LARGE MARGIN SEMI-TIED COVARIANCE TRANSFORMS FOR DISCRIMINATIVE TRAINING

George Saon, Daniel Povey and Hagen Soltau

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
E-mail: {gsaon,dpovey,hsoltau}@us.ibm.com

ABSTRACT

We discuss the applicability of large margin techniques to the problem of estimating linear transforms for discriminative training of a semi-tied covariance (STC) model. Since STC models are good proxies for full-covariance (FC) Gaussian models, the idea is to combine the benefit of the latest discriminative training techniques and the modeling advantage of FC Gaussians at a much lower computational cost. We study the interaction of these transforms with feature-space and model-space discriminative training on state-of-the-art speaker adapted systems built for a large-scale Arabic broadcast news transcription task.

Index Terms— speech recognition, covariance matrices, discriminative training

1. INTRODUCTION

Nowadays, most modern ASR systems use some form of discriminative training of the acoustic model and/or of the features that are fed into the recognizer. While the type of discriminative training may differ, there is a wide-spread agreement in the structure of the acoustic model: the observations within an HMM state are usually modeled with mixtures of diagonal covariance Gaussians.

With increased computational power available, it makes sense to revisit the diagonal covariance assumption for large-scale speech recognition tasks. One step in the direction of full-covariance modeling is given by the STC formulation [1], where each individual Gaussian has its own diagonal covariance but shares a common linear transformation with other Gaussians from the same regression class which "rotates" the covariance. As the number of regression classes increases, one can move closer and closer to a full-covariance model which is the limit when one transform is used for each component.

The other approach that is frequently encountered in the literature is to decorrelate the dimensions of the observation vectors at the feature level by applying feature-space transformations such as the maximum likelihood linear transform (MLLT) [2] which is equivalent to a single inverse STC transform applied to the features and to the Gaussian means. This approach of using a single feature-space transformation can only go so far and the argument has been made that a model-based solution to the dimension decorrelation problem which uses multiple transforms such as STC should be superior [1].

A couple of years ago, there has been a surge of interest in more accurate modeling of the Gaussian precision matrices (inverse covariances). Among the leading approaches, we can mention the extended MLLT (or EMLLT) model [3] and the subspace precision and mean (or SPAM) model [4]. It was found that large-scale discriminative estimation is more complex to implement for these models [5] and their advantage over the STC approach is unclear especially in the case of speaker-adapted systems with feature-space discriminative training.

In the original formulation, STC transforms are estimated in a maximum likelihood framework. It is a simple conceptual leap to consider other objective functions for estimating the transforms. In particular, discriminative training criteria which make use of correct and competing paths should be interesting to explore.

The idea of discriminative linear transforms (DLT) is not new. In [6], the author introduces mean and variance transforms for supervised and unsupervised speaker adaptation estimated using either MMI or MPE. Discriminative mean transforms for speaker adaptation are also discussed in [7, 8]. A more closely related work to ours is [9], where the authors use the MMI criterion for STC transform estimation during speaker adaptive training.

What differentiates our work is that the variance transforms are estimated using a large margin objective function. We also discuss the interaction with feature-space discriminative training and various forms of speaker adaptation which was something lacking in the prior art. We were inspired by the work of [10], where the authors use DLT's as a criterion mapping function from ML to MPE. The transforms are estimated to maximize the MPE criterion over the entire training data using MLLR-adapted speaker models. The distinction with [10] is that we drop the MLLR step at training time and train full variance DLT's (as opposed to mean and diagonal transforms) using the large margin objective function introduced in [11, 12].

At the outset, we expect this approach to work on top of a state-of-the-art discriminatively trained system because the observation model is richer. Had we tried to train mean and diagonal variance scaling transforms with large margin, it would have been hard to improve over an acoustic model trained directly with the same criterion. This is because parameter training through transforms is, in principle, subsumed by direct parameter estimation (modulo tying and smoothing issues). Indeed, experiments not reported here seem to confirm that training means and diagonal variances with DLT's, while competitive, is less efficient than direct estimation.

