Discriminative Estimation of Subspace Constrained Gaussian Mixture Models for Speech Recognition – 3/8/2004 17:29

Scott Axelrod, *Member, IEEE,* Vaibhava Goel, *Member, IEEE,* Ramesh Gopinath, *Member, IEEE,* Peder Olsen, *Member, IEEE,* and Karthik Visweswariah, *Member, IEEE*

Abstract

In this paper we study discriminative training of acoustic models for speech recognition under two criteria: maximum mutual information (MMI) and a novel "error-weighted" training technique. We present a general proof that the standard MMI training technique is valid for arbitrary model types, that is for any kind of acoustic models with any kind of parameter tying. We report experimental results for subspace constrained Gaussian mixture models (SCGMMs), where the exponential model weights of all Gaussians are required to belong to a common "tied" subspace, as well as for SPAM models which impose separate subspace constraints on the precision matrices (i.e. inverse covariance matrices) and means. It has been shown previously that SCGMMs and SPAM models generalize and yield significant error rate improvements over previously considered model classes such as diagonal models, models with semi-tied covariances, and EMLLT (extended maximum likelihood linear transformation) models. We show here that MMI and error weighted training each individually result in over 20% relative reduction in word error rate on a digit task over maximum likelihood (ML) training. We also show that a gain of as much as 28% relative can be achieved by combining these two discriminative estimation techniques.

Discriminative Estimation of Subspace Constrained Gaussian Mixture Models for Speech Recognition – 3/8/2004 17:29

I. INTRODUCTION

In most of the state-of-the-art speech recognition systems, hidden Markov models (HMMs) are used to estimate the likelihood of an acoustic observation given a word sequence. One of the ingredients of the HMM models is a probability distribution $p(x|s)$ for the acoustic vector $x \in \mathbf{R}^d$ at a particular time, conditioned on an HMM state s. Typically, $p(x|s)$ is taken to be a Gaussian mixture model (GMM). The mean and covariance of each Gaussian in the mixture belongs in general to $\mathbf{R}^d$ and the space of symmetric positive definite $d \times d$ matrices, respectively. In practice, however, constraints are needed, especially on covariances, to allow for robust estimation, efficient storage, and efficient computations. The most common constraint is to restrict Σ to the space of diagonal positive definite matrices. Other recently proposed model types yield significant speed and accuracy gains by placing a subspace constraint on the inverse covariance matrices (also called precision matrices). In order of least to most general, such models include: semi-tied covariance [1] or maximum likelihood linear transformation (MLLT) [2] models for which the basis in which the precision matrices are diagonal is trained; extended MLLT (EMLLT) [3], [4] models which constrain the precision matrices to be a linear combination of rank one matrices; affine EMLLT [5] models which add a constant (the affine basepoint) to the precision matrix for all Gaussians; SPAM [6] models which place a general subspace constraint on the precision matrices and means of all the Gaussians; and finally subspace constrained Gaussian mixture models (SCGMMs) [7] in which the exponential models parameters (i.e. the mean and precision matrix combined into a single vector) for all of the Gaussians are required to lie in a common affine subspace of $\mathbf{R}^{d+d(d+1)/2}$. A comprehensive review of all these models types including their relationship to linear discriminant analysis and experiment results on large and small vocabulary tasks was presented in [5]. In related work, a special case of the SPAM model in which the precision matrices are constrained to be a linear combination of tied positive definite matrices was considered in [8].

The parameters of the state dependent Gaussian mixture models are commonly obtained using maximum likelihood (ML) estimation which aims at selecting model parameters that results in the highest likelihood of acoustic training data given its labeled word sequence. In other words, if (X^*, W^*) denotes the acoustic training data and its labeled word sequence, then the ML estimate is $\Omega_{ml} = \operatorname{argmax} P(X^*|W^*; \Omega)$ where the model parameters Ω are varied over a set of admissible values. ML estimation is *generative* in nature - it selects a parameter value that best explains the observed acoustics as generated by the labeled word sequence. It does not take into account the likelihood of the observation under other word sequences. It is typically carried out using the Baum-Welch

or expectation-maximization (EM) algorithm [9] in which parameters are trained iteratively, with the parameter update at each iteration guaranteed to improve the ML objective function. An in-depth review of ML estimation for different types of subspace constrained Gaussian models was presented in [5]. In particular, that reference discusses training of both the "tied" parameters, which specify the constraining subspace, and the "untied" parameters, which specify the location of each Gaussian within the common subspace (as well as the Gaussian priors).

An alternative to ML training is *discriminative* estimation which refers to a set methods that consider likelihood of observed data under the labeled as well as alternative word sequences. Many references have shown that models trained discriminatively are more accurate than those trained by ML (at fixed computational speed and memory cost). Many variants of the discriminative objective functions and the training procedure have been considered including: (i) maximum mutual information (MMI) estimation [10] whereby the parameters are selected so as to maximize the mutual information between the acoustic training data and the labeled word sequence ($\mathrm{argmax}\ P(X^*, W^*; \Omega)/P(X^*; \Omega)P(W^*; \Omega)$); (ii) conditional maximum likelihood (CML) training [11] whereby the conditional likelihood of labeled word sequence is maximized given the acoustic training data ($\mathrm{argmax}\ P(W^*|X^*; \Omega)$); (iii) minimum classification error (MCE) training [12] which attempts to directly reduce the classification errors on the training set; and (iv) corrective training [13] that, qualitatively speaking, carries out ML estimation with more emphasis on parts of the training data that are in error. We note that in case the distribution $P(W^*)$ is not a function of the parameters, the CML and MMI criteria differ only be a constant and therefore lead to identical parameter estimates. In such cases CML and MMI are often used interchangeably in the literature. We refer the reader to Dan Povey's Ph. D. thesis [14] for a wealth of information about discriminative training.

In this paper we apply discriminative estimation to subspace constrained Gaussian mixture models (SCGMMs). In particular we focus on two methods - MMI training and a procedure that is quite similar to corrective training [13] which we call *error weighted training* [15]. This paper extends our previous work on discriminative estimation of SPAM models [15] to the most general form of subspace constrained Gaussian mixture models.

MMI estimation was first proposed for HMMs with discrete distributions by Bahl et.al. [16] and a popular implementation, based on a growth transform due to Baum and Eagon [17], was proposed by Gopalakrishnan, Kanevsky, Nadas and Nahamoo [18]. Extension of MMI estimation to continuous distributions was carried out by Normandin [19]; his approach was to discretize the continuous distributions, carry out the iterative estimation procedure of Gopalakrishnan and Byrne [18] in discrete domain, and derive the estimates for parameters of the continuous distributions as a limiting case of the discrete updates. It is these estimates and their heuristic variants that are commonly used to update GMM parameters [10]. However, a shortcoming of Normandin's derivations is that the validity of the estimates is not rigorously established.

An alternate MMI parameter estimation method that combines ideas from [18] and auxiliary function approach of EM was recently proposed by Gunawardana and Byrne [20]. Their procedure results in estimates that are identical to the updates given by Normandin, but it offers a significant advantage - the auxiliary function constructed for MMI estimation is much like the auxiliary function for ML estimate. Consequently, the MMI and ML estimation

can be carried out using similar optimization procedures. As with Normandin's approach, Gunawardana's approach introduces an auxiliary parameter D and provide intuition, but no rigorous proof, that the MMI update formula obtained when D is large does indeed increase the MMI objective function. Recently, Kanevsky [21] has proved that these update formulas do in fact rigorously guarantee an increase of the MMI objective function for sufficiently large D. In related work, Jebara and Pentland [22] have given alternative update formulas for discriminative estimation of mixtures of exponential models used for binary classification.

For MMI estimation of GMMs, we follow here a variant of the auxiliary function based method of Gunawardana and Byrne [20]. One of the main contributions of this paper is to give a proof of validity of the update formula; that is to say we prove that maximizing the auxiliary function increases the MMI objective function value. Our proof is more general than the one in [21], which applies only to diagonal GMMs, although with a slightly generalized notion of objective function. Also our proof avoids the need to perform detailed estimates. We are thus able to obtain and prove validity of the update formula for the tied subspace. Furthermore, the form of the auxiliary "Q" function is identical to that used in the EM algorithm for ML training, so that the final step of training, maximizing the Q function, is identical as in the ML case, which has already been discussed in detail in [5].

Our proof of validity of the MMI update formula relies (as in [21]) on the choice of a large enough value for an auxiliary parameter D. In practice in our experiments, we make a choice of D similar to the one found useful in [10] and verify after the fact that the the choice works reasonably well.

For comparison to MMI training, we also present results for error weighted training which is similar to ML training except that the training sentences in which the model is in error are weighted more heavily than other sentences. We also show that the best results can be obtained experimentally by a combination of MMI and error weighted training.

A. Outline

In section II we present our notation for HMM based speech recognition and review the definition of general SCGMMs and the special case of SPAM models. In section III we introduce the auxiliary functions for ML training and summarize how it may be optimized. In section IV we see that error weighted training may be performed with the aid of an auxiliary function of the same form as for ML training. Section V contains: a derivation of the MMI auxiliary function for completely general models and for subspace constrained Gaussian mixture models (section V-A); further discussion of the auxiliary function and update rules (section V-B); and some heuristics that are helpful in practical implementation (section V-C). Our experimental results and conclusion are presented in sections VI and VII. A detailed and very general proof that the auxiliary function for MMI training is valid is given in the appendix.

II. HMM AND SUBSPACE CONSTRAINED GMM DEFINITIONS

We consider speech recognition systems consisting of the following components: a *frontend* which processes a raw input acoustic waveform into a time series of acoustic feature vectors $X = (x_1, x_2, ..., x_T)$, where x_t is a vector in

$\mathbf{R}^d$ called the acoustic data vector at time frame t; a *language model* which provides a prior probability distribution $P(W)$ over possible word sequence $W = (w_1, w_2, ..., w_N)$ that the user may utter; an *acoustic model* which gives a conditional probability distribution $P(X|W)$ for the acoustic data given the word sequence; and a *search strategy* that finds the word sequence W that (approximately) maximizes the joint likelihood $P(X, W) = P(X|W)P(W)$.

An HMM based acoustic model provides a set of states $\mathcal{S}$, a probability distribution $p(S|W)$ over possible state sequences $S = (s_1, ..., s_T)$ produced by a word sequence W; and probability density functions $p(x|s)$ associated with a state $s \in \mathcal{S}$ and an acoustic vector $x \in \mathbf{R}^d$. The state sequence model $P(S|W)$ for an HMM has a particular form allowing efficient calculation, but everything we say in this paper applies with an arbitrary state sequence model. The conditional distribution $P(X|W)$ is written as a sum over hidden state sequences $S = (s_1, ..., s_T)$ that may be associated with the word sequence W:

$$P(X|W) = \sum_S P(X|S)P(S|W) \quad (1)$$

$$P(X|S) = \prod_{t=1}^{T} p(x_t|s_t) \ . \quad (2)$$

We take the distributions $p(x|s)$ for each s to be a Gaussian mixture model

$$p(x|s) = \sum_{g \in \mathcal{G}(s)} \pi_g \mathcal{N}(x; \mu_g, \Sigma_g) \ , \quad (3)$$

where π_g is the prior for Gaussian g, $\mathcal{N}(x; \mu_g, \Sigma_g)$ is a Gaussian distribution with mean μ_g and covariance Σ_g, and $\mathcal{G}(s)$ is the set of Gaussians associated with state s. For definiteness we assume that the set of Gaussians for distince states are disjoint. This allows us the convenience of talking about the state $s(g)$ which a given Gaussian g is associated with.

