ADAPTING SEMI-TIED FULL-COVARIANCE MATRIX HMMS

M.J.F. Gales

CUED/F-INFENG/TR 298

July 1997

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

Email: mjfg@eng.cam.ac.uk

Abstract

There is normally a simple choice made in the form of the covariance matrix to be used with HMMs. Either a diagonal covariance matrix is used, with the underlying assumption that elements of the feature vector are independent, or a full or block-diagonal matrix is used, where all or some of the correlations are explicitly modelled. Unfortunately when using full or block-diagonal covariance matrices there tends to be a dramatic increase in the number of parameters per Gaussian component, limiting the number of components which may be robustly estimated. This paper investigates a recently introduced form of covariance matrix, the semi-tied full-covariance matrix. This allows a few "full" covariance matrices to be shared over many distributions, whilst each distribution maintains its own "diagonal" covariance matrix. In current systems it is essential to be able to rapidly adapt the acoustic models to a particular speaker or new acoustic environment. This paper examines two linear-transformation speaker-adaptation schemes that may be applied to these semi-tied models. Both yield maximum likelihood estimates of the transform, but differ in the domains in which the transforms are estimated. A large-vocabulary speaker-independent speech-recognition task was used to assess the performance of the techniques. Both the adaptation schemes showed gains in performance. Depending on the semi-tied model set used and the adaptation scheme improvements over the unadapted models ranged from 3% to 11% relative. Furthermore, a 9% relative reduction in word error rate was achieved over the standard model set adapted using maximum likelihood linear regression.

1 Introduction

There is normally a simple choice made in the form of the covariance matrix to be used with hidden Markov models (HMMs). Either a diagonal covariance matrix is used, with the underlying assumption that each element of the feature vector is independent, or a full or block-diagonal matrix is used, where all or some of the correlations are explicitly modelled. Unfortunately when using full or block-diagonal covariance matrices there tends to be a dramatic increase in the number of parameters per Gaussian component, limiting the number of components which may be robustly estimated. To overcome this problem multiple diagonal-covariance Gaussian distributions may be used. In addition to being able to model non-Gaussian distributions they can model correlations between elements of the feature vector. Unfortunately only a limited number of components may be robustly estimated for each state. If there were some way of effectively decorrelating the data associated with a particular state, or group of states, improved speech recognition performance should be possible, either using fewer components per state, or allowing non-Gaussian distributions associated with the state to be better modelled. This lead to the development of *semi-tied full-covariance* matrices[1].

Semi-tied full-covariance matrices are a natural extension of the state-specific rotation scheme of [10]. Instead of estimating the decorrelating transform independently of the specific components associated with it, the transform is estimated in a maximum-likelihood (ML) fashion given the current model parameters. This form of covariance modelling was first introduced in [3]. On a large-vocabulary, speaker-independent speech recognition task it was shown to give a 10% reduction in word error rate over a similar diagonal covariance model set. Furthermore, the performance of a 12-components per state system was achieved with 6-components per state. Though in themselves useful reductions in word error rate or number of parameters, most state-of-the-art systems are required to make extensive use of both speaker and environmental adaptation schemes to yield good recognition performance. For semi-tied full-covariance matrices to be useful for speech recognition effective adaptation techniques are essential.

Many adaptation schemes have been examined for HMMs. Two popular schemes are maximum a-posteriori (MAP) adaptation [6] and linear transformation of the model parameters, for example maximum likelihood linear regression (MLLR) [9]. If standard MAP estimation is used then this may be simply applied to adapt the means and diagonal element of the covariance matrix. The difference being that instead of applying the formulae in the cepstral domain, they may be applied in the *normalised domain* (the domain where the covariance matrices are assumed to be diagonal)[2]. Alternatively linear transformation schemes may be used. These linear adaptation schemes have been found to be powerful tools for speaker and environmental adaptation [4]. When using these linear transformation techniques there are typically very few transforms. It may therefore be necessary for the same transform to be shared over components from different semi-tied classes. This adds an additional degree of complexity when adapting these semi-tied full-covariance models. Two possible forms of adapting these models with linear transforms are described in this paper. The first, *full-covariance MLLR*, is a simple extension of the standard MLLR scheme to adapting models with, effectively, full covariance matrices. The second, *normalised-domain MLLR*, performs the linear transformation in a domain determined by the semi-tied full-covariance matrix. The schemes have different computational requirements and typically give different performance. However under certain constraints, the two schemes result in equivalent transforms.

The next section describes the basic concept of the semi-tied full-covariance matrix. In section 3, two adaptation schemes are described. This section includes a discussion of the computational loads involved and the constraints for the two schemes to be equivalent. Results are then presented for speaker adaptation on a large-vocabulary speaker-independent speech recognition task.

---

[1] The form of the matrices is not actually restricted to be full. For example in the work presented here, block-diagonal matrices are used.

[2] If extensions to the standard MAP scheme such as [12] are used then modifications to the standard formulae may be required. This is also true if the full element of the covariance matrix is to be adapted.

2 Semi-Tied Full-Covariance Matrices

Semi-tied full-covariance matrices [3] are a simple extension to the standard diagonal, block-diagonal, or full covariance matrices used with HMMs. Instead of having a distinct covariance matrix for every component in the recogniser, each component consists of two elements, a component specific diagonal covariance element, $\boldsymbol{\Sigma}^{(m)}_{\text{diag}}$, and a *semi-tied* class dependent, non-diagonal matrix[3] , $\mathbf{A}^{(r)\prime}$. The form of the covariance matrix is

$$\boldsymbol{\Sigma}^{(m)} = \mathbf{A}^{(r)\prime}\boldsymbol{\Sigma}^{(m)}_{\text{diag}}\mathbf{A}^{(r)\prime T} \tag{1}$$

$\mathbf{A}^{(r)\prime}$ may be tied over a set of components, for example all those associated with the same state of a particular context-independent phone. Possible clustering schemes are discussed in [3].

