DISCRIMINATIVE TRAINING FOR FULL COVARIANCE MODELS

Peder A. Olsen, Vaibhava Goel and Steven J. Rennie

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

ABSTRACT

In this paper we revisit discriminative training of full covariance acoustic models for automatic speech recognition. One of the difficult aspects of discriminative training is how to set the constant D that appears in the parameter updates. For diagonal covariance models, this constant D is set based on knowing the smallest value of D, D*, for which the resulting covariances remain positive definite. In this paper we show how to compute D* analytically, and show empirically that knowing this smallest value is important. Our baseline speech recognition models are state of the art broadcast news systems, built using the boosted Maximum Mutual Information criterion and feature space Maximum Mutual Information for feature selection. We show that discriminatively built full covariance models outperform our best diagonal covariance models. Moreover, full covariance models at optimal performance can be obtained by only a few discriminative iterations starting with a diagonal covariance model. The experiments also show that systems utilizing full covariance models are less sensitive to the choice of the number of gaussians.

Index Terms— Full Covariance Modeling, Maximum Mutual Information, Discriminative Training, Quadratic Eigenvalue Problem.

1. INTRODUCTION

A number of researchers have shown that full covariance models can outperform the performance of diagonal covariance for maximum likelihood trained speaker independent systems, [1, 2, 3, 4]. However, few state of the art systems actually use full covariance models. The best full covariance models have a very large number of parameters and are easy to over-train. It has also been observed that diagonal covariance models ([5, 6]) benefit more from techniques such as feature space minimum phone error (fMPE), [7], and discriminative training, [8, 9, 10] than do full covariance models, [1]. Concerning the number of parameters in full covariance models, there are several methods that compactly represent inverse covariances with little loss in performance, [11, 12, 13, 10, 14]. In this paper we address the issue of overtraining discriminatively trained full covariance models. The constant D that appears in the discriminative parameter update controls model smoothing, and is crucial to making discriminative training work. For diagonal covariance models, this constant D is set based on knowing the smallest value of D, D*, for which the resulting updated model has positive definite covariances. Traditionally D is chosen without regard for the real value of D*, either by using the D* from the diagonally constrained model, or by replacing D* by a rough estimate (for example, by doubling the diagonal covariance D* until reaching a positive definite matrix). In this paper we show how D* can be computed analytically by solving a quadratic eigenvalue problem. We show results on a state of the art broadcast news system that makes use of Boosted Maximum Mutual Information (BMMI) for discriminative training and feature space MMI (fMMI) for feature selection. The resulting full covariance models outperform the best diagonal covariance models. This is not always the case as very large diagonal covariance model can match the performance of the best full covariance models. The resulting full covariance models contain many more parameters than the best performing diagonal covariance models. However, we see that the best performance can be reached for a much wider range of full covariance systems than for diagonal covariance systems. It is also noted that the knowledge of the critical value D* helped improve the results.

2. ANATOMY OF A FULL COVARIANCE MODEL

Let the features generated by the front-end of the speech recognizer be denoted by $\mathbf{x} \in \mathbb{R}^d$. A corresponding full covariance model can then be written

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \quad (1)$$

Each Hidden Markov Model (HMM) state s in an acoustic model is normally modeled as a mixture of gaussians $\mathcal{G}_s$ as follows:

$$p(\mathbf{x}|s) = \sum_{g \in \mathcal{G}_s} \pi_g \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g). \quad (2)$$

Maximum Likelihood (ML) estimation of the parameters $\{\pi_g,\ \boldsymbol{\mu}_g,\ \boldsymbol{\Sigma}_g\}$ is typically done by iteratively applying the Expectation Maximization algorithm. Given observations $\{\mathbf{x}_t\}_{t=1}^T$, posterior probabilities $\gamma_g(\mathbf{x}_t)$ are computed using

the forward–backward algorithm, and the model parameters are updated according to the equations

$$\hat{\pi_g} = \frac{1}{T}\sum_{t=1}^{T}\gamma_g(\mathbf{x}_t) \quad (3)$$

$$\hat{\boldsymbol{\mu}}_g = \frac{1}{T\hat{\pi}_g}\sum_{t=1}^{T}\gamma_g(\mathbf{x}_t)\mathbf{x}_t \quad (4)$$

$$\hat{\boldsymbol{\Sigma}}_g = \frac{1}{T\hat{\pi}_g}\sum_{t=1}^{T}\gamma_g(\mathbf{x}_t)(\mathbf{x}_t-\hat{\boldsymbol{\mu}}_g)(\mathbf{x}_t-\hat{\boldsymbol{\mu}}_g)^T. \quad (5)$$