A. The Gaussian as an Exponential Model

To describe the model types and estimation procedures considered in this paper, we rewrite a Gaussian distribution,

$$\mathcal{N}(x; \mu, \Sigma) = \det\left(\frac{\Sigma^{-1}}{2\pi}\right)^{1/2} e^{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)} \ , \quad (4)$$

in the form of an exponential model. To do so, we first write the Gaussian as

$$\mathcal{N}(x; \mu, \Sigma) = e^{-1/2 x^T P x + \psi^T x + K(P,\psi)} \ , \quad (5)$$

where

$$P = \Sigma^{-1} \quad (6)$$

$$\psi = P\mu \quad (7)$$

$$2\,K(P, \psi) = -d \log 2\pi + \log\det(P) - \psi^T P^{-1} \psi \ . \quad (8)$$

The inverse covariance matrix P is called the *precision matrix* and we will refer to ψ simply as the *linear weights*. Next we define the feature vector $f(x)$ and weight vector θ, which are both column weight vectors of size $d(d+3)/2$:

$$f(x) = \begin{bmatrix} -1/2\,\mathrm{vec}(xx^T) \\ x \end{bmatrix} \qquad \theta = \begin{bmatrix} \mathrm{vec}(P) \\ \psi \end{bmatrix} \ . \tag{9}$$

Here, for any $d \times d$ symmetric matrix S, $\mathrm{vec}(S)$ is a column vector whose entries are the $d(d+1)/2$ upper triangular elements of S written in some fixed order, with the off diagonal elements multiplied by $\sqrt{2}$. This map is defined so as to preserve inner product, i.e. so that $\mathrm{vec}(S_1)^T\mathrm{vec}(S_2)$ equals $\mathrm{Tr}(S_1^T S_2)$ for any symmetric matrices S_1 and S_2. For convenience, we may also write column vectors as pairs, e.g. $\theta = (\mathrm{vec}(P), \psi)$.

Now we may write (5) in standard exponential model format

$$\mathcal{N}(x; \mu, \Sigma) \quad = \quad e^{\theta^T f(x) + K(P,\psi)} \ . \tag{10}$$

We will interchangeably use (P, ψ) and θ as is convenient.

B. Subspace Constrained GMMs

We now describe two models types for which discriminative estimation is discussed in this paper. The first one is the Subspace Constrained Gaussian Mixture Model (SCGMM) [5], [7]. An SCGMM requires that the θ_g in (10) belong to a common F-dimensional affine subspace of the space of all parameters. Letting $b_0 \in \mathbf{R}^{d(d+3)/2}$ be a basepoint of the affine space and B be a matrix of size $d(d+3)/2 \times F$ whose columns form a basis of the subspace, we may write:

$$\theta_g = b_0 + B\lambda_g \ . \tag{11}$$

The parameters b_0 and B are shared across Gaussians; these are referred to as *tied* model parameters. The Gaussian specific parameters, λ_g, are referred to as *un-tied* model parameters. Gaussian priors, π_g, also form part of the un-tied SPAM model parameters.

Note that the distribution (10) with the constraint (11) may be regarded as an exponential distribution with $F+1$ dimensional features $f_B(x)$,

$$f_B(x) \quad = \quad [b_0 \ B]^T f(x) \ . \tag{12}$$

From this point of view, the choice of constraining subspace may be viewed as a selection problem for the features of the exponential model.

It was discussed in [5] how several well known model types are special cases of SCGMMs. In this paper, we restrict focus to general SCGMMs and the more restricted SPAM model class which imposes separate subspace constraints on the precisions and means. SPAM models requires the precision matrices and linear weights to be in D and L dimensional affine subspaces, respectively. D is free to range from 0 (or 1 if no affine shift term is included) to the full covariance value of $d(d+1)/2$ while L ranges from 0 to d. For the SPAM models, we may

write:

$$P_g = S_0 + \sum_{k=1}^{D} \lambda_g^k S_k \tag{13}$$

$$\psi_g = l_0 + \sum_{k=1}^{L} \lambda_g^{k+D} l_k \ . \tag{14}$$

This corresponds to taking B in (11) to be block diagonal,

$$\theta_g = \begin{bmatrix} \mathrm{vec}(P_g) \\ \psi_g \end{bmatrix} = \begin{bmatrix} \mathrm{vec}(S_0) \\ l_0 \end{bmatrix} + \begin{bmatrix} B_{11} & 0 \\ 0 & B_{22} \end{bmatrix} \lambda_g \ , \tag{15}$$

where B_{11} is the $d(d+1)/2 \times D$ matrix whose k'th column is $\mathrm{vec}(S_k)$ and B_{22} is the $d \times L$ matrix whose k'th columns is l_k.

III. Maximum Likelihood Parameter Estimation

The technique we use for parameter estimation is to update parameters so as to maximize an auxiliary function associated with a utility function. In the next subsection we give a definition of auxiliary function as well as an example which is the archetype for the ML, error weighted, and MMI auxiliary functions, discussed in sections III-B, IV, and V, respectively. Since the auxiliary functions in all of these cases have the same form, they may optimized by the same method, which is summarized in section III-C.

A. Auxiliary Functions

Given a smooth "utility" function $F(\Omega)$ of a parameter Ω, an auxiliary function (with scaling) associated to F is a smooth function $Q(\Omega, \Omega^0)$ satisfying

$$Q(\Omega, \Omega^0) - Q(\Omega^0, \Omega^0) \leq S(\Omega^0) \ \left[F(\Omega) - F(\Omega^0)\right] \ , \tag{16}$$

where $S(\Omega^0)$ is a positive function which we call the scaling function. For a fixed choice of Ω^0, any choice of Ω such that $Q(\Omega, \Omega^0)$ is greater than $Q(\Omega^0, \Omega^0)$ will satisfy $F(\Omega) > F(\Omega^0)$. A strategy to maximize the function F is to update the parameter Ω iteratively by the update formula $\Omega^{n+1} = \mathrm{argmax}_\Omega \, Q(\Omega, \Omega^n)$. The sequence $F(\Omega^n)$ is guaranteed to be monotonically increasing. Although $F(\Omega^n)$ is not guaranteed in general to converge to a global maximum, it is guaranteed that the value of $F(\Omega^n)$ will strictly increase unless a fixed point of the update formula is reached. This can only happen at a critical point of F because (16) implies that the gradient with respect to Ω of $Q(\Omega, \Omega^0)$, equals $S(\Omega^0)$ times the gradient of $F(\Omega)$, when evaluated at $\Omega = \Omega^0$. Generally the critical point of F in the last sentence will be a local maximum, although it is possible to converge to a saddle point of F.

One example of an auxiliary function is the function $Q_1(p, p_0) = p_0 \log p$, which is an auxiliary function (with identity scaling) for the identity function $F_1(p) = p$ of a single positive real variable p. This follows directly from the concavity of the log function, which implies that $\log x \leq x - 1$ for any positive x. We can obtain more general

utility and auxiliary functions by taking positive linear combinations of F_1 and Q_1. Specifically, given a positive function $q^{gen}(y)$ of a parameter y in a measure space Y and a positive function $\mathcal{P}(y;\Omega)$, we define the functions

$$F^{gen}(\Omega) = \int_y q^{gen}(y)\mathcal{P}(y;\Omega) \tag{17}$$

and

$$Q^{gen}(\Omega,\Omega_0) = \int_y q^{gen}(y)\mathcal{P}(y;\Omega_0)\log\mathcal{P}(y;\Omega) \ . \tag{18}$$

To see that Q^{gen} is an auxiliary function for F^{gen}, we need only show the following function is always positive:

$$\Delta^{gen}(\Omega) \ = \ \left[F^{gen}(\Omega) - F^{gen}(\Omega^0)\right] - \left[Q^{gen}(\Omega,\Omega^0) - Q^{gen}(\Omega^0,\Omega^0)\right] \ . \tag{19}$$

Positivity follows by rewriting,

$$\Delta^{gen}(\Omega) \ = \ \int_y q^{gen}(y)\mathcal{P}(y;\Omega_0)\mathcal{H}(y;\Omega) \tag{20}$$

$$\mathcal{H}(y,\Omega) \ = \ h(\frac{\mathcal{P}(y;\Omega)}{\mathcal{P}(y;\Omega^0)} - 1) \tag{21}$$

$$h(s) \ = \ s - log(1+s) > 0 \ , \tag{22}$$

and using the fact that each factor under the integral sign in (20) is positive.

B. Auxiliary function for ML Training

Let (X^*, W^*) denote the totality of given acoustic training data and its word sequence label; it may be formed by a concatenation of several smaller utterances. The goal of maximum likelihood (ML) training is to maximize the total likelihood,

$$L(\Omega) \ = \ P(X^*|W^*;\Omega) \ , \tag{23}$$

of "generating" X^* given W^*. Here Ω denotes the set of parameters involved in the definition of distributions $p(x|s)$.

Since the utility function $L(\Omega)$ is expensive to evaluate, requiring running through all of the training data, the Baum-Welch or Expectation Maximization procedure was developed. It provides an auxiliary function Q_{ml} that can be evaluated cheaply in terms of statistics collected in a single pass over the training data and whose optimization guarantees an increase of the target utility function. Q_{ml} is actually a special case of (18), as can be seen by dropping the second and third terms in (47) and proceeding as in the remainder of the derivation of Q_{mmi} in section V-A. The function Q_{ml} is:

$$Q_{ml}(\Omega,\Omega^0) \ = \ \sum_g \sum_t \gamma_{ml}(t,g)\log\pi_g \, p(x_t|g;\Omega), \tag{24}$$

where

$$\gamma_{ml}(t,g) \ = \ P(g(t) = g|X^*, W^*;\Omega^0) \tag{25}$$

is the conditional probability of observing Gaussian g at time t given the training acoustic data and reference word scripts.

Q_{ml} as well as the error weighted auxiliary functions defined in section IV and the MMI auxiliary function described in Section V may all be written in the form:

$$Q(\Omega, \Omega^0) \quad = \quad \sum_g \int_x \kappa(g,x) \, \log \pi_g \, p(x_t|g;\Omega) \tag{26}$$

For the ML auxiliary function, the "weight function" κ takes the form:

$$\kappa_{ml}(g,x) \quad = \quad \sum_t \delta(x - x_t) \gamma_{ml}(t,g) \ . \tag{27}$$

C. Optimization of the Auxiliary Function

The techniques for optimization the EM auxiliary function (24) were developed in [6], [7], [23], [24] and reviewed in detail in [5]. The techniques apply generally to optimizing any auxiliary function of the form (26); and so apply directly to the auxiliary functions presented below for MMI and error weighted training. We now summarize the training methodology.

