ADAPTATION OF PRECISION MATRIX MODELS ON LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

K.C. Sim and M.J.F. Gales

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {kcs23,mjfg}@eng.cam.ac.uk

ABSTRACT

Recently, structured precision matrix models were found to outperform the conventional diagonal covariance matrix models. Minimum phone error discriminative training of these models gave very good unadapted performance on large vocabulary continuous speech recognition systems. To obtain state-of-the-art performance, it is important to apply adaptation techniques efficiently to these models. In this paper, simple row-by-row iterative formulae are described for both MLLR mean and constrained MLLR transform estimations of these models. These update formulae are derived within the standard expectation maximisation framework and are guaranteed to increase the likelihood of the adaptation data. Efficient approximate schemes for these adaptation methods are also investigated to further reduce the computation. Experimental results are presented based on the MPE trained Subspace for Precision and Mean models, evaluated on both broadcast news and conversational telephone speech English tasks.

1. INTRODUCTION

The Hidden Markov Model (HMM) is the most popular acoustic model for continuous speech recognition tasks. The output probability distribution associated to each state is typically represented by a multivariate Gaussian Mixture Model (GMM). An issue when using multivariate Gaussians is how to model correlation efficiently in the feature in that increasing the feature dimension dramatically increases the number of model parameters and computational cost (curse of dimensionality). Typically, a diagonal covariance matrix approximation is employed to circumvent this problem. Gaussian components are used to model the correlation implicitly. Recently, efficient forms of explicit correlation modelling have been achieved using structured precision matrix approximations. For examples, the Semi-tied Covariance (STC) [1], Extended Maximum Likelihood Linear Transform (EMLLT) [2] and Subspace for Precision and Mean (SPAM) [3] models have been found to yield good performance gains on LVCSR systems using both Maximum Likelihood (ML) [4, 5, 6] and Minimum Phone Error (MPE) [6] training criteria. In general, these structured precision matrix models can be expressed in a general form of basis superposition:

$$\boldsymbol{P}_m = \sum_{i=1}^{n} \lambda_{ii}^{(m)} \boldsymbol{S}_i = \sum_{i=1}^{n} \lambda_{ii}^{(m)} \sum_{r=1}^{R} \lambda_{rr} \boldsymbol{a}_{ir}' \boldsymbol{a}_{ir} \qquad (1)$$

where $\boldsymbol{S}_i$ is the ith basis matrix and $\lambda_{ii}^{(m)}$ is the corresponding basis coefficient. $\boldsymbol{S}_i$ is a symmetric matrix with an arbitrary rank, R, which can be further decomposed into a superposition of R basis vectors, $\boldsymbol{a}_{ir}$. $\boldsymbol{P}_m$ is constrained to be positive-definite. If, $\boldsymbol{S}_i$ is rank-1 ($R = 1$), equation 1 becomes a STC model when $n = d$ and an EMLLT model when $d < n \leq \frac{d}{2}(d+1)$. Removing the rank-1 constraint gives the SPAM model (with unconstrained mean), which gave the best performance on LVCSR systems [6].

Previously, MLLR speaker adaptation and Speaker Adaptive Training (SAT) techniques have been applied to EMLLT [4] and SPAM models [5]. However, the estimation of the adaptation transforms proposed in the papers does not have an efficient closed-form solution and was achieved using standard numerical optimisation techniques. This paper presents efficient forms of speaker adaptation and adaptive training of these precision matrix models, focusing primarily on the MPE discriminatively trained SPAM models. Iterative row-by-row update formulae are derived within the Expectation Maximisation (EM) framework for both MLLR mean and constrained MLLR adaptations. Efficient approximate schemes to further reduce computational cost are also discussed.

The rest of this paper is organised as follows. Section 2 describes a general form of row-by-row iterative update approach, which forms the basic foundation for the EM-based transform estimation formulae for both the MLLR mean and constrained MLLR adaptations. The derivation of these update formulae are given in Sections 3 and 4 respectively. Section 5 then presents a more compact statistics required for SPAM models by exploiting the precision matrix structure. Finally, experimental results on Broadcast News (BN) and Conversational Telephone Speech (CTS) English tasks are given in Section 6.

