STATE OF THE ART DISCRIMINATIVE TRAINING OF SUBSPACE CONSTRAINED GAUSSIAN MIXTURE MODELS IN BIG TRAINING CORPORA

Jing Huang, Peder A. Olsen, Vaibhava Goel

IBM, TJ Watson Research Center
{jghg,pederao,vgoel}@us.ibm.com

ABSTRACT

Discriminatively trained full-covariance Gaussian mixture models have been shown to outperform its corresponding diagonal-covariance models on large vocabulary speech recognition tasks. However, the size of full-covariance model is much larger than that of diagonal-covariance model and is therefore not practical for use in a real system. In this paper, we present a method to build a large discriminatively trained full-covariance model with large (over 9000 hours) training corpora and still improve performance over the diagonal-covariance model. We then reduce the size of the full-covariance model to the size of its baseline diagonal-covariance model by using subspace constrained Gaussian mixture model (SCGMM). The resulting discriminatively trained SCGMM still retains the performance of its corresponding full-covariance model, and improves 5% relative over the same size diagonal-covariance model on a large vocabulary speech recognition task.

Index Terms— Discriminative Training, Full Covariance Modeling, Subspace Constrained Gaussian Mixture Model, Large Corpora

1. INTRODUCTION

Full-covariance models have been shown to outperform the diagonal-covariance models for maximum likelihood trained systems [1, 2, 3, 4], as well as for discriminative trained models [5]. In [5] full-covariance discriminatively trained model improved performance over the best diagonal covariance models. This is not always the case as very large diagonal covariance model could match the performance of the best full covariance models, especially when training data is plenty.

Most of the full-covariance work we have seen only deal with training corpora of modest size (around a couple of hundred hours) and not the larger training corpora currently used to train state of the art diagonal models. In this paper we discriminatively train full-covariance models that beat the best large diagonal covariance model. Then we find a compact representation no larger than that of the diagonal model that still retains the improved performance.

There are several methods that compactly represent inverse covariances with little loss in performance [6, 7, 8, 9, 10, 11, 12, 13, 14], and we used the subspace constrained Gaussian mixture model (SCGMM) representation in [7, 8, 11, 12]. For discriminative training we used the Minimum Phone Error (MPE) criterion [15], but a number of other competitive discriminative training criteria exist. For example Maximum Mutual Information (MMI) [16], Minimum Classification Error (MCE) [17], Minimum Bayes Risk (MBR) [18], Boosted MMI (BMMI) [15], and large margin training [19] are some potential alternatives.

1.1. Overview of our approach

A full covariance gaussian $f(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ with mean $\boldsymbol{\mu}_g$ and covariance $\boldsymbol{\Sigma}_g$ can be written in the form of exponential family

$$f(\mathbf{x}) = \frac{e^{\boldsymbol{\theta}_g^\top \boldsymbol{\phi}(\mathbf{x})}}{Z_{\text{fc}}(\boldsymbol{\theta}_g)}, \text{ where } \boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} \mathbf{x} \\ \text{vec}(\mathbf{x}\mathbf{x}^\top) \end{pmatrix}, \quad (1)$$

and

$$\boldsymbol{\theta}_g = \begin{pmatrix} \boldsymbol{\psi}_g \\ \mathbf{p}_g = \text{vec}(\mathbf{P}_g) \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{\Sigma}_g^{-1}\boldsymbol{\mu}_g \\ \text{vec}(\boldsymbol{\Sigma}_g^{-1}) \end{pmatrix} \quad (2)$$

and

$$Z_{\text{fc}}(\boldsymbol{\theta}_g) = \frac{1}{2}\psi_g^\top \mathbf{P}_g^{-1}\psi_g - \frac{1}{2}\log\det\mathbf{P}_g + \frac{d}{2}\log(2\pi) \quad (3)$$

is the partition function that normalizes the distribution. The SCGMM model then constructs a new exponential family by further constraining the parameters $\boldsymbol{\theta}_g$ to be in a subspace; $\boldsymbol{\theta}_g = \mathbf{B}\boldsymbol{\lambda}_g$, where $\mathbf{B} \in \mathbb{R}^{D\times\delta}$, $\delta \ll D$ represents the subspace that is shared among all gaussians. This is similar to the subspace GMM approach in [13], but as we require the corresponding covariance to be positive definite, our methods are more complex.