The Expectation Maximization algorithm alternates updating the gaussian parameters and the posteriors. This is the de facto standard for building acoustic models for gaussian mixture models (GMMs) with the maximum likelihood criterion. It has the advantage that the priors are greater than zero and the covariances of the model are positive definite if there is sufficient data. In contrast, discriminative training algorithms need an additional smoothing term to enforce these constraints.

2.1. Exponential Family Formulation

Let us also formulate the full covariance model as an exponential family since this is sometimes more convenient than the canonical formulation.
Let

$$\mathrm{vec}(\mathbf{X}) = \sqrt{2}\left(\begin{array}{ccccccc}\frac{X_{11}}{\sqrt{2}} & X_{12} & \frac{X_{22}}{\sqrt{2}} & X_{13} & \cdots & \frac{X_{dd}}{\sqrt{2}}\end{array}\right)^T. \quad (6)$$

denote the vector of elements formed from the lower triangular part of the symmetric matrix $\mathbf{X}$. This operation satisfies the property that $\mathrm{vec}(\mathbf{A})^T\mathrm{vec}(\mathbf{B}) = \mathrm{trace}(\mathbf{A}^T\mathbf{B})$. We denote the inverse operation by $\mathrm{mat}(\mathrm{vec}(\mathbf{X})) = \mathbf{X}$. With this notation we define the exponential model parameters

$$\mathbf{P} = \boldsymbol{\Sigma}^{-1} \quad (7)$$

$$\mathbf{p} = \mathrm{vec}(\mathbf{P}) \quad (8)$$

$$\boldsymbol{\psi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \quad (9)$$

$$\boldsymbol{\theta} = \left(\begin{array}{c}\boldsymbol{\psi}\\ -\frac{1}{2}\mathbf{p}\end{array}\right). \quad (10)$$

The gaussian probability density function (pdf) can then be written

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{e^{\boldsymbol{\theta}^T\boldsymbol{\phi}(\mathbf{x})}}{Z(\boldsymbol{\theta})}, \qquad \boldsymbol{\phi}(\mathbf{x}) = \left(\begin{array}{c}\mathbf{x}\\ \mathrm{vec}(\mathbf{x}\mathbf{x}^T)\end{array}\right),$$

and $Z(\boldsymbol{\theta})$ is the partition function, which is log convex, and is given by

$$\log Z(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\psi}^T\mathbf{P}^{-1}\boldsymbol{\psi} - \frac{1}{2}\log\det\mathbf{P} + \frac{d}{2}\log(2\pi). \quad (11)$$

The derivative of the log-partition function is

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log Z(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\boldsymbol{\phi}(\mathbf{x})] = \left(\begin{array}{c}\boldsymbol{\mu}\\ \mathrm{vec}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\end{array}\right). \quad (12)$$

3. DISCRIMINATIVE ESTIMATION

In this section we build on [8], [10]. For the two discriminative training criteria Maximum Mutual Information and Minimum Phone Error, there is an auxiliary function Q that guarantees an increase in the objective function for sufficiently large values of the control parameter D. This auxiliary function is of the same form as the ML objective function. Define the auxiliary statistics

$$\mathbf{s}_{\mathrm{num}} = \sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t)\boldsymbol{\phi}(\mathbf{x}_t)$$

$$\mathbf{s}_{\mathrm{den}} = \sum_t \gamma_{\mathrm{den}}(\mathbf{x}_t)\boldsymbol{\phi}(\mathbf{x}_t)$$

where $\gamma_{\mathrm{num}}(\mathbf{x}_t)$ and $\gamma_{\mathrm{den}}(\mathbf{x}_t)$ are posterior counts corresponding to the numerator and denominator HMM state lattices, respectively. Here the numerator lattice corresponds to a recognition against the reference transcript, and the denominator lattice to a recognition against the task grammar or language model. The auxiliary function is given by

$$Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) = (\mathbf{s}_{\mathrm{num}} - \mathbf{s}_{\mathrm{den}} + DE_{\hat{\boldsymbol{\theta}}}[\boldsymbol{\phi}(\mathbf{x})])^T\boldsymbol{\theta} - ((\sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t) - \gamma_{\mathrm{den}}(\mathbf{x}_t)) + D)\log Z(\boldsymbol{\theta}). \quad (13)$$

