INITIALIZING SUBSPACE CONSTRAINED GAUSSIAN MIXTURE MODELS

Peder A. Olsen, Karthik Visweswariah and Ramesh Gopinath

IBM T.J. Watson Research Center

{pederao,kv1,rameshg}@us.ibm.com

ABSTRACT

A recent series of papers [1, 2, 3, 4] introduced Subspace Constrained Gaussian Mixture Models (SCGMMs) and showed that SCGMMs can very efficiently approximate Full Covariance Gaussian Mixture Models (FCGMMs); a significant reduction in the number of parameters is achieved with little loss in the accuracy of the model. SCGMMs were arrived at as a sequence of generalizations of diagonal covariance GMMs. As an artifact of this process the initialization of SCGMM parameters in that work is complex i.e., relies on best parameter settings of less general models. This paper overcomes this problem by showing how an FCGMM can be used to give a simple and direct initialization of an SCGMM. The initialization scheme is powerful enough that as the number of parameters in an SCGMM approaches that of an FCGMM (i.e., large SCGMMs) further training of the SCGMM is unnecessary.

1. INTRODUCTION

In most state-of-the-art speech recognition systems, hidden Markov models (HMMs) are used to estimate likelihood of an acoustic observation given a word sequence. One of the key ingredients of the HMM models is a probability distribution $p(\mathbf{x}|s)$ for the acoustic vector $\mathbf{x} \in \mathbb{R}^d$ at particular time, conditioned on an HMM state s. Typically, $p(\mathbf{x}|s)$ is taken to be a Gaussian mixture model (GMM), or more generally, a mixture of exponential models:

$$P(\mathbf{x}|s) = \sum_{g \in s} \pi_g \mathcal{E}(\mathbf{x}; \boldsymbol{\theta}_g, \mathbf{f}), \tag{1}$$

where

$$\mathcal{E}(\mathbf{x}; \boldsymbol{\theta}, \mathbf{f}) = \frac{e^{\boldsymbol{\theta}^\top \mathbf{f}(\mathbf{x})}}{Z(\boldsymbol{\theta})}, \tag{2}$$

is the general exponential model and $Z(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} e^{\boldsymbol{\theta}^\top \mathbf{f}(\mathbf{x})} \mathrm{d}\mathbf{x}$ is the normalizer for the exponential distribution. From computational and storage considerations most speech recognition systems take $\mathcal{E}(\mathbf{x}; |\boldsymbol{\theta}_g, \mathbf{f})$ to be a diagonal Gaussian distribution. A recent series of papers [1, 2, 3, 4] introduced Subspace Constrained Gaussian Mixture Models (SCGMMs) that provide an efficient "slider" between Diagonal Covariance GMMs (DCGMMs) and Full Covariance GMMs (FCGMMs). In that work SCGMMs were arrived at via a sequence of generalizations of DCGMMs and hence, for historical reasons, the initialization of the parameters of SCGMMs was complex i.e., relied on the best available parameter settings of less general models. Effectively with that approach one needed to have training software for less general models in order to arrive at an initialization for an SCGMM. This paper overcomes that problem by providing a simple and direct method to initialize parameters of an SCGMM model.

A full covariance gaussian

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \tag{3}$$

can be written in form of an exponential model as follows:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \mathcal{E}(\mathbf{x}; \boldsymbol{\theta}_{\mathrm{fc}}, \mathbf{f}_{\mathrm{fc}}) = \frac{e^{\boldsymbol{\theta}_{\mathrm{fc}}^\top \mathbf{f}_{\mathrm{fc}}(\mathbf{x})}}{Z_{\mathrm{fc}}(\boldsymbol{\theta}_{\mathrm{fc}})}, \tag{4}$$

where we define the full covariance features $\mathbf{f}_{\mathrm{fc}}$ to be

$$\mathbf{f}_{\mathrm{fc}}(\mathbf{x}) = (\mathbf{x}^\top, -\mathrm{vec}(\mathbf{x}\mathbf{x}^\top)^\top)^\top, \tag{5}$$

and vec is an operator on symmetric matrices defined as a vector containing the elements of the lower triangular portion with the diagonal scaled by $1/\sqrt{2}$

$$\mathrm{vec}(\mathbf{X}) = (\frac{X_{11}}{\sqrt{2}}, X_{12}, \frac{X_{22}}{\sqrt{2}}, X_{13}, \ldots, \frac{X_{dd}}{\sqrt{2}})^\top. \tag{6}$$