To begin, we decompose the auxiliary function (26) into a sum of a piece that depends only on the set of all priors $\pi = \{\pi_g\}$ and a piece that depends on the set $\Theta = \{\theta_g\}$ specifying the individual mixture components. That is, we write $\Omega = (\pi, \Theta)$ and

$$Q(\Omega; \Omega^0) \quad = \quad Q^\pi(\pi, \Omega^0) + Q^\Theta(\Theta, \Omega^0) \tag{28}$$

$$Q^\pi(\pi, \Omega^0) \quad = \quad \sum_g n(g) \log \pi_g \tag{29}$$

$$Q^\Theta(\Theta, \Omega^0) \quad = \quad \sum_g \int_x k(g,x) \log p(x|g;\Omega) \tag{30}$$

$$= \quad \sum_g n(g) K(\theta_g) + \theta_g^T \, \hat{f}_g \tag{31}$$

$$n(g) \quad = \quad \int_x k(g,x) \tag{32}$$

$$\hat{f}_g \quad = \quad \int_x k(g,x) f(x) \ . \tag{33}$$

For notation brevity, we have left implicit the dependence of k,n, and $\hat{f}$ on Ω^0. The second expression for Q^Θ above comes from writing the probability distribution in exponential model format (10):

$$\log p(x|g;\Omega) \quad = \quad \theta_g^T \, f(x) + K(\theta_g). \tag{34}$$

We refer to $n(g)$ as the total count for Gaussian g and $\hat{f}_g$ as the total feature sum for Gaussian g. The statistics for the ML case are

$$n^{ml}(g) \quad = \quad \sum_t \gamma_{ml}(t,g), \text{ and} \tag{35}$$

$$\hat{f}_g^{ml} \quad = \quad \sum_t \gamma_{ml}(t,g) f(x_t) \ . \tag{36}$$

The term in the auxiliary function depending only on the priors may be written

$$Q^{\pi}(\pi, \Omega^0) = \sum_s N(s) \left[\sum_{g;s(g)=s} \tilde{\pi}_g \log \pi_g \right] \quad (37)$$

$$N(s) = \sum_{g;s(g)=s} n(g), \quad (38)$$

$$\tilde{\pi}_g = n(g)/N(s(g)). \quad (39)$$

In the above, $s(g)$ is the state that Gaussian g is associated with (i.e. the state so that $g \in \mathcal{G}(s)$), $N(s)$ is the total count for state s. Optimizing Q^{π} in the standard way, the prior π_g of the new model equals $\tilde{\pi}_g$. Note that in the ML case, $\tilde{\pi}_g$ is the probability, under the old model Ω^0, of seeing Gaussian g given that the state is $s(g)$.

Optimization for $\Theta = \{\theta_g = b_0 + B\lambda_g\}$ in Q^{Θ} (31) uses a quasi-Newton search strategy. An important first step is finding an appropriate seed value for the search. One successful approach we found was to seed with the solution of certain quadratic approximations to the exact Q function. Given the seed, optimization is performed by alternating between optimization of the tied and untied parameters. The optimal untied parameters λ_g for a specific Gaussian g are found by maximizing

$$(\tilde{f}_g^T B)\lambda_g + K(b_0 + B\lambda_g) \ . \quad (40)$$

This optimization may be performed extremely efficiently because the line search step of the quasi-Newton search may be computed very quickly by writing the restriction of (40) to the line being searched as a simple sum. Also, the statistics gathering for the untied parameters optimization can be sped up using the fact that only the statistics of $f(x)^T B$ need be gathered. The optimal tied parameters b_0 and B ($\{l_k\}$ and $\{S_k\}$ if b_0 and B take the restricted SPAM form (13)-(15)) is also accomplished by quasi-Newton with efficient line searches, although the procedure is more painful because it does not break down into a separate optimization for each Gaussian.

IV. Error Weighted Training

A number of well known training procedures, such as corrective training [13] and boosting [25] are motivated by the idea of paying extra attention to the training cases where the current model is making a mistake. Here we introduce a simple implementation of this idea which we call "error weighted" training.

Unlike ML and MMI training, the utility function for error weighted training depends on a reference model Ω^{ref}. It also depends on a positive weighting parameter α. The first step is to determine the training sentences X^*_{err} that are in error under the model Ω^{ref}. The utility function is the same as the maximum likelihood utility function except that the sentence in error receive a weight of $(1+\alpha)$, $P(X^*|W^*, \Omega) + \alpha P(X^*_{err}, W^*_{err}, \Omega)$. The auxiliary function for this take the form (28)-(33) using statistics

$$n^{\alpha}(g) = n^{ml}(g) + \alpha n^{err}(g)$$

$$\hat{f}_g^{\alpha} = \hat{f}_g^{ml} + \alpha \hat{f}_g^{err} \quad (41)$$

where n^{ml} and $\hat{f}^{ml}$ are the ML statistics (36) and n^{err} and $\hat{f}^{err}$ are the same statistics but with the sum restricted to the sentences containing errors.

Parameters are optimized, as in the ML case, by the technique of section III-C but using the statistics $n^{\alpha}(g)$ and $\tilde{f}_g^{\alpha}$. In the experiments here, we restrict to a single error weighted training update seeded by an ML trained model Ω^0 and we take the reference model for finding the error sentences to be the seed model.

The error weighted procedure described above has the advantage that it is exceedingly simple to implement. On the other hand, it is expected to work best when the training sentence error rate is already fairly small. In addition to being close to corrective training, we also view it as a "cheap" version of boosting, which would provide a more complex hierarchical structure for the weighting of error data. We also view error weighted training as discriminative in nature because its focus on the sentences with errors is an an attempt to minimize recognition error rate.

V. MMI Estimation of Parameters

In section V-A we present a variant of the method of Gunawardana et.al. [20] for deriving an auxiliary function $Q_{mmi}(\Omega, \Omega^0; D)$, depending on an additional parameter D, which aids in maximizing the MMI objective function $R(\Omega)$ defined below. Q_{mmi} is is of the form (18) except that the quantity $q(y)$ is not guaranteed to be positive.

In the appendix we give a rigorous proof that the update formula obtained using Q_{mmi} does indeed lead to an increase of the objective function, provided D is chosen large enough. Our proof does not give a specific description of what D values are allowable. In section V-C we present some heuristics for choosing D.

A. An Auxiliary Function for MMI Estimation

The MMI objective function is

$$R(\Omega) = \frac{P(X^*, W^*; \Omega)}{P(X^*; \Omega)P(W^*)} \tag{42}$$

$$= \frac{P(X^*|W^*; \Omega)}{P(X^*; \Omega)} \tag{43}$$

$$= \frac{L(\Omega)}{M(\Omega)} \ . \tag{44}$$

The numerator $L(\Omega)$ may be written

$$\begin{aligned} L(\Omega) &= \sum_S P(X^*|S;\Omega)P(S|W^*) \\ &= \sum_S \sum_{G \in \mathcal{G}(S)} P(X^*, G|S;\Omega)P(S|W^*) \\ &= \sum_G P(X^*, G|S(G);\Omega)P(S(G)|W^*) \ . \end{aligned} \tag{45}$$

In the above, $\mathcal{G}(S)$ is the set of Gaussian sequences that can be made by picking one Gaussian each from the states that belong to S and $S(G)$ is the state sequence determined by the Gaussian sequence G. The only state sequences we consider in the remainder of the paper are of the form $S(G)$, which we will henceforth abbreviate simply as S.

The denominator $M(\Omega)$ may be written

$$M(\Omega) \quad = \quad \sum_G P(X^*, G|S;\Omega)P(S) \tag{46}$$

Note that in (46) we use $P(S) = \sum_W P(S,W)$ as opposed to $P(S|W^*)$ as in (45) because the denominator includes a language model whereas the numerator does not.

Given a starting parameter value Ω^0, The derivation of the auxiliary function for $R(\Omega)$ proceeds in two steps. First, following Gopalakrishnan et.al. [18], we define

$$F(\Omega, \Omega^0) \quad = \quad L(\Omega) - R(\Omega^0)M(\Omega) + C(\Omega) \ , \tag{47}$$

where $C(\Omega)$ is a constant function of Ω to be specified momentarily. Note that $F(\Omega^0, \Omega^0) = C(\Omega)$. More importantly, observe that if Ω^1 has the property that $F(\Omega^1, \Omega^0) > F(\Omega^0, \Omega^0)$, then $R(\Omega^1) > R(\Omega^0)$. Thus F is an auxiliary function for R.

Motivated by the expansions (45) and (46) and by (17) and (18), we select

$$C(\Omega) \quad = \quad C(\Omega; D) = \sum_G \int_X D_G P(X, G|S;\Omega)P(S) \tag{48}$$

where D_G is a constant associated to the Gaussian sequence G. and D is shorthand for the set $\{D_G\}$ of all such constants. Notice that $C(\Omega = (\pi, \Theta); D)$ is a constant function of Θ, for fixed π. As a function of π, it is only constant if D_G is independent of G. Thus when training only the parameters Θ specifying the Gaussian distributions we may allow D_G to vary with G, but when training the priors D_G must be constant in G.

The function F can now be re-written as

$$F(\Omega, \Omega^0) \quad = \quad \sum_G \int_X q_{mmi}(X, G; \Omega^0, D)P(X, G|S;\Omega) \ , \tag{49}$$

$$\begin{aligned} q_{mmi}(X, G; \Omega^0, D) \quad &= \quad \delta(X - X^*)P(S|W^*) \\ &\quad - \delta(X - X^*)R(\Omega^0)P(S) \\ &\quad + D_G P(S) \ , \end{aligned} \tag{50}$$

where $\delta(X - X^*)$ is the Dirac delta function centered at X^*. The above is a special case of the generic function F^{gen} (17), as can be seen by letting y be shorthand for the pair (X, G) and setting

$$q^{gen}(y) \quad = \quad q_{mmi}(X, G; \Omega^0, D) \tag{51}$$

$$\mathcal{P}(y;\Omega) \quad = \quad P(X, G|S(G);\Omega) \ . \tag{52}$$

Note that the integration over y in (17) is now shorthand for summation over G combined with integration over X.

Let Q_{mmi} be the auxiliary function Q^{gen} defined in (51) with q^{gen} and $\mathcal{P}$ as in (51) and (52). By taking the constants D_G large enough, it would appear that the function $q_{mmi}(X, G; \Omega^0, D)$ can be made everywhere positive, so that Q_{mmi} would be a valid auxiliary function associated with F, and therefore also with R (since F is an auxiliary function for R). This reasoning is essentially equivalent to that in reference [20]. However there is a

fallacy: due to the presence of the $\delta(X - X^*)$ functions, D_G values that make q everywhere positive will in general be impossible to find. We remedy this in the appendix by giving a proof that, although q need not be positive, it is always possible to find D_G values that guarantee validity of Q_{mmi} as an auxiliary function.

