

Gaussian Mixture Models Reduction by Variational Maximum Mutual Information

Yossi Bar-Yosef and Yuval Bistriz, *Fellow, IEEE*

Abstract—Gaussian mixture models (GMMs) are widely used in a variety of classification tasks where it is often important to approximate high order models by models with fewer components. The paper proposes a novel approach to this problem based on a parametric realization of the maximum mutual information (MMI) criterion and its approximation by a closed-form expression named *variational*-MMI (VMMI). The maximization of the VMMI can be carried out in an analytically tractable manner and it aims at improving the discrimination ability of the reduced set of models, a goal that was not targeted in previous approaches that simplify each class-related GMM independently. Two effective algorithms are proposed and studied for the optimization of the VMMI criterion. One is a steepest descent type algorithm, and the other, called *line search A-functions* (LSAF), uses concave associated functions. Experiments held in two speech related tasks, phone recognition and language recognition, demonstrate that the VMMI-based parametric model reduction algorithms significantly outperform previous non-discriminative methods. According to these experiments, the EM-like LSAF-based algorithm requires less iterations and converges to a better value of the objective function compared to the steepest descent algorithm.

Index Terms—Continuous-discrete MMI, discriminative learning, Gaussian mixture models reduction, hierarchical clustering.

I. INTRODUCTION

GAUSSIAN MIXTURE MODELS (GMMs) have been successfully used to model complex density functions for a variety of classification tasks. The adequate representation of large amounts of data with complex distributions requires high order models with large number of Gaussian components. A need to replace these high order models by models with reduced number of Gaussian components arises when the classification tool is imported to simpler computational platforms or has to meet real time processing constraints. Several studies in recent years have developed parametric methods to produce lower order models that approximate the given high order models [1]–[9]. These methods, that sometimes are also regarded as learning mixture hierarchies (after [1] that considered

learning the mixture parameters of level l from the knowledge of the parameters of the lower level $l + 1$), offer an effective and computationally less consuming alternative to learning the reduced models anew from the data. These methods also offer a good solution to situations where accessing the original data is no longer an option (e.g. [10], [11]), and the only remaining alternative is the expensive learning of models from Monte Carlo simulated samples.

The problem that is being considered involves classification of N classes, where each class c is presented by a mixture of multivariate Gaussian density functions of dimension d (the length of vector x) and of order (the number of Gaussian components) M_c ,

$$F_c(x) = \sum_{i=1}^{M_c} \alpha_{ci} f_{ci}(x) \quad , \quad c = 1, \dots, N \quad , \quad (1)$$

where $f_{ci}(x) \sim \mathcal{N}(m_{ci}, V_{ci})$, $i = 1, \dots, M_c$, and α_{ci} presenting the weights. The goal of the parametric approximation problem is to find a new set of models of reduced order $R_c < M_c$,

$$G_c(x) = \sum_{j=1}^{R_c} \beta_{cj} g_{cj}(x) \quad , \quad c = 1, \dots, N \quad , \quad (2)$$

with $g_{cj}(x) \sim \mathcal{N}(\mu_{cj}, \Sigma_{cj})$, and weights β_{cj} , that approximate the given set of high order models in some effective sense of similarity.

In [12] we have introduced the idea of treating the above problem by simplifying high order mixture models in a manner that aims at improving the discrimination capabilities of the reduced order GMMs. The *maximum correct association* (MCA) criterion that was proposed there, tries to fit each component in the set of high order models to its “correct” reduced model. All studies preceding [12], including those in [1]–[9], consider the maximization of a certain similarity measure between each reduced model $G_c(x)$ and its corresponding high-order model $F_c(x)$ without attending to the discrimination quality of the resulting set of reduced models. The MCA scheme was found to provide better reduced model sets in tasks of phone classification [12] and language authentication [13].

In this paper we address the above idea of discriminative model reduction of GMM classes from the perspective of maximum mutual information (MMI). Namely, we require that the mutual information between the *data* (represented by the set of high order models (1)) and the relevant *classes* (modeled by the set of reduced models (2)) is maximized. The MMI approach that was proposed for discriminative training of GMMs in [14]

Manuscript received August 11, 2014; revised January 08, 2015; accepted January 20, 2015. Date of publication January 30, 2015; date of current version February 18, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Prakash Ishwar.