Each component, m, has the following parameters: component weight $c^{(m)}$, component mean $\mu^{(m)}$ and the diagonal element of the semi-tied full covariance matrix $\boldsymbol{\Sigma}^{(m)}_{\text{diag}}$. In addition it is associated with a semi-tied class, which has an associated matrix $\mathbf{A}^{(r)\prime}$. This is used to generate the components covariance matrix as described in equation 1. It is very complex to optimise these parameters directly so an expectation-maximisation approach is adopted [1]. Optimising many of the parameters is closely related to the standard parameter optimisation problem. Here, rather than dealing with$\mathbf{A}^{(r)\prime}$, it is simpler to deal with its inverse, $\mathbf{A}^{(r)}$ [3], thus

$$\mathbf{A}^{(r)} = \mathbf{A}^{(r)\prime -1} \tag{2}$$

The re-estimation formulae for the component weights and transition probabilities are identical to the standard HMM cases [11]. The mean and diagonal elements of the covariance matrix are given by

$$\mu^{(m)} = \frac{\sum_{\tau=1}^{T}\gamma_m(\tau)\mathbf{o}(\tau)}{\sum_{\tau=1}^{T}\gamma_m(\tau)} \tag{3}$$

and

$$\sigma^{(m)2}_{\text{diag}i} = \frac{\sum_{\tau=1}^{T}\gamma_m(\tau)\left(o_i^{(r)}(\tau) - \mu_i^{(rm)}\right)^2}{\sum_{\tau=1}^{T}\gamma_m(\tau)} \tag{4}$$

where

$$\mu^{(rm)} = \mathbf{A}^{(r)}\mu^{(m)} \tag{5}$$

$$\mathbf{o}^{(r)}(\tau) = \mathbf{A}^{(r)}\mathbf{o}(\tau) \tag{6}$$

and

$$\gamma_m(\tau) = p(q_m(\tau)|\mathcal{M}, \mathbf{O}_T) \tag{7}$$

where $q_m(\tau)$ indicates component m at time τ, $\mathbf{O}_T$ is the complete set of training data and $\mathbf{o}(\tau)$ is the observation at time τ. Throughout this paper various subsets of the complete set of components, M, will be considered. Thus $M^{(r)}$ is the subset of components that share the same full covariance element $\mathbf{A}^{(r)}$ or *semi-tied* class, $M^{(s)}$ the set of components that use the same transformation matrix $\mathbf{W}^{(s)}$ or *regression* class, and $M^{(rs)}$ specifies the subset that uses both $\mathbf{A}^{(r)}$ and $\mathbf{W}^{(s)}$, $M^{(r)} \cap M^{(s)}$. The semi-tied classes are numbered 1 to R and the regression classes from 1 to S.

[3]For this work $\mathbf{A}^T$ represents the transpose of matrix $\mathbf{A}$, the use of $\mathbf{A}'$ is simply to indicate a modified domain or related matrix.

The ML estimate of $\mathbf{A}^{(r)}$ is more complicated and an iterative scheme is required. The solution given here is described in [2] (an alternative scheme is given in [3]). It is shown that

$$\mathbf{a}_i^{(r)} = \mathbf{c}_i \mathbf{G}^{(ri)-1} \sqrt{\left(\frac{\sum_{m \in M^{(r)}} \sum_{\tau=1}^{T} \gamma_m(\tau)}{\mathbf{c}_i \mathbf{G}^{(ri)-1} \mathbf{c}_i^T} \right)} \tag{8}$$

where

$$\mathbf{G}^{(ri)} = \sum_{m \in M^{(r)}} \frac{1}{\sigma_{\text{diag}i}^{(m)2}} \sum_{\tau=1}^{T} \gamma_m(\tau) \left(\mathbf{o}(\tau) - \mu^{(m)} \right) \left(\mathbf{o}(\tau) - \mu^{(m)} \right)^T \tag{9}$$

$\sigma_{\text{diag}i}^{(m)2}$ is element i of the leading diagonal of $\mathbf{\Sigma}_{\text{diag}}^{(m)}$ and $\mathbf{c}_i$ is the i^{th} row vector of the cofactors of $\mathbf{A}^{(r)}$. The optimisation described is an iterative one over rows, since each row is related to the other rows by the cofactors. Thus each row is optimised given the current estimate of all the other rows.

One problem with the optimisation described is that the estimates of the diagonal and full sections of the covariance matrix are related in a rather complicated fashion to one another. Moreover when simultaneously optimised there are an infinite number of possible solutions [8][4]. There is also the added problem that it is necessary to store a full matrix for each component to be able to generate $\mathbf{G}^{(ri)}$ as $\mathbf{\Sigma}_{\text{diag}}^{(m)}$ varies. This is both impractical in terms of memory requirements and computational speed for many large vocabulary systems. These problems are overcome by updating the means and either the diagonal or full-element of the covariance matrix at each iteration. It is now only necessary to store either $\mathbf{G}^{(ri)}$ for each semi-tied class, or $\sum_{\tau=1}^{T} o_i^{(r)2}$ for each component in addition to the standard statistics for the means, component weights and transition probabilities.

One of the major advantages of semi-tied full covariance is their computational efficiency during recognition. The likelihood calculation is based on

$$\mathcal{L}\left(\mathbf{o}(\tau); \mu^{(m)}, \mathbf{\Sigma}^{(m)} \right) = \mathcal{N}\left(\mathbf{o}^{(r)}(\tau); \mathbf{A}^{(r)}\mu^{(m)}, \mathbf{\Sigma}_{\text{diag}}^{(m)} \right) + \frac{1}{2}\log\left(|\mathbf{A}^{(r)}|^2 \right) \tag{10}$$

where $m \in M^{(r)}$. Thus by storing $\mathbf{A}^{(r)}\mu^{(m)}$ instead of $\mu^{(m)}$ the cost of calculating the likelihoods associated with semi-tied full covariance matrices is that of one matrix vector multiplication per-semi-tied class and an addition[5].

3 Linear-Transformation Adaptation

Linear transformations have been shown to be a powerful tool for both speaker and environmental adaptation. One currently popular scheme is MLLR [9]. Unfortunately this scheme relies on the covariance matrix being diagonal, so is not directly applicable to semi-tied full-covariance matrices. Models similar to semi-tied full-covariance models, generated using state-specific rotations [10], have previously been adapted using linear transformations [7]. The scheme used is related to the normalised-domain MLLR scheme described below, though the transforms were generated in a least-squares fashion rather than using maximum likelihood. Two modified MLLR-like schemes are proposed here. The same general form of the transformation is used.

$$\hat{\mu} = \mathbf{W}\xi \tag{11}$$

where ξ is the extended mean vector $\begin{bmatrix} 1 & \mu^T \end{bmatrix}^T$. The two schemes only differ in terms of the domain in which the transform is estimated and applied.

[4]By altering the columns of $\mathbf{A}^{(r)}$ and the elements of the means and variances appropriately there are an infinite number of solutions all yielding the same likelihood, but with different scaling factors.