It can be verified that in terms of these features the model parameters can be written $\boldsymbol{\theta}_{\mathrm{fc}} = (\boldsymbol{\psi}^\top, \mathbf{p}^\top)^\top \in \mathbb{R}^{(d+1)(d+2)/2}$, where the model parameters $\boldsymbol{\psi} \in \mathbb{R}^d$ and $\mathbf{p} \in \mathbb{R}^{d(d+1)/2}$ corresponds to the linear and quadratic features. In terms of the precision matrix $\mathbf{P} = \boldsymbol{\Sigma}^{-1}$ the quadratic model parameters $\mathbf{p}$ are

$$\mathbf{p} = \mathrm{vec}(\mathbf{P}) \tag{7}$$

and the linear model parameters are

$$\boldsymbol{\psi} = \mathbf{P}\boldsymbol{\mu}. \tag{8}$$

The normalizer for the full covariance model in terms of $\mathbf{P}$ and $\boldsymbol{\psi}$ is

$$2\log(Z(\boldsymbol{\theta})) = \log\det\left(\frac{\mathbf{P}}{2\pi}\right) - \boldsymbol{\psi}^\top \mathbf{P}^{-1}\boldsymbol{\psi}\ . \tag{9}$$

A Subspace Constrained Gaussian, [1], is an exponential model with features $\boldsymbol{\Phi}\mathbf{f}_{\mathrm{fc}}(\mathbf{x}) \in \mathbb{R}^D$, where $\boldsymbol{\Phi} \in \mathbb{R}^{D \times (d+1)(d+2)/2}$ and $\boldsymbol{\lambda} \in \mathbb{R}^D$. $\mathcal{S}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\Phi})$ denotes the Subspace Constrained Gaussian and satisfies the relation

$$\mathcal{S}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\Phi}) = \mathcal{E}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\Phi}\mathbf{f}_{\mathrm{fc}}). \tag{10}$$

This paper will describe how, with minimal computational effort, one can obtain a good initial value for the basis matrix $\boldsymbol{\Phi}$ and the exponential model parameters $\boldsymbol{\lambda}_g$, $g \in s$ from an FCGMM. As the name suggests the SCGMM can be viewed as an FCGMM, where the full covariance exponential model parameters are constrained to be in a subspace, i.e. $\boldsymbol{\theta}_g = \boldsymbol{\lambda}_g{}^\top \boldsymbol{\Phi} = \sum_{k=1}^D \lambda_{gk}\boldsymbol{\phi}_k$, where $\boldsymbol{\phi}_k$ is the vector corresponding to the kth row of $\boldsymbol{\Phi}$.

Section 2 describes how to find a good initial basis $\boldsymbol{\Phi}$, Section 3 describes how to initialize $\boldsymbol{\lambda}_g$ and Section 4 gives experimental results with the new initialization scheme.

1.1. Previous Methods to Initialize SCGMM models

In [1] the SCGMM model was initialized with a Subspace Precision And Mean (SPAM) model. A SPAM model is a special case of an SCGMM model, where the basis matrix $\mathbf{\Phi}$ is block diagonal of the form

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_{22} \end{pmatrix},$$

where $\mathbf{\Phi}_{11} \in \mathbb{R}^{D_1 \times d}$ corresponds to the linear features and $\mathbf{\Phi}_{22} \in \mathbb{R}^{D_2 \times d(d+1)/2}$ corresponds to the quadratic features and $D_1 + D_2 = D$. The SPAM model used to initialize the SCGMM model was trained with the methods described in [5]. This SPAM model was initialized using a method that used a quadratic approximation of the log likelihood function as described in [2]. In this paper we avoid the use of a SPAM model altogether and use the ideas of [2, 4] to directly initialize the SCGMM model.