The paper is organized as follows: in section 2 we briefly revisit the STC formulation and introduce the modifications for the large margin estimation case. Section 3 describes the experiments and results and section 4 provides a final discussion.

2. LARGE MARGIN STC

Let $\theta = (\mathbf{A}, \{\mu_j\}_{1 \leq j \leq N}, \{\Sigma_j\}_{1 \leq j \leq N})$ be a shorthand notation for an N-state HMM with transition probability matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$,

We would like to acknowledge the support of DARPA under Grant HR0011-06-2-0001 for funding part of this work.

Gaussian mixture component means $\mu_j \in \mathbb{R}^n$ and covariances $\Sigma_j \in \mathbb{R}^{n\times n}$, $\Sigma_j = diag(\sigma_{j1}^2 \dots \sigma_{jn}^2)$. We assume an HMM with a single Gaussian component per state in order to simplify the subsequent notations. This is without loss of generality since a multiple emission densities per state HMM can be turned into a single emission density per state HMM by increasing the number of states accordingly.

Let the training data be $(\mathbf{X}, W^r)$ where $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_T$, $\mathbf{x}_t \in \mathbb{R}^n$ represents the acoustic observation sequence and $W^r = w_1^r, \dots, w_m^r$, represents the correct word sequence. Without loss of generality, we represent the entire training data as $(\mathbf{X}, W^r)$ even when it consists of independent utterances.

2.1. STC model

The STC model with a single semi-tied transform is comprised of the Gaussian-specific diagonal covariances Σ_j and a common matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ which gets applied to the covariances in the following way:

$$\hat{\Sigma}_j^{-1} = \mathbf{H}^T \Sigma_j^{-1} \mathbf{H} \tag{1}$$

that is, we transform the precision matrices Σ_j^{-1} instead of the actual covariances. Again, for the sake of simplicity, we make the assumption of a single STC transform in order to avoid a cluttered notation which might obscure the arguments. The log-likelihood of an observation $\mathbf{x}$ for state j under this model is:

$$\begin{aligned}&\log p(\mathbf{x}|q_j, \mathbf{H}) = \log|\mathbf{H}| - \\ &-\frac{1}{2}\left(\log|2\pi\Sigma_j| + Tr\left\{\mathbf{H}^T\Sigma_j^{-1}\mathbf{H}(\mathbf{x}-\mu_j)(\mathbf{x}-\mu_j)^T\right\}\right)\end{aligned} \tag{2}$$

2.2. Large margin objective function

In [12], we have introduced the following objective function for discriminative training of the acoustic model parameters:

$$\max_{\theta,\rho}\left\{\rho + \frac{1}{\lambda}\log\frac{p_\theta(\mathbf{X}|W^r)P(W^r)}{\sum_W p_\theta(\mathbf{X}|W)P(W)e^{\rho H(W^r,W)}}\right\} \tag{3}$$

where $\rho \geq 0$ is a margin scale parameter and $H(W^r, W)$ is the frame-based Hamming distance between the Viterbi state sequences corresponding to the reference word sequence W^r and the competing word sequence W. $P(W)$ is the language model probability of W which we assume to be constant for the purpose of this discussion. $p_\theta(\mathbf{X}|W)$ represents the likelihood of the acoustic sequence given the word sequence and depends on the HMM parameters θ. $\lambda > 0$ controls the trade-off between margin maximization and constraint violation. This objective function is a variant of the boosted MMI criterion introduced in [11] and is motivated by turning the constrained margin maximization problem

$$\begin{aligned}&\max \rho \\ &s.t.\ \log p(\mathbf{X}, W^r) - \log p(\mathbf{X}, W) \geq \rho H(W^r, W),\ \forall W\end{aligned} \tag{4}$$

into an unconstrained problem via a penalty function. The penalty function (3) balances margin maximization and constraint violation through the parameter λ. Additional steps have to do with collapsing the constraints into a maximum constraint and with using the "softmax trick" to obtain a differentiable expression [12].