Here $\hat{\boldsymbol{\theta}}$ are the current parameters, and $\boldsymbol{\theta}$ the new parameters to be determined. The corresponding maximum can be computed by solving the maximum likelihood problem for the normalized statistics

$$\frac{\mathbf{s}_{\mathrm{num}} - \mathbf{s}_{\mathrm{den}} + DE_{\hat{\boldsymbol{\theta}}}[\boldsymbol{\phi}(\mathbf{x})]}{\sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t) - \gamma_{\mathrm{den}}(\mathbf{x}_t) + D}. \quad (14)$$

The resulting statistics must correspond to a positive definite covariance matrix and a positive count. The corresponding statistics for the mean and covariance are:

$$\begin{array}{llll}\mathbf{m}_{\mathrm{num}} = \sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t)\mathbf{x}_t, & \mathbf{S}_{\mathrm{num}} = \sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t)\mathbf{x}_t\mathbf{x}_t^T,\\ \mathbf{m}_{\mathrm{den}} = \sum_t \gamma_{\mathrm{den}}(\mathbf{x}_t)\mathbf{x}_t, & \mathbf{S}_{\mathrm{den}} = \sum_t \gamma_{\mathrm{den}}(\mathbf{x}_t)\mathbf{x}_t\mathbf{x}_t^T,\\ \hat{\boldsymbol{\mu}} = E_{\hat{\boldsymbol{\theta}}}[\mathbf{x}], & \mathbf{S} = E_{\hat{\boldsymbol{\theta}}}[\mathbf{x}\mathbf{x}^T].\end{array}$$

It has been common practice to choose D to be given by $D = \max\{C_1(\sum_t \gamma_{\mathrm{den}}(\mathbf{x}_t)), C_2D^*\}$, where D^* is the smallest value of D for which the covariance is positive definite and C_1 and C_2 are constants. To find this value of D we must effectively solve a quadratic eigenvalue problem. The resulting covariance estimate can be seen to be

$$\boldsymbol{\Sigma} = \frac{\mathbf{S}_{\mathrm{num}} - \mathbf{S}_{\mathrm{den}} + D\mathbf{S}}{(\sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t) - \gamma_{\mathrm{den}}(\mathbf{x}_t)) + D} - \frac{(\mathbf{m}_{\mathrm{num}} - \mathbf{m}_{\mathrm{den}} + D\hat{\boldsymbol{\mu}})(\mathbf{m}_{\mathrm{num}} - \mathbf{m}_{\mathrm{den}} + D\hat{\boldsymbol{\mu}})^T}{((\sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t) - \gamma_{\mathrm{den}}(\mathbf{x}_t)) + D)^2} \quad (15)$$

$$= \frac{\mathbf{A}_0 + \mathbf{A}_1D + \hat{\boldsymbol{\Sigma}}D^2}{((\sum_t \gamma_{\mathrm{num}}(\mathbf{x}_t) - \gamma_{\mathrm{den}}(\mathbf{x}_t)) + D)^2}, \quad (16)$$

where the matrices $\mathbf{A}_0$ and $\mathbf{A}_1$ are given by

$$\begin{aligned}\mathbf{A}_0 &= \mathbf{S}_{\text{num}} - \mathbf{S}_{\text{den}} + (\sum_t \gamma_{\text{num}}(\mathbf{x}_t) - \gamma_{\text{den}}(\mathbf{x}_t))\mathbf{S} \\ &\quad -(\mathbf{m}_{\text{num}} - \mathbf{m}_{\text{den}})\hat{\boldsymbol{\mu}}^T - \hat{\boldsymbol{\mu}}(\mathbf{m}_{\text{num}} - \mathbf{m}_{\text{den}})^T \\ \mathbf{A}_1 &= (\mathbf{S}_{\text{num}} - \mathbf{S}_{\text{den}})(\sum_t \gamma_{\text{num}}(\mathbf{x}_t) - \gamma_{\text{den}}(\mathbf{x}_t)) \\ &\quad -(\mathbf{m}_{\text{num}} - \mathbf{m}_{\text{den}})(\mathbf{m}_{\text{num}} - \mathbf{m}_{\text{den}})^T.\end{aligned}$$

