Diagonal Priors for Full Covariance Speech Recognition

Peter Bell [1], Simon King [2]

Centre for Speech Technology Research, University of Edinburgh
Informatics Forum, 10 Crichton St, Edinburgh, EH8 9AB, UK
[1] Peter.Bell@ed.ac.uk
[2] Simon.King@ed.ac.uk

***Abstract*—We investigate the use of full covariance Gaussians for large-vocabulary speech recognition. The large number of parameters gives high modelling power, but when training data is limited, the standard sample covariance matrix is often poorly conditioned, and has high variance. We explain how these problems may be solved by the use of a diagonal covariance smoothing prior, and relate this to the shrinkage estimator, for which the optimal shrinkage parameter may itself be estimated from the training data. We also compare the use of generatively and discriminatively trained priors. Results are presented on a large vocabulary conversational telephone speech recognition task.**

I. Introduction

HMM-based systems for automatic speech recognition (ASR) typically model the acoustic features using mixtures of multivariate Gaussians (GMMs). Whilst the Gaussians are most commonly restricted to have diagonal covariance matrices, it has been shown [1], [2], [3] that increasing the number of covariance parameters generally improves recognition performance, above the improvement that may be achieved simply by increasing the number of Gaussians. Various schemes have been proposed for increasing the number of covariance parameters per Gaussian, which may vary between d, in the diagonal case, and $\frac{1}{2}d(d+1)$ in the full covariance case (where d is the size of the acoustic feature vector). These schemes most commonly control the number of free parameters by representing the inverse covariance matrices as a linear combination of some set of basis elements: examples of such schemes include Semi-Tied Covariance Matrices (STC) [4], Extended Maximum Likelihood Linear Transforms [5] and Subspace for Precision and Mean (SPAM) [1] – a review can be found in [6].

In this paper we investigate the use of full covariance models, where each Gaussian has the maximum number of free covariance parameters. In a large vocabulary recognition system, such models present some practical issues with parameter storage, and the cost of evaluating the Gaussian likelihoods, but these requirements do not pose any hard limits on the models that may be used. However, the use of full covariance matrices requires that two major problems be addressed: firstly, the covariance matrices must be well-conditioned, in order to avoid amplifying numerical errors when evaluating the Gaussian log likelihoods; secondly, models trained on limited training data must generalise well to unseen test data, despite the very large number of free parameters.

In [3], Povey demonstrated that full covariance models can outperform SPAM models, but noted that it is essential to smooth the off-diagonal elements. In this paper we explore this technique further in the context of the two problems outlined above, and relate it to our previous investigation of the "shrinkage estimator" [7]. We consider the choice of prior, and go on to discuss the estimation of the optimal shrinkage parameter. We present results on a conversational telephone speech task.

II. Full covariance estimation

A. Background

In what follows, we generally take a generative approach and assume that each Gaussian m has true, fixed parameters, a mean μ_m and covariance Σ_m, that could, in principle, be perfectly inferred were there are infinite training data available. (We will discuss some problems with this approach in a later section).

Given training data observations $\mathbf{x}(t)$ and weights $\gamma_m(t)$, the sample covariance matrix is computed as:

$$S_m = \frac{\sum_t \gamma_m(t)(x(t)-\hat{\mu}_m)(x(t)-\hat{\mu}_m)^T}{\beta_m} \tag{1}$$

where $\beta_m = \sum_t \gamma_m(t)$, a measure of the total amount of data available for the Gaussian m. ($\hat{\mu}_m$ is the sample estimate of the mean). In the context of the EM algorithm, the weights $\gamma_m(t)$ are set to the posterior probabilities of the observations $x(t)$, given some previous parameter set, and using S_m as the estimate of the covariance matrix that maximises a lower bound on the log likelihood of the training data. S_m is guaranteed to be positive semidefinite.