Denoting the dependence of Q_{mmi} on D explicitly, we have

$$\begin{aligned} Q_{mmi}(\Omega, \Omega^0; D) &= \int_y q^{gen}(y) \mathcal{P}(y; \Omega^0) \log \mathcal{P}(y; \Omega) \\ &= \sum_G \int_X q_{mmi}(X, G; \Omega^0, D) P(X, G|S; \Omega^0) \\ &\quad \times \log P(X, G|S; \Omega) \end{aligned} \tag{53}$$

To massage Q_{mmi} into the form the form (26) which Q_{ml} took, we let Q^S_{mmi} be Q_{mmi} divided by $P(X^*|W^*; \Omega^0)$ (so that Q^S_{mmi} will be an auxiliary function with scale factor). Using (50) to expand $q_{mmi}(X, G; \Omega^0, D)$ in (53) and expanding $\log P(X, G|S; \Omega)$ as a sum over time, one can see that Q^S_{mmi} does indeed have the form (26). Explicitly, we have:

$$Q^S_{mmi}(\Omega, \Omega^0; D) = \sum_g \int_x \kappa_{mmi}(g, x) \, \log \pi_g \, p(x|g; \Omega), \tag{54}$$

where

$$\kappa_{mmi}(g, x) = \kappa_{ml}(g, x) - \kappa_{den}(g, x) + D_g \, p(x|g; \theta^0_g). \tag{55}$$

The three components of κ_{mmi} arise from the three terms in q. The first component, due to the numerator $L(\Omega)$ in F, is simply $\kappa_{ml}(g, x)$ as defined in (27). The second component, arising due to the "denominator" term $R(\Omega^0) M(\Omega)$ in F, is

$$\kappa_{den}(g, x) = \sum_t \delta(x - x_t) \gamma_{den}(t, g) \tag{56}$$

$$\gamma_{den}(t, g) = P(g(t) = g|X^*, \Omega^0) \ . \tag{57}$$

This is identical to κ_{ml} except that the conditional probability of observing Gaussian g at time t in the definition of γ_{den} is not conditioned on W^*. Finally, the third component of κ_{mmi} arises due to the constant term $C(\Omega)$ in F. The quantity D_g in that term is

$$D_g = \sum_t \sum_W \sum_{G; G(t)=g} \frac{D_G P(G|S, \Omega^0) P(S, W)}{P(X^*|W^*; \Omega^0)} \ . \tag{58}$$

Written in terms of exponential model sufficient statistics as in (28)-(33), the scaled MMI auxiliary function is:

$$\begin{aligned} Q^S_{mmi}(\Omega, \Omega^0; D) &= \sum_g n^{mmi}(g) \left(\log \pi_g + K(\theta_g) \right) \\ &\quad + \theta_g^T \, \hat{f}^{mmi}_g \end{aligned} \tag{59}$$

$$n^{mmi}(g) = n^{ml}(g) - n^{den}(g) + D_g \tag{60}$$

$$\hat{f}^{mmi}_g = \hat{f}^{ml}_g - \hat{f}^{den}_g + D_g \, E_{\theta^0_g}(f(x)) \tag{61}$$

Here n^{den} and $\hat{f}^{den}$ are defined as in the ML statistics (36) except that γ_{den} is used instead of γ_{ml}. $E_{\theta_g^0}$ is the expectation operator with respect to the conditional distribution for Gaussian g and the unupdated model; so

$$E_{\theta_g^0}(f(x)) \quad = \quad \int_x p(x|g;\theta_g^0)\, f(x) \ . \tag{62}$$

As stated previously, since the form of this auxiliary function is the same as that of $Q_{ml}(\Omega, \Omega^0)$, the two can be maximized using common numerical optimization procedures. Note that our formulation naturally allows for a Gaussian dependent D_g value (when updating Θ). This has empirically been found to be of value in MMI estimation [10]. We observe that any (positive) values of $\{D_g\}$ can be arrived at from some positive choice of $\{D_G\}$. To see this, it suffices to take D_G of the form $D_G = \epsilon + a_{G(1)}\delta(G = constant)$, where ϵ is a G independent constant and the delta function is one if all components of G are equal and zero otherwise.

B. Discussion of the auxiliary function and why it is valid

1) Large D implies small update: One may think of $D = \{D_g\}$ as providing an upper bound on the step size in going from Ω^0 to Ω by an update obtained by approximately maximizing Q_{mmi}. Informally, this follows from the fact that the only dependence of $Q_{mmi}(\Omega, \Omega^0; D)$ on D is in the linear term

$$\sum_g D_g \int_x p(x|g;\theta_g^0) \log p(x|g,\theta_g) \ . \tag{63}$$

This term has a global maximum when θ_g equals θ_g^0 and decays rapidly for θ_g far from θ_g^0. Taking D_g large enough, therefore forces the maximum of Q_{mmi} to be near θ_g^0. To make this argument rigorous, we need to show that the other terms in Q_{mmi}, specifically the denominator terms, don't increase at a faster rate than the D_g term decreases. This can be done easily enough given a concrete formula for the probability distributions $p(x|g,\theta_g)$. In the proof of the theorem of the next section, we prove the result quite generally with a simple topological argument.

2) Comparison of Q_{mmi} maximization to gradient search: Since the gradients with respect to Ω of $R(\Omega)$ and $Q_{mmi}(\Omega, \Omega^0; D)$ agree when $\Omega = \Omega^0$, gradient assent with small step size is a special case of an update which increases Q_{mmi}. One difficulty of gradient search is choosing appropriate values for the step size. The approach of maximizing Q_{mmi} replaces this problem with the problem of choosing acceptable values of D.

Note that the small step gradient search merely increases $Q_{mmi}(\Omega, \Omega^0)$ over it's value when Ω equals Ω^0. The update rule which maximize Q_{mmi} has the advantage that it has access to higher order terms in the Taylor series for (the lower bound Q_{mmi} to) $R(\Omega)$. To see this explicitly, it is instructive to look at the special case when the covariance matrices Σ_g and the precision matrices $P_g = \Sigma_g^{-1}$ are held fixed and only the unconstrained means μ_g are being trained. In that case, the auxiliary function Q_{mmi} may be written (up to irrelevant constants) as:

$$Q_{mmi} \quad = \quad \sum_g \mu_g^T P_g \left[-0.5\, n^{mmi}(g)\mu_g + n^{ml}(g)\mu_g^{ml} - n^{den}(g)\mu_g^{den} + D_g\mu_g^0 \right] \ , \tag{64}$$

where μ_g^{ml} and μ_g^{den} are the means from the numerator and denominator statistics ($\hat{f}^{ml}$ and $\hat{f}^{den}$) alone. The

maximum of (64) occurs when

$$\mu_g = \frac{n^{ml}(g)\mu_g^{ml} - n^{den}(g)\mu_g^{den} + D_g\mu_g^0}{n^{ml}(g) - n^{den}(g) + D_g} \quad (65)$$

$$= \mu_g^0 + \frac{1}{n^{mmi}(g)}\left[n^{ml}(g)(\mu_g^{ml} - \mu_g^0) - n^{den}(g)(\mu_g^{den} - \mu_g^0)\right] \ . \quad (66)$$

The value at $\Omega = \Omega^0$ of the gradient with respect to μ_g of the MMI objective function $R(\Omega)$ is:

$$\nabla_{\mu_g} R(\Omega)|_{\Omega=\Omega^0} = \nabla_{\mu_g} Q_{mmi}(\Omega, \Omega^0; D)|_{\Omega=\Omega^0} \quad (67)$$

$$= n^{ml}(g)P_g^0(\mu_g^{ml} - \mu_g^0) - n^{den}(g)P_g^0(\mu_g^{den} - \mu_g^0) \quad (68)$$

Thus the MMI update for μ_g is

$$\mu_g = \mu_g^0 + \frac{1}{n^{mmi}(g)}\Sigma_g^0\nabla_{\mu_g} R(\Omega_0) \ . \quad (69)$$

Note that this looks like gradient ascent with a step size of $1/n^{mmi}(g)$ except for the factor Σ_g^0. The factor $-n^{mmi}(g)P_g^0$ (i.e. the inverse of $-\Sigma_g^0/n^{mmi}(g)$) is the Hessian of Q_{mmi} with respect to μ_g. That factor is also an approximation to (actually a lower bound to) the Hessian H of $R(\Theta)$ at Θ^0 (obtained by dropping terms involving derivatives of $\gamma_{ml}(t,g)$ and $\gamma_{den}(t,g)$ and adding the term D_g to $n^{mmi}(g)$). Thus the update formula (69) obtained by maximizing Q_{mmi} is best viewed not as simple gradient ascent, but as an approximation to the quadratic Newton update

$$\mu_g = \mu_g^0 - H^{-1}\nabla_{\mu_g} R(\Theta) \ , \quad (70)$$

which generally leads to more rapid convergence than simple gradient search.

3) Why the auxiliary function is valid for large D: We already know that $F(\Omega, \Omega^0)$. is an auxiliary function for $R(\Omega)$. So it suffices to show that $Q_{mmi}(\Omega, \Omega^0)$ is an auxiliary function for $F(\Omega, \Omega^0)$. Recall that Q_{mmi} is an example of the generic function Q^{gen} (18) and F is an example of the generic function F^{gen} (17) when the weighting function q^{gen} equals q_{mmi} (50). The condition that Q^{gen} is an auxiliary function for F^{gen} is equivalent to the statement that the function Δ^{gen} (20) is positive. This in turn would be guaranteed if q_{mmi} were positive. Unfortunately, the negative "denominator" delta function in the second term of q_{mmi} violates this positivity. Fortunately, the negative contribution to Δ^{gen} due to the denominator term is drowned out, for large D_G, by the positive contribution due to the $D_G P(S)$ term in q_{mmi}. This can be seen by letting $D_G^{(1)}$ be fixed positive constants and setting $D_G = D_G^1/\epsilon$, where ϵ is a small positive parameter. The negative term has an extra factor of ϵ and so it is indeed much smaller than the positive term for small ϵ.

4) Technicalities in proof: In the next section we will formally prove the statement just made that the positive terms in $\Delta^{gen}(\Omega)$ dominate the negative terms for small ϵ. Care must be taken to show that this domination is uniform in Ω (i.e. for small enough ϵ, the domination occurs for *all* Ω. For the case of completely general acoustic probability distribution $p(x;\theta)$, we need to make a few technical assumptions, all of which are valid for the special case of tied exponential models (such as SCGMMs).

First, we need to assume that the set of possible parameters is effectively compact (closed and bounded). In the general case, we simply assume compactness. For tied exponential models, we use the fact that the set of parameters Ω for which $Q_{mmi}(\Omega, \Omega^0; D)$ is greater than $Q_{mmi}(\Omega^0, \Omega^0; D)$ is compact. We also need to assume that the distributions are everywhere positive and that they are analytic functions of θ. The compactness and positivity assumptions are required to avoid degeneracies such as distributions becoming infinitely peaked or else zero at some of the training data points. The analyticity assumption is actually quite a bit stronger than what we actually need, all we really need is that the any non-zero functions that arise in the proof have the property that they have a non-zero leading term in their Taylor series.

The need for these technical assumptions is best illustrated by concrete counterexamples to the theorem when the assumptions are not valid. For brevity, we will not present such counterexamples here, but we encourage the interested reader to construct such counterexamples for a system with vocabulary consisting of the two words A and B with equal language model probability, each represented by a single state, with one Gaussian per state; with two dimensional feature vectors lying in the unit disk; and with the entire training data X^* consisting of a single frame equal to the zero vector with transcript A.

C. Heuristics in Discriminative Estimation Procedures

Although we show in the appendix that MMI training by maximizing the function Q_{mmi} does indeed improve the MMI objective function provided that the constants D_g in (58) are chosen large enough; our proof does not give a concrete formula specifying how large D_g has to be. Instead we will resort to values found useful based on experimental evidence. In fact, several heuristic choices are needed in practice to make discriminative estimation, in particular MMI training [10], yield improvements in error rates. In this section we present these details as they apply to our implementation of MMI and error weighted training. In particular we discuss selection of D_g values, update strategies for mixture weights π_g in MMI and error weighted training, and a likelihood scale that is used in computation of $\gamma_{den}(g,t)$ defined in (57). Our choices are based largely on experimental evidence of Woodland et.al. [10].

