matrix is found by

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

where α satisfies a simple quadratic expression.

After the transform \bar{T}_r has been obtained, the Gaussian parameters are reestimated as

$$\bar{\mu}_s = \frac{\sum_{\tau} \gamma'_{s,\tau} \bar{T}_r \zeta_{\tau} + D_s \mu_s}{\sum_{\tau} \gamma'_{s,\tau} + D_s} \quad (4)$$

$$\bar{\Sigma}_s = \frac{\sum_{\tau} \gamma'_{s,\tau} \bar{T}_r \zeta_{\tau} (\bar{T}_r \zeta_{\tau})^T + D_s (\Sigma_s + \mu_s \mu_s^T)}{\sum_{\tau} \gamma'_{s,\tau} + D_s} - \bar{\mu}_s \bar{\mu}_s^T \quad (5)$$

Note that $\gamma'_{s,\tau}$ has been estimated using the updated values of the parameters $(\bar{T}_r, \mu_s, \Sigma_s)$. The constants D_s are determined by the accumulators [14]. This defines a two-step, iterative procedure. Transforms are estimated via equation (3), after which MMI parameter estimates are found via equations (4) and (5).

2.1. Effective DLLT Estimation

On inspection of the definition of G_i it can be seen that, even for small values of D_s , the resulting transform will have dominant diagonal terms when the covariance Σ_s in J_s is diagonal. This behavior is greatly exaggerated when D_s is set to the large values required by MMI. In these situations, the resulting transform is effectively identity. We note that MLLT does not have the problem since it has no D_s or J_s terms. To address this problem 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 in the training test.

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

$$q(o|s; \theta) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_s|}} e^{-\frac{1}{2}(o - T_{\mathcal{R}(s)}^{(k)} \xi_s)' \Sigma_s^{-1} (o - T_{\mathcal{R}(s)}^{(k)} \xi_s)}.$$

Here, $T_r^{(k)}$ is the extended transformation matrix, $[b_r^{(k)} \ A_r^{(k)}]$ associated with the r th regression class and speaker k ; and ξ_s is the extended mean vector $[1 \ \mu_s^T]^T$. The augmented parameter set is defined as $\theta = (T_r^{(k)}, \mu_s, \Sigma_s)$.

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.

With the HMM parameters fixed, the estimate of each speaker dependent transform is found by differentiating the logarithm of the emission density q with respect to $\bar{T}_r^{(k)}$ and substituting the result in equation (1) :

$$\begin{aligned} \bar{T}_r^{(k)} : \sum_s \left(\sum_{\tau} \gamma_{s,\tau}^{(k)} \Sigma_s^{-1} o_{\tau} + D_s^{(k)} \Sigma_s^{-1} T_r^{(k)} \xi_s \right) \xi_s^T \\ = \sum_s \left(\sum_{\tau} \gamma_{s,\tau}^{(k)} + D_s^{(k)} \right) \Sigma_s^{-1} \bar{T}_r^{(k)} \xi_s \xi_s^T. \end{aligned} \quad (6)$$

Here, $\gamma_{s,\tau}^{(k)}$ is estimated for each state per speaker under the initial parameters $(T_r^{(k)}, \mu_s, \Sigma_s)$. Given the new estimate of the speaker dependent transform $\bar{T}_r^{(k)}$, speaker independent means are then reestimated as

$$\begin{aligned} \bar{\mu}_s = \left(\sum_{\tau,k} \left(\gamma_{s,\tau}^{(k)} + D_s^{(k)} \right) \bar{A}_r^{(k)T} \Sigma_s^{-1} \bar{A}_r^{(k)} \right)^{-1} \times \\ \sum_{\tau,k} \bar{A}_r^{(k)T} \Sigma_s^{-1} \left(\gamma_{s,\tau}^{(k)} (o_{\tau} - \bar{b}^{(k)}) + D_s^{(k)} \bar{A}_r^{(k)} \mu_s \right). \end{aligned} \quad (7)$$

The constants $D_s^{(k)}$ are set on a per speaker basis. They are determined by the accumulators [14] and guarantee that the first term in the right-hand side 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 (6), after which MMI mean estimates are found via equation (7). The SI covariance matrices can be similarly estimated.