In [8], Povey suggested smoothing the off-diagonal elements of the covariance matrix and, in [3], showed that this results in significant performance gains over the unsmoothed matrix. His suggestion was to reduce the off-diagonal elements by a factor $\frac{\tau}{\tau+\beta_m}$, where τ is a smoothing constant, set to an empirically-determined value of 100. Defining D_m to be a diagonal matrix consisting of the diagonal elements of S_m, and writing the smoothed matrix as U_m, we see trivially that

$$U_m = \frac{\beta_m}{\tau + \beta_m} S_m + \frac{\tau}{\tau + \beta_m} D_m \tag{2}$$

$$\equiv (1 - \lambda) S_m + \lambda D_m \tag{3}$$

Equation (2) formulates the smoothed matrix as a maximum a-posteriori (MAP) estimate, where the prior is chosen to be the diagonal matrix with prior variance τ. Equation (3) reformulates the matrix as a shrinkage estimator [9] with shrinkage parameter λ. Both formulations yield benefits in the analysis of the smoothed matrix. In the following sections we consider the properties of the model U_m with reference to three criteria: matrix conditioning, generalisation and discriminative power.

B. Matrix conditioning

Since each S_m is positive semidefinite, we may define the condition number to be the ratio of the ratio of the largest and smallest eigenvalues:

$$\kappa(S_m) = \frac{\lambda_{max}(S_m)}{\lambda_{min}(S_m)} \tag{4}$$

The amplification of errors when inverting the matrix – required for the log-likelihood computation during decoding – is directly proportional to the condition number, and a matrix is said to be well-conditioned when κ is small. In the extreme case when the number of observations for which $\gamma_m(t)$ is non-zero is less than d, S_m is guaranteed to be non-invertible, and the condition number is infinite. In ASR systems, d is typically 39, and could even be 52 – so this is a practical consideration in systems with relatively large numbers of Gaussians and small amounts of data.

Moreover, it is shown in [9] that even when $n > d$, S_m will be, on average, less well-conditioned than the true matrix Σ_m. This follows from the mathematical result that the eigenvalues of any symmetric matrix are the most dispersed diagonal elements that can be obtained by rotation: that is to say, the diagonal elements of the matrix $R^T SR$, for any rotation R, are maximally dispersed when R is the matrix of eigenvectors. The eigenvectors of S_m and Σ_m are not equal, in general, despite the fact that $\mathrm{E}(S_m) = \Sigma_m$. Importantly, the result also shows that the eigenvalues of the diagonal matrix, D_m (which are of course, just the diagonal elements themselves) must be less dispersed than the eigenvectors of S_m, and of Σ. It follows that the smoothed U_m is always better-conditioned than S_m.

C. Generalisation

In the context of statistical learning, *generalisation* refers to the ability of a model whose parameters are estimated from a finite set of training data to perform well when applied to unseen test data. Considering the training data to be randomly sampled from the true underlying distribution, we can view the estimated covariance matrix U_m as a random variable. Since the test data are sampled from the same distribution, and assuming again the existence of a true matrix Σ_m, a covariance model U_m will generalise well if, on average, it is close to Σ_m: we therefore seek to minimise the convex error function

$$\mathrm{E}\|U_m - \Sigma_m\|^2 = \mathrm{E}\|U_m - \mathrm{E}(U_m)\|^2 + \|\mathrm{E}(U_m) - \Sigma_m\|^2 \tag{5}$$

where the expectation denotes the fact that U_m is a function of the random training data. For the matrix norm, we use the Frobenius norm, given by

$$\|A\|_F = \sqrt{trA^T A} = (\sum_i \sum_j |A_{ij}|^2)^{\frac{1}{2}} \tag{6}$$

which arises from the inner product $\langle A, B \rangle = \mathrm{tr}\, A^T B$.

Equation (5) decomposes the error function into two terms: a variance term, and bias term respectively. This illustrates the trade-off between a model that is too complex, which will have a high variance (ie. it over-fits to the observed training data), and a model that is too simple, which produces estimates that, on average, deviate from the true parameter. Importantly, the error function is strictly convex with respect to the parameter λ from (3): by using a linear combination of S_m and D_m, the expected error is reduced below a simple weighting of errors from the two estimators.

D. Discrimination

The optimality properties of generative modelling for classification depend upon the assumption of model correctness, which does not hold in practice for the HMM-GMM. Training model parameters according to the explicitly discriminative MMI criterion instead has been shown to yield performance improvements [10]. The MMI criterion can be viewed as the sum, over all training utterances, of the margin between the model-based log likelihoods of the correct transcription and the closest competing transcription. In the case of infinite training data, it has been shown [11] that the criterion provides an upper bound on the model-free expected error rate. In the case of finite training data, however, the resulting estimators may still have high variance.