1) Selecting D_g: In our experiments we follow a D_g selection procedure that is analogous to a method described by Woodland et.al.:

$$D_g = \max\left(C_1 n^{den}(g), C_2 D_g^*\right) \tag{71}$$

where C_1 and C_2 are constants and D_g^* is the smallest value such that when a full covariance matrix (diagonal in [10]) is estimated from the MMI stats, it comes out to be positive definite. Although this choice of D_g is not guaranteed to be large enough to guarantee that Q_{mmi} is a valid auxiliary function, we shall see that the choice is practically useful.

Let us make the definition of D_g^* more explicit. To do so, we first write down the MMI update of a full covariance matrix, given some D_g values:

$$\hat{\Sigma}_g = \frac{\hat{f}_g^{ml(F)} - \hat{f}_g^{den(F)} + D_g(\Sigma_g^0 + \mu_g^0 \mu_g^{0T})}{n^{ml}(g) - n^{den}(g) + D_g} - \hat{\mu}_g \hat{\mu}_g^T \tag{72}$$

Here $\hat{f}_g^{ml(F)}$ denotes, with a slight abuse of notation, a matrix constructed from the $\mathrm{vec}(xx^T)$ portion of $\hat{f}_g^{ml}$ statistics (cf. (9) and (36)). Similarly, $\hat{f}_g^{den(F)}$ is a matrix constructed from the denominator statistics. Furthermore,

$$\hat{\mu}_g = \frac{\hat{f}_g^{ml(L)} - \hat{f}_g^{den(L)} + D_g \mu_g^0}{n^{ml}(g) - n^{den}(g) + D_g} \tag{73}$$

where superscript (L) denotes the x portion of the statistics (cf. (9)). (Note that the above equation is simply a reformulation of (65).) By substituting $\hat{\mu}_g$ from (73) into (72), one can see that $\hat{\Sigma_g}$ is singular (i.e. has a 0 eigenvector) precisely when there is a vector y such that

$$0 = \left(A_0 + D_g A_1 + D_g^2 A_2\right) y \tag{74}$$

$$A_0 = c_g \tilde{\Sigma}_g - \tilde{\mu}_g \tilde{\mu}_g^T \tag{75}$$

$$A_1 = c_g(\Sigma_g^0 + \mu_g^0 \mu_g^{0T}) - \mu_g^0 \tilde{\mu}_g^T - \tilde{\mu}_g \mu_g^{0T} + \tilde{\Sigma}_g \tag{76}$$

$$A_2 = \Sigma_g^0 \tag{77}$$

$$c_g = n^{ml}(g) - n^{den}(g) \tag{78}$$

$$\tilde{\mu_g} = \hat{f}_g^{ml(L)} - \hat{f}_g^{den(L)} \tag{79}$$

$$\tilde{\Sigma}_g = \hat{f}_g^{ml(F)} - \hat{f}_g^{den(F)} \tag{80}$$

Equation (74) is called a quadratic eigenvalue equation. D_g^* equals the largest positive real D_g for which there exists a y solving the quadratic eigenvalue equation.

2) Updating Mixture Weights: The MMI update of mixture weights can be obtained by maximizing (59) with respect to π_g, i.e. by maximizing

$$\sum_s \sum_{g;s(g)=s} \left(n^{ml}(g) - n^{den}(g) + D_g\right) \log \pi_g$$

subject to constraints $\sum_{g;s(g)=s} \pi_g = 1 \; \forall s$. The theorem proved in the appendix only guarantees, however, that this update will result in an increase of the MMI objective function if all the D_g are equal to some large enough constant D.

An alternative update that uses both numerator and denominator statistics and has been reported to result in larger MMI objective function increase [10] is obtained by maximizing

$$\sum_s \sum_{g;s(g)=s} n^{ml}(g) \log \pi_g - \frac{n^{den}(g)}{\pi_g^0} \pi_g \tag{81}$$

For MMI estimation we experiment with both these methods, as well as with the ML updates of the mixture weights with only the "numerator counts" $n^{ml}(g)$.

For error weighted training, the mixture weights can be updated with error weighted counts $n^{\alpha}(g)$; or a simple ML update using $n^{ml}(g)$ can be performed. As it is mentioned in the experimental results (Section VI-B), the different mixture weight update strategies did not matter for MMI and for this reason we chose only the ML update for error weighted training.

3) Likelihood Scaling: Another parameter that plays a crucial role in MMI training is the scaling of the acoustic likelihoods $P(X|S)$ in computing the denominator statistics. As discussed in the next section, for the experiments we consider we find it best to use the identity scaling. For this reason we have not included a scaling constant in any of the formulas above; although we point out to the reader that in general one should be included.

VI. Experimental Results

In this section we report on results of MMI and error weighted training of SPAM and SCGMM model parameters. Our experimental plan is as follows. First, in a set of experiments on a SPAM model, we gauge the sensitivity of discriminative estimation to the heuristics mentioned in Section V-C. The parameter values and strategies learned are then applied to discriminative estimation of SCGMM models. In experiments with SCGMMs, we also analyze the effect of the SCGMM subspace dimension on discriminative estimation by building models with different subspace sizes and by discriminatively updating a full covariance model. Following these, we apply a combination of error weighted training and MMI estimation [15] to SCGMMs. The application of this method to the SPAM model was presented in [15], we duplicate those results here for completeness.

In our final set of experiments, we study the effect of discriminatively estimating a full covariance model in 117 dimensions, seeding with a full covariance model built in 52 dimensional space. This is an effort to complement and understand previous results [5] where we examined the degredation in performance of ML trained models when going from 52 dimensional LDA features to the full 117 dimensional unprojected feature space.

A. Data Sets and System Description

The testing was carried out on a data set contained digit strings of constrained length (seven and ten). These strings were recorded in a car under three conditions : idling, moving at about 30 miles per hour, and moving at about 60 miles per hour. There were altogether 10298 sentences and 71084 words in the test set.

Bootstrap models are built using a "full" training data set consisting of 462K sentences which are a mix of digit and non-digit word strings. The models that we report on here are trained on a digit specific subset of the full training set which was comprised of 66608 sentences. This was collected in cars under the three conditions described above; Although the majority of the digit data was collected under the idling condition.

The acoustic feature vectors were obtained by first computing 13 Mel-frequency cepstral coefficients (including energy) for each time slice under a 25 msec. window with a 15 msec. shift. The front end also performs adaptive mean and energy normalization. Nine 13-dimensional vectors were concatenated to form 117 dimensional features which were then projected to a 52 dimensional space using LDA. All of the acoustic models, with the exception of full covariance models described in Section VI-E, were built on the LDA projected 52-dimensional features. The full covariance models of Section VI-E were built on 117 dimensional spliced features. The phone set was a digit specific phone set containing 37 phones and 2 silence phones. Each phone was modeled with a three state left to right HMM, resulting in a total of 117 context independent states.

We experimented with two types of acoustic models differing in the number of Gaussian components. The SPAM models of Section VI-B were built with a mixture of 15 Gaussians for each digit phone state and 100 Gaussians for each silence state; resulting in a total of 2265 Gaussians. In contrast, the SCGMMs and full covariance models of Sections VI-C through VI-E were built with a total of 4701 Gaussians determined using the Bayesian information criterion [26].

All experiments use the same grammar based language models, HMM state transition probabilities, and Viterbi decoder which is passed state dependent probabilities for each frame vector which are obtained by table lookup [27] based on the ranking of probabilities obtained with a constrained Gaussian mixture model.

B. Sensitivity of Discriminative Estimation to Variation in Heuristic Parameters

Our first set of experiments were carried out on SPAM models on the $d = 52$ dimensional LDA features with $D = 13$ dimensional tied precision subspace and $L = 13$ dimensional tied mean subspace. These results have previously been reported in [15]. A bootstrap model from which the baseline model was obtained was a SPAM model built on all the training data, following the ML training procedure specified in [7], using full covariance statistics collected on all of the training data. This bootstrap model had a test set word/sentence error rate of 2.14/12.64%. This model was then specialized on digits using several iterations of Baum-Welch training on the digit specific subset of the training data. The resulting model was used as the baseline SPAM model for subsequent discriminative training. This baseline model had word/sentence error rates of 1.78/10.41% on the full test set.

For error weighted training, a recognition pass was carried out on the digit specific subset of the training data with the baseline SPAM model. This resulted in a *training* set word/sentence error rate of 1.36/10.19%. The 10.19% sentences with errors were then used to collect statistics, in addition to statistics collected over the entire digit training data, to carry out error weighted training.

baseline word/sentence error rate : 1.78/10.41							
optimized parameters		α values					
		0	8	64	128	256	1024
untied	wer	1.79	1.55	1.46	1.45	1.45	1.45
	ser	10.46	9.37	8.97	8.97	9.00	9.03
all	wer	1.78	1.54	1.40	1.39	1.39	1.38
	ser	10.40	9.16	8.56	8.50	8.49	8.44

TABLE I

ERROR WEIGHTED TRAINING WITH DIFFERENT VALUES OF α. REPORTED ARE WORD AND SENTENCE ERROR RATES FOR SPAM MODELS ON 52 DIMENSIONAL LDA FEATURE SPACE, WITH D=13 DIMENSIONAL PRECISION SUBSPACE AND L=13 DIMENSIONAL MEAN SUBSPACE. ALL MODELS WERE OBTAINED FROM THE BASELINE MODEL BY ONE ROUND OF ERROR WEIGHTED TRAINING, FOR THE CASE WHEN ONLY THE UNTIED PARAMETERS ARE UPDATED AS WELL AS THE CASE WHEN BOTH THE TIED AND UNTIED PARAMETERS ARE UPDATED.

The error rate performance of this training method is presented in Table I. We experimented with updating just the untied parameters as well as both tied and untied parameters. There are several things to be noted from this table. First of all, we note that even though $\alpha = 0$ means no contribution from the error statistics, the error rates may be different from the baseline numbers because of one additional EM iteration of error weighted training. Updating both tied and untied parameters is consistently better than updating just the untied parameters. With increasing weight on the statistics gathered from the sentences decoded incorrectly, the error rates drop strikingly - the largest word error rate gain, obtained at $\alpha = 1024$, is over 22% relative over the baseline, and the corresponding sentence error rate improvement is about 19% relative. The experiments were stopped at $\alpha = 1024$ because the changes in error rates were marginal.

Note that even when we weighted the 10.19% of training data that was in error by 1024, the test set error rate did not increase. This may be due to the fact that we carry out only one EM iteration using the error weighted statistics (so that the model retains a strong "memory" of the baseline model). By contrast, as we see in Section VI-C, for SCGMMs of higher dimension the test error rate does not always decrease monotonically with α, the error does start to increase with increased weights on error statistics.

To gather MMI statistics we chose an acoustic scaling [10] of 1.0. This choice was motivated by the observation that error weighted training results in a large improvements when the statistics from sentences in error are weighted heavily ($\alpha = 1024$). Intuitively, a larger scale (when it multiplies the acoustic model likelihoods) would lead to a cancellation of numerator and denominator statistics gathered from correctly recognized sentences and hence would tend to focus more on sentences with errors. This intuition was corroborated by some preliminary experiments where acoustic scale of 1.0 was found to be significantly better than other smaller values we tried. Since the three different mixture weight update methods discussed in Section V-C.2 were within 0.5% relative of each other, we restrict to reporting results when the ML update of mixture weights $\{\pi_g\}$ was used.