2.3. Weak-sense auxiliary function for optimization

Assuming a fixed margin scale ρ, we can write (3) while only considering the terms which depend explicitly on $\mathbf{H}$:

$$\begin{aligned}&\mathcal{F}(\mathbf{H}) = \mathcal{F}^{num}(\mathbf{H}) - \mathcal{F}^{den}(\mathbf{H}) = \\ &\log p_\theta(\mathbf{X}|W^r, \mathbf{H}) - \log\sum_W p_\theta(\mathbf{X}|W, \mathbf{H})P(W)e^{\rho H(W^r,W)}\end{aligned} \tag{5}$$

where $\mathcal{F}^{num}$ and $\mathcal{F}^{den}$ correspond respectively to the numerator term[1] (the log-likelihood of the observations given the correct transcription) and to the denominator term (the log of the average likelihood of the competing paths "boosted" by the frame-based Hamming distance to the reference state sequence). Following the definition from [13], $\mathcal{G}(\mathbf{H}, \overline{\mathbf{H}})$ is a weak-sense auxiliary function for $\mathcal{F}(\mathbf{H})$ around $\overline{\mathbf{H}}$, if

$$\left.\frac{\partial\mathcal{G}(\mathbf{H}, \overline{\mathbf{H}})}{\partial\mathbf{H}}\right|_{\mathbf{H}=\overline{\mathbf{H}}} = \left.\frac{\partial\mathcal{F}(\mathbf{H})}{\partial\mathbf{H}}\right|_{\mathbf{H}=\overline{\mathbf{H}}} \tag{6}$$

A valid weak-sense auxiliary function can be constructed as the difference of two strong-sense auxiliary functions[2] corresponding to the numerator and to the denominator part plus a smoothing term:

$$\begin{aligned}&\mathcal{G}(\mathbf{H}, \overline{\mathbf{H}}) = \mathcal{G}^{num}(\mathbf{H}, \overline{\mathbf{H}}) - \mathcal{G}^{den}(\mathbf{H}, \overline{\mathbf{H}}) + \mathcal{G}^s(\mathbf{H}, \overline{\mathbf{H}}) \\ &= \sum_{\mathbf{q}} p_\theta(\mathbf{q}|\mathbf{X}, W^r, \mathbf{H}) \log p_\theta(\mathbf{X}, \mathbf{q}|\overline{\mathbf{H}}) - \\ &- \sum_{\mathbf{q}}\sum_W \frac{p_\theta(\mathbf{q}, W, \mathbf{X}|\mathbf{H})e^{\rho H(W^r,W)}}{\sum_{W'} p_\theta(W', \mathbf{X}|\mathbf{H})e^{\rho H(W^r,W')}} \log p_\theta(\mathbf{X}, \mathbf{q}|\overline{\mathbf{H}}) + \\ &+\mathcal{G}^s(\mathbf{H}, \overline{\mathbf{H}})\end{aligned} \tag{7}$$

where the first summation is taken over all possible state sequences $\mathbf{q}$ of length T. In the auxiliary function for the denominator the state sequence posteriors are computed using joint word and observation sequence likelihoods which are scaled by the exponentiated Hamming distances. The previous expression can be re-written in terms of posterior state occupancies as follows:

$$\begin{aligned}&\mathcal{G}(\mathbf{H}, \overline{\mathbf{H}}) = \sum_{j=1}^N\sum_{t=1}^T \left[\gamma_t^{num}(j) - \gamma_t^{den}(j)\right] \log p(\mathbf{x}_t|q_j, \overline{\mathbf{H}}) + \\ &+\sum_{j=1}^N D_j \int_{\mathbb{R}^n} p(\mathbf{x}|q_j, \mathbf{H}) \log p(\mathbf{x}|q_j, \overline{\mathbf{H}}) d\mathbf{x}\end{aligned} \tag{8}$$

where $\gamma_t^{num}(j)$ refers to the posterior probability of being in state q_j at time t given the observation sequence and the correct transcription, while $\gamma_t^{den}(j)$ is the posterior of being in state q_j at time t given the observation sequence calculated over all word hypotheses and D_j is a state-dependent smoothing constant. The above expression for the smoothing term $\mathcal{G}^s$ was suggested in [7] and a proof that it is a weak-sense auxiliary function can be found in [6].