The authors are with the School of Electrical Engineering, Tel Aviv University, Tel Aviv 69978, Israel (e-mail: bistriz@tau.ac.il; yossibaryosef@gmail.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSP.2015.2398844

provides a most effective method to train GMMs from empirical data. Our derivation enhances the MMI approach with a new dressing in showing that it implies a hierarchical discriminating criterion for deriving reduced order models, depending only on the knowledge of higher order models parameters. The resulting continuous-discrete MMI criterion produces an objective function that cannot be maximized by an analytically tractable algorithm. To overcome this difficulty, we use the variational approach [15] to approximate the MMI objective function by a closed-form objective function that we call *variational maximal mutual information* (VMMI).

In its second part, the paper considers efficient ways to optimize the VMMI criterion. Since, the VMMI function is not a convex (or concave) objective function, a straight forward approach is to optimize it by some gradient algorithm. Here we adopted to our need the generalized probabilistic descent (GPD) algorithm [16] that obtained good results in our experiments. The second optimization approach that we propose for the VMMI criterion is based on associating it with certain concave functions, which leads to a first-order optimization scheme that follows the Line Search Associated Function (LSAF) approach in [17]. The resulting algorithm provides an EM-like parameter estimation procedure that is shown empirically to provide an efficient and robust optimization scheme that compares favorably with the GPD based optimization.

The third part of the paper evaluates the above two optimization algorithms for the VMMI objective function by experiments held in two speech related tasks, phone recognition and language recognition. In these experiments, both algorithms outperform the advanced non-discriminative parametric model reduction methods proposed recently in [3], [4], [6], as well as low order models of comparable size trained directly from the original data by traditional expectation-maximization (EM).

Actually, as it will become apparent, the VMMI criterion provides an information theoretic interpretation (and some generalization) to the MCA criterion in [12], [13] and the LSAF-based optimization algorithm provides a more stable procedure that better optimizes the classification ability of the reduced GMMs than the optimization algorithms used previously in these references.

The paper is constructed as follows. The next section brings important background and details on non discriminative GMM reduction. Section III derives the VMMI criterion and generalizes it. Section IV describes its optimization by the GPD algorithm. Section V derives (with supplements brought in the Appendix) the LSAF algorithm for VMMI. Experimental results are presented in Section VI, and the paper ends with concluding remarks.

II. NON DISCRIMINATIVE GMM REDUCTION

Assume we have a single high order source model, $F(x) = \sum_i \alpha_i f_i(x)$, with M Gaussian components $f_i(x) \sim \mathcal{N}(m_i, V_i)$, and we seek a reduced target model $G(x) = \sum_j \beta_j g_j(x)$, with R components of $g_j(x) \sim \mathcal{N}(\mu_j, \Sigma_j)$, such that $R < M$. A reasonable approach to obtain the reduced GMM is to minimize the distance between the two probability density functions. The distance between two probability densities $F(x)$ and $G(x)$ is usually

measured by their *relative entropy* also known as the Kullback-Leibler (KL) divergence, and is given by the expression

$$KL(F||G) = \int F(x) \log \frac{F(x)}{G(x)} dx. \quad (3)$$

The divergence between two multivariate d -dimensional Gaussian distributions $f(x) \sim \mathcal{N}(m, V)$ and $g(x) \sim \mathcal{N}(\mu, \Sigma)$, has the following closed-form expression:

$$KL(f||g) = \frac{1}{2} \log \frac{|\Sigma|}{|V|} + \frac{1}{2} \text{tr}(\Sigma^{-1}V) - \frac{d}{2} + \frac{1}{2} (m - \mu)^T \Sigma^{-1} (m - \mu). \quad (4)$$

Unfortunately, no similar closed-form expression is available for the KL-divergence between two mixtures of Gaussians.

Several studies addressed the simplification of Gaussian mixture models, where most of them suggested some adequate closed-form approximation for the KL-divergence. Goldberger and Roweis [2] introduced a component grouping algorithm that minimizes an approximation of the KL-divergence based on Gaussian matching. Later, a soft version of the Gaussian-matching clustering appeared in [3] and in [4]. Bruneau *et al.* [5], used a parametric EM algorithm based on the variational-Bayes method for component reduction. Dognin *et al.* [6] followed the variational KL approximation, introduced by Hershey *et al.* [15], to suggest a so called *variational expectation-minimization* (varEM) algorithm for GMM simplification in the framework of speech recognition. Goldberger *et al.* too presented in [4] an EM algorithm based on the unscented-transform approximation of the KL-divergence. Variations of Bregman divergence were investigated by Nielsen *et al.* [7] in a k -means clustering approach, and later by Garcia *et al.* [8] with extensions to exponential families and soft Bregman clustering. Zhang and Kwok [9] obtained model simplification by minimizing an upper bound of the approximation error using the L_2 distance.