3.1. The Quadratic Eigenvalue Problem

For the covariance to be positive definite we need the matrix $\mathbf{X}(D) = \mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2$ to be positive definite. Let $\{e_i(D)\}_{i=1}^d$ be the eigenvalues of $\mathbf{X}(D)$, then

$$\det(\mathbf{X}(D)) = \prod_{i=1}^{d} e_i(D) = \det(\mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2). \quad (17)$$

$\mathbf{X}(D)$ is positive definite if all its eigenvalues are positive. Since $\mathbf{X}(D)$ is a continuous function of D, and $\hat{\boldsymbol{\Sigma}}$ is assumed to be positive definite, it follows that $\mathbf{X}(D) = D^2(\hat{\boldsymbol{\Sigma}} + \mathbf{A}_1/D + \mathbf{A}_0/D^2) = D^2(\hat{\boldsymbol{\Sigma}} + \mathcal{O}(1/D))$ is positive definite for sufficiently large values of D. In the interior of the region of positive definite matrices it follows that $\det(\mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2) > 0$. At the boundary at least one of the eigenvalues of $\mathbf{X}(D)$ becomes zero, and thus we have $\det(\mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2) = 0$. Let D^* be the largest solution to the equation $\det(\mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2) = 0$. By continuity, it follows that $\det(\mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2) > 0$ for all $D > D^*$. If $\det(\mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2) = 0$ for some value $D = D_j$, then D_j is a *quadratic eigenvalue* corresponding to the quadratic eigenvalue problem

$$\mathbf{A}_0\mathbf{y} + D\mathbf{A}_1\mathbf{y} + D^2\hat{\boldsymbol{\Sigma}}\mathbf{y} = 0, \quad (18)$$

where $\mathbf{y}$ is a *quadratic eigenvector*. Since $\det(\mathbf{A}_0 + \mathbf{A}_1 D + \hat{\boldsymbol{\Sigma}} D^2)$ is a polynomial of degree $2d$ in D, there will be a total of $2d$ quadratic eigenvalues D_j. The matrix $\mathbf{X}(D)$ turns from positive semi-definite to strictly positive definite for D greater than the largest quadratic eigenvalue $D^* = \max\{D_j \mid \text{Im}(D_j) = 0, j = 1, \ldots, 2d\}$.

We can solve the quadratic eigenvalue problem by introducing the auxiliary eigenvector $\mathbf{z} = \lambda\mathbf{y}$ and instead solve the linear eigensystem

$$\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{A}_0 & -\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{A}_1 \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}. \quad (19)$$

The largest positive real eigenvalue for the linear eigenvalue problem gives us the critical value D^*.

3.2. I–smoothing

A common technique to mitigate the effects of overtraining is to impose a Bayesian prior on the parameters of the model. The Bayesian technique known as *I–smoothing*, [8], uses the KL-divergence of the model from the previous EM iteration as a penalty term.

$$\begin{aligned}&\tau D(\mathcal{N}(x; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \| \mathcal{N}(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ &= \tau \boldsymbol{\theta}^T E_{\hat{\boldsymbol{\theta}}}[\boldsymbol{\phi}(\mathbf{x})] - \tau \log Z(\boldsymbol{\theta}) + K(\hat{\boldsymbol{\theta}}),\end{aligned}$$

where the constant $K(\hat{\boldsymbol{\theta}})$ does not depend on $\boldsymbol{\theta}$. Adding this penalty to the auxiliary function simply increases the value of D by τ. We therefore consider choosing D by $D = \tau + \max\{C_1(\sum_t \gamma_{\text{den}}(\mathbf{x}_t)), C_2 D^*\}$ in the auxiliary function (13). Note that although the values $\sum_t \gamma_{\text{den}}(\mathbf{x}_t)$ and D^* are gaussian dependent, the constants τ, C_1 and C_2 are shared among all gaussians. When training full covariance models τ, C_1 and C_2 are the parameters that we need to tune to get the best performance.