[5]If only one semi-tied class is used then $\log(|\mathbf{A}|)$ does not discriminate between the models so may be ignored. Also, this addition may be removed if desired by appropriately scaling $\mathbf{A}^{(r)}$ and $\mathbf{\Sigma}_{\text{diag}}^{(m)}$.

1. **Full-covariance MLLR**: This is an extension to MLLR, so that the full-covariance matrix case may be handled [4][6]. Thus the transformation is estimated using

$$\mathrm{vec}\left(\mathbf{Z}^{(s)}\right) = \left(\sum_{m \in M^{(s)}} \mathrm{kron}(\mathbf{V}^{(m)}, \mathbf{D}^{(m)})\right) \mathrm{vec}\left(\mathbf{W}^{(s)}\right) \tag{12}$$

where vec(.) converts a matrix to a vector ordered in terms of the rows, kron(.) is the Kronecker product,

$$\mathbf{V}^{(m)} = \sum_{\tau=1}^{T} \gamma_m(\tau) \boldsymbol{\Sigma}^{(m)-1} \tag{13}$$

$$\mathbf{Z}^{(s)} = \sum_{m \in M^{(s)}} \sum_{\tau=1}^{T} \gamma_m(\tau) \boldsymbol{\Sigma}^{(m)-1} \mathbf{o}(\tau) \xi^{(m)T} \tag{14}$$

and

$$\mathbf{D}^{(m)} = \xi^{(m)} \xi^{(m)T} \tag{15}$$

$\xi^{(m)}$ is the extended mean vector. If implemented directly this may be computationally expensive. Typically in large vocabulary systems it is not possible to store $\boldsymbol{\Sigma}^{(m)-1}$, as this would require a full symmetric matrix for each component, so it must be calculated on-the-fly. Furthermore the Kronecker product is required for every component observed in the adaptation data. An alternative efficient implementation is described in appendix A.

Adapting the means in this form requires that the recognition likelihood is based on

$$\mathcal{L}\left(\mathbf{o}(\tau); \mu^{(m)}, \boldsymbol{\Sigma}^{(m)}, \mathbf{W}^{(s)}\right) = \mathcal{N}\left(\mathbf{o}^{(r)}(\tau); \mathbf{A}^{(r)} \mathbf{W}^{(s)} \xi^{(m)}, \boldsymbol{\Sigma}_{\mathrm{diag}}^{(m)}\right) + \frac{1}{2} \log\left(|\mathbf{A}^{(r)}|^2\right) \tag{16}$$

2. **Normalised-domain MLLR**: This scheme is based on the normalised-domain MLLR scheme originally described in [5]. Here rather than adapting the models in the cepstral domain, the models are adapted in a domain determined by the semi-tied full-covariance matrix, i.e. using $\mathbf{A}^{(r)}\mathbf{o}(\tau)$ rather than $\mathbf{o}(\tau)$ (assuming that that the component of interest is in the set $M^{(r)}$). In appendix B it is shown that

$$\mathbf{z}_i'^{(s)} = \mathbf{w}_i^{(s)} \mathbf{G}'^{(si)} \tag{17}$$

where

$$\mathbf{Z}'^{(s)} = \sum_{r=1}^{R} \sum_{m \in M^{(rs)}} \sum_{\tau=1}^{T} \gamma_m(\tau) \boldsymbol{\Sigma}_{\mathrm{diag}}^{(m)-1} \mathbf{o}^{(r)}(\tau) \xi^{(rm)T} \tag{18}$$

$$\mathbf{G}'^{(si)} = \sum_{r=1}^{R} \sum_{m \in M^{(rs)}} \frac{1}{\sigma_{\mathrm{diag}i}^{(m)2}} \xi^{(rm)} \xi^{(rm)T} \sum_{\tau=1}^{T} \gamma_m(\tau) \tag{19}$$

$\xi^{(rm)}$ is the extended transformed mean vector. The likelihood is based on

$$\mathcal{L}\left(\mathbf{o}(\tau); \mu^{(m)}, \boldsymbol{\Sigma}^{(m)}, \mathbf{W}^{(s)}\right) = \mathcal{N}\left(\mathbf{o}^{(r)}(\tau); \mathbf{W}^{(s)} \xi^{(rm)}, \boldsymbol{\Sigma}_{\mathrm{diag}}^{(m)}\right) + \frac{1}{2} \log\left(|\mathbf{A}^{(r)}|^2\right) \tag{20}$$

[6]The notation described here was suggested by Olivier Cappé of ENST (see http://sig.enst.fr/ cappe/others.html)

This estimation is far simpler than the full-covariance MLLR case, however it makes two additional assumptions. First, that the normalised-domains for all components, $m \in M^{(s)}$, are "similar". Furthermore, that the scaling of each dimension introduced by the full section of the covariance matrix is similar. As previously mentioned there is a degree of flexibility in estimating the full and diagonal elements of the covariance matrix. For all the work here only one of either the full or diagonal elements of the semi-tied full-covariance matrix is updated at each iteration. Thus the estimation process does not have the degree of flexibility to select arbitrary scalings. Nonetheless, this paper also considers a simple scaling routine that prevents excess differences in scaling. Here, where scaling is used, it is selected such that

$$\|\mathbf{a}_i^{(r)}\|^2 = 1 \tag{21}$$

$\mathbf{\Sigma}_{\text{diag}}^{(m)}$ is then scaled appropriately. There is no difference in performance in the unadapted systems, this only affects the generation and use of the transformation matrices $\mathbf{W}^{(s)}$ for normalised-domain MLLR.

The estimation of the two transforms have very different computational loads. For an n-dimensional feature vector, full-covariance MLLR involves the inversion of an $n(n+1) \times n(n+1)$ matrix, which requires $\mathcal{O}(n^6)$ operations. In contrast normalised-domain MLLR requires n inversions of $(n+1) \times (n+1)$ matrices, costing $\mathcal{O}(n^4)$ operations[7]. The cost has ignored the accumulation of the statistics which, particularly for the full-covariance MLLR, may be expensive. This is discussed in appendix A, where provided that the number of components is far greater than the number of semi-tied classes full-covariance MLLR is approximately n times as expensive as normalised-domain MLLR which has about the same computational cost as MLLR.

Due to the difference in computational load it is of interest what constraints there are for the two techniques to yield the same result. Examining equations 16 and 20, it is simple to see that the two schemes will yield equivalent transforms provided the linear transformation clustering does not "cross" semi-tied full-covariance clusters i.e.

$$\text{if } M^{(rs)} \neq \emptyset \text{ then } \quad M^{(r)} \cap M^{(s)} = M^{(s)}, \quad r = 1, \ldots, R \quad s = 1, \ldots, S \tag{22}$$

and the transform is sufficiently flexible (for many tasks the structure of the transformation matrix is constrained, for example it may have a block-diagonal structure). The required flexibility is that the structure of the transform must include both $\mathbf{A}^{(r)}$ and $\mathbf{A}^{(r)-1}$. The proof of these constraints is straightforward (see appendix C).