In the remaining of this section we consider the variational EM method described in [6], shown there to maximize a variational approximation of the likelihood measure between a source GMM given a desired target GMM. The resulting algorithm also coincides with previous reduction algorithms that were presented in [3] and in [4] (where it was named GMAC). In the following, we detail the variational approach because the introduced expressions provide relevant background for the further adoption of this technique to our proposed discriminative criterion. The varEM algorithm was shown to present well the best performance attainable in a speech recognition application [6] (approaching performance of a standard maximum-likelihood training from raw data). Therefore, we shall also take its performance as a baseline to which we compare our new results.

The cross-entropy (also the expected log-likelihood) between the source and target models is defined by

$$L(F||G) := \int F(x) \log G(x) dx = \sum_i \alpha_i \int f_i(x) \log \sum_j \beta_j g_j(x) dx. \quad (5)$$

As described in [15], a variational approximation for $L(F\|G)$ can be obtained by introducing variational distribution parameters, $q_{(j|i)} > 0$, subjected to $\sum_j q_{(j|i)} = 1$. By Jensen's inequality we obtain a lower bound for (5) as follows

$$\begin{aligned} L(F\|G) &= \sum_i \alpha_i \int f_i(x) \log \sum_j q_{(j|i)} \frac{\beta_j g_j(x)}{q_{(j|i)}} dx \\ &\geq \sum_i \alpha_i \int f_i(x) \sum_j q_{(j|i)} \log \frac{\beta_j g_j(x)}{q_{(j|i)}} dx \\ &= \sum_i \alpha_i \sum_j q_{(j|i)} \left[\log \frac{\beta_j}{q_{(j|i)}} + \int f_i(x) \log g_j(x) dx \right] \\ &= \sum_i \alpha_i \sum_j q_{(j|i)} \left[\log \frac{\beta_j}{q_{(j|i)}} - KL(f_i\|g_j) - H(f_i) \right] \\ &:= L_q(F\|G), \end{aligned} \quad (6)$$

where $KL(f_i\|g_j)$ is the KL-divergence between the two Gaussian components, f_i and g_j , and $H(f_i)$ is the entropy of Gaussian f_i . The maximum of the lower bound $L_q(F\|G)$ with respect to the variational parameters $q_{(j|i)}$ occurs at the values (see e.g. [18, Ch. 10])

$$\hat{q}_{(j|i)} = \frac{\beta_j e^{-KL(f_i\|g_j)}}{\sum_{j'} \beta_{j'} e^{-KL(f_i\|g_{j'})}}. \quad (7)$$

Setting the above value into $L_{\hat{q}}(F\|G)$ gives the maximum value for this lower bound

$$L_{\hat{q}}(F\|G) = \sum_i \alpha_i \log \sum_j \beta_j e^{-KL(f_i\|g_j) - H(f_i)}. \quad (8)$$

The expression (8) provides a closed-form approximation for the cross-entropy $L(F\|G)$ because the KL-divergence between each pair of multivariate d -dimensional Gaussians has the closed-form expression (4). It can therefore be used to approximate the KL-divergence (3), via $KL(F\|G) = L(F\|F) - L(F\|G)$, by

$$\begin{aligned} \widetilde{KL}(F\|G) &= L_{\hat{q}}(F\|F) - L_{\hat{q}}(F\|G) \\ &= \sum_i \alpha_i \log \frac{\sum_{i'} \alpha_{i'} e^{-KL(f_i\|f_{i'})}}{\sum_j \beta_j e^{-KL(f_i\|g_j)}}. \end{aligned} \quad (9)$$

The varEM algorithm in [6] searches for the reduced model G that maximizes $L_{\hat{q}}(F\|G)$ (8), which presents a lower bound for the cross-entropy in (5). The algorithm performs expectation-maximization iterations over the parameters $\{\beta_j, \mu_j, \Sigma_j\}$ of G as follows. The E-step (expectation step) computes the optimal values $\hat{q}_{(j|i)}$ in (7), based on the current parameters of G , that maximize $L_q(F\|G)$ in (6). Then the M-step (maximization step) maximizes $L_{\hat{q}}(F\|G)$ in (8) with respect to the parameters of G , using the following update formulas:

$$\hat{\beta}_j = \sum_i \alpha_i \hat{q}_{(j|i)}; \quad (10)$$

$$\hat{\mu}_j = \frac{\sum_i \alpha_i \hat{q}_{(j|i)} m_i}{\sum_i \alpha_i \hat{q}_{(j|i)}}; \quad (11)$$

$$\hat{\Sigma}_j = \frac{\sum_i \alpha_i \hat{q}_{(j|i)} [V_i + (m_i - \hat{\mu}_j)(m_i - \hat{\mu}_j)^T]}{\sum_i \alpha_i \hat{q}_{(j|i)}}. \quad (12)$$

The algorithm alternates between the E-step (7) and the M-step (10)–(12) until the convergence of $L_{\hat{q}}(F\|G)$.