2. SUBSPACE APPROXIMATIONS

The parameters of a GMM are trained typically to maximize the likelihood of the training data $\{\mathbf{x}_k\}_{k=1}^N$. One starts with an initial value for the parameters and then iteratively updates them using the Expectation Maximization algorithm [6] with each sweep of the data. While the initialization problem for DCGMMs and FCGMMs is trivial, for SCGMMs it is complicated by the fact that one requires an initial choice for both the basis $\mathbf{\Phi}$ and the exponential model parameters $\boldsymbol{\lambda}_g$. Our approach is to find a suitable quadratic approximation to the likelihood function and then to explicitly optimize this quadratic function to obtain an initial value for the SCGMM parameters. Additionally, in our approach the initialization process does not go through the data. Instead, one starts with a full-trained FCGMM model and constructs an SCGMM model by a quadratic approximation. More precisely, if $\boldsymbol{\theta}_g$ represents a Gaussian g in the FCGMM, then we can solve:

$$\min_{\boldsymbol{\lambda}_g, \mathbf{\Phi}} \sum_g w_g \|\boldsymbol{\theta}_g - \sum_{k=1}^D \lambda_{gk} \boldsymbol{\phi}_k\|_2^2. \tag{11}$$

The basis vectors, $\{\boldsymbol{\phi}_k\}_{k=1}^D$, minimizing this problem are the principal components of $\boldsymbol{\theta}_g$. That is $\boldsymbol{\phi}_k$, $k = 1, \ldots, D$ corresponds to the eigenvectors with the top D largest eigenvalues of the covariance matrix of $\{\boldsymbol{\theta}_g\}_{g=1}^G$ with weights w_g, $1 = \sum_g w_g$. Since we know how to solve this problem efficiently we can use this solution as an initial choice for optimizing the likelihood over the data. Notice that if the norm in (11) is replaced by another norm, e.g. $\|\boldsymbol{\theta}\|_{\mathbf{A}}^2 = \boldsymbol{\theta}^\top \mathbf{A} \boldsymbol{\theta}$, where $\mathbf{A}$ is a positive definite symmetric matrix, the solution can once again be found by noting that the transform $T(\boldsymbol{\theta}_g) = \mathbf{A}^{1/2} \boldsymbol{\theta}_g$ changes the norm back to the Euclidean distance, i.e. $\|\boldsymbol{\theta}\|_{\mathbf{A}}^2 = \|T(\theta_g)\|_2^2$. This flexibility in the choice of $\mathbf{A}$ suggests that we should choose it so that the likelihood criterion is well-approximated by the quadratic function implied by (11).

The covariance matrix whose eigenvalues and eigenvectors we are seeking is

$$\mathbf{\Sigma}_\theta = \sum_g w_g (T(\boldsymbol{\theta}_g) - \boldsymbol{\mu}_\theta)(T(\boldsymbol{\theta}_g) - \boldsymbol{\mu}_\theta)^T, \tag{12}$$

where

$$\boldsymbol{\mu}_\theta = \sum_g w_g T(\boldsymbol{\theta}_g). \tag{13}$$

Note that the size of the covariance matrix $\mathbf{\Sigma}_\theta \in \mathbb{R}^{(d+1)(d+2)/2 \times (d+1)(d+2)/2}$ is growing quadratically in d and it is difficult to handle the eigenvalue problem for values $d > 200$. In the case we consider in this paper $d = 40$ and so we are okay.