The MMI numerator and denominator statistics were combined to form the auxiliary function of (59) with a D_g value selected according to (71). The values of constants C_1 and C_2 in (71) were searched over sequentially, starting from their recommended values of 1.0 and 2.0, respectively [10]. We first found the optimal value of C_1 keeping C_2 fixed at 2.0, and then C_2 was searched over with C_1 fixed at this optimal value. These results are presented in Table II.

We note that even though MMI has a lot more heuristic parameters than error weighted training, the optimal MMI performance of 1.36/8.25% is only marginally better than the optimal error weighted performance of 1.38/8.44%.

C. Discriminative Estimation of SCGMMs and Effect of Subspace Dimensionality

Our next set of experiments were to study the effect of discriminative estimation on our most general form of subspace constrained GMMs, namely the SCGMM models described in Section II-B. In these experiments we also explored the effect of subspace dimensionality on discriminative estimation.

Four baseline SCGMMs were built with subspace dimensionality, F, of 16, 26, 40, and 78. These models were built in 52 dimensional LDA feature space using a procedure similar to one used for the SPAM model of Section VI-

baseline word/sentence error rate : 1.78/10.41							
optimized parameters		C_1 values, $C_2 = 2.0$					
		1.0	0.5	0.25	0.1	0.05	0.01
untied	wer	1.69	1.62	1.56	1.44	1.41	1.43
	ser	10.0	9.68	9.37	8.76	8.59	8.67
all	wer	1.68	1.60	1.55	1.40	1.39	1.41
	ser	9.97	9.58	9.37	8.54	8.41	8.59
optimized parameters		C_2 values, $C_1 = 0.05$					
		2.0	1.5	1.1	1.05	1.01	1.005
untied	wer	1.41	1.36	1.37	1.36	1.36	1.36
	ser	8.59	8.31	8.46	8.41	8.41	8.41
all	wer	1.39	1.36	1.36	1.36	1.36	1.36
	ser	8.41	8.25	8.31	8.33	8.35	8.35

TABLE II

ERROR RATES FOR MMI UPDATE OF BASELINE SPAM($d = 52, D = 13, L = 13$) MODEL WITH DIFFERENT C_1 AND C_2 VALUES, WHEN ONLY THE UNTIED PARAMETERS ARE UPDATED AS WELL AS WHEN ALL PARAMETERS ARE UPDATED.

B. First, four bootstrap SCGMMs (one of each subspace size) were built on all the training data, following the procedure specified in [7], using full covariance statistics collected on all of the training data. These models were then specialized on digits using two iterations of Baum-Welch training on the digit specific subset of the training data. The resulting models had test set word/sentence error rates of 1.66/10.57%, 1.11/7.07%, 0.90/5.86%, and 0.71/4.76%, respectively.

For discriminative estimation of these models, we use the parameter values and strategies that were found to be optimal for the smaller SPAM model of Section VI-B. In particular, we use - likelihood scaling of 1.0 for gathering denominator MMI statistics, ML update of mixture weights for both MMI and error weighted training, $\alpha = 1024$ for error weighted training, and $C1 = 0.05, C2 = 1.5$. We update both tied and untied parameters since that was found to be consistently better than updating only the untied ones. The word and sentence error rates for these four models are presented in Table III.

From Table III we note that discriminative estimation of SCGMMs consistently yields improved accuracies, except error weighted training for $F = 78$ where the sentence error rate sees a slight degradation. The gains due to MMI estimation range from around 14% relative for the smallest model size to around 8% relative for the largest one. Furthermore, note that with growing model size the baseline error rates decrease and at the same time the gains that we get from discriminative estimation goes down. This is in accordance with our intuition that as the number of parameters in the model grows the choice of their placement strategy, as long as taken from a set of reasonable alternatives, becomes less important.

Since the parameters for these experiments were based on the results under a SPAM($d = 52, D = 13, L = 13$)

	optimized parameters		subspace dimension (F)			
			16	26	40	78
baseline	all	wer	1.66	1.11	0.90	0.71
		ser	10.57	7.07	5.86	4.76
error weighted	all	wer	1.45	1.00	0.85	0.71
training		ser	9.36	6.54	5.52	4.82
MMI	all	wer	1.41	0.95	0.77	0.65
		ser	9.02	6.16	5.07	4.36

TABLE III

MMI AND ERROR WEIGHTED TRAINING OF SUBSPACE CONSTRAINED GAUSSIAN MIXTURES MODELS WITH VARYING SUBSPACE DIMENSION F. ALL MODELS ARE ON 52 DIMENSIONAL LDA FEATURE SPACE AND ARE OBTAINED BY ONE ROUND OF UPDATING BOTH TIED AND UNTIED PARAMETERS TO MAXIMIZE THE AUXILIARY FUNCTION. FOR MMI, TRAINING D_g WAS CHOSEN USING (71) WITH $C1 = 0.05$ AND $C2 = 1.5$. FOR ERROR WEIGHTED TRAINING, $\alpha = 1024$.

model, these may be suboptimal for SCGMMs and it may be for this reason that we see a degradation in case of error weighted training of $F = 78$ model. To analyze this we present in Table IV the results of varying the α parameter for error weighted training of the four SCGMMs. From these results we note that for $F = 78$ there are alpha values for which error weighted training results in an smaller word/sentence errors. However, the optimal α value varies with subspace size and seems to get smaller with increase in subspace dimensionality.

F		base-	error weighted training, α values					
		line	0	8	64	128	256	1024
16	wer	1.66	1.63	1.50	1.46	1.45	1.45	1.45
	ser	10.57	10.33	9.62	9.39	9.37	9.38	9.36
26	wer	1.11	1.11	1.04	1.00	1.00	1.00	1.00
	ser	7.07	7.03	6.70	6.51	6.55	6.54	6.54
40	wer	0.90	0.90	0.83	0.84	0.84	0.84	0.85
	ser	5.86	5.86	5.44	5.49	5.48	5.50	5.52
78	wer	0.71	0.71	0.67	0.70	0.71	0.71	0.71
	ser	4.76	4.72	4.52	4.68	4.75	4.80	4.82

TABLE IV

EFFECT OF α ON ERROR WEIGHTED TRAINING OF SCGMMS. RESULTS ARE COMPARABLE TO ERROR WEIGHTED RESULT IN TABLE III WHICH GIVES VALUES FOR $\alpha = 1024$ ONLY.

To see whether our choice of parameters for MMI actually resulted in an objective function improvement, we computed the MMI objective function (44) for the four MMI updated SCGMMs and compared those with their respective baseline values. These quantities are presented in Table V. In that table, "num" refers to the numerator

log-likelihood, $\log L(\Omega)$, computed over the entire training data, "den" is the denominator log-likelihood, $\log M(\Omega)$, over the training data, and, "diff" is the log of the objective function, $\log R(\Omega)$, obtained by subtracting denominator from numerator.

		subspace dimension (F)			
		16	26	40	78
baseline	num	-1.2134e9	-1.1874e9	-1.1703e9	-1.1534e9
	den	-1.2148e9	-1.1889e9	-1.1718e9	-1.1549e9
	diff	1.418e7	1.458e7	1.483e7	1.507e7
MMI	num	-1.2137e9	-1.1878e9	-1.1706e9	-1.1538e9
updated	den	-1.2152e9	-1.1893e9	-1.1722e9	-1.1554e9
	diff	1.506e7	1.530e7	1.548e7	1.568e7

TABLE V

MMI OBJECTIVE FUNCTION VALUES FOR BASELINE AND MMI MODELS IN TABLE III. "NUM" REFERS TO THE NUMERATOR $\log L$; "DEN" IS THE DENOMINATOR $\log M$; AND 'DIFF" IS $\log R = \log L - \log M$.

From Table V we note that, as expected, the baseline likelihood of the training data under both numerator and denominator increases with subspace dimensionality. Note that the baseline MMI objective function also increases with increasing subspace size, as we should expect since the model is becoming more accurate. MMI estimation leads to an increase in MMI objective function, even with all the heuristic choices that were made. It is interesting to note that the increase in objective function is achieved by reducing both numerator and denominator likelihoods, the latter more than the former. The better error rate performance of MMI estimated models, even though they have a lower numerator likelihood than the baseline models, says that the MMI objective function is better correlated with error rate than the ML objective function.

In a related set of experiments we carry out MMI estimation of a full covariance model; this model corresponds to a subspace dimension of $F = 1430$. A baseline full covariance model was built by first converting the full covariance statistics collected on the entire training data into a bootstrap model and then carrying out one iteration of Baum-Welch training using the digit only data. This baseline model had a test set word/sentence error rate of 0.66/4.44%. Comparing this with the SCGMMs of four subspace sizes above we see that while there is a significant gain in going from subspace size 16 to 78, the gain is relatively much smaller from 78 to full covariance model. However, as also observed in [5], with the amount of training data we are using, even the full covariance model does not seem to be over-trained.

MMI estimation of the baseline full covariance model was carried out analogously to that of SCGMMs, using the same values of the heuristic parameters. This resulted in word/sentence error rates of 0.65/4.22%; a negligible improvement over the baseline. However, it may be useful to observe that as with ML estimation of the full covariance model, even discriminative estimation does not seem to be overtraining with the amount of data used.

D. Combination of Error Weighted Training and MMI

In this section we discuss a simple method of combining MMI and error weighted training that results in lower error rates than either method alone. This method was presented for the SPAM model in [15], we first reproduce those results and then extend its application to SCGMMs.

Comparing the error rate performance of SPAM untied parameter estimation under the two discriminative criteria (Tables I and II), it appears that MMIE is significantly better than error weighted training at estimating these parameters. However when both tied and untied parameters are estimated the error weighted training is quite close in performance to MMIE, suggesting that error weighted training may be better at estimating the tied model parameters. To confirm this hypothesis, we combined the two estimation procedures by taking the tied model from error weighted training ($\alpha = 1024$) and updated the untied parameters with MMI using optimal parameter values from Table II. This resulted in our best SPAM model with a word/sentence error rate of 1.27/8.13% which is a word error rate improvement of over 28% relative over the baseline 1.78/10.41%.

		subspace dimension (F)			
		16	26	40	78
baseline	wer	1.66	1.11	0.90	0.71
	ser	10.57	7.07	5.86	4.76
error weighted	wer	1.45	1.00	0.83	0.67
cheating	ser	9.36	6.51	5.44	4.52
MMI	wer	1.41	0.95	0.77	0.65
	ser	9.02	6.16	5.07	4.36
combination	wer	1.31	0.92	0.76	0.64
method	ser	8.61	6.05	5.00	4.31

TABLE VI

COMBINATION OF MMI AND ERROR WEIGHTED TRAINING FOR SCGMMS. SAME SETUP AS TABLE III BUT NOW ERROR WEIGHTED TRAINING USE OPTIMAL α AND ADDITIONAL COMBINATION METHOD USES THE TIED SUBSPACE FROM ERROR WEIGHTED TRAINING AND MMI TRAINING FOR THE UNTIED PARAMETERS.

To apply this combination method to SCGMM models discusses in Section VI-C, we first update the tied parameters with error weighted training using the optimal α for each model size (cheating), i.e. $\alpha = 1024$ for $F = 16$, $\alpha = 64$ for $F = 26$, and $\alpha = 8$ for $F = 40$ and $F = 78$, and then update the untied parameters with MMI using parameters as described in Section VI-C. The resulting models had error rates as presented in Table VI. As with the SPAM model case, the combination method for SCGMMs yields lower error rates than MMI or error weighted training by itself, although the improvement only has some significance for small subspace size.