In this paper we experiment with more than 9000 hours of training data, and the baseline diagonal model has $400,000$ Gaussians. Discriminative full covariance models are first obtained by a few discriminative iterations starting with the baseline diagonal covariance model. The "magic constant" D^* is determined according to the approach outlined in [5].

Once a full covariance model is trained, the basis of the SCGMM are estimated using this model. The projection coefficients of SCGMM are then discriminatively trained. Our discriminatively trained SCGMM proves to be effective: retains the performance of its corresponding full-covariance model, and improves 5% relative over the same size diagonal-covariance model on a large vocabulary speech recognition task.

The rest of the paper is organized as follows: full-covariance discriminative training is briefly reviewed in Section 2. Section 3 describes the basis training and discriminative training of subspace constrained Gaussian mixture model. The experimental setup and results are reported in Section 4, and conclusions are presented in Section 5.

2. DISCRIMINATIVE TRAINING OF THE FULL-COVARIANCE MODEL

Recall that the maximum likelihood estimation of a FC model is equivalent to maximizing the strictly convex objective function

$$Q(\boldsymbol{\theta}_g) = \mathbf{s}_g^\top \boldsymbol{\theta}_g - c_g \log Z_{\text{fc}}(\boldsymbol{\theta}_g), \tag{4}$$

where $\mathbf{s}_g = \sum_t \gamma_g(\mathbf{x}_t)\boldsymbol{\phi}(\mathbf{x}_t)$ is the gaussian statistics weighted by the posterior counts $\gamma_g(\mathbf{x}_t)$, and $c_g = \sum_t \gamma_g(\mathbf{x}_t)$ is the count for the gaussian. For discriminative modeling we arrive at an auxilliary objective function that has the same form as (4):

$$\begin{aligned} Q(\boldsymbol{\theta}_g, \hat{\boldsymbol{\theta}}_g) &= (\mathbf{s}_{\text{num},g} - \mathbf{s}_{\text{den},g} + D_g E_{\hat{\boldsymbol{\theta}}_g}[\boldsymbol{\phi}(\mathbf{x})])^T \boldsymbol{\theta}_g \\ &\quad -(\sum_t \gamma_{\text{num},g}(\mathbf{x}_t) - \gamma_{\text{den},g}(\mathbf{x}_t) + D_g) \log Z_{\text{fc}}(\boldsymbol{\theta}_g), \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\theta}}_g$ is the present value of the parameters. Unfortunately, the statistics and counts are not well-formed for all values of D_g. Therefore we need to judisciously choose D_g so that the effective count is positive and the effective covariance is positive definite. D_g is usually set using the following formula:

$$D_g = \tau + \max\{C_1 \sum_t \gamma_{\text{den},g}(\mathbf{x}_t), C_2 D_g^*\}, \tag{6}$$

where τ, C_1, C_2 are global variables that do not depend on g. The "magic constant' D_g^* is the smallest value for which D_g yields a positive definite covariance statistics. It was shown in [5] how to exactly determine D_g^* by solving a quadratic eigenvalue problem. For $C_1 > 1$ the term $C_1 \sum_t \gamma_{\text{den},g}(\mathbf{x}_t)$ guarantees that counts will be positive and the second term $C_2 D_g^*$ guarantees a positive definite covariance for $C_2 > 1$. τ can therefore be thought of as a smoothing count. After the constants τ, C_1 and C_2 are chosen, and D_g estimated, the full covariance model is simply computed from the full covariance statistics in the usual way.