4. EXPERIMENTAL RESULTS

4.1. System Description

The baseline system is a speaker independent continuous mixture density, tied state, gender-independent, HMM sys-

	MLLT		DLLT-1		DLLT-2	
	Swb1	Swb2	Swb1	Swb2	Swb1	Swb2
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 Swb1 and Swb2 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.

tem. 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 [2]. The acoustic models used were crossword triphones with 4000 tied triphone states with 6 Gaussian components per state. Lattice rescoring experiments were performed using the AT&T Large Vocabulary Decoder [8].

The recognition tests were carried out on a subset of the 2000 Hub-5 Switchboard-1 evaluation set (Swb1) and the 1998 Hub-5 Switchboard-2 evaluation set (Swb2) [2]. The Swb1 test set was composed of 866 utterances from 22 conversation sides, and the Swb2 test set was composed of 913 utterances 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 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 word lattices generated on the training data. Our approach is based on the MMI training procedure developed by Woodland and Povey [14]. However, rather than accumulating statistics via the Forward-Backward procedure at the word level, we use the Viterbi procedure over triphone segments. These 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 regression 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 Swb1 ML baseline WER 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 Swb2. These experiments show that DLLT can yield better improvement than MLLT 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 (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 (DLLT-2 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%). These experiments show that DLLT performs best when seeded by MLLT.

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 4 iterations (33.4%/44.2%). For comparison we present results with further iterations of ML-SAT (34.1%/44.8%). 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. DISCUSSION

This paper describes the integration of discriminative linear transforms into MMI estimation for robust ASR. We find that DLTs can be used in conjunction with MMI in both speaker adaptive training and feature normalization.

We have reported results on the Switchboard corpus. We have found that DLLT can be used to remove some broad corpus based differences as shown by the gains found in applying DLLT trained models to both Swb1 and Swb2. Similar performance gains were found using DSAT models. This behavior can be explained by noticing that the CML

	ML-SAT		DSAT	
	Swb1	Swb2	Swb1	Swb2
0	35.9	47.0	*	*
1	35.8	45.9	34.2	44.9
2	35.3	45.4	33.8	44.5
3	35.0	45.2	33.8	44.3
4	34.8	45.1	33.4	44.2
5*	34.5	45.1	33.6	44.2
6	34.2	44.9		
7	34.3	45.0		
8	34.1	44.8		

Table 2: Word Error Rate (%) of systems trained with ML-SAT and DSAT estimation and evaluated on Swb1 and Swb2 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.

training objective used for DSAT is closer to the criterion used in recognition. The two different modeling approaches suggest that DLLT and DSAT may provide complementary gains when used together if in fact they are capturing different acoustic phenomena.

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 FSM tools for MMI estimation.

6. REFERENCES

- [1] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *ICSLP*, 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. *Comp. Spch. & Lang.*, 12, 1998.
- [4] M. J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Trans. Spch. & Aud. Proc.*, 7(3), 1999.
- [5] R. A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *ICASSP*. IEEE, 1998.
- [6] A. Gunawardana. Maximum mutual information estimation of acoustic hmm emission densities. Technical Report CLSP Research Note No. 40, CLSP, JHU, 2001.
- [7] A. Gunawardana and W. Byrne. Discriminative speaker adaptation with conditional maximum likelihood linear regression. In *Eurospeech*, 2001.
- [8] A. Ljolje. The AT&T LVCSR-2001 system. In *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.

- [9] J. McDonough, T. Schaaf, and A. Waibel. On maximum mutual information speaker-adapted training. In *ICASSP*. IEEE, 2002.
- [10] Y. Normandin. Maximum mutual information estimation of hidden Markov models. In C.-H. Lee, F. Soong, and K. Paliwal, editors, *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer, 1996.
- [11] S. Tsakalidis, V. Doumptotis, and W. Byrne. Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation. Technical Report CLSP Research Note No. 47, CLSP, JHU, 2002.
- [12] L. F. Uebel and P. C. Woodland. Discriminative linear transforms for speaker adaptation. In *Proc. ITRW ASR*. ISCA, 2001.
- [13] L. F. Uebel and P. C. Woodland. Improvements in linear transforms based speaker adaptation. In *ICASSP*. IEEE, 2001.
- [14] P. C. Woodland and D. Povey. Large scale discriminative training for speech recognition. In *Proc. ITRW ASR*. ISCA, 2000.
- [15] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book, Version 3.0*, July 2000.