2.1. The Expectation Maximization Algorithm

To determine a good choice for the metric matrix $\mathbf{A}$ we shall approximate the likelihood with a quadratic function of the parameters. To do this we need to review how the parameters are estimated in the Expectation Maximization (EM) algorithm, [6]. The EM algorithm introduces an auxilliary function $Q(\Theta, \hat{\Theta})$; where Θ, $\hat{\Theta}$ denotes model parameters $\{\pi_g, \boldsymbol{\theta}_g\}_g$ and $\{\hat{\pi}_g, \hat{\boldsymbol{\theta}}_g\}_g$ respectively. The auxilliary function satisfies $Q(\Theta, \Theta) = 0$ and $L(\Theta) - L(\hat{\Theta}) \geq Q(\Theta, \hat{\Theta})$ where $L(\Theta) = \sum_t \log p(\mathbf{x}_t | s_t)$ is the log likelihood of the training data. The auxilliary function is given by

$$\begin{aligned} Q(\Theta, \hat{\Theta}) &= \sum_t \sum_{g \in s_t} \gamma_{tg} \log \frac{\pi_g \mathcal{E}(\mathbf{x}_t; \boldsymbol{\theta}_g, \mathbf{f})}{\hat{\pi}_g \mathcal{E}(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_g, \mathbf{f})} \\ &= -\sum_g n(g) \ell_g(\Theta), \end{aligned} \tag{14}$$

where γ_{tg} are the occupation counts

$$\gamma_{tg} = \begin{cases} \frac{\hat{\pi}_g \mathcal{E}(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_g, \mathbf{f})}{\sum_{g^* \in s_t} \hat{\pi}_{g^*} \mathcal{E}(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_g^*, \mathbf{f})} & \text{if} \quad g \in s_t \\ 0 \quad \text{otherwise,} \end{cases}$$

$n(g) = \sum_t \gamma_{tg}$ and

$$\ell_g(\Theta) = -\frac{1}{n(g)} \sum_t \gamma_{tg} \log \frac{\pi_g \mathcal{E}(\mathbf{x}_t; \boldsymbol{\theta}_g, \mathbf{f})}{\hat{\pi}_g \mathcal{E}(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_g, \mathbf{f})}. \tag{15}$$

To improve the likelihood $L(\Theta) > L(\hat{\Theta})$ it is sufficient to maximize the auxilliary function $Q(\Theta, \hat{\Theta})$ with respect to Θ. The maximum value with respect to the priors, means and variances for an FCGMM is given by:

$$\begin{aligned} \tilde{\pi}_g &= \frac{n(g)}{\sum_{g^* \in s} n(g^*)}, \\ \boldsymbol{\mu}_g &= \frac{1}{n(g)} \sum_t \gamma_{tg} \mathbf{x}_t \qquad \text{and} \\ \mathbf{\Sigma}_g &= \frac{1}{n(g)} \sum_t \gamma_{tg} (\mathbf{x}_t - \boldsymbol{\mu}_g)(\mathbf{x}_t - \boldsymbol{\mu}_g)^T. \end{aligned} \tag{16}$$

Dropping terms that are independent of $\boldsymbol{\theta}_g$ and optimizing with respect to π_g in (15) we get

$$\ell_g(\Theta) = \boldsymbol{\theta}_g^\top \frac{1}{n(g)} \sum_t \gamma_{tg} \mathbf{f}(\mathbf{x}_t) - \tilde{\pi}_g \log Z(\boldsymbol{\theta}_g), \tag{17}$$

where $\tilde{\pi}_g$ is given in (16).

2.2. Determining the Frobenius Norm

As we iterate the update formulas of the EM algorithm to convergence the gradient of (17) approaches zero. Thus for a value