3. SUBSPACE CONSTRAINED GAUSSIAN MIXTURE MODEL

In determining the SCGMM coefficients $\boldsymbol{\lambda}_g$ and the subspace matrix $\mathbf{B}$ there are a number of problems to tackle, but the theory of exponential families [20] is a great help. The maximum likelihood objective function

$$Q_g(\boldsymbol{\lambda}_g) = \mathbf{s}_g^\top \boldsymbol{\lambda}_g - c_g \log Z(\boldsymbol{\lambda}_g), \tag{7}$$

where $Z(\boldsymbol{\lambda}_g) = Z_{\text{fc}}(\mathbf{B}\boldsymbol{\lambda}_g)$, is strictly convex when the statistics and counts are well formed. Therefore the values of $\boldsymbol{\lambda}_g$ can be determined using any convex optimization package *if* we know an initial value of $\boldsymbol{\lambda}_g$ and $\mathbf{B}$ such that the corresponding full covariance parameters $\boldsymbol{\theta}_g = \mathbf{B}\boldsymbol{\lambda}_g$ is well-formed. Similarly, we can find $\mathbf{B}$ by optimizing the convex auxilliary objective

$$Q(\mathbf{B}) = \sum_g \mathbf{s}_g^\top \boldsymbol{\lambda}_g - c_g \log Z(\boldsymbol{\lambda}_g), \tag{8}$$

This problem of finding good initial well formed values for the coefficients $\boldsymbol{\lambda}_g$ and the basis matrix $\mathbf{B}$ is the most difficult problem. Luckily this problem has been dealt with for us in [21], as briefly discussed in the following.

3.1. Initialization

Given a valid initial choice of the parameters $\boldsymbol{\lambda}_g$ and $\mathbf{B}$ the parameters $\boldsymbol{\lambda}_g$ can be determined using convex optimization if all other parameters are fixed. On the other hand, if we fix $\boldsymbol{\lambda}_g$ for all g then $\mathbf{B}$ can also be determined using convex optimization. The key problem is to determine an initial choice for the parameters.

The basic idea for finding a good initial value is to realize that we simply wish to have $\mathbf{B}\boldsymbol{\lambda}_g \approx \boldsymbol{\theta}_g$. We can do this by simply minimizing $\sum_g \|\boldsymbol{\theta}_g - \mathbf{B}\boldsymbol{\lambda}_g\|_{\mathbf{H}}$, where $\|\mathbf{x}\|_{\mathbf{H}} = \mathbf{x}^\top \mathbf{H}\mathbf{x}$ for some matrix $\mathbf{H}$. For $\mathbf{H} = \mathbf{I}$ the quadratic objective function is minimized when $\mathbf{B}$ consists of the top δ eigenvalues of the covariance of $\boldsymbol{\theta}_g$, and $\boldsymbol{\lambda}_g$ is the minimum least square solution given the eigenvalues, i.e. $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sum_g \pi_g \boldsymbol{\theta}_g \boldsymbol{\theta}_g^\top - \boldsymbol{\mu}_{\boldsymbol{\theta}}^2$, $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_g \pi_g \boldsymbol{\theta}_g$ and $\pi_g \propto c_g$.

In [21], it was argued that a better choice for $\mathbf{H}$ was the Hessian for the full covariance objective function at $\boldsymbol{\mu}_{\boldsymbol{\theta}}$. With this choice of Hessian it was shown that the following transformation

$$\begin{pmatrix} \boldsymbol{\psi}_g \\ \mathbf{p}_g \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\psi_g - \mathbf{P}_g \boldsymbol{\mu}_{\boldsymbol{\theta}}) \\ \text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2} \mathbf{P}_g \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2}) \end{pmatrix} \tag{9}$$

brings the problem back to the $\mathbf{H} = \mathbf{I}$ case, where we know how to find $\mathbf{B}$ and $\boldsymbol{\lambda}_g$. Then the basis vectors in the transformed space is transformed back into the original space giving the optimal solution to the quadratic problem. The coefficients $\boldsymbol{\lambda}_g$ are simply given by

$$\boldsymbol{\lambda}_g = (\mathbf{B}^\top \mathbf{H} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{H} \boldsymbol{\theta}_g. \tag{10}$$