The MMI criterion motivates a simple modification to the smoothing technique. In the full covariance case, we would expect MMI training to yield only small improvements over ML training due to the very large number of parameters (diminishing the importance of model correctness), whilst continuing to suffer from high variance: in the diagonal case, we would expect the converse. We therefore propose replacing the prior D_m with a discriminatively-trained equivalent, whilst using the standard full sample covariance matrix as before.

III. The Shrinkage Parameter

A. Optimisation of the shrinkage parameter

An interesting consideration is the method for choosing the optimal prior weight τ, or shrinkage constant λ. Whilst this could, in practice, be chosen heuristically, with reference to some held-back development data, it is worthwhile to investigate whether a suitable constant may in fact, be obtained analytically. As in our previous work [7], we adopt the approach of [9], generalised in [12]. In what follows, we omit the dependence on m for clarity. Consider the formulation in (3). Using (5), we minimise

$$\mathrm{E}\|U-\Sigma\|^2 = \mathrm{E}\|\lambda(D-\Sigma)+(1-\lambda)(S-\Sigma)\|^2 \tag{7}$$

$$\begin{aligned} &= \lambda^2\mathrm{E}\|D-\Sigma\|^2 + (1-\lambda)^2\mathrm{E}\|S-\Sigma\|^2 \\ &\quad + 2\lambda(1-\lambda)\mathrm{E}\langle D-\Sigma, S-\Sigma\rangle \end{aligned} \tag{8}$$

Differentiating with respect to λ and setting the result equal to zero, we obtain

$$\begin{aligned} &\mathrm{E}\|S-\Sigma\|^2 - \mathrm{E}\langle D-\Sigma, S-\Sigma\rangle \\ &= \lambda[\mathrm{E}\|D-\Sigma\|^2 + \mathrm{E}\|S-\Sigma\|^2 - 2\mathrm{E}\langle D-\Sigma, S-\Sigma\rangle] \end{aligned} \tag{9}$$

$$= \lambda\mathrm{E}\|(S-\Sigma)-(D-\Sigma)\|^2 \tag{10}$$

We decompose Σ into its diagonal and off-diagonal elements: $\Sigma = \Sigma^{\mathsf{diag}} + \Sigma^{\mathsf{od}}$. Since $\mathrm{E}S = \Sigma$, $\mathrm{E}\langle\Sigma^{\mathsf{od}}, S-\Sigma\rangle = 0$, and we add this to the second term on the left-hand side, giving

$$\mathrm{E}\langle D-\Sigma^{\mathsf{diag}}, S-\Sigma\rangle = \mathrm{E}\|D-\Sigma^{\mathsf{diag}}\|^2 \tag{11}$$

since the off-diagonal terms then vanish from the inner product. We therefore obtain

$$\lambda = \frac{\mathrm{E}\|S-\Sigma\|^2 - \mathrm{E}\|D-\Sigma^{\mathsf{diag}}\|^2}{\mathrm{E}\|S-D\|^2} \tag{12}$$

When D consists simply of the diagonal elements of S, then the numerator in (12) becomes

$$\sum_{i\neq j}\mathrm{E}(S_{ij}-\Sigma_{ij})^2 \tag{13}$$

whilst the denominator becomes

$$\sum_{i\neq j}\mathrm{E}S_{ij}^2 \tag{14}$$

As presented, the calculations are not invariant to arbitrary scaling of feature dimensions. To remedy this we adopt the approach of [12], dividing each element S_{ij} by $\sqrt{S_{ii}S_{jj}}$. (The diagonal elements themselves are not changed by the smoothing process).

B. Estimating the parameter from data

The numerator and denominator terms above are unknown, but may be estimated from data. In this analysis, we fix the number and weighting of observations for each Gaussian (ie. the $\gamma(t)$ and β), but assume that the actual observations vary randomly according to the true distribution.