$\tilde{\Theta}$ close to the optimal Θ the following quadratic approximation holds by virtue of the Taylor series of $\ell_g(\tilde{\Theta})$:

$$\ell_g(\tilde{\Theta}) \approx -(\boldsymbol{\theta}_g - \tilde{\boldsymbol{\theta}}_g)^\top \mathbf{H}_{\boldsymbol{\theta}_g} (\boldsymbol{\theta}_g - \tilde{\boldsymbol{\theta}}_g), \tag{18}$$

where $\mathbf{H}_{\boldsymbol{\theta}_g}$ is the Hessian of $-\log Z(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_g$. Setting $\tilde{\boldsymbol{\theta}}_g$ to be the subspace constrained gaussians approximation to the full covariance gaussians we see that the auxilliary function can be approximated by the quadratic form $-\sum_g \tilde{\pi}_g \|\tilde{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\|^2_{\mathbf{H}_{\boldsymbol{\theta}_g}}$. Ignoring the sign the preceeding quadratic form fails to conform to (11) and so we make an additional approximation by replacing $\mathbf{H}_{\boldsymbol{\theta}_g}$ with $\mathbf{H}_{\bar{\boldsymbol{\theta}}}$, where $\bar{\boldsymbol{\theta}}$ is some representative "best" approximation to $\boldsymbol{\theta}_g$. To choose the representative $\bar{\boldsymbol{\theta}}$ in a principled fashion we choose it so as to minimize the average Kullback Leibler divergence $\min_{\bar{\boldsymbol{\theta}}} \sum_g \pi_g D(\boldsymbol{\theta}_g \| \bar{\boldsymbol{\theta}})$, where

$$D(\boldsymbol{\theta}_g \| \bar{\boldsymbol{\theta}}) = \log\left(\frac{Z(\bar{\boldsymbol{\theta}})}{Z(\boldsymbol{\theta}_g)}\right) + (\boldsymbol{\theta}_g - \bar{\boldsymbol{\theta}}) E_{\boldsymbol{\theta}_g}[\mathbf{f}(\mathbf{x})]$$

and

$$E_{\boldsymbol{\theta}_g}[\mathbf{f}(\mathbf{x})] = \begin{pmatrix} \boldsymbol{\mu}_g \\ \mathrm{vec}(\boldsymbol{\Sigma}_g + \boldsymbol{\mu}_g \boldsymbol{\mu}_g^\top) \end{pmatrix}.$$

The solution to this problem is the $\bar{\boldsymbol{\theta}}$ that corresponds to the total mean and covariance of the gaussians, i.e.

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= \sum_g \pi_g \boldsymbol{\mu}_g \\ \bar{\boldsymbol{\Sigma}} &= \sum_g \pi_g (\boldsymbol{\Sigma}_g + \boldsymbol{\mu}_g \boldsymbol{\mu}_g^\top) - \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}^\top. \end{aligned}$$

We will write $\bar{\mathbf{P}} = \bar{\boldsymbol{\Sigma}}^{-1}$ and $\bar{\boldsymbol{\psi}} = \bar{\mathbf{P}} \bar{\boldsymbol{\mu}}$ for the corresponding exponential model parameters.

Now the Hessian $\mathbf{H}_{\bar{\boldsymbol{\theta}}}$ is a matrix of size $(d+1)(d+2)/2 \times (d+1)(d+2)/2$. Thus the computation cost of computing $\mathbf{H}_{\bar{\boldsymbol{\theta}}}^{1/2}$ will have a cost similar to the cost of computing the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. However we shall see that the computation of $\mathbf{H}_{\bar{\boldsymbol{\theta}}}^{1/2} \boldsymbol{\theta}_g$ can be simplified considerably.

2.3. Finding a Square Root of the Hessian

For a symmetric matrix $\mathbf{X}$ with eigenvalues $e_i > -1$ we have the following relation

$$\begin{aligned} \log\det(\mathbf{I} + \mathbf{X}) &= \sum_i \log(1 + e_i) \\ &= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \sum_i e_i^k \\ &= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \mathrm{trace}(\mathbf{X}^k). \end{aligned}$$