E. Discriminative Lifting of FC Models from 52 *to* 117 *Dim*

In our final set of experiments, we study the effect of discriminatively estimating a full covariance model in 117 dimensions, seeding with a full covariance model built in 52 dimensional space. This is an effort to complement previous results [5] where we examined the degredation in performance of ML trained models when going from 52 dimensional LDA features to the full 117 dimensional unprojected feature space.

In Table VII we report word and sentence error rate results for experiments with full covariance models on the 52 dimensional LDA feature space we have used so far, as well as on the 117 dimensional unprojected space. The first column of the table reports results for a seed model. In the 52 dimensional case, the seed model is the same as the ML trained model mentioned at the end of section VI-C. In the 117 dimensional case, the seed model is obtained by extending the 52 dimensional full covariance seed Gausian mixture model to 117 dimensions using the total mean and covariance matrix in the dimensions complementary to the first 52 LDA dimensions. More concretely, the last $117 - 52$ components of the means for all Gaussians in the 117 dimensional seed model are set to those components of the total data mean and the 117 dimensional covariance matrices are taken to be block diagonal with a Gaussian dependent 52×52 block and a Gaussian independent $(117 - 52) \times (117 - 52)$ block coming from the total data covariance matrix. Note that the 52 and 117 dimensional seed models result in the same speech recognition output because the extension to the additional dimensions is Gaussian independent.

The second column of Table VII gives the results for one round of ML training using the EM algorithm. The third column reports the results for one round of MMI training using the same parameter settings as in Tables III, V, and VI. The effect of ML training in 52 dimensions is minimal since the seed model is already trained by ML to a fair degree of convergence. However, the 117 dimensional seed model undergoes significant degredation when subjected to one round of ML training. This is consistent with the results in [5] and the point of view that including too many LDA dimensions causes the models to focus too much on non-informative noise dimensions. By contrast, MMI training in 117 dimensions does not result in degredation because the training is based on an information criterion.

The astute reader may remark that the lack of degredation mentioned in the previous sentence may be caused by the fact that the parameters D_g are so large that the 117 dimensional MMI training is simply not moving the model. This is not the case, however, as can be seen by the fact that there is a significant word/sentence "error rate" of $0.33\%/2.37\%$ between the 117 dimensional MMI model and the seed model. Similarly the word/sentence error rate between the 52 dimensional MMI model and the seed model is $0.25\%/1.85\%$, which is again significant.

VII. CONCLUSION

Subspace constrained Gaussian mixture models provide a powerful and very flexible class of acoustic models for speech recognition. In a series of papers [6], [7], [23], [24] we developed techniques for ML training of such models and explored the relationship between subspace constrained modeling and dimensional reduction through the LDA technique. The experiments in those papers were applied only to small vocabulary tasks. In [28] we applied subspace constrained modeling to unlimited resource large vocabulary tasks in a speaker adaptive setting and we

	seed	ML	MMI
52	0.66/4.44	0.68/4.60	0.65/4.22
117	0.66/4.44	1.02/6.30	0.68/4.34

TABLE VII

WORD/SENTENCE ERROR RATES FOR FULL COVARIANCE MODELS IN 52 AND 117 DIMENSIONS. THE FIRST COLUMN GIVES RESULTS FOR A SEED MODEL 52 DIMENSIONAL MODEL AND ITS "LIFT" TO A 117 DIMENSIONAL MODEL WHOSE COMPONENTS COMPLEMENTARY TO THE FIRST 52 LDA DIMENSIONS ARE GAUSSIAN INDEPENDENT. THE SECOND AND THIRD COLUMN GIVE RESULTS AFTER ONE ROUND OF ML AND MMI TRAINING.

saw reasonable error rate reductions. All of these and other results were presented in a comprehensive way in [5]. In [15] we began an exploration of discriminative training of subspace constrained models. We introduced error weighted training in that paper and applied both MMI and error weighted training to SPAM models. In this paper, which may be viewed as a companion to [5], we have tried to present a systematic approach to discriminative training of subspace constrained models.

We have seen that both MMI and error weighted training yield improvements for SPAM models as well as general SCGMM models. We found gains both from training the untied parameters as well as the tied subspace. The best overal gain is found when the tied subspace is trained by error weighted training and the untied parameters are trained by MMI. We have seen that the relative improvement is best when the subspace dimension is low so that there is a smaller number of parameters to train. In the unconstrained case, i.e. for full covariance models, our experiments showed that MMI training yields no significant improvement, but on the other hand it avoids the significant degradation of ML training in the 117 dimensional case. It will be very interesting to see how these results carry over to other applications besides the noisy digit application to which all of the experiments here are restricted.

We have provided a general proof that the technique of optimizing an auxiliary function yields an improvement of the MMI objective function, assuming that the Gaussian dependent algorithm parameters D_g are chosen large enough. Our proof applies for any type of acoustic modeling with arbitrary parameter tying, in particular to general subspace constrained Gaussian mixture models and the special cases of SPAM, EMLLT, MLLT, and diagonal models. Our general update equation reduces to the standard one in the diagonal case.

Although our proof is very general, we do not provide a specific formula for the minimal allowable D_g. In our experiments we are forced to resort to a generalization of the heuristic given in [10] for choosing D_g. One of our motivation for being careful to be precise about the technical qualifications (analyticity, positivity, and effective compactness of the parameter space) on the class of acoustic models is the hope that the detailed analysis here will motivate future research at arriving at a useful formula for the lower bound for D_g. We close by remarking that, although the detailed proof here may appear rather technical, the underlying idea is a very simple and powerful application of power series analysis.

REFERENCES

[1] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[2] R. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. ICASSP*, 1998.

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

[4] ——, "Modeling inverse covariance matrices by basis expansion," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 37–46, 2004.

[5] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariah, "Subspace constrained gaussian mixture models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, 2003, submitted.

[6] S. Axelrod, R. A. Gopinath, P. Olsen, and K. Visweswariah, "Dimensional reduction, covariance modeling, and computational complexity in ASR systems," in *Proc. ICASSP*, 2003.

[7] K. Visweswariah, S. Axelrod, and R. Gopinath, "Acoustic modeling with mixtures of subspace constrained exponential models," in *Proc. Eurospeech*, 2003, to appear.

[8] V. Vanhoucke and A. Sankar, "Mixtures of inverse covariances," in *Proc. ICASSP*, 2003.

[9] 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.

[10] P. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–47, 2002.

[11] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 4, pp. 814 – 817, 1983.

[12] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043 – 3054, 1992.

[13] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Estimating hidden Markov model parameters so as to maximize speech recognition accuracy," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 77 – 83, 1993.

[14] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, March 2003.

[15] V. Goel, S. Axelrod, R. Gopinath, P. Olsen, and K. Visweswariah, "Discriminative estimation of subspace precision and mean (SPAM) models," in *Proc. Eurospeech*, 2003, to appear.

[16] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, 1986.

[17] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of the American Mathematical Society*, vol. 73, pp. 360–363, 1967.

[18] P. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Transactions on Information Theory*, vol. 37, pp. 107 – 113, 1991.

[19] Y. Normandin, "Hidden Markov models, maximum mutual information estimation and the speech recognition problem," Ph.D. dissertation, McGill University, Montreal, 1991.

[20] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. ICASSP*, 2002.

[21] D. Kanevsky, "Extended Baum transformations for general functions," in *Proc. ICASSP*, 2004, to appear.

[22] T. Jebara and A. Pentland, "On reversing Jensen's inequality," in *Proceedings of NIPS*, 2000.

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

[24] K. Visweswariah, P. Olsen, R. A. Gopinath, and S. Axelrod, "Maximum likelihood training of subspaces for inverse covariance modeling," in *Proc. ICASSP*, 2003.

[25] Y. Freund and R. Schapire, "Decision theoretic generalization of on-line learning and an application to boosting," in *Second European Conference on Computational Learning Theory*, 1995.

[26] S. Chen and R. A. Gopinath, "Model selection in acoustic modeling," in *Proc. Eurospeech*, 1999.

[27] L. Bahl, S. Balakrishnan-Aiyer, M. Franz, P. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan, and S. Roukos, "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task," in *Proc. ICASSP*, 1995.

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

APPENDIX I
PROOF THAT THE AUXILIARY FUNCTION FOR MMI TRAINING IS VALID

In this appendix we prove that $Q_{mmi}(\Omega, \Omega^0; D)$ (53) is an auxiliary function for the MMI objective function $R(\Omega)$ (44) provided D is chosen large enough. We also prove that a value Ω of Ω^0 for which $Q_{mmi}(\Omega, \Omega^0; D)$ is greater than $Q_{mmi}(\Omega^0, \Omega^0; D)$ is necessarily near to Ω^0 if D is large enough. The theorem is proved for any type of component distributions $p(x|g; \theta_g)$, where there may be any sort of parameter tying across the components, which is encoded in the choice of the set $\mathcal{K}$ of allowable parameters. SCGMMs are just one special kind of model to which this theorem can be applied. For simplicity, we hold the priors fixed in the theorem, although the same technique can be used to prove that Q_{mmi} is a valid auxiliary function even for varying priors, as long as the constants D_g are independent of g (see comment below (48)).

Theorem. *Let (X^*, W^*) be training data for an HMM based speech recognizer with state distribution of the form*

$$p(x|s; \Omega) = \sum_{g \in \mathcal{G}(s)} \pi_g p(x|g; \theta_g) \ , \tag{82}$$

where the distribution $p(x|g; \theta_g)$ is a nowhere vanishing *probability distribution conditioned on the component number g which depends analytically on x and some parameters θ_g. The set of all component parameters $\Theta = \{\theta_g\}$ is required to belong to a subset $\mathcal{K}$ of Euclidean space. The priors $\pi = \{\pi_g\}$ are to be held fixed, so that a choice of $\Theta \in \mathcal{O}$ determines all the parameters $\Omega = (\pi, \Theta)$ of the full state model. Assume given and fixed the HMM joint distribution $P(S, W)$ over state and word sequence as well as a "base" model $\Omega^0 = (\pi, \Theta^0)$.*

Fix positive constants $D_G^{(1)}$ depending on a component sequence $G = (g(1), ..., g(T))$, where T is the total number of time frames in X^*. For any $\epsilon > 0$, let $D_G = D_G^{(1)}/\epsilon$ and

$$D \ = \ \{D_G\} = \{D_G^{(1)}/\epsilon\} = D^{(1)}/\epsilon \ . \tag{83}$$

Also let $R(\Omega) = P(X^*|W; \Omega)/P(X^*; \Omega)$ be the MMI objective function, y denote the pair (X, G), S be the state sequence determined by G, and:

$$q_0(y) \ = \ D_G^{(1)} P(S), \tag{84}$$

$$q'(y) \ = \ \delta(X - X^*) \left[P(S|W^*) - R(\Omega^0) P(S) \right], \tag{85}$$

$$q_\epsilon(y) \ = \ q_0(y) + \epsilon q'(y), \tag{86}$$

$$\mathcal{P}(y; \Omega) \ = \ P(X, G|S; \Omega), \tag{87}$$

$$F_\epsilon(\Omega) \ = \ \int_y q_\epsilon(y) \left[\mathcal{P}(y; \Omega) - \mathcal{P}(y; \Omega^0) \right], \ \textit{and} \tag{88}$$

$$Q_\epsilon(\Omega) \ = \ \int_y q_\epsilon(y) \mathcal{P}(y; \Omega^0) \log \frac{\mathcal{P}(y; \Omega)}{\mathcal{P}(y; \Omega^0)} \tag{89}$$

So

$$F_\epsilon(\Omega)/\epsilon \quad = \quad F(\Omega,\Omega^0) - F(\Omega^0,\Omega^0), \text{ and} \tag{90}$$

$$Q_\epsilon(\Omega)/\epsilon \quad = \quad Q_{mmi}(\Omega,\Omega^0;D) - Q_{mmi}(\Omega^0,\Omega^0;D) \tag{91}$$

where F and Q_{mmi} on the right hand side above are as defined in the section V-A. In particular, $Q_\epsilon(\Omega^0)=0$.