2. ADAPTATION OF PRECISION MATRIX MODELS

An important aspect of any form of improved acoustic models is the applicability of adaptation techniques to these models. This paper considers the Maximum Likelihood Linear Regression (MLLR) mean [7] and constrained (CMLLR) [8] adaptation schemes for the structured precision matrix models. Estimating the MLLR mean transforms for full covariance matrix systems using the direct closed-form solution [9] is computationally expensive and in some cases results in numerical stability issues. In this paper, an efficient row-by-row iterative update approach is presented. Diagonal precision matrix approximation is used to initialise the transforms. The results given later shows that such initialisation scheme provides a very good approximation that subsequent iterative updates may be safely omitted. Also, previous work on SAT training of EMLLT [4] and SPAM [5] models was realised based on numerical optimisation techniques. Again, an efficient row-by-row

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

In *Proc. ICASSP 2005*, Volume I, pp. 97–100.

iterative closed-form solution is derived in this paper.

The row-by-row iterative update formulae for MLLR mean and CMLLR transformation matrices are derived within the standard Expectation Maximisation (EM) framework. The general form of the auxiliary function to be maximised is given by

$$\mathcal{Q}(\boldsymbol{W}^r) = K + \eta\beta \log|\boldsymbol{W}^r| - \frac{1}{2}\sum_{m=1}^{M_r} \mathrm{Tr}(\boldsymbol{P}_m \boldsymbol{X}^{(mr)}) \qquad (2)$$

where $\boldsymbol{W}^r$ is the transformation matrix, K subsumes terms independent of $\boldsymbol{W}^r$, η is a selector variable that is set to 0 and 1 for MLLR mean and CMLLR respectively, $\beta = \sum_{m=1}^{M_r}\sum_{t=1}^{T}\gamma_m(t)$, M_r is the number of component in regression class r, $\gamma_m(t)$ is the posterior of component m at time t and

$$\boldsymbol{X}^{(mr)} = \sum_{t=1}^{T}\gamma_m(t)(\boldsymbol{x}_{mt} - \boldsymbol{W}^r\boldsymbol{y}_{mt})(\boldsymbol{x}_{mt} - \boldsymbol{W}^r\boldsymbol{y}_{mt})'$$

$\boldsymbol{x}_{mt}$ and $\boldsymbol{y}_{mt}$ can either be the observation vector or the mean vector depending on the adaptation scheme. $\boldsymbol{y}_{mt}$ is the vector to be adapted. Differentiating equation (2) with respect to $\boldsymbol{w}_i^r$, the ith row of $\boldsymbol{W}^r$, yields

$$\frac{\partial\mathcal{Q}(\boldsymbol{W}^r)}{\partial\boldsymbol{w}_i^r} = \eta\beta\frac{\boldsymbol{c}_i}{\boldsymbol{c}_i\boldsymbol{w}_i^{r\prime}} - \boldsymbol{w}_i^r\boldsymbol{G}^{(rii)} + \boldsymbol{k}^{(ri)} \qquad (3)$$

where $\boldsymbol{c}_i$ is the cofactors of the ith row of $\boldsymbol{W}^r$ and

$$\begin{aligned}\boldsymbol{G}^{(rij)} &= \sum_{m=1}^{M_r} p_m(i,j)\boldsymbol{G}_m && (4)\\ \boldsymbol{k}^{(ri)} &= \sum_{m=1}^{M_r}\boldsymbol{p}_m(i)\boldsymbol{K}_m - \sum_{j=1,j\neq i}^{d} w_j^r\boldsymbol{G}^{(rij)} && (5)\end{aligned}$$

$p_m(i,j)$ and $\boldsymbol{p}_m(i)$ denotes the (i,j)th element and ith row of $\boldsymbol{P}_m$ respectively. The component level statistics are given by

$$\boldsymbol{G}_m = \sum_{t=1}^{T}\gamma_m(t)\boldsymbol{y}_{mt}\boldsymbol{y}'_{mt} \quad \text{and} \quad \boldsymbol{K}_m = \sum_{t=1}^{T}\gamma_m(t)\boldsymbol{x}_{mt}\boldsymbol{y}'_{mt}$$

Next, row-by-row estimation formulae for MLLR mean and CMLLR adaptations are derived in Section 3 and 4 respectively.