III. DISCRIMINATIVE REDUCTION OF GMM CLASSES

Our goal in this section is to find a *target* set $\{G_c\}_{c=1}^N$ of reduced GMMs (2) that best approximates a given *source* set $\{F_c\}_{c=1}^N$ of high order GMMs (1) in the sense of maximum mutual information. We first derive a maximum mutual information (MMI) criterion to this parametric framework. Then, we use the variational technique to approximate the resulting MMI criterion by a closed-form objective function to be called *variational*-MMI (VMMI). Finally, we generalize the latter objective function such that it admits interpolation between the discriminative VMMI function and the non-discriminative function considered in the previous section.

A. Maximum Mutual Information (MMI)

The mutual information between two continuous random vectors X and Y with joint pdf $f(x, y)$ is defined by

$$I(X; Y) = \int_{\mathbf{x}, \mathbf{y}} f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy. \quad (13)$$

We wish to maximize the mutual information between X regarded as a continuous variable that presents the data, and Y presenting the discrete set $C = \{1, \dots, N\}$ of N classes, having a-priori probabilities p_c for each class $c = 1, \dots, N$. Expressing $f(x, y)$ by the conditional distribution $f(x|y)$, it is possible to write the above mutual information expression as follows

$$I(X; Y) = \int f(y) f(x|y) \log \left(\frac{f(x|y)}{\int f(y') f(x|y') dy'} \right) dx dy. \quad (14)$$

Next, to achieve the discretization with respect to Y , we set into the integrands the marginal distribution of y , $f(y) = \sum_{c=1}^N p_c \delta(y - c)$, where $\delta(y)$ is the delta of dirac, and p_c marks the discrete prior class probability of class c . We use the familiar property of this operator when integrated over a vicinity of $y = 0$. This leads to the following expression for the mutual information, that we denote by $\mathcal{I}(X; C)$, to stress that we actually reached a new entity—the *mutual information for a mixture of discrete-continuous random variables*,

$$\mathcal{I}(X; C) = \sum_{c=1}^N p_c \int f(x|c) \log \left(\frac{f(x|c)}{\sum_{k=1}^N p_k f(x|k)} \right) dx. \quad (15)$$

To pull the above general formulation toward our goal we need two assumptions regarding the calculation of the expectation in (15):

1) The expected value that is calculated in (15) should be based on the new model parameters. That is, we force the conditional densities inside the logarithm term to be subjected to models parameters $\theta_c \sim \{\beta_{cj}, \mu_{cj}, \Sigma_{cj}\}$ such that we now use $f(x|c, \theta_c) = G_c(x)$.

2) From data perspective, we assume that data x of class c is distributed according to its high order model $F_c(x)$ in (1). Therefore, the integration is done with respect to a weighting pdf $f(x|c)$ that is set to $F_c(x)$.

Eventually, a new realization for the MMI criterion between the discrete variable C (representing the classes) and the continuous random vector X that is distributed according to the high order models $\{F_c\}$, and is modeled by the reduced order models $\{G_c\}$, is formed by

$$\hat{I}_\theta(X; C) = \sum_{c=1}^N p_c \int F_c(x) \log \left(\frac{G_c(x)}{\sum_{k=1}^N p_k G_k(x)} \right) dx, \quad (16)$$

and more explicitly is given by

$$\sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \int f_{ci}(x) \log \left(\frac{\sum_{j=1}^{R_c} \beta_{cj} g_{cj}(x)}{\sum_{k=1}^N p_k \sum_{j=1}^{R_k} \beta_{kj} g_{kj}(x)} \right) dx. \quad (17)$$

Finally, if we define

$$G_w(x) = \sum_{c=1}^N p_c G_c(x), \quad (18)$$

then, the mutual information realization in (16) between the continuous random vector of ‘data’ and the discrete random variable of ‘classes’, can be presented also in the form

$$\Lambda = \sum_{c=1}^N p_c [L(F_c \| G_c) - L(F_c \| G_w)]. \quad (19)$$

B. Variational Approximation of MMI

The parametric mutual information, Λ (19), that we wish to maximize still does not offer a convenient access to the parameters that need to be optimized. To handle this difficulty, we apply the variational technique (exactly as in [15]) to approximate it by a closed-form expression. By applying the variational approximation for $L(F_c \| G_c)$ and for $L(F_c \| G_w)$ in (19) (using (8)) we receive the following *variational*-MMI (VMMI) objective function

$$\begin{aligned} \mathcal{J} &= \sum_{c=1}^N p_c [L_{\hat{q}}(F_c \| G_c) - L_{\hat{w}}(F_c \| G_w)] \\ &= \sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \log \left(\frac{p_c \sum_{j=1}^{R_c} \beta_{cj} e^{-KL(f_{ci} \| g_{cj})}}{\sum_{k=1}^N p_k \sum_{j=1}^{R_k} \beta_{kj} e^{-KL(f_{ci} \| g_{kj})}} \right). \quad (20) \end{aligned}$$