The following are true assuming either that $\mathcal{K}$ is compact or that $p(x|g,\theta_g)$ is an exponential model:

C1. *For any open neighborhood $\mathcal{O}'$ of Θ^0 in $\mathcal{K}$, there is a positive ϵ_0 such that, for any ϵ smaller than ϵ_0, $Q_\epsilon(\Omega)>0$ implies that Θ belongs to $\mathcal{O}'$.*

C2. *There is a positive ϵ_1 such that, for any ϵ smaller than ϵ_1,*

$$Q_\epsilon(\Omega) \quad \leq \quad \epsilon \, \left[R(\Omega) - R(\Omega^0)\right] \tag{92}$$

for any $\Omega=(\pi,\Theta)$ with $\Theta\in\mathcal{K}_\epsilon$. Here $\mathcal{K}_\epsilon$ equals $\mathcal{K}$ when $\mathcal{K}$ is compact; otherwise $\mathcal{K}_\epsilon$ is the set of $\Theta\in\mathcal{K}$ for which $Q_\epsilon(\pi,\Theta)$ is positive.

Proof of **C1** *assuming $\mathcal{K}$ is compact:*

First note that

$$Q_0(\Omega) \quad = \quad \sum_G D_G^{(1)} P(G) \int_X P(X|G;\Omega^0) \log \frac{P(X|G;\Omega)}{P(X|G;\Omega^0)} \ .$$

We have written the probabilities inside the logarithm as conditioned on G, which we are able to do since the prior terms cancel. From this it follows that the global maximum of $Q_0(\Omega)$ occurs when $P(X|G;\Omega)$ equals $P(X|G;\Omega^0)$. Identifying parameter values that yield the same distribution, we can say simply that the global maximum occurs when Ω equals Ω^0.

Conclusion **C1** now follows from some simple topological reasoning, which we now spell out for the sake of completeness. The reader is encouraged to draw some pictures to verify that the argument is actually quite simple. To begin, we let $H:[0,1]\times\mathcal{K}\mapsto[0,1]\times\mathbf{R}$ be the map

$$H(\epsilon,\Theta) = (\epsilon, Q_\epsilon(\pi,\Theta)) \ . \tag{93}$$

For $r\leq 1$, let $B_r=[0,r)\times(-r,\infty)$ and $\mathcal{B}_r=H^{-1}(B_r)$. Note that $\mathcal{B}_r$ is open in $[0,1]\times\mathcal{K}$ for $r>0$ (because B_r is open within $[0,1]\times\mathbf{R}$ and H is continuous). Since B_r shrinks with r and Ω^0 is the global maximum of Q_0, we see that $\mathcal{B}_r$ shrinks with r to the point $(0,\Theta^0)$, i.e.

$$\mathcal{B}_r \quad \subseteq \quad \mathcal{B}_{r'} \qquad \text{for } r<r', \tag{94}$$

$$\mathcal{B}_0 \quad = \quad \cap_{r>0}\mathcal{B}_r = \{(0,\Theta^0)\} \ . \tag{95}$$

But a family of open sets shrinking to a point within a compact set must eventually be contained within any given open neighborhood of that point. Thus, for any neighborhood $\mathcal{O}'$ about Θ^0 in $\mathcal{K}$, there is some ϵ_0, such that

$$\mathcal{B}_{\epsilon_0} \subset [0,\epsilon_0)\times\mathcal{O}' \ . \tag{96}$$

This says that, if $\epsilon < \epsilon_0$ and $Q_\epsilon(\pi, \Theta)$ is greater than $-\epsilon$, then Θ belongs to $\mathcal{O}'$. This is (slightly stronger than) the desired result **C1**.

Proof of **C1** *for exponential models:*

For exponential models, $Q_\epsilon(\Omega)$ can be written in terms of the sufficient statistics (61), which we write here as a linear function of ϵ:

$$\hat{f}^{mmi}_{g,\epsilon} = \hat{f}^{mmi}_{g,0} + \epsilon \hat{f}'_g \tag{97}$$

$$\hat{f}^{mmi}_{g,0} = \hat{f}^{ml}_g - \hat{f}^{den}_g \tag{98}$$

$$\hat{f}'_g = D^{(1)}_g \, E_{\theta^0_g}(f(x)) \ . \tag{99}$$

Discarding irrelevant constants,

$$Q_\epsilon(\Omega) = \sum_g Q_g(\Theta_g) - Q_g(\Theta^0_g) \tag{100}$$

$$Q_g(\Theta_g) = \left[\theta^T_g \hat{f}^{mmi}_{g,0} + K(\theta_g)\right] + \epsilon \theta^T_g \hat{f}'_g \ . \tag{101}$$

The term in square brackets is concave with a maximum at θ^0_g. This implies that it falls off faster than $-\epsilon_0 ||\theta_g - \theta^0_g||$ for some ϵ_0. Hence, for $\epsilon < \epsilon_0$, the set of Θ for which $Q_\epsilon(\Omega)$ is positive is contained within some compact set $\mathcal{K}$. The proof of **C1** for $\mathcal{K}$ compact now applies.

Proof of **C2** *assuming $\mathcal{K}$ is compact:*

Since we already know from the section V-A that $F(\Omega, \Omega^0)$ is an auxiliary function for $R(\Theta)$, it suffices to show that for small ϵ, the difference

$$\Delta_\epsilon(\Omega) = F_\epsilon(\Omega) - Q_\epsilon(\Omega) \tag{102}$$

is positive for $\Theta \in \mathcal{K}$. We write the difference as

$$\Delta_\epsilon(\Omega) = \int q_\epsilon(y) \, \mathcal{P}(y; \Omega^0) \, \mathcal{H}(y, \Omega), \tag{103}$$

$$\mathcal{H}(y, \Omega) = h(\frac{\mathcal{P}(y; \Omega)}{\mathcal{P}(y; \Omega^0)} - 1), \text{ where} \tag{104}$$

$$h(s) = s - log(1 + s) > 0 \ . \tag{105}$$

Since h is always positive, positivity of q_ϵ would guarantee positivity of Δ_ϵ. Unfortunately, q_ϵ is not positive, even for small ϵ. But, and this is the heart of the proof, we can break up Δ_ϵ as

$$\Delta_\epsilon(\Omega) = \Delta_0(\Omega) + \epsilon \, \Delta'(\Omega) \tag{106}$$

$$\Delta'(\Omega) = \int q'(y) \, \mathcal{P}(y; \Omega^0) \, \mathcal{H}(y, \Omega), \tag{107}$$

$$= \sum_G \left[P(X^*, G | W^*; \Omega^0) - R(\Omega^0) P(X^*, G; \Omega^0)\right] \times \mathcal{H}(X^*, G; \Omega) \ .$$

Because Δ' get multiplied by a factor of ϵ, bounding of $|\Delta'|$ by a fixed (Θ independent) multiple of the strictly positive function Δ_0 would guarantee that Δ_ϵ is positive for ϵ small, which is what we need to show.

It suffices to bound the individual terms $P(X^*, G; \Omega^0)\mathcal{H}(X^*, G; \Omega)$ appearing in Δ' (107) by a fixed multiple of the term $\int_X P(X, G; \Omega^0)\mathcal{H}(X, G; \Omega)$ of Δ_0 ((103) with $\epsilon = 0$).

Now fix G and let f be the positive and analytic function:

$$f(X, \Theta) \quad = \quad P(X, G; \Omega^0)\mathcal{H}(X, G; \Omega) \ . \tag{108}$$

We need to show that there is a constant $C > 0$ such that

$$\int_X f(X, \Theta) \geq C \ f(X^*, \Theta) \tag{109}$$

We prove with a Taylor series argument below that, for $\tilde{\Theta}$ a point with

$$f(X^*, \tilde{\Theta}) = 0 \ , \tag{110}$$

there is an open ball about $\tilde{\Theta}$ for which (109) is true for some C. By compactness of $\mathcal{K}$ we can use a finite number of such balls to show that (109) is true for Θ in an open neighborhood $\mathcal{O}_0$ of the "zero set" (consisting of all $\tilde{\Theta}$ satisfying (110)).

The complement $\bar{\mathcal{O}}_0$ of the open set $\mathcal{O}_0$ is a compact subset of $\mathcal{K}$ on which $f(X^*, \Theta)$ is strictly positive and continuous. Since continuous functions are uniformly continuous on compact sets, there is a ball B about X^* so that $f(X, \Theta) \geq f(X^*, \Theta)/2$ for $X \in B$ and $\Theta \in \bar{\mathcal{O}}_0$. We conclude that (109) is true on $\bar{\mathcal{O}}_0$ for some C. Combined with the fact that (109) is true on $\mathcal{O}_0$ for some C, we have shown (109) holds for some C and any Θ.

Taylor Series Argument

The proof of **C2** in the case when $\mathcal{K}$ is compact is now complete, except for the promised Taylor series argument that surrounding any $\tilde{\Theta}$ satisfying (110) there is an open ball in $\mathcal{K}$ upon which (109) holds for some constant C. Of course, (109) is true even with $C = 0$ for $\Theta = \tilde{\Theta}$ because the right hand side is zero. What we need to do is verify that the left hand side does not vanish as $\Theta \mapsto \tilde{\Theta}$ faster than the right hand side.

By compactness of $\mathcal{K}$ and analyticity, we can choose a small ball B about X^* and a neighborhood $\mathcal{O}_{\tilde{\Theta}}$ of $\tilde{\Theta}$ upon which the Taylor series with remainder for f of any desired order is valid uniformly. We need only consider the Taylor series up to the leading non-vanishing term which is of some order m. (If all terms vanish, then by analyticity, f vanishes on a ball, about $(X^*, \tilde{\Theta})$ and we are done). Then $\int_X f(X, \Theta)$ is bounded below by

$$N(\Theta) \quad = \quad \int_{X \in B} f(X, \Theta) enspace. \tag{111}$$

Now $N(\Theta)$ vanishes to order m as $\Theta \mapsto \tilde{\Theta}$. On the other hand $f(X^*, \Theta)$ vanishes at least that fast, so we are done.

Proof of **C2** *for exponential models:*

Let $\mathcal{O}'$ be a small open neighborhood of Θ^0, i.e. a neighborhood whose closure, $\mathcal{K}'$, is compact. By **C1**, $\mathcal{K}_\epsilon$ is contained in $\mathcal{O}'$, and so $\mathcal{K}'$, for ϵ smaller than some constant ϵ_0. Then the proof of **C2** for compact sets applies

to show that, for ϵ smaller than some ϵ_1, (92) is valid for $\Theta \in \mathcal{K}'$. Therefore, (92) is valid for ϵ smaller than $\min(\epsilon_0, \epsilon_1)$ and $\Theta \in \mathcal{K}_\epsilon$.