This section has only addressed the problem of linear transformations for adapting the means. It is also possible to adapt the variances [4, 2]. There are a variety of options, see appendix D. When the performing this in a normalised-domain fashion, thus

$$\hat{\mathbf{\Sigma}}^{(m)-1} = \mathbf{A}^{(r)T}\mathbf{H}^T\mathbf{\Sigma}_{\text{diag}}^{(m)-1}\mathbf{H}\mathbf{A}^{(r)} \tag{23}$$

it is trivial to show that the re-estimation formulae are identical to those derived in [2], other than using $\mathbf{o}^{(r)}(\tau)$ and $\mu^{(rm)}$ instead of $\mathbf{o}(\tau)$ and $\mu^{(m)}$ (the proof follows the same lines as that described in appendix B and is discussed in appendix D). For the full-covariance MLLR version of variance adaptation, where

$$\hat{\mathbf{\Sigma}}^{(m)-1} = \mathbf{H}^T\mathbf{A}^{(r)T}\mathbf{\Sigma}_{\text{diag}}^{(m)-1}\mathbf{A}^{(r)}\mathbf{H} \tag{24}$$

it is necessary to modify the re-estimation formulae. Details of this optimisation are given in appendix D. Variance adaptation is not further considered in this paper, but has been shown to be useful for both speaker and particularly environmental adaptation [4].

[7] It is worth comparing these costs to the original normalised-domain MLLR [5]. Here the inverse covariance matrix is decomposed into its Choleski factors, and the transforms estimated in the normalised-domain determined by these factors. Since in this normalised domain all covariance matrices are equal (the identity matrix), only a single matrix inversion is required (in the same fashion as least-squares linear regression) so the cost is $\mathcal{O}(n^3)$.

4 Experiments and Results

An initial investigation of the use of semi-tied full-covariance matrices was carried out on a large-vocabulary speaker-independent continuous-speech recognition task. All recognition experiments were carried out on the 1994 ARPA Hub 1 data. The H1 task is an unlimited vocabulary task with approximately 15 sentences per speaker. This data was recorded in a clean[8] environment.

4.1 Recognition System

The baseline system used for the recognition task was a gender-independent cross-word-triphone mixture-Gaussian tied-state HMM system. This was the same as the "HMM-1" model set used in the HTK 1994 ARPA evaluation system [13]. The speech was parameterised into 12 MFCCs, C_1 to C_{12}, along with normalised log-energy and the first and second differentials of these parameters. This yielded a 39-dimensional feature vector, to which cepstral mean normalisation was applied. The acoustic training data consisted of 36493 sentences from the SI-284 WSJ0 and WSJ1 sets, and the LIMSI 1993 WSJ lexicon and phone set were used. The standard HTK system was trained using decision-tree-based state clustering to define 6399 speech states. The number of components per-state was increased using *mixing-up* [14] until there were 12 components in each speech state. This standard model-set will be referred to as *Standard*. For the H1 task a 65k word list and dictionary was used with the trigram language model described in [13]. All decoding used a dynamic-network decoder.

Two semi-tied full-covariance systems were investigated. For both systems $\mathbf{A}^{(r)}$ was constrained to be block-diagonal, with separate blocks for the static, delta and delta-delta elements of the feature vector. All components of the same context-independent phone were clustered together into the same semi-tied class. Both systems had 12-components per speech state with the same state clustering as the standard system.

1. **System 1**: This was built from scratch using mixing-up. The procedure was as follows starting with the single component standard system: perform two iterations of Baum-Welch updating the means and diagonal covariance elements; update the full covariance element; perform two iterations of Baum-Welch updating the means and diagonal covariance elements; mix-up and repeat as required. This scheme allows full advantage of the semi-tied full-covariance matrices to be made. This model set will be referred to as *Semi-Tied (1)*.

2. **System 2**: Here the standard 12-component system is used as the initial model. Using this model set the full element of the covariance matrix is estimated and then two iterations of Baum-Welch updating the means and diagonal covariance elements was performed. This model set will be referred to as *Semi-Tied (2)*.

Two modes of adaptation were examined. The first, *incremental* adaptation, is where the models are continually updated as more adaptation data (recognised sentences) become available. The results are thus generated causally. For incremental adaptation experiments the regression classes were determined dynamically using a regression class tree and a minimum occupancy threshold. The second mode, *unsupervised static* adaptation, is where the transcription of the adaptation data is unknown and all the adaptation data is available in one block. Given the unadapted system a transcription is hypothesised, then given this hypothesised transcription the models adapted. This may be repeated many times, though for this work only it was only performed once. In this mode two fixed regression classes were used.

4.2 Results

Table 1 shows the baseline performance of the three systems. As previously noted [3], semi-tied systems can give around 10% reduction in word error rate when trained from scratch, *Semi-Tied (1)*, and around 5% when trained as a "second" pass, *Semi-Tied (2)*.

[8] Here the term "clean" refers to the training and test conditions being from the same microphone type with a high signal-to-noise ratio.

System	Error Rate (%)		
	H1 Dev	H1 Eval	Average
Standard	9.57	9.20	9.38
Semi-Tied (1)	8.62	8.12	8.36
Semi-Tied (2)	9.00	8.59	8.78

Table 1: Baseline and semi-tied covariance matrices results on H1 development and evaluation data [3]

System	Adaptation Scheme	Error Rate (%)		
		H1 Dev	H1 Eval	Average
Standard	MLLR	8.06	8.13	8.10
Semi-Tied (1)	Full-Covariance	7.62	7.21	7.40
	Normalised-Domain	8.33	7.89	8.10
Semi-Tied (2)	Full-Covariance	8.18	7.70	7.93
	Normalised-Domain	8.11	7.54	7.81

Table 2: Incremental adaptation results on H1 development and evaluation data

Table 2 shows the performance of the two linear transformation adaptation schemes used in incremental adaptation mode. In all cases block-diagonal transforms were used with the regression classes determined using a regression class tree (same minimum occupancies were used in all cases, the value of the threshold was empirically determined on similar tasks). For the *Semi-Tied (1)* system both adaptation schemes reduced the word error rate. However it is clear that the full-covariance MLLR scheme yielded better performance with around a 9% gain over the standard system with MLLR. The limitations of the normalised-domain MLLR are clearly illustrated. Thus the assumptions behind normalised-domain MLLR are not valid for the *Semi-Tied (1)* system. This is not really surprising since it is not necessary that the elements of $\mathbf{o}^{(r)}(\tau)$ be related to one another across semi-tied classes.