[1] Note that $P(W^r)$ has been dropped since it does not depend on $\mathbf{H}$.

[2] $\mathcal{K}$ is said to be a strong-sense auxiliary function for $\mathcal{F}$ around $\overline{\mathbf{H}}$ iff $\mathcal{K}(\mathbf{H}, \overline{\mathbf{H}}) - \mathcal{K}(\overline{\mathbf{H}}, \overline{\mathbf{H}}) \leq \mathcal{F}(\mathbf{H}) - \mathcal{F}(\overline{\mathbf{H}})$ (cf. [13]).

2.4. Sufficient statistics for STC transform estimation

By plugging (2) into (8) we get after some manipulations

$$\mathcal{G}(\mathbf{H},\overline{\mathbf{H}}) = \sum_{j=1}^{N}\sum_{t=1}^{T}[\gamma_t^{num}(j)-\gamma_t^{den}(j)][log|\overline{\mathbf{H}}|- \\ -\frac{1}{2}\left(log|2\pi\Sigma_j| + Tr\left\{\overline{\mathbf{H}}^T\Sigma_j^{-1}\overline{\mathbf{H}}(\mathbf{x}_t-\mu_j)(\mathbf{x}_t-\mu_j)^T\right\}\right)\Big] \\ +\sum_{j=1}^{N} D_j\left[log|\overline{\mathbf{H}}| - \frac{1}{2}\left(log|2\pi\Sigma_j| + Tr\left\{\overline{\mathbf{H}}^T\Sigma_j^{-1}\overline{\mathbf{H}}\hat{\Sigma}_j\right\}\right)\right] \tag{9}$$

where we have used the fact that

$$\int_{\mathbb{R}^n} p(\mathbf{x}|q_j,\mathbf{H})(\mathbf{x}-\mu_j)(\mathbf{x}-\mu_j)^T d\mathbf{x} = \hat{\Sigma}_j = (\mathbf{H}^T\Sigma_j^{-1}\mathbf{H})^{-1} \tag{10}$$

Taking the derivative of (9) with respect to $\overline{\mathbf{H}}$ and setting it to zero leads to[3]:

$$\gamma\overline{\mathbf{H}}^{-T} = \sum_{j=1}^{N}\Sigma_j^{-1}\overline{\mathbf{H}}[D_j\hat{\Sigma}_j + \\ +\sum_{t=1}^{T}[\gamma_t^{num}(j)-\gamma_t^{den}(j)](\mathbf{x}_t-\mu_j)(\mathbf{x}_t-\mu_j)^T\Big] \tag{11}$$

where $\gamma = \sum_{j=1}^{N} D_j + \sum_{t=1}^{T}[\gamma_t^{num}(j)-\gamma_t^{den}(j)]$. The sufficient statistics for STC transform estimation are defined by the total count γ and the dimension-specific matrices

$$G_{(i)} = \sum_{j=1}^{N}\frac{1}{\sigma_{ji}^2}[D_j\hat{\sigma}_{ji}^2 + \\ +\sum_{t=1}^{T}[\gamma_t^{num}(j)-\gamma_t^{den}(j)](\mathbf{x}_t-\mu_j)(\mathbf{x}_t-\mu_j)^T\Big] \tag{12}$$

Given these statistics, equation (11) can be solved for $\overline{\mathbf{H}}$ iteratively as described in [1].

2.5. Smoothing techniques

The selection of the smoothing constants D_j is critical to the maximization. The best choice is similar to the one used in standard discriminative training, namely

$$D_j = E\sum_{t=1}^{T}\gamma_t^{den}(j) \tag{13}$$

with $E = 2.0$ in practice. A second smoothing technique which was found to be beneficial is given by the H-criterion [14] and consists in multiplying the state denominator posteriors by a fraction $H \in [0,1]$. In our experiments, the best results were obtained with $H = 0.5$. In Figure 1 we show the influence of the two smoothing parameters on the recognition results (DEV'07 test-set, feature and model-space discriminatively trained system, no MLLR adaptation). Note that $H = 0$ corresponds to maximum likelihood STC estimation.

[3]Notation $(\cdot)^{-T}$ means transposed inverse.