3. MLLR MEAN ADAPTATION

MLLR adaptation of the mean vector [7] can be written as

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{A}^r\boldsymbol{\mu}_m + \boldsymbol{b}^r = \boldsymbol{W}^r\boldsymbol{\xi}_m \qquad (6)$$

where $\boldsymbol{A}^r$ and $\boldsymbol{b}^r$ are the $d \times d$ linear transformation matrix and the bias vector respectively associated to the regression class, r ($m \in r$). $\boldsymbol{\mu}_m$ and $\hat{\boldsymbol{\mu}}_m$ denote the original and adapted mean vectors respectively for component m. $\boldsymbol{W}^r = [\boldsymbol{A}^r \mid \boldsymbol{b}^r]$ and $\boldsymbol{\xi}_m = [\boldsymbol{\mu}'_m \; 1]'$ are the augmented transformation matrix and mean vector respectively. These parameters can be estimated by solving equation (3) with $\boldsymbol{x}_{mt} = \boldsymbol{o}_t$ and $\boldsymbol{y}_{mt} = \boldsymbol{\xi}_m$. Thus,

$$\boldsymbol{G}_m = \beta_m\boldsymbol{\xi}_m\boldsymbol{\xi}'_m \quad \text{and} \quad \boldsymbol{K}_m = \boldsymbol{u}_m\boldsymbol{\xi}'_m$$

where the sufficient statistics are given by

$$\beta_m = \sum_{t=1}^{T}\gamma_m(t) \qquad \boldsymbol{u}_m = \sum_{t=1}^{T}\gamma_m(t)\boldsymbol{o}_t$$

With $\eta = 0$, equating (3) to zero and solve for $\boldsymbol{w}_i^r$ gives the ML update as

$$\boldsymbol{w}_i^r = \boldsymbol{G}^{(rii)-1}\boldsymbol{k}^{(ri)}$$

This update formula is dependent on the other rows through the term $\boldsymbol{k}^{(ri)}$ in equation (5). Hence, an initial estimate of $\boldsymbol{W}^r$ is required and an iterative approach used. Although $\boldsymbol{W}^r$ can be initialised as an identity matrix, a better starting value may be found by using a diagonal precision matrix approximation, where $p_m(i,j) = 0$ for $j \neq i$. Equation (5) simplifies to that of a diagonal covariance matrix system [7]. In fact, the results presented later indicates that subsequent row-by-row iterations yield very little gain in terms of likelihood and the diagonal precision matrix approximation itself gives good estimates.

4. CONSTRAINED MLLR ADAPTATION

A simple way to achieve variance adaptation for structured precision matrix models is to train speaker-dependent basis matrices. The efficiency of this kind of variance adaptation depends on the computational cost of the basis matrix update of the precision matrix model. This approach is computationally inefficient for EMLLT and SPAM models. Alternatively, feature space CMLLR transforms may be used, where a single transformation matrix is estimation for both the mean vector and the covariance matrix. This can also be viewed as a feature-based speaker normalisation [10] technique where speaker-dependent feature transform is estimated. In CMLLR, a linear feature transformation matrix, $\boldsymbol{W}^r = [\boldsymbol{A}^r \mid \boldsymbol{b}^r]$, is estimated for each regression class, r such that

$$\hat{\boldsymbol{\zeta}}_t = \boldsymbol{A}^r\boldsymbol{o}_t + \boldsymbol{b}^r = \boldsymbol{W}^r\boldsymbol{\zeta}_t \qquad (7)$$

where $\boldsymbol{\zeta}_t$ and $\hat{\boldsymbol{\zeta}}_t$ are the augmented vectors of the original and adapted observation respectively. Again, equation (2) is maximised to obtain the ML estimate of $\boldsymbol{W}^r$, but now with $\boldsymbol{x}_{mt} = \boldsymbol{\mu}_m$ and $\boldsymbol{y}_{mt} = \boldsymbol{\zeta}_t$. Thus,

$$\boldsymbol{G}_m = \sum_{t=1}^{T}\gamma_m(t)\boldsymbol{\zeta}_t\boldsymbol{\zeta}'_t \quad \text{and} \quad \boldsymbol{K}_m = \boldsymbol{\mu}_m\boldsymbol{u}'$$

The sufficient statistics are β, $\boldsymbol{G}_m$ and $\boldsymbol{u}_m = \sum_{t=1}^{T}\gamma_m(t)\boldsymbol{\zeta}_t$. Setting equation (3) to zero with $\eta = 1$ yields the ML update for each row of $\boldsymbol{W}^r$ as

$$\boldsymbol{w}_i^r = \alpha\left(\boldsymbol{c}_i + \lambda\boldsymbol{k}^{(ri)}\right)\boldsymbol{G}^{(rii)-1} \qquad (8)$$

Equation (8) is similar to the update formula derived for the case of diagonal covariance matrix [10], differed by the term $\boldsymbol{k}^{(ri)}$, which also depends on other rows in this case. α is found by solving a quadratic equation as described in [10]. It is easy to see that when $p_m(i,j) = 0$ for $j \neq i$, equation (8) simplifies to the case of diagonal covariance matrix systems.