Again, for the *Semi-Tied (2)* system both adaptation schemes yielded gains in performance. However what is interesting is that the normalised-domain scheme gave marginally better than the full-covariance scheme. This is because the full-section of the covariance matrix is added in a second pass, hence the ability of the system to give markedly different normalised domains is limited. This is illustrated by the reduced gain in performance compared to *Semi-Tied (1)* in table 1. The slight gain in performance of normalised-domain MLLR over full-covariance MLLR may result from numerical inaccuracy due to the inversion of large matrices, in this case a 182 by 182 matrix (for this work singular value decomposition was used for the inversion).

The normalised domain performance figures given in table 2 used no scaling. When scaling, as described in section 3, was used little change in performance was observed. The average performance over the two test sets of the *Semi-Tied (1)* system was 8.03% and for the *Semi-Tied (2)* system 7.88%. The indication from this is that the difference in recognition performance of the *Semi-Tied (1)* system using full-covariance MLLR and normalised-domain MLLR was not due to large differences in scaling.

Static unsupervised adaptation was also examined on the same task. Rather than using a minimum-threshold criterion to select the regression classes for adaptation, two fixed regression classes were used. Furthermore only a single adaptation run was used, i.e. the transcription from the unadapted system was used as the hypothesised transcription for adaptation. Similar trends to those in the incremental adaptation case were observed. The standard system gave an average word error rate over the two tests of 8.30%. The best performance was obtained with the *Semi-Tied (1)* system using the full-covariance MLLR adaptation, 7.56%, a 9% reduction in word error rate compared to the standard system using MLLR. It is interesting to compare this result with the performance when an efficient block-diagonal variance transform is used on the standard system [2]. This variance transform is closely related to semi-tied full-covariance matrices. It may be viewed as estimating speaker dependent semi-tied matrices, though the number of transforms that may be robustly estimated is limited by the amount of adaptation data. Using a block-diagonal variance transform gave a word error rate of 7.90%, still 4% worse than the adapted semi-tied system[9].

5 Conclusions

This paper has investigated applying linear transformations, trained in a maximum likelihood sense, to HMMs with semi-tied full-covariance matrices. Two possible schemes are described. The first, full-covariance MLLR, uses a full-covariance matrix version of MLLR. The second, normalised-domain MLLR, generates transforms in a normalised domain in which the covariance matrices for the components are assumed to be, and modelled as, diagonal. In terms of computational load normalised domain has the same cost as standard MLLR, whereas full-covariance MLLR is more expensive. Under certain limited circumstances the two transforms yield the same effective results. The two schemes, using two possible methods of building semi-tied systems, were compared with a standard large vocabulary speech recognition system adapted using MLLR on a speaker adaptation task. The use of the full-covariance MLLR scheme with a model trained using semi-tied full-covariance matrices throughout was found to give around a 10% gain over the standard scheme with MLLR, and 20% over an unadapted standard system. Though the adaptation schemes were only investigated on a speaker adaptation task, the same performance gains are expected when using the adaptation schemes for environment adaptation.

Acknowledgements

Mark Gales is funded as a Research Fellow at Emmanuel College, Cambridge.

[9] The block-diagonal variance adaptation was only performed in static adaptation mode, due to the amount of data required to robustly estimate the transform.

A Efficient Full-Covariance MLLR

This appendix describes a method of accumulating the statistics for full-covariance MLLR which under many circumstances dramatically reduces the computational load. The majority of the additional computational load is spent generating the accumulate $\sum_{m\in M^{(s)}} \text{kron}\left(\mathbf{V}^{(m)}, \mathbf{D}^{(m)}\right)$. Using the fact that

$$\sum_{m\in M^{(s)}} \text{kron}\left(\mathbf{V}^{(m)}, \mathbf{D}^{(m)}\right) = \sum_{r=1}^{R} \sum_{m\in M^{(rs)}} \text{kron}\left(\mathbf{A}^{(r)T}\mathbf{\Sigma}_{\text{diag}}^{(m)-1}\mathbf{A}^{(r)}\sum_{\tau=1}^{T}\gamma_m(\tau), \mathbf{D}^{(m)}\right) \tag{25}$$

and as by definition $\mathbf{\Sigma}_{\text{diag}}^{(m)-1}$ is diagonal, it is possible to write

$$\mathbf{A}^{(r)T}\mathbf{\Sigma}_{\text{diag}}^{(m)-1}\mathbf{A}^{(r)} = \mathbf{A}^{(r)T}\left(\sum_{i=1}^{n}\frac{1}{\sigma_{\text{diag}i}^{(m)2}}\mathbf{B}^{(ri)}\right) = \sum_{i=1}^{n}\frac{1}{\sigma_{\text{diag}i}^{(m)2}}\mathbf{B}^{(ri)T}\mathbf{B}^{(ri)} \tag{26}$$

where

$$\mathbf{b}_j^{(ri)} = \begin{cases} \mathbf{a}_i^{(r)}, & (i=j) \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{27}$$

Hence

$$\begin{aligned}
\sum_{m\in M^{(s)}} \text{kron}\left(\mathbf{V}^{(m)}, \mathbf{D}^{(m)}\right) &= \sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{i=1}^{n}\text{kron}\left(\mathbf{B}^{(ri)T}\left(\frac{1}{\sigma_{\text{diag}i}^{(m)2}}\mathbf{B}^{(ri)}\right)\sum_{\tau=1}^{T}\gamma_m(\tau), \mathbf{D}^{(m)}\right) \\
&= \sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{i=1}^{n}\text{kron}\left(\mathbf{B}^{(ri)T}\mathbf{B}^{(ri)}, \frac{1}{\sigma_{\text{diag}i}^{(m)2}}\mathbf{D}^{(m)}\sum_{\tau=1}^{T}\gamma_m(\tau)\right) \\
&= \sum_{r=1}^{R}\sum_{i=1}^{n}\text{kron}\left(\mathbf{B}^{(ri)T}\mathbf{B}^{(ri)}, \sum_{m\in M^{(rs)}}\frac{1}{\sigma_{\text{diag}i}^{(m)2}}\mathbf{D}^{(m)}\sum_{\tau=1}^{T}\gamma_m(\tau)\right)
\end{aligned} \tag{28}$$