Using the above relation up to order 2 in $\mathbf{X}^k$ yields the expression

$$\log\det(\mathbf{P} + \Delta\mathbf{P}) \approx \log\det\mathbf{P} + \mathrm{trace}\mathbf{X} - \frac{1}{2}\mathrm{trace}\mathbf{X}^2,$$

where $\mathbf{X} = \mathbf{P}^{-1/2} \Delta\mathbf{P} \mathbf{P}^{-1/2}$. The Hessian can now be computed by identifying second order terms in the Taylor expansion of $\log Z(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})$. We find

$$\Delta\boldsymbol{\theta}^\top \mathbf{H}_{\boldsymbol{\theta}} \Delta\boldsymbol{\theta} = \frac{1}{2}\mathrm{trace}\mathbf{X}^2 + \|\mathbf{P}^{-1/2}\Delta\boldsymbol{\psi} - \mathbf{X}\mathbf{P}^{-1/2}\boldsymbol{\psi}\|^2.$$

Thus the linear transform T,

$$T(\Delta\theta_{\mathrm{fc}}) = \begin{pmatrix} \mathbf{P}^{-1/2}(\Delta\boldsymbol{\psi} - \Delta\mathbf{P}\mathbf{P}^{-1}\boldsymbol{\psi}) \\ \mathrm{vec}(\mathbf{P}^{-1/2}\Delta\mathbf{P}\mathbf{P}^{-1/2}) \end{pmatrix},$$

has the property that $\|\Delta\boldsymbol{\theta}\|_{\mathbf{H}_{\boldsymbol{\theta}}} = \|T(\Delta\boldsymbol{\theta})\|_2$, and computing this transform only involves computing the square root, inverting matrice and simple multiplication involving matrices of size $d \times d$. This is considerably cheaper than the direct computation of $\mathbf{H}_{\bar{\boldsymbol{\theta}}}^{1/2}$.

2.4. The Basis Selection Algorithm

For a subspace constrained gaussian to be well defined, the precision parameters must correspond to a positive definite matrix. This however, is not automatically the case in the suggested algorithm. By experimental verification we verified that even the top eigenvector in the above scheme did *not* correspond to a positive definite precision matrix. Motivated by this we consider instead an SCGMM with an affine basis $\boldsymbol{\theta}_0 + \sum_{k=1}^{D} \lambda_{gk} \boldsymbol{\theta}_k$ with an offset $\boldsymbol{\theta}_0$ corresponding to a positive definite covariance matrix. We use $\boldsymbol{\theta}_0 = \bar{\boldsymbol{\theta}}$.

In summary the algorithm for basis selection becomes

1. Compute $\bar{\boldsymbol{\theta}}$ and the related quantities $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{\Sigma}}$, $\bar{\boldsymbol{\psi}}$, $\bar{\mathbf{P}}$, $\bar{\mathbf{P}}^{1/2}$ and $\bar{\mathbf{P}}^{-1/2}$.
2. Transform the full covariance parameters $\boldsymbol{\theta}_g = ({\boldsymbol{\psi}_g}^\top, {\mathbf{p}_g}^\top)^\top$ by the formula

$$T(\boldsymbol{\theta}_g) = \theta_g^1 = \begin{pmatrix} \boldsymbol{\psi}_g^1 \\ \mathbf{p}_g^1 \end{pmatrix}$$

where

$$\begin{pmatrix} \boldsymbol{\psi}_g^1 \\ \mathbf{p}_g^1 \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{P}}^{-\frac{1}{2}}(\boldsymbol{\psi}_g - \mathbf{P}_g \bar{\boldsymbol{\mu}}) \\ \mathrm{vec}(\bar{\mathbf{P}}^{-1/2} \mathbf{P}_g \bar{\mathbf{P}}^{-1/2}) \end{pmatrix}. \tag{19}$$