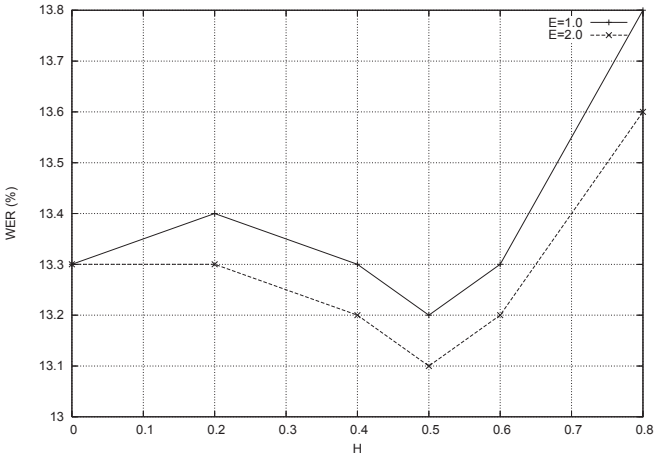

Fig. 1. Evolution of the word error rate as a function of H for $E = 1.0$ and $E = 2.0$ on DEV'07.

Finally, the third smoothing technique that we looked at is I-smoothing [15] which consists in adding τ points of statistics for each Gaussian to the numerator counts. The problem is that I-smoothing would require storing outer-product statistics for each Gaussian which is prohibitive in terms of memory[4].

3. EXPERIMENTS AND RESULTS

We report some experimental results on a large scale Arabic broadcast news transcription task which is part of the DARPA GALE program. The training data consists of 1400 hours of manually transcribed Arabic broadcast news and broadcast conversation shows and is provided by the LDC. Results are given on two test-sets: DEV'07 and EVAL'07 each having approximately 3 hours of speech.

The acoustic features are 40-dimensional vectors obtained via an HDA+MLLT projection of 9 consecutive spliced frames of 13-dimensional VTLN-warped PLP features which are mean and variance normalized on a per speaker basis. Additionally, the features are transformed through feature-space MLLR at both training and test time. Additionally, at test time we apply MLLR adaptation through a regression tree with at most 16 mean and diagonal variance scaling transforms with a minimum count of 3000 frames per transform.

The baseline system uses unvowelized (or graphemic) acoustic models with a pentaphone cross-word acoustic context. The size of the acoustic model is 5000 context-dependent HMM states and 400K 40-dimensional Gaussians. More details about the vocabulary and the language model used can be found in [16].

We report results on three sets of models: one set trained with maximum likelihood and two sets estimated with our most up-to-date discriminative training scheme which uses the large margin (or boosted MMI) criterion [11, 12]. The discriminative training is applied either only to the models (for set 2) or to both the features and the models (for set 3).

For STC transform estimation, we clustered the 400K Gaussians using k-means into 2048 regression classes corresponding to

[4]Another option would be to add $\tau/\sum_t \gamma_t^{num}(j)$ counts to (12) for Gaussian j for every frame. This requires knowing $\sum_t \gamma_t^{num}(j)$ beforehand.

Model	LM-STC	DEV07	EVAL07
ML	no	17.1%	19.6%
ML	yes	16.2%	18.5%
mBMMI	no	14.2%	16.4%
mBMMI	yes	13.9%	16.0%
fBMMI+mBMMI	no	12.7%	14.9%
fBMMI+mBMMI	yes	12.7%	14.8%

Table 1. Word error rates for different configurations on DEV'07 and EVAL'07. All decodings use VTLN, FMLLR and MLLR. Note that the baselines are already trained with a single ML STC as part of the HDA+MLLT transform. LM-STC stands for large margin (multiple) STC transforms. fBMMI and mBMMI stand for feature and model-space boosted MMI trained models.

the same number of transforms. In both training and decoding, we use an efficient hierarchical Gaussian likelihood evaluation scheme described in [17]. The top-level hierarchy is given by the same 2048 cluster centers. At run-time, we only evaluate the Gaussians which map to the top-N (say N=300) centers. Since there is a one-to-one correspondence between clusters and transforms, we use at most N STC transforms per frame. This improves the memory use and the transform application speed by a factor of 7 versus applying all the transforms on every frame. To avoid costly full-covariance likelihood evaluations, the transforms are applied at run-time to the features and off-line to the model means (instead of the precisions). The transformed features are computed on-demand and are cached for future use.