After this process we will have $\mathbf{B}\boldsymbol{\lambda}_g \approx \boldsymbol{\theta}_g$ for most g and a very good choice for the matrix $\mathbf{B}$. However, there will still be a few values for which $\mathbf{B}\boldsymbol{\lambda}_g$ is not positive definite. For these values we found a valid value for $\boldsymbol{\lambda}_g$ by a method that iteratively project onto two convex sets. This method is known as a generalized Bregman iteration [22]. Let $\boldsymbol{\theta}_g^{(1)} = \mathbf{B}\boldsymbol{\lambda}_g$. Then project $\boldsymbol{\theta}_g^{(1)}$ onto the set of slightly positive definite matrices

$$\mathcal{P}_\epsilon = \left\{ \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\psi} \\ \mathrm{vec}(\mathbf{P}) \end{pmatrix} : \mathbf{P} \succ \epsilon \mathbf{I} \right\} \tag{11}$$

simply by setting all eigenvalues below ϵ to ϵ. Then we project this value onto the linear subspace giving $\boldsymbol{\lambda}_g^{(1)} = (\mathbf{B}^\top \mathbf{H} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{H} \boldsymbol{\theta}_g^{(1)}$. Since the linear subspace is known to contain positive definite matrices there exists an ϵ so that the set $\mathcal{P}_\epsilon$ intersects with the linear subspace. The iterative projection scheme which we showed one iteration of will eventually converge to a point in the intersection of the two convex sets. We use the first value k such that $\mathbf{B}\boldsymbol{\lambda}_g^{(1)}$ is in $\mathcal{P}_{\epsilon/2}$ as an initial value.

After finding the basis matrix $\mathbf{B}$ and valid initial coefficients $\boldsymbol{\lambda}_g$ we simply use convex optimization to maximize (7) to refine the values for $\boldsymbol{\lambda}_g$ and to maximize (8) to refine the value for $\mathbf{B}$. This is equivalent to doing maximum likelihood training of the SCGMM parameters given the full covariance stats.

3.2. Discriminative Training of SCGMM

Training the SCGMM parameters discriminatively boils down to iteratively solving maximum likelihood problems, with the exception that we have to decide the value of the discriminative smoothing parameters D_g. Finding D_g^* is a very difficult problem for the SCGMM model. On the other hand if $\mathbf{B}\boldsymbol{\lambda}_g \approx \boldsymbol{\theta}_g$ then it is reasonable to expect that D_g^* is very close to the "magic constant" D_g^* in the full covariance case. That is exactly the approach we took. We first collected the full covariance statistics $\mathbf{s}_{\mathrm{num}}$, $\mathbf{s}_{\mathrm{den}}$; computed D_g^* in the full covariance case, from which we computed the total full covariance discriminative statistics

$$\mathbf{s} = \mathbf{s}_{\mathrm{num},g} - \mathbf{s}_{\mathrm{den},g} + D_g E_{\hat{\boldsymbol{\theta}}_g}, \tag{12}$$

and correspondingly for the counts. The corresponding SCGMM statistics is then simply $\mathbf{B}^\top \mathbf{s}$, and the parameters is then learned by convex optimization using the projected full covariance statistics. For a more thorough discussion of discriminative training of the SCGMM parameters we refer the readers to [12].

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Experimental Setup

We used an internal US English speech recognition task for our experiments. The training set consists of 9300 hours of recordings. The training transcripts are obtained through decoding from an existing large vocabulary continuous speech recognition (LVCSR) system. Recordings that are deemed (by the recognizer) to be all silence or noise, or sentences decoded with very low confidence are excluded from the training data. There are a total of 300 test speakers with around $2,400$ testing utterances and $24,000$ words.

Acoustic features were constructed from 12 dimensional Mel-frequency Cepstra coefficients and their first, second and third derivative, followed by a Linear Discriminant Analysis (LDA) projection to 32 dimension.

Our baseline acoustic model is an HMM with $35,000$ context-dependent states and with about $400,000$ gaussian components. The acoustic model was train with maximum likelihood followed by fMPE (feature-space minimum phone error rate) training and discriminative MPE (Minimum Phone Error) training, as described in [23].