We define

$$w_{ij}(t) = (x_i(t)-\hat{\mu}_i)(x_j(t)-\hat{\mu}_j) \tag{15}$$

The sample covariance matrix S is the sample mean of these observations:

$$S_{ij} = \frac{\sum_t \gamma(t)w_{ij}(t)}{\beta} \tag{16}$$

The (i,j)th term of the numerator can be estimated by

$$\frac{\sum_t \gamma(t)^2}{\beta^2}.\frac{1}{\beta}\sum_t \gamma(t)(w_{ij}-S_{ij})^2 \tag{17}$$

$$= \frac{\sum_t \gamma(t)^2}{\beta^2}\left[\frac{\sum_t \gamma(t)w_{ij}^2}{\beta} - S_{ij}^2\right] := \frac{\delta}{\beta}\alpha_{ij} \tag{18}$$

where $\alpha_{ij} = \frac{\sum_t \gamma(t)w_{ij}^2}{\beta} - S_{ij}^2$ and $\delta = \frac{\sum_t \gamma(t)^2}{\beta}$ are estimated constants that we would expect to be independent of the sample count β. The numerator (13) is

$$\frac{\delta}{\beta}\alpha := \frac{\delta}{\beta}\sum_{i\neq j}\alpha_{ij} \tag{19}$$

δ can be seen as a correction term to allow for the increased variance when samples from nearby Gaussians "overlap" in feature space.

We now consider the estimation of the $\mathrm{E}S_{ij}^2$ terms in the denominator. [12] suggest simply replacing the expectation by the sample values S_{ij}. We have, however, observed that this leads to considerable error for small β. Decomposing

$$\mathrm{E}S_{ij}^2 = (\mathrm{E}S_{ij})^2 + \operatorname{var} S_{ij} \tag{20}$$

we see that the expression consists of a expectation term which we would expect to be constant with β and a variance which reduces with $\frac{1}{\beta}$. S_{ij} has a Wishart distribution, and we observe that the total variance can be approximated by

$$\sum_{i\neq j}\operatorname{var} S_{ij} \approx \frac{2\delta\alpha}{\beta} \tag{21}$$

We can then estimate a third constant representing the bias term,

$$C = \sum_{i\neq j} S_{ij}^2 - \frac{2\delta\alpha}{\beta} \tag{22}$$

The shrinkage parameter is then given by

$$\lambda = \frac{\alpha\delta/\beta}{C + 2\alpha\delta/\beta} \tag{23}$$

$$= \frac{\alpha\delta/C}{\beta + 2\alpha\delta/C} \tag{24}$$

Comparing to (2) we see that this is similar to using a prior with weight $\alpha\delta/C$, except the weighting is doubled in the denominator. This suggests that in the limit as the quantity of training data is reduced towards zero, the off-diagonal elements are reduced by half, rather than vanishing to zero. α, δ and C are all independent of the number of samples per Gaussian, and since all matrices are scale free, it is possible to pool the constants across Gaussians. In the results presented here, we average the estimates of α and C, but compute δ for each Gaussian. (For interest, we found $\alpha = 740$, $C = 3.1$ and a mean $\delta = 0.75$, giving an average shrinkage parameter $\lambda = 0.23$).

C. Practical considerations

We briefly discuss the practical issues when estimating the shrinkage parameter from data. The estimation is computationally inexpensive since the computational cost is dominated by the computing of the $\gamma(t)$ during the E-step of the EM algorithm, which is required anyway for estimation of the other parameters. Another issue is the storage of the statistics: computing δ for each Gaussian requires the sums of w_{ij}^2 to be stored, which could potentially require $O(d^2)$ memory, equivalent to storing an additional covariance matrix. However, this can be avoided by the summing over i and j on the fly and subtracting the S_{ij} terms after all the statistics have been accumulated.

IV. Experiments

A. Setup

We performed large-vocabulary speech recognition experiments with full covariance models on the NIST Hub 5 *Eval01* data, comprising around 6 hours of conversational telephone speech from 60 male and 60 females speakers. Our system was loosely based on the 2005 AMI recogniser [13].

Cross-word triphone acoustic models were trained on 277 hours of speech from the Switchboard-1, Switchboard-2 and Call Home corpora. The acoustic feature vector contained 12 PLP plus energy coefficients, their delta and double deltas, with CMN and CVN applied on a per-utterance basis. The baseline system consisted of approximately 120,000 diagonal covariance Gaussians. The feature vector was extended to include third-differential coefficients, and a single HLDA projection was applied to reduce the dimensionality to 39. For decoding, HTK's HDecode tool was used with a bigram language model to generate lattices for the test utterances. These were then rescored with a trigram language model to produce a one-best transcription.

Using these transcriptions, speaker adaptation was performed using CMLLR with 32 regression classes per speaker. We did not apply VTLN, though we would expect it to give further improvements on the results shown here. Table I shows the results from the baseline systems.