Using this expression has two advantages over the standard form. Firstly it is unnecessary to calculate the "full" covariance matrix for each component. Second in terms of the statistics required to be accumulated per-component, this has now becomes the need to store $\mathbf{D}'^{(rsi)}$, where

$$\mathbf{D}'^{(rsi)} = \sum_{m\in M^{(rs)}}\frac{1}{\sigma_{\text{diag}i}^{(m)2}}\mathbf{D}^{(m)}\sum_{\tau=1}^{T}\gamma_m(\tau) \tag{29}$$

The final equality used is thus

$$\sum_{m\in M^{(s)}} \text{kron}\left(\mathbf{V}^{(m)}, \mathbf{D}^{(m)}\right) = \sum_{r=1}^{R}\sum_{i=1}^{n}\text{kron}\left(\mathbf{B}^{(ri)T}\mathbf{B}^{(ri)}, \mathbf{D}'^{(rsi)}\right) \tag{30}$$

Only the common situation that the number of semi-tied classes, R, is dramatically smaller than the number of components, M will be considered. Thus the majority of the computational load is in the accumulation of the statistics from the various components. For standard MLLR, normalised-domain MLLR, and full-covariance MLLR, the cost per component is summarised in the table 3. All cases use is made of the fact that both $\mathbf{D}^{(m)}$ and $\mathbf{V}^{(m)}$ are symmetric matrices. For the standard case the generation of the inverse covariance matrix is not included. From the table the use of the new efficient accumulation scheme, though more expensive than the standard MLLR and normalised-domain MLLR schemes, is $\mathcal{O}(n)$ times more efficient than the standard accumulation scheme. There is additional overhead after accumulating the statistics over all the components. However provided the number of components is far larger than the number of semi-tied classes this will not be a significant cost.

Adaptation Scheme	Cost per Component
MLLR	$\frac{1}{2}(n+1)(n+2)$
Normalised-Domain MLLR	$\frac{1}{2}(n+1)(n+2)$
Full-Covariance MLLR (1)	$\frac{1}{4}n(n+1)^2(n+2)$
Full-Covariance MLLR (2)	$\frac{1}{2}n(n+1)(n+2)$

Table 3: Accumulate computational cost per-component (*Full-Covariance (1)* uses equation 12, *Full-Covariance (2)* uses equation 28)

B Normalised-Domain MLLR

The problem is to obtain the ML estimate of $\mathbf{W}^{(s)}$ when the likelihood is calculated using

$$\mathcal{L}\left(\mathbf{o}(\tau);\mu^{(m)},\mathbf{\Sigma}^{(m)},\mathbf{W}^{(s)}\right)=\mathcal{N}\left(\mathbf{o}^{(r)}(\tau);\mathbf{W}^{(s)}\xi^{(rm)},\mathbf{\Sigma}^{(m)}_{\text{diag}}\right)+\frac{1}{2}\log\left(|\mathbf{A}^{(r)}|^2\right) \tag{31}$$

Setting up the auxiliary equation for this case gives

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})=-\frac{1}{2}\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\Big[K^{(m)}+\log\left(\left|\mathbf{\Sigma}^{(m)}_{\text{diag}}\right|\right)- \\ \log\left(\left|\mathbf{A}^{(r)}\right|^2\right)+\left(\mathbf{A}^{(r)}\mathbf{o}(\tau)-\mathbf{W}^{(s)}\xi^{(rm)}\right)^T\mathbf{\Sigma}^{(m)-1}_{\text{diag}}\left(\mathbf{A}^{(r)}\mathbf{o}(\tau)-\mathbf{W}^{(s)}\xi^{(rm)}\right)\Big]\end{aligned} \tag{32}$$

where $K^{(m)}$ is the standard normalising constant. Differentiating this with respect to $\mathbf{W}^{(s)}$ and equating to zero

$$\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\mathbf{\Sigma}^{(m)-1}_{\text{diag}}\mathbf{A}^{(r)}\mathbf{o}(\tau)\xi^{(rm)T}=\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\mathbf{\Sigma}^{(m)-1}_{\text{diag}}\mathbf{W}^{(s)}\xi^{(rm)}\xi^{(rm)T} \tag{33}$$

This is an identical expressions to the standard MLLR case, other than that $\xi^{(rm)}$ is used as the extended mean and $\mathbf{A}^{(r)}\mathbf{o}(\tau)$ as the observation. Thus it is trivial to state that

$$\mathbf{z}_i^{\prime(s)}=\mathbf{w}_i^{(s)}\mathbf{G}^{\prime(si)} \tag{34}$$

where

$$\mathbf{Z}^{\prime(s)}=\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\mathbf{\Sigma}^{(m)-1}_{\text{diag}}\mathbf{o}^{(r)}(\tau)\xi^{(rm)T} \tag{35}$$

$$\mathbf{G}^{\prime(si)}=\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\frac{1}{\sigma^{(m)2}_{\text{diag}i}}\xi^{(rm)}\xi^{(rm)T}\sum_{\tau=1}^{T}\gamma_m(\tau) \tag{36}$$

$\xi^{(rm)}$ is the extended transformed mean vector, $\begin{bmatrix} 1 & \mu^{(rm)T}\end{bmatrix}^T$ where $\mu^{(rm)}=\mathbf{A}^{(r)}\mu^{(m)}$.

C Transform Equivalence

This appendix gives the simple proof that when all the components belonging to the same regression class also belong to the same semi-tied class, both full-covariance MLLR and normalised-domain

MLLR yield the same effective transform, provided the transformation matrix is sufficiently powerful. For the full-covariance MLLR case the following auxiliary function is maximised with respect to $\mathbf{W}^{(s)}$

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = -\frac{1}{2}\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\Big[K^{(m)} + \log\left(\left|\boldsymbol{\Sigma}_{\text{diag}}^{(m)}\right|\right) - \\ \log\left(\left|\mathbf{A}^{(r)}\right|^2\right) + \left(\mathbf{o}(\tau) - \mathbf{W}^{(s)}\xi^{(m)}\right)^T\boldsymbol{\Sigma}^{(m)-1}\left(\mathbf{o}(\tau) - \mathbf{W}^{(s)}\xi^{(m)}\right)\Big]\end{aligned} \tag{37}$$