4.2. Results

The baseline diagonal model has a word error rate (WER) of 14.1% on the test data. Table 1 shows the performance of the discriminatively trained full-covariance model at each iteration, starting from the diagonal-covariance model. The final FC model improves 5% relative to the baseline diagonal model. The gain comes mostly from the first iteration.

iter	C_1	WER
1	4.0	13.6%
2	8.0	13.5%
3	20.0	**13.4%**

Table 1. Word error rates (WERs) for each iteration of discriminative training of FC models with MPE.

We then take the FC model at iteration 3 and train SCGMM as described in Section 3.1. If we set the dimension of the basis as 64, then the resulting SCGMM has the same number of parameters as the baseline diagonal model; If we set the dimension of the basis as 54, then the resulting SCGMM is 15% smaller than the baseline diagonal model. Table 2 shows the performance of the discriminatively trained SCGMMs at each iteration. From Table 3 we show that the 64-dim SCGMM has the same performance as the FC model it trained from, and is 5% better than the baseline diagonal model. Even with 54-dim SCGMM the performance is still better than the baseline model with 15% less model parameters.

4.3. Related Work

SCGMM is just one of many ways to create compact models. MIC [10] (mixtures of inverse covariances), SPAM (subspace for precision and mean model) and recent SGMM (subspace

iter	64-dim	54-dim
1	13.9	14.3
2	13.6	14.1
3	13.5	13.9
4	13.5	13.9
5	13.4	13.8

Table 2. Word error rates (WERs) for each iteration of discriminative training of SCGMMs with MPE.

model	WER
baseline	14.1
FC	13.4
SCGMM-64dim	**13.4**
SCGMM-54dim	13.8

Table 3. Comparison of Word error rates (WERs) for different models.

Gaussian mixture models) [13, 24, 25] are also well-known techniques. SCGMM is a direct generalisation of SPAM, which in turn is a generalization of MIC. The basis in MIC are typically positive definite inverse covariance matrices, while we do not have this restriction in SCGMM; in SGMM the means and weights are represented in a shared subspace, and covariances are shared among all states and better modeled with full covariances. [2] presented a complete and explicit formula for efficient optimization of SPAM basis and coefficients, and showed that SPAM gave better results than diagonal models and close to performance of smoothed full covariance models on a 80-hour large vocabulary training data.

Yet another approach to creating compact full covariance representations is to share covariances across several gaussians. Parameter estimation for tied full covariance models is described in [26].

With limited training data, over-training would be a concern for full covariance models. In [4] a diagonal covariance smoothing prior was used to smooth off-diagonal elements in the full covariances as suggested in [2]. An effective analytic approach was presented to estimate the shrinkage weight parameter directly from the data and this was shown to have better results than the diagonal models. In [14] sparse inverse covariance matrices were used to address the limited data problem. When the training data is only a couple of hours then sparse inverse covariances have better results than the full covariances. However, the focus of our work is on really large amount of training data and on comparing the full covariance model with state of the art diagonal covariance models containing hundreds of thousands Gaussians. This is in contrast with most of the full-covariance work that we are aware of where training corpora and models of modest sizes are used.

5. CONCLUSIONS AND DISCUSSIONS

In this paper we present a successful recipe of discriminatively training a subspace constrained Gaussian mixture model that retains the performance of discriminative trained full-covariance model, while reducing the size of the full-covariance model to the size of the diagonal-covariance model. We show that even with the large training data and the large size of the diagonal-covariance model, this recipe brings about 5% relative improvement over the large diagonal model. We have not looked into comparing the computational cost of SCGMM with that of the diagonal model.

6. REFERENCES

[1] S. Axelrod, R. Gopinath, P. Olsen, and K. Visweswariah, "Dimensional reduction, covariance modeling, and computational complexity in ASR systems," in *Proceedings of ICASSP*, Hong Kong, April 2003, vol. 1, pp. 915–915.

[2] Daniel Povey, "SPAM and full covariance for speech recognition," in *Proceedings of Interspeech*, Pittsburgh, PA, September 2006, pp. 2338–2341.