Unlike the case of MLLR mean, diagonal precision matrix approximation does not work for constrained MLLR because the estimated transforms operates on both the mean vectors and the precision matrices. However, the CMLLR transforms estimation process for SPAM models can be approximated using a diagonal covariance matrix model. For good approximation, this model should be the starting point used to train the SPAM model.

5. SUFFICIENT STATISTICS FOR SPAM MODELS

The required statistics associated to each regression class r for both MLLR mean and CMLLR adaptations are $\boldsymbol{G}^{rij}$ for $1 \leq i \leq d$; $1 \leq j \leq i$ and $\boldsymbol{k}^{ri}$ for $1 \leq i \leq d$, as given by equations (4) and (5) respectively. The number of parameters to be stored for these statistics are $[\frac{d}{2}(d+1)]^2 + d^2$, which is dominated by $\boldsymbol{G}^{(rij)}$. For structured precision matrix models, the memory requirement can be reduced by exploiting the basis superposition structure. Substituting equation (1) into equation (4) yields

$$\boldsymbol{G}^{(rij)} = \sum_{b=1}^{n} s_b(i,j)\boldsymbol{G}^{(rb)} \quad \text{and} \quad \boldsymbol{G}^{(rb)} = \sum_{m=1}^{M_r} \lambda_{bb}^{(m)}\boldsymbol{G}_m$$

where $s_b(i,j)$ denotes the (i,j)th element of the bth basis matrix, $\boldsymbol{S}_b$ and $1 \leq b \leq n$. So, instead of storing $\frac{d}{2}(d+1)$ terms of $\boldsymbol{G}^{(rij)}$, only n terms of $\boldsymbol{G}^{(rb)}$ are needed. Thus, the required memory is reduced from the order $\mathcal{O}(d^4)$ to $\mathcal{O}(nd^2)$. These statistics are directly related to those presented in [5] where $\boldsymbol{G}_1^k$ and $\boldsymbol{G}_4^k$ are the same as $\boldsymbol{G}^{(rb)}$ for MLLR mean and CMLLR cases respectively. The notation k used in [5] has the same meaning as b used in this paper. Also, $\boldsymbol{G}_3^k$ relates to $\boldsymbol{K}^{(rb)} = \sum_{m=1}^{M_r} \lambda_{bb}^{(m)}\boldsymbol{K}_m$.

6. EXPERIMENTAL RESULTS

Experimental results are presented based on two LVCSR English tasks: Broadcast News (BN) and Conversational Telephone Speech (CTS). 12 PLP coefficients were used with the `C0` energy term, first, second and third derivatives to form a 52-dimensional feature vector. Side-based vocal tract length, cepstral mean and cepstral variance normalisations were only used in the CTS task. Systems were built using triphone models with approximately 6000 distinct states, within the 39-dimensional HLDA subspace. CMLLR transforms were used for building SAT models. Instead of training the SAT+SPAM system from the SPAM system, the training approach described in [5] was adopted, where a speaker adaptively trained diagonal covariance matrix system (SAT+DIAGC) was used as the starting point. In other words, the SPAM precision matrix modelling was performed within the SAT feature space. In testing, MLLR mean transforms for the SPAM models were estimated using two row-by-row iterations as described in Section 3 (`mllr`) or simply approximated using the diagonal precision matrix assumption (`mllr+`). Similarly, the CMLLR transforms were estimated either using the exact method (`cmllr`) as described in Section 4 or approximated using a SAT+DIAGC system (`cmllr+`).

Figure 1 illustrates the change in the average log likelihood of one speaker with increasing number of iterations for both MLLR mean and CMLLR adaptations. On each iteration, the component alignment was recomputed based on the transforms estimated in the previous iteration. The average log likelihood was found to increase upon every iteration. In Figure 1(a), there is very little difference between the `mllr` and `mllr+` methods for MLLR mean transform estimation. For CMLLR, the log likelihood gain from using the `cmllr` method is about twice that of the approximated method, `cmllr+`, as depicted in Figure 1(b).