An important observation is that we only estimate the transforms not the diagonal variances. Otherwise stated, the diagonal variances remain unchanged after the STC estimation step (estimated using either ML or BMMI). This means that, in the ML case, the gains coming from discriminative variance estimation could be further increased.

In Table 1 we show the results on DEV'07 and EVAL'07 for the various configurations that we have investigated: ML-trained versus model-space BMMI training versus both feature and model-space BMMI-trained systems with and without discriminative STC modeling. One can conclude from these results that discriminative STC modeling helps on top of systems that have not been trained with discriminative feature-space transforms. This suggests that there is an overlap in functionality between the STC transforms and the discriminative feature space mapping which leads to gains that are not additive. Indeed, our feature space boosted-MMI mapping can be construed as a region-dependent transform (or RDT) [18] trained to maximize the large margin criterion. The feature-space application of the STC transforms is also region-dependent and is optimized according to the same criterion. Conversely, large margin STC modeling can be thought of as a "poor man's" discriminative feature space transform because it is easier to implement but it also leads to inferior gains.

4. CONCLUSION

The contribution of this paper is two-fold: we formulate the estimation of precision matrix transforms using a large margin objective function and test it on a large-scale Arabic broadcast news transcription task. We deliberately made it hard on ourselves and compared the results with our most up-to-date discriminatively trained speaker-adapted systems. Our findings suggest that large margin STC modeling is useful for, either a maximum likelihood trained system, or for a model-space only discriminatively trained system. In the case of both feature and model-space training, the benefit from this type of modeling is greatly reduced. Future work will address the joint large margin training of STC transforms and model parameters.

5. REFERENCES

[1] M.J.F. Gales, "Maximum likelihood linear transformation for HMM-based speech recognition," in *Computer Speech and Language*, 1998.

[2] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *ICASSP-98*, 1998.

[3] P. Olsen and R. Gopinath, "Modeling inverse covariances by basis expansion," in *ICASSP-02*, 2002.

[4] S. Axelrod, P. Olsen, and R. Gopinath, "Modeling with a subspace constraint on inverse covariance matrices," in *ICASSP-02*, 2002.

[5] K.C. Sim and M. Gales, "Minimum phone error training of precision matrix models," in *IEEE Trans. on Speech and Audio Processing*, 2006.

[6] L. Wang, *Discriminative linear transforms for adaptation and adaptive training*, Ph.D. thesis, Cambridge University, 2006.

[7] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Eurospeech-01*, 2001.

[8] L. Uebel and P. Woodland, "Discriminative linear transforms for speaker adaptation," in *ISCA ITRW Adaptation Methods for Automatic Speech Recognition*, 2001.

[9] J. McDonough and A. Waibel, "Maximum mutual information speaker adapted training with semi-tied covariance matrices," in *ICASSP-03*, 2003.

[10] K. Yu, M. Gales, and P. Woodland, "Unsupervised discriminative adaptation using discriminative mapping transforms," in *ICASSP-08*, 2008.

[11] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature space discriminative training," in *ICASSP-08*, 2008.

[12] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," in *Interspeech-08*, 2008.

[13] D. Povey, *Discriminative Training for Large Voculabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[14] P. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo, and M. Picheny, "Decoder selection based on cross-entropies," in *ICASSP-88*, 1988.

[15] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP-02*, 2002.

[16] H. Soltau, G. Saon, B. Kingsbury, J. Kuo, L. Mangu, D. Povey, and G. Zweig, "The IBM 2006 GALE Arabic ASR system," in *ICASSP-07*, 2007.

[17] G. Saon, D. Povey, and G. Zweig, "Anatomy of an extremely fast LVCSR decoder," in *Interspeech-05*, 2005.

[18] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region dependent feature transforms for speech recognition," in *ICASSP-06*, 2006.