4. EXPERIMENTS

We evaluated the discriminative full covariance modeling on a Broadcast News Large Vocabulary Speech Recognition (LVCSR) task. The acoustic model training set comprises 50 hours of data from the 1996 and 1997 English Broadcast News Speech corpora (LDC97S44 and LDC98S71), and was created by selecting entire shows at random. The EARS Dev-04f set (dev04f), a collection of 3 hours of audio from 6 shows from November 2003, is used for testing the models.

The acoustic features are obtained by first computing 13-dimensional perceptual linear prediction (PLP) features with speaker-based mean, variance, and vocal tract length normalization. Nine such features were concatenated and projected to a 40 dimensional space using Linear Discriminant Analysis (LDA). An fMMI transform [9] was estimated to arrive at the final feature space in which acoustic models were trained. The acoustic models consist of 44 phonemes with each phoneme modeled as three-state, left-to-right HMMs with no skip states. Mixtures of exponential distributions are used to model each state, with the overall model having 50K components.

The baseline (fMMI only) acoustic models were built using first maximum likelihood training and then the boosted MMI [9] estimation process. These models had a word error rate of 19.4% on the dev04f test set.

4.1. Choosing D

To determine what values of C_1, C_2 and τ are suitable we ran a number of experiments starting with the baseline system. These numbers can be seen in Table 1. Notice that the best results were obtained by taking C_2 close to 1. The results on the first line of Table 1, with $C_1 = C_2 = \tau = \infty$ correspond to a diagonal covariance model with a total of 50,000 gaussians. This diagonal covariance model is the initial model

used for all the full covariance builds. Interestingly, the lowest word error rate is obtained for $C_2 = 1.06$ for full covariance models, thus underlining the importance of exactly knowing D^*. This is surprising as a much larger value of C_2 ($C_2 \approx 2$) is known to be best for discriminative training of diagonal covariance models. The finding is consistent with the values of C_2 found for subspace mean and precision models (SPAM), [10]. We retrained diagonal acoustic models of different sizes along with corresponding full covariance models as well. These can be seen in Table 2. We did not train the 200K model with full covariances, as it is computationally costly, and did not seem likely to change our assessment. It can be seen in the table that the 200K diagonal model performs on par with a 20K diagonal covariance model. These two models have roughly the same number of parameters, so there appears to be little advantage to training diagonal covariance models in terms of number of parameters. The best discriminatively trained full covariance models in Table 2 were all trained using multiple iterations (usually 2 or 3 iterations), and with different parameter settings for C_1, C_2 and τ.

C_1	C_2	τ	WER	NER
∞	∞	∞	19.4%	4387
2	2	500	18.8%	4262
2	1.5	500	18.7%	4237
2	2	250	18.7%	4220
1	1.5	250	18.6%	4196
0.5	1.5	250	18.6%	4202
0.25	1.5	250	18.6%	4201
1	1.25	250	18.5%	4193
1	1.12	250	18.5%	4182
1	**1.06**	**250**	**18.5%**	**4174**
1	1.03	250	18.5%	4181

Table 1. Word error rates (WERs) and number of errors (NER) for one iteration of discriminative training with the BMMI criterion.

nGauss	Diagonal Covariance		Full Covariance	
	WER	NER	WER	NER
10K	20.5%	4627	19.4%	4379
20K	19.6%	4433	18.9%	4281
30K	19.3%	4357	18.5%	4192
40K	19.0%	4290	18.5%	4195
50K	19.1%	4311	18.4%	4173
100K	18.8%	4245	**18.4%**	**4164**
150K	**18.6%**	**4211**	18.5%	4180
200K	18.9%	4277		

Table 2. Diagonal covariance and full covariance models of different sizes.

5. CONCLUSION

We have shown in this paper how to determine the "magic constant" D^* for full covariance models. We have demonstrated that we can beat state of the art diagonal covariance systems with discriminatively trained full covariance models. Full covariance models ranging from 30,000 to 150,000 gaussians all get within 0.1% of the best word error rate, all better than the best diagonal system.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] Ramesh A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proceedings of ICASSP*, Seattle, Washington, May 1998, vol. II, pp. 661–664.

[6] Mark J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[7] 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.

[8] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, Cambridge University, 2003.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.