Word Error Rate (WER) performance was also examined. For BN task, 16-component models were trained using 374 hours of `bnetrain04sub` training data. This consists of 143 hours of carefully annotated data and 231 hours of lightly supervised data. Adaptation experiments were conducted based on three 3-hour test sets: `eval03`, `dev04` and `dev04f`. 4-gram rescoring lattices were generated using an adapted HLDA system[1]. Rescoring results are summarised in Table 1. For MLLR mean adaptation, a gender dependent (GD) DIAGC system was chosen as the baseline. This system gave WERs of 10.7%, 13.2% and 20.0% on the three test sets. The exceptionally poor performance on `dev04f` is due to the large mismatch between the training and the test data. Both `mllr+` and `mllr` configurations yielded the same performance, which is 0.1% absolute better than the baseline on `eval03` and `dev04`. The gain on `dev04f` is larger, 0.5% absolute. This shows that MLLR mean adaptation can be efficiently approximated with the diagonal precision matrix assumption for the SPAM models and other forms of precision matrix models such as EMLLT.

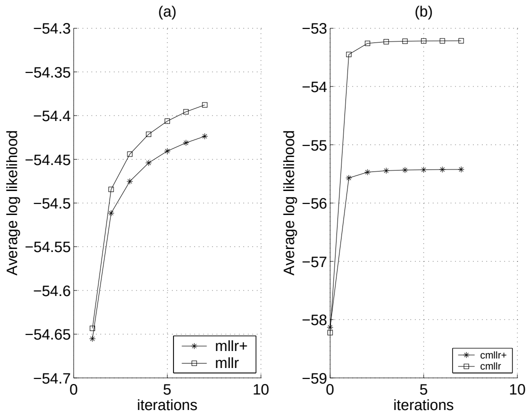

Fig. 1. Change in average log likelihood of one speaker on CTS with increasing number of MLLR iterations for (a) MLLR mean and (b) CMLLR, for 28-component SPAM model

System	Adapt Config	Test Set WER (%)		
		eval03	dev04	dev04f
DIAGC	mllr	10.7	13.2	20.0
SPAM	mllr+	10.6	13.1	19.5
	mllr	10.6	13.1	19.5
SAT+DIAGC	cmllr	10.6	13.1	19.5
SAT+SPAM	cmllr+	10.2	12.7	18.6
	cmllr	10.2	12.8	18.8

Table 1. Comparisons of MLLR mean and CMLLR adaptations for 16-comp DIAGC and SPAM models on BN system

Also, two forms of CMLLR adaptation for SAT+SPAM models were compared using the SAT+DIAGC system as the baseline. This system has the same WER performance as the MLLR mean adapted SPAM system. The `cmllr+` configurations gained 0.4% absolute on the first two test sets and 0.9% on `dev04f`. Again, there is a large gain from the adapted SPAM models due to the mismatch between the training and test sets. Similar performance was obtained on `eval03` using the exact `cmllr` configuration. Surprisingly, 0.1% and 0.2% degradations were observed on `dev04` and `dev04f` although the likelihood of the test data given these transforms was higher than those approximated using `cmllr+`. Apart from the gains from the `mllr+` and `mllr` SPAM models on

[1] Similar to the `P2` stage of the CU-HTK evaluation system

eval03 and dev04, all the gains shown in Table 1 were found to be statistically significant[2].

Similar comparisons were made on the CTS task. 28-component models were trained using 400 hours of Fisher data (fsh2004sub) and evaluated on two test sets. eval03 consists of two parts, Switchboard (s25) and Fisher (fsh), 3 hours each. dev04, on the other hand, is a 3 hours test set, containing only Fisher data. Table 2 summarises the results of various adaptation configurations on CTS. The WERs of the baseline DIAGC system after MLLR adaptation were 22.3% and 18.4% on eval03 and dev04 respectively. SPAM model with diagonal precision matrix approximated MLLR adaptation gave 0.4-0.5% gains, although a large proportion of the gain on eval03 came from s25 (0.6%). Performing two additional row-by-row iterations, although improved the likelihood, degraded the WER performance by 0.1% on the fsh part of eval03 and dev04. The SAT+DIAGC system is about 0.3%-0.5% absolute better than the non-SAT baseline on both test sets. Using this model to estimate the CMLLR transforms for the SAT+SPAM system (cmllr+) improved the WERs by 0.5% and 0.3% absolute on eval03 and dev04 respectively. Again, the gain on s25 dominated for the eval03 test set. Exact implementation using the cmllr method gave a consistent improvement of 0.1% on all test sets.