As in [7], we initialised the full covariance models directly from the final set of diagonal-covariance Gaussians. We found that the estimation of the full-covariance models was quick to converge, so the models used for the results presented below were ML-trained using just one iteration with full covariance, keeping the Gaussian means fixed. To reduce the computational cost of decoding with the full covariance models, we instead used lattice rescoring of the baseline bigram lattices, again applying a trigram language model to obtain the final transcription.

The question of speaker adaptation of full covariance models has been considered in [14]. For speaker adaptation of a full-covariance system, CMLLR has the advantage that it can be formulated as feature-space transform rather than a model-space transform, so it is not necessary to recompute full covariance matrices. We used the diagonal-covariance approximation suggested in [14] to obtain the transforms, but found that results were little improved over simply using the original CMLLR transforms obtained for the diagonal models, and we present results using the latter.

B. Experiments with a diagonal prior

We investigated the effects of off-diagonal smoothing on the full covariance recognition performance. As discussed in Section II-B, a sample covariance matrix based on fewer than d samples will be non-invertible when no off-diagonal smoothing is applied, preventing it being used for likelihood computation. The simplest way of avoiding this is to use the full covariance matrix when the number of samples is sufficient, and back off to the diagonal matrix otherwise. We call this the "naive" full covariance system. We obtained results using a range of values of the prior parameter τ, and also with analytically obtained shrinkage parameters: selected results are shown in Table II, and graphically in Figure 1, using the convention $\tau = 0$ for the naive system. All results used CMLLR speaker adaptation – we show unadapted results with $\tau = 100$ for comparison.

TABLE I
Diagonal covariance WER results on Hub5 Eval01 with bigram and trigram language models

System	Bigram	Trigram
Baseline	40.3%	37.2%
HLDA	38.5%	35.5%
HLDA + CMLLR	35.6%	33.3%

TABLE II
Selected WER results with full covariance models, using a trigram LM

System	WER
Diagonal	33.3%
Naive fullcov	32.1%
$\tau = 10$	31.3%
$\tau = 20$	31.0%
$\tau = 40$	30.7%
$\tau = 100$	30.5%
$\tau = 200$	30.8%
$\tau = 400$	31.3%
Shrinkage	30.6%
Unadapted, $\tau = 100$	31.8%

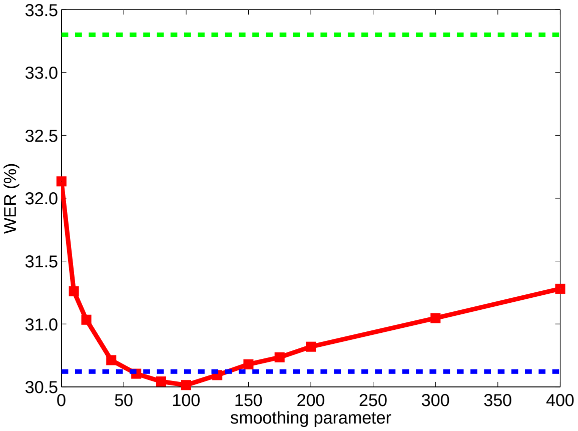

Fig. 1. WER of ML-trained full covariance models, with varying smoothing parameter τ (red) compared with diagonal models (dashed green) and shrinkage estimate (dashed blue)

C. Experiments with a discriminatively-trained prior

In the second set of experiments we trained the full covariance models with ML, but used a discriminatively-trained diagonal covariance prior. The priors were initialised with the standard diagonal covariance Gaussians, with both mean and variance parameters re-estimated to maximise the MMI criterion, using four iterations of the EBW algorithm [8]. Speaker adaptation was performed non-discriminatively using CMLLR as before. Full covariance models were initialised from the diagonal-MMI system, and a single EM iteration was used to update the covariance parameters. The mean parameters were again kept fixed. As above, we investigated the effect of varying the prior weight, using MMI-trained diagonal matrices as a prior. For comparison, results are shown using the ML-trained priors initialised from the same diagonal-MMI system. We compare the results with those using the analytically-obtained shrinkage parameters. Selected results are shown in Table III, and graphically in Figure 2.