Rearranging this yields

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = -\frac{1}{2}\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\Big[K^{(m)} + \log\left(\left|\boldsymbol{\Sigma}_{\text{diag}}^{(m)}\right|\right) - \\ \log\left(\left|\mathbf{A}^{(r)}\right|^2\right) + \left(\mathbf{A}^{(r)}\mathbf{o}(\tau) - \mathbf{A}^{(r)}\mathbf{W}^{(s)}\xi^{(m)}\right)^T\boldsymbol{\Sigma}_{\text{diag}}^{(m)-1}\left(\mathbf{A}^{(r)}\mathbf{o}(\tau) - \mathbf{A}^{(r)}\mathbf{W}^{(s)}\xi^{(m)}\right)\Big]\end{aligned} \tag{38}$$

Expanding the expression containing $\mathbf{W}^{(s)}$

$$\begin{aligned}\mathbf{A}^{(r)}\mathbf{W}^{(s)}\xi^{(m)} &= \mathbf{A}^{(r)}\left(\mathbf{H}^{(s)}\mu^{(m)} + \mathbf{b}^{(s)}\right) \\ &= \mathbf{A}^{(r)}\mathbf{H}^{(s)}\mu^{(m)} + \mathbf{A}^{(r)}\mathbf{b}^{(s)}\end{aligned} \tag{39}$$

If there is only one semi-tied class for all components of the regression class in question and the transformation matrix is sufficiently flexible, such that it is possible to model $\mathbf{A}^{(r)}\mathbf{H}^{(s)}\mathbf{A}^{(r)-1}$ then it is possible to write

$$\begin{aligned}\mathbf{A}^{(r)}\mathbf{W}^{(s)}\xi^{(m)} &= \mathbf{H}'^{(s)}\mathbf{A}^{(r)}\mu^{(m)} + \mathbf{A}^{(r)}\mathbf{b}^{(s)} \\ &= \mathbf{W}'^{(s)}\xi^{(rm)}\end{aligned} \tag{40}$$

Substituting this back into equation 38 gives

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = -\frac{1}{2}\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\Big[K^{(m)} + \log\left(\left|\boldsymbol{\Sigma}_{\text{diag}}^{(m)}\right|\right) - \\ \log\left(\left|\mathbf{A}^{(r)}\right|^2\right) + \left(\mathbf{A}^{(r)}\mathbf{o}(\tau) - \mathbf{W}'^{(s)}\xi^{(rm)}\right)^T\boldsymbol{\Sigma}_{\text{diag}}^{(m)-1}\left(\mathbf{A}^{(r)}\mathbf{o}(\tau) - \mathbf{W}'^{(s)}\xi^{(rm)}\right)\Big]\end{aligned} \tag{41}$$

This is identical to equation 32. Thus the two transforms will yield equivalent performance. The relation ship between the transforms is

$$\mathbf{H}'^{(s)} = \mathbf{A}^{(r)}\mathbf{H}^{(s)}\mathbf{A}^{(r)-1} \tag{42}$$

and

$$\mathbf{b}'^{(s)} = \mathbf{A}^{(r)}\mathbf{b}^{(s)} \tag{43}$$

where the $'$ indicates the normalised-domain transform.

D Variance Adaptation

This appendix considers the problem of variance adaptation when semi-tied full-covariance matrices are used. As usual the optimisation is performed in two stages [5] where initially, if necessary, the means are compensated to give $\hat{\mu}^{(m)}$. The variance transform is then obtained. As for the means there are a variety of ways that the variance may be transformed[10].

[10] For ease of notation the regression class indicator is dropped, thus $\mathbf{H}$ really represents the transform for regression class s, $\mathbf{H}^{(s)}$.

D.1 Choleski and Normalised-Domain Transforms

The simplest scheme is based on the Choleski factors as described in [5] where

$$\hat{\mathbf{\Sigma}}^{(m)-1} = \mathbf{L}^{(m)T}\mathbf{H}\mathbf{L}^{(m)} \tag{44}$$

where $\mathbf{L}^{(m)}$ is the Choleski factor of $\mathbf{\Sigma}^{(m)-1}$. Here the formulae detailed in [5] may be directly applied. Though easy to estimate this dramatically increases the recognition time computational cost [4].

Alternatively a normalised-domain version may be used. Here the inverse covariance matrix is given by

$$\hat{\mathbf{\Sigma}}^{(m)-1} = \left(\mathbf{H}\mathbf{A}^{(r)}\right)^T \mathbf{\Sigma}^{(m)-1}_{\text{diag}} \left(\mathbf{H}\mathbf{A}^{(r)}\right) \tag{45}$$

Using this form of transformation the auxiliary function may be written as

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = -\frac{1}{2}\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\Big[K^{(m)} + \log\left(\left|\mathbf{\Sigma}^{(m)}_{\text{diag}}\right|\right) - \quad (46)\\ \log\left(\left|\mathbf{A}^{(r)}\right|^2\right) - \log\left(|\mathbf{H}|^2\right) + \left(\mathbf{A}^{(r)}\mathbf{o}(\tau) - \hat{\mu}^{(rm)}\right)^T \mathbf{H}^T\mathbf{\Sigma}^{(m)-1}_{\text{diag}}\mathbf{H}\left(\mathbf{A}^{(r)}\mathbf{o}(\tau) - \hat{\mu}^{(rm)}\right)\Big]\end{aligned}$$

It is simple to see that this has the same form as the efficient full variance transform described in [2], other than working in the normalised domain. this has little increase in computational cost over the standard semi-tied models[11] as instead of transforming by $\mathbf{A}^{(r)}$ the data is transformed by $\mathbf{H}\mathbf{A}^{(r)}$. Note if the transform is constrained to be diagonal both the Choleski and normalised-domain transforms will yield the same transform.

D.2 Full Transform

The final option for the variance transform is to use

$$\hat{\mathbf{\Sigma}}^{(m)-1} = \left(\mathbf{A}^{(r)}\mathbf{H}\right)^T \mathbf{\Sigma}^{(m)-1}_{\text{diag}} \left(\mathbf{A}^{(r)}\mathbf{H}\right) \tag{47}$$

This has no simple equivalence to previously defined transforms. It is of interest as the previous two transforms make assumptions similar in many ways to the assumptions for normalise-domain MLLR, whereas no such assumptions made with this form. The following expression must be maximised