System	Adapt Config	eval03			dev04
		s25	fsh	Avg	Avg
DIAGC	mllr	26.1	18.1	22.3	18.4
SPAM	mllr+	25.5	17.9	21.9	17.9
	mllr	25.5	18.0	21.9	18.0
SAT+DIAGC	cmllr	25.8	17.8	21.9	17.9
SAT+SPAM	cmllr+	25.0	17.6	21.4	17.6
	cmllr	24.9	17.5	21.3	17.5

Table 2. Comparisons of MLLR mean and CMLLR adaptations for 28-comp DIAGC and SPAM models on CTS system

Finally, a state-of-the-art SAT+SPAM system was trained using the 2180 hours fsh2004h5etrain03b training data. This training data comprises both Fisher (1820 hours fsh2004) and Switchboard (360 hours h5etrain03b) data. This system was evaluated on both eval03 and dev04 test sets and compared with the SAT+DIAGC system.

System	Adapt Config	eval03			dev04
		s25	fsh	Avg	Avg
SAT+DIAGC	cmllr	22.7	15.5	19.2	16.1
SAT+SPAM	cmllr+	22.1	15.0	18.6	15.7
	cmllr	22.1	15.0	18.7	15.5

Table 3. Comparisons of CMLLR adapted 36-comp SAT+DIAGC and SAT+SPAM models on state-of-the-art CTS

In Table 3, the WER performance of the baseline SAT+DIAGC system was 19.2% and 16.1% on eval03 and dev04 respectively. As before, the difference between cmllr and cmllr+ for SAT+SPAM is small. Comparing to SAT+DIAGC, the SAT+SPAM system gained about 0.5-0.6% and 0.4-0.6% absolute on eval03 and dev04 respectively. These gains were found to be statistically significant. Similar gains were also found with more complex adaptation techniques [11].

7. CONCLUSIONS

This paper has examined the linear adaptation of structured precision matrix models combining the speaker adaptive training, SPAM precision matrix modelling and MPE discriminative training in state-of-the-art large vocabulary continuous speech recognition systems. In contrast to the previous work, this paper presented simple iterative row-by-row update formulae for both MLLR mean and constrained MLLR adaptation of structured precision matrix models which guanratees to increase the likelihood of the adaptation data. Further approximations of these adaptation schemes to reduce computational cost were found to yield similar performance. Experimental results were presented based on MPE discriminatively trained SPAM models for broadcast news and conversational telephone speech English tasks. The SAT+SPAM system gave the best performance gain of approximately 0.5% absolute over the SAT+DIAGC system.

8. REFERENCES

[1] M. 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.

[2] P. Olsen and R. A. Gopinath, "Modelling inverse covariance matrices by basis expansion," in *Proc. ICASSP*, 2002.

[3] S. Axelrod, R. Gopinath, and P. Olsen, "Modeling with a subspace constraint on inverse covariance matrices," in *Proc. ICSLP*, 2002.

[4] J. Huang, V. Goel, R. Gopinath, B. Kingsbury, P. Olsen, and K. Visweswariah, "Large vocabulary conversational speech recognition with the extended maximum likelihood linear transformation (EMLLT) model," in *Proc. ICSLP*, 2002.

[5] S. Axelrod, V. Goel, B. Kingsbury, K. Visweswariah, and R. A. Gopinath, "Large vocabulary conversational speech recognition with a subspace constraint on inverse covariance matrices," in *Proc. Eurospeech*, 2003.

[6] K C Sim and M J F Gales, "Precision matrix modelling for large vocabulary continuous speech recognition," Tech. Rep. CUED/F-INFENG/TR485, Cambridge University, 2004.

[7] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression speaker adaptation of contiuous density HMMs," *Computer Speech and Languages*, 1997.

[8] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep. CUED/F-INFENG/TR291, Cambridge University, 1997.

[9] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Languages*, vol. 10, pp. 249–264, 1996.

[10] M. J. F. Gales, "Maximum likelihood multiple projection schemes for hidden Markov models," Tech. Rep. CUED/F-INFENG/TR365, Cambridge University, 1999.

[11] X. Liu, M. J. F. Gales, K. C. Sim, and K. Yu, "Investigation of acoustic modelling techniques of LVCSR systems," *submited to ICASSP*, 2005.

[2] Significance tests were carried out using the NIST Scoring Toolkit.