TABLE III
SELECTED WER RESULTS WITH FULL COVARIANCE MODELS, INITIALISED FROM MMI-TRAINED DIAGONAL SYSTEM, USING A TRIGRAM LM

System	ML prior	MMI prior
Diagonal	-	31.2%
Naive fullcov	31.5%	31.1%
$\tau = 10$	30.7%	30.7%
$\tau = 20$	30.5%	30.5%
$\tau = 40$	30.3%	30.2%
$\tau = 100$	30.3%	30.1%
$\tau = 200$	30.6%	30.1%
$\tau = 400$	30.9%	30.2%
Shrinkage	30.4%	30.1%

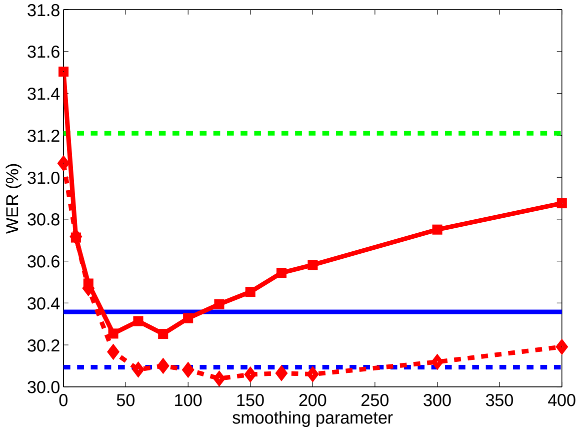

Fig. 2. WER of ML-trained full covariance models initialised from MMI-trained diagonal models, with varying smoothing parameter τ: using an ML prior (solid red) and an MMI prior (dashed red); compared with diagonal MMI-trained models (dashed green) and shrinkage estimates with ML and MMI priors (solid and dashed blue respectively)

V. CONCLUSION

The results demonstrate that off-diagonal smoothing is essential for good performance with full covariance models: the reduction in WER over diagonal models is more than doubled, compared to the naive full covariance systems, when the optimal prior weight is used. The analytic method for obtaining a shrinkage parameter – which can be viewed as a form of prior weight – directly from the data was shown to be effective, achieving close to the best performance obtained by tuning τ on the test set. Using an MMI-trained diagonal prior was also shown to be effective, yielding performance gains over both an MMI-trained diagonal system, and an ML-trained full covariance system with conventional smoothing.

It appears that we can make good use of the additional statistics of the training data in computing the shrinkage parameter. However, our analytically obtained solution uses a generative model: we would expect the effectiveness to be increased if the parameter itself were optimised with respect to some explicitly discriminative criterion. Furthermore, the performance could perhaps be improved by computing several different values of the constants α and C, but it is not yet clear how they would be best tied across Gaussians. We will address these questions in future work.

ACKNOWLEDGEMENT

We would like to thank members of the AMI Project, and Giulia Garau in particular, for their help in the construction of our baseline system.

REFERENCES

[1] S. Axelrod, V. Goel, R. A. Gopinath, P. A. Olsen, and K. Visweswariah, "Subspace constrained Gaussian mixture models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1144–1160, Nov. 2005.

[2] M. Varjokallio and M. Korimo, "Comparison of subspace methods for Gaussian mixture models in speech recognition," in *Proc. Interspeech*, 2007.

[3] D. Povey, "Spam and full covariance for speech recognition," in *Proc. ICSLP*, 2006.

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

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

[6] K. Sim and M.J.F.Gales, "Precision matrix modelling for large vocabulary continuous speech recognition," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.485, Jun. 2004.

[7] P. Bell and S. King, "A shrinkage estimator for speech recognition with full covariance HMMs," in *Proc. Interspeech*, 2008.

[8] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University Engineering Department, 2003.

[9] O. Ledoit and M. Wolf, "A well-conditioned estimator for large covariance matrices," *Journal of Multivariate Analysis*, vol. 88, pp. 365–411, 2004.

[10] V. Valtchev, J. Odell, P. Woodland, and S. Young, "Mmie training of large vocabulary recognition systems," *Speech Communication*, vol. 16, no. 4, pp. 303–314, 1997.

[11] R. Schlüter and H. Ney, "Model-based mce bound to the true Bayes' error," *IEEE Signal Processing Letters*, vol. 8, no. 5, 2001.

[12] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[13] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation*, 2005.

[14] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance gaussians," in *Proc. ICSLP*, 2006.