3. Compute the mean, $\boldsymbol{\mu}_\theta$, and the covariance, $\boldsymbol{\Sigma}_\theta$, of the transformed full covariance parameters. Then compute the eigenvectors $\mathbf{e}_k = (\boldsymbol{\psi}_k^e, \mathbf{p}_k^e)$, $k = 1, \ldots, (d+1)(d+2)/2$, and associated eigenvalues e_k.
4. Invert the transform T on the eigenvectors to compute the SCGMM basis:

$$\begin{aligned} \boldsymbol{\phi}_1 &= \boldsymbol{\theta}_0 \\ \boldsymbol{\phi}_{k+1} &= T^{-1}(\mathbf{e}_k) = \begin{pmatrix} \boldsymbol{\psi}_k^{e2} \\ \mathbf{p}_k^{e2} \end{pmatrix}, \end{aligned}$$

where

$$\begin{pmatrix} \boldsymbol{\psi}_k^{e2} \\ \mathbf{p}_k^{e2} \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{P}}^{1/2}(\boldsymbol{\psi}_k^e + \mathbf{P}_k^e \bar{\mathbf{P}}^{1/2} \bar{\boldsymbol{\mu}}) \\ \mathrm{vec}\left(\bar{\mathbf{P}}^{1/2} \mathbf{P}_k^e \bar{\mathbf{P}}^{1/2}\right) \end{pmatrix}.$$

2.4.1. Feature Transform Interpretation

An interesting property of the model transform for the gaussians given by (19) is that it corresponds to the data transform $\mathbf{x}_t \rightarrow \bar{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}_t - \bar{\boldsymbol{\mu}})$. This transform normalizes the data to have zero mean and unit covariance. Such a normalization is a good method to avoid many numerical problems. If we normalize the data in this way then steps 2 and 4 in the above procedure becomes unnecessary!

3. INITIALIZING SCGMM COEFFICIENTS

The value of $\boldsymbol{\lambda}_g$ minimizing the Frobenius norm $\|\boldsymbol{\theta}_g - \boldsymbol{\lambda}_g \boldsymbol{\Phi}\|_{\mathbf{H}_{\bar{\theta}}}$ is given by

$$\boldsymbol{\lambda}_g = (\boldsymbol{\Phi}^\top \mathbf{H}_{\bar{\theta}} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{H}_{\bar{\theta}} \boldsymbol{\theta}_g,$$

where $\boldsymbol{\Phi}$ is the SCGMM basis matrix consisting of the rows $\boldsymbol{\phi}_k$. Let $\mathbf{E}$ be the matrix consisting of the rows $\mathbf{e}_k = T(\boldsymbol{\phi}_k)$, $k = 1, \ldots, D$. The coefficients $\boldsymbol{\lambda}_g$ can then be efficiently computed by the formula

$$\boldsymbol{\lambda}_g = (\mathbf{E}^\top \mathbf{E})^{-1} \mathbf{E}^\top T(\boldsymbol{\theta}_g). \tag{20}$$

However minimizing the Frobenius norm may lead to an approximate model $\tilde{\boldsymbol{\theta}}_g = \sum_{k=1}^{D} \lambda_{gk} \boldsymbol{\phi}_k$ for which the corresponding precision matrix is *not positive definite*. Such a model would not correspond to a well defined distribution, and so any such value of $\boldsymbol{\lambda}_g$ must be discarded. We propose two "back-off" methods to find alternative values of $\boldsymbol{\lambda}_g$ that are guaranteed to be good. The first method is to only project onto the first basis element, whose precision is the inverse of the total covariance and as such is positive definite (unless the data is very sparse or degenerate). As this first method is somewhat crude, we also propose using a method known as Projection Onto Convex Sets (POCS). The POCS algorithm is described in the appendix of [4] and consists of alternatingly projecting onto $\mathcal{S} = \{\mathbf{P} : \mathbf{P} = \sum_{k=1}^{D} \lambda_k \boldsymbol{\phi}_k\}$ and $\mathcal{P}_t = \{\mathbf{P} : \mathbf{P} \geq t\mathbf{I}\}$ for some chosen fixed value $t > 0$. Projecting onto $\mathcal{P}_t$ is a simple matter of changing all eigenvalues below t to t. Since this algorithm may in general converge slowly we backed off to the first method if after a small modest number of initial iterations the method did not yield an element in $\mathcal{S} \cap \mathcal{P}_0$.