[3] Peter Bell and Simon King, "A shrinkage estimator for speech recognition with full covariance HMMs," in *Proceedings of Interspeech 2008*, Brisbane, Australia, September 2008, pp. 910–913.

[4] Peter Bell and Simon King, "Diagonal priors for full covariance speech recognition," in *Proceedings IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009, pp. 113–117.

[5] P. Olsen, Goel V., and Rennie S., "Discriminative Training for Full Covariance Models," in *Proceedings of ICASSP*, 2011.

[6] Jeff A. Bilmes, "Factored sparse inverse covariance matrices," in *Proceedings of ICASSP*, Istanbul, Turkey, June 2000, vol. 2, pp. 1009–1012.

[7] S. Axelrod, V. Goel, B. Kingsbury, K. Visweswariah, and R. Gopinath, "Large Vocabulary Conversational Speech Recognition with a Subspace Constraint on Inverse Covariance Matrices," in *Proceedings of Eurospeech*, 2003.

[8] V. Goel, S. Axelrod, R. Gopinath, P. Olsen, and K. Visweswariah, "Discriminative Estimation of Subspace Precision & Mean (SPAM) Models," in *Proceedings of Eurospeech*, 2003.

[9] Peder A. Olsen and Ramesh A. Gopinath, "Modeling inverse covariance matrices by basis expansion," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 1, pp. 37–46, 2004.

[10] Vincent Vanhoucke and Ananth Sankar, "Mixtures of inverse covariances," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 3, pp. 250–264, 2004.

[11] Scott Axelrod, Vaibhava Goel, Ramesh Gopinath, Peder A. Olsen, and Karthik Visweswariah, "Subspace constrained gaussian mixture models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1144–1160, 2005.

[12] Scott Axelrod, Vaibhava Goel, Ramesh Gopinath, Peder A. Olsen, and Karthik Visweswariah, "Discriminative estimation of subspace constrained gaussian mixture models for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 172–189, 2006.

[13] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N.K. Goel, M. Karafiát, A. Rastrow, et al., "Subspace gaussian mixture models for speech recognition," in *ICASSP*, 2010, pp. 4330–4333.

[14] W. Zhang and Fung P., "Low resource speech recognition with automatically learned sparse inverse covariance matrices," in *Proceedings of ICASSP*. IEEE, 2012, pp. 4737–44740.

[15] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proceedings of ICASSP*. IEEE, 2008, pp. 4057–4060.

[16] Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *ICASSP*, 1986, vol. II, pp. 49–52.

[17] Biing-Hwang Juang, Wu Hou, and Chin-Hui Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[18] Vaibhava Goel and William J Byrne, "Minimum Bayes-risk automatic speech recognition," *Computer Speech and Language*, vol. 14, no. 2, pp. 115–135, 2000.

[19] Fei Sha and Lawrence K Saul, "Large margin hidden Markov models for automatic speech recognition," *Advances in neural information processing systems*, vol. 19, pp. 1249, 2007.

[20] Lawrence D. Brown, *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, vol. 9 of *Lecture Notes - Monograph Series*, Institute of Mathematical Statistics, Hayward, California, 1986.

[21] Peder A. Olsen, Karthik Visweswariah, and Ramesh Gopinath, "Initializing subspace constrained gaussian mixture models," in *ICASSP*, Philadelphia, Pennsylvania, March 2005, vol. I, pp. 661–664.

[22] L. Bregman, "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex optimization," *USSR Comput. Math. and Math. Phys.*, vol. 7, pp. 200–217, 1967.

[23] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proceedings of ICASSP*, Philadelphia, Pennsylvania, April 2005, vol. 1, pp. 961–964.

[24] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, et al., "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4334–4337.

[25] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, et al., "The subspace gaussian mixture modela structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[26] Wolfgang Macherey, Ralf Schlüter, and Hermann Ney, "Discriminative training with tied covariance matrices," in *ICSLP*, 2004, vol. I, pp. 681–684.