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = -\frac{1}{2}\sum_{r=1}^{R}\sum_{m\in M^{(rs)}}\sum_{\tau=1}^{T}\gamma_m(\tau)\Big[K^{(m)} + \log\left(\left|\mathbf{\Sigma}^{(m)}_{\text{diag}}\right|\right) - \quad (48)\\ \log\left(\left|\mathbf{A}^{(r)}\right|^2\right) - \log\left(|\mathbf{H}|^2\right) + \left(\mathbf{o}(\tau) - \hat{\mu}^{(m)}\right)^T \mathbf{H}^T\mathbf{A}^{(r)T}\mathbf{\Sigma}^{(m)-1}_{\text{diag}}\mathbf{A}^{(r)}\mathbf{H}\left(\mathbf{o}(\tau) - \hat{\mu}^{(m)}\right)\Big]\end{aligned}$$

Using the same equalities as described in [2] and only considering the first row of the transform $\mathbf{H}$, $\mathbf{h}_i$, (and ignoring all elements independent it)

$$\frac{\partial\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})}{\partial\mathbf{h}_i} = \beta\frac{\mathbf{c}_i}{\mathbf{c}_i\mathbf{h}_i^T} - \mathbf{h}_i\mathbf{G}^{(ii)} - \sum_{j\neq i}\mathbf{h}_j\mathbf{G}^{(ij)} \tag{49}$$

where

$$\mathbf{G}^{(ij)} = \sum_{m\in M^{(s)}} v_{ij}^{(m)}\sum_{\tau=1}^{T}\gamma_m(\tau)\mathbf{D}^{(m)}(\tau) \tag{50}$$

[11] This assumes that all the components from the same semi-tied class use the same variance transform.

$$\mathbf{D}^{(m)}(\tau) = \left(\mathbf{o}(\tau) - \hat{\mu}^{(m)}\right)\left(\mathbf{o}(\tau) - \hat{\mu}^{(m)}\right)^T \tag{51}$$

$$\mathbf{V}^{(m)} = \mathbf{\Sigma}^{(m)-1} \tag{52}$$

and

$$\beta = \sum_{m=1}^{M}\sum_{\tau=1}^{T}\gamma_m(\tau) \tag{53}$$

Only considering this row and equating to zero gives

$$\mathbf{c}_i\mathbf{h}_i^T\mathbf{h}_i = \left(\beta\mathbf{c}_i - \mathbf{c}_i\mathbf{h}_i^T\sum_{j\neq i}\mathbf{h}_j\mathbf{G}^{(ij)}\right)\mathbf{G}^{(ii)-1} \tag{54}$$

It is interesting to consider the various solutions of this equation. Only the first two are applicable to the semi-tied models, but the other two illustrates the relationships with formulae previously derived.

1. **Full transform, full variance**: This form is very similar to the constrained model-space optimisation described in [2]. Thus $\mathbf{h}_i$ may be written as

$$\mathbf{h}_i = \alpha\left(\mathbf{c}_i + \lambda\sum_{j\neq i}\mathbf{h}_j\mathbf{G}^{(ij)}\right)\mathbf{G}^{(ii)-1} \tag{55}$$

   This expression may then be substituted into equation 54 and equations for α and λ obtained (see [2] for details). This is again an iterative process as each row of $\mathbf{H}$ is dependent on the other rows according to the cofactors.

2. **Diagonal transform, full variance**: A simplified solution is possible when $\mathbf{H}$ is constrained to be diagonal. In this case

$$\frac{\partial\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})}{\partial h_{ii}} = \beta\frac{1}{h_{ii}} - \sum_{j=1}^{n}h_{jj}g_{ij}^{(ij)} \tag{56}$$

   Equating to zero and rearranging yields

$$h_{ii}^2 g_{ii}^{(ii)} + h_{ii}\sum_{j\neq i}h_{jj}g_{ij}^{(ij)} - \beta = 0 \tag{57}$$

   This may be easily solved iteratively, one element at a time.

3. **Full transform, diagonal variance**: For this case

$$\mathbf{G}^{(ij)}\begin{cases}\mathbf{G}^{(ii)}, & (i=j)\\ \mathbf{0}, & \text{otherwise}\end{cases} \tag{58}$$

   Therefore equation 55 simplifies to

$$\mathbf{h}_i = \alpha\mathbf{c}_i\mathbf{G}^{(ii)-1} \tag{59}$$

   The final solution is given by [2]

$$\mathbf{h}_i = \mathbf{c}_i\mathbf{G}^{(ii)-1}\sqrt{\left(\frac{\beta}{\mathbf{c}_i\mathbf{G}^{(ii)-1}\mathbf{c}_i^T}\right)} \tag{60}$$

4. **Diagonal transform, diagonal variance**: Equation 56 simplifies to

$$\frac{\partial \mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}})}{\partial h_{ii}} = \beta \frac{1}{h_{ii}} - h_{ii} g_{ii}^{(ii)} \tag{61}$$

Equating to zero and rearranging yields

$$h_{ii}^2 = \frac{\beta}{g_{ii}^{(ii)}} \tag{62}$$

In this case the new variance is given by

$$\hat{\sigma}_i^{(m)2} = \frac{\sigma_i^{(m)2}}{h_{ii}^2} \tag{63}$$

It is easy to see that this has the form given in [4].

References

[1] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[2] M J F Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report CUED/F-INFENG/TR291, Cambridge University, 1997. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.

[3] M J F Gales. Semi-tied full-covariance matrices for hidden Markov models. Technical Report CUED/F-INFENG/TR287, Cambridge University, 1997. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.

[4] M J F Gales and P C Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264, 1996.

[5] M J F Gales and P C Woodland. Variance compensation within the MLLR framework. Technical Report CUED/F-INFENG/TR242, Cambridge University, 1996. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.

[6] J L Gauvain and C H Lee. Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions Speech and Audio Processing*, 2:291–298, 1994.

[7] D M Hindle, A Ljolje, and M D Riley. Recent improvements in the AT&T speech-to-text (STT) system. In *Proceedings ARPA Speech Recognition Workshop*, 1996.

[8] N Kumar. *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD thesis, John Hopkins University, (to be published) 1997.

[9] C J Leggetter and P C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9:171–186, 1995.

[10] A Ljolje. The importance of cepstral parameter correlations in speech recognition. *Computer Speech and Language*, 8:223–232, 1994.

[11] L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, February 1989.

[12] B M Shahshahani. A Markov random field approach to bayesian speaker adaptation. In *Proceedings ICASSP*, pages 697–670, 1996.

[13] P C Woodland, J J Odell, V Valtchev, and S J Young. The development of the 1994 HTK large vocabulary speech recognition system. In *Proceedings ARPA Workshop on Spoken Language Systems Technology*, pages 104–109, 1995.

[14] S J Young, J Jansen, J Odell, D Ollason, and P Woodland. *The HTK Book (for HTK Version 2.0)*. Cambridge University, 1996.