Initializing $\boldsymbol{\lambda}_g$ like this works reasonably well for large values of D ($D \gg d$) as the initial estimate given by 20 would then yield a $\tilde{\boldsymbol{\theta}}_g$ that is close to $\boldsymbol{\theta}_g$. However for smaller values of D the method should only be considered as a technique to initialize $\boldsymbol{\lambda}_g$ to yield a positive definite precision. But, as the techniques for training $\boldsymbol{\lambda}_g$ described in [1, 4] converges quite fast, this is not a serious problem.

4. EXPERIMENTS

The SCGMM basis initialisation introduced in [1, 4] was quite circuitous and required an initial Subspace Precision And Mean (SPAM) model. The method proposed in section 3 has the advantage that it does not require a SPAM model at all and is essentially no more complicated than initializing a SPAM model. To measure the power of the SCGMM initialization scheme we computed the WER for: 1) The initial SCGMM model, 2) $\boldsymbol{\lambda}_g$ coefficients retrained by maximum likelihood and 3) SCGMM coefficients and basis elements retrained by maximum likelihood.

Neither of the three models above saw any training data, and maximum likelihood training was done against statistics inferred from the baseline full covariance model. The experiments used the same IBM internal training and test corpora that [1, 4] reported results on. The baseline 40 dimensional 10,000 component full and diagonal covariance acoustic models have error rates that are respectively 1.23% and 2.28%. These acoustic models are substantially better than the comparable acoustic models reported on in [1, 4]. We can see in Table 1 how the baseline systems compare to SCGMM models of various basis sizes.

For $D = 80$ the fully trained SCGMM model gives a word error rate of 1.68% and substantially outperforms the diagonal covariance SCGMM model which has a comparable number of free parameters. Also, the comparable SCGMM system that only retrained the SCGMM coefficients yields a word error rate comparable to the fully trained SCGMM model for all $D > 40$. For $D = 640$ no training at all appears to be necessary!

Basis size	ML retrained parameters		
D	none	$\boldsymbol{\lambda}_g$	$\boldsymbol{\lambda}_g$ and basis
10	79.11%	13.87%	5.17%
20	21.79%	4.23%	2.96%
40	9.63%	2.26%	2.08%
80	5.41%	1.73%	1.68%
160	2.40%	1.44%	1.43%
320	1.61%	1.34%	1.32%
640	1.31%	1.28%	1.26

Table 1. Word Error rates for various types of SCGMM models

5. CONCLUSION

We have proposed a simple method to initialize SCGMM basis and coefficient parameters. The method is simple to implement and is in the example $d = 40$ experiment by itself sufficient for $D = 640$ ($D \gg d$). For all values $D \geq 80$ the basis training itself may be avoided altogether with only a small loss in performance. And only the SCGMM coefficients need to be trained in this case.

6. REFERENCES

[1] Karthik Visweswariah, Scott Axelrod, and Ramesh Gopinath, "Acoustic modeling with mixtures of subspace constrained exponential model," in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, September 2003, vol. 3, pp. 2613–2616.

[2] Scott Axelrod, Ramesh Gopinath, and Peder Olsen, "Modeling with a subspace constraint on inverse covariance matrices," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, September 2002, vol. 3, pp. 2177–2180.

[3] Scott Axelrod, Vaibhava Goel, Ramesh A. Gopinath, Peder A. Olsen, and Karthik Visweswariah, "Constrained gaussian mixture models for speech recognition," *Transactions in Speech and Audio Processing*, 2003, submitted.

[4] Scott Axelrod, Vaibhava Goel, Ramesh A. Gopinath, Peder A. Olsen, and Karthik Visweswariah, "Subspace constrained gaussian mixture models for speech recognition," *Transactions in Speech and Audio Processing*, 2004, to appear.

[5] Scott Axelrod, Ramesh Gopinath, Peder Olsen, and Karthik Visweswariah, "Dimensional reduction, covariance modeling and computational complexity in asr systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003, IEEE, vol. I, pp. 912–915.

[6] 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*, vol. 39, no. B, pp. 1–38, 1977.