Note that the referenced $L_{\hat{q}}(F_c \| G_c)$ and $L_{\hat{w}}(F_c \| G_w)$ terms are now obtained by the corresponding optimal variational parameters ((7)): the first set of parameters is defined by

$$\hat{q}_{c(j|i)} = \frac{\beta_{cj} e^{-KL(f_{ci} \| g_{cj})}}{\sum_{j'=1}^{R_c} \beta_{cj'} e^{-KL(f_{ci} \| g_{cj'})}} \quad (21)$$

and the second set of parameters $\hat{w}_{(kj|ci)}$ is defined by

$$\hat{w}_{(kj|ci)} = \frac{p_k \beta_{kj} e^{-KL(f_{ci} \| g_{kj})}}{\sum_{k'=1}^N p_{k'} \sum_{j'=1}^{R_{k'}} \beta_{k'j'} e^{-KL(f_{ci} \| g_{k'j'})}}. \quad (22)$$

Let us recapitulate what we achieved in this section. Assuming that the distribution of data X is given by the high order models $F_c(x)$, but is modeled by the reduced order models $G_c(x)$, we showed that the maximization of the mutual information between X and the discrete random variable representing the classes C amounts to the maximization of the parametric realization in (17). Then we showed that the non closed-form expression (17) can be approximated by the closed-form VMMI objective function \mathcal{J} (20). Thus the VMMI objective paves the way for analytically tractable algorithms to derive low order models that approximately maximize the referenced mutual information. Two effective algorithms to optimize \mathcal{J} will be presented in the next two sections.

C. Relation to Maximum Correct Association

As mentioned, in [12] we have introduced a discriminative approach for parametric mixture model reduction, where it was termed ‘‘maximum correct association’’ (MCA), and further studied it in [13]. The term correct association stems from a possible interpretation of the above optimal variational parameters $\hat{w}_{(kj|ci)}$, as presenting the probability of associating component i of the source model c (i.e. f_{ci}) with component j of the reduced model k (i.e. g_{kj}). Consequently, the probability for the ‘‘correct association’’ of f_{ci} with class c can be given by $p(c|f_{ci}) = \sum_{j=1}^{R_c} \hat{w}_{(cj|ci)}$. The expectation of the logarithm of this probability of correct association is given by

$$\sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \log p(c|f_{ci}) = \sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \log \left(\sum_{j=1}^{R_c} \hat{w}_{(cj|ci)} \right). \quad (23)$$

The expression (23) coincides with \mathcal{J} (20) and it reduces to the MCA objective function in [12] when all a-priori class probabilities are assumed to be equal by $p_c = 1/N$, $\forall c = 1, \dots, N$.

D. A Generalized Criterion

Consider next the modification of the MMI criterion (19) into the form

$$\Lambda_h = \sum_{c=1}^N p_c [L(F_c \| G_c) - hL(F_c \| G_w)]. \quad (24)$$

Namely, assume that an interpolation parameter h is added to the basic criterion such that Λ_h coincides with Λ in (19) for $h = 1$, presenting a ‘‘full’’ MMI criterion, while for $h = 0$ the objective degenerates to the non-discriminative maximum

likelihood (ML) criterion. Such interpolation, named the h -criterion, has been suggested as an improvement to MMI in the context of speech recognition [19], [20], and was shown in [21] to be useful in reducing over-fitting effects. Later on, the authors in [22] used a similar interpolation principle to investigate the convex optimization of an approximated MMI criterion by using a sufficiently small interpolation coefficient.

It can be readily realized that the variational approximation of the generalization (24) can be given by the following corresponding modification of (20)

$$\mathcal{J}_h = \sum_{c=1}^N p_c [L_{\hat{q}}(F_c \| G_c) - h L_{\hat{w}}(F_c \| G_w)]. \quad (25)$$

The objective function \mathcal{J}_h enhances \mathcal{J} with a controllable tradeoff between the VMMI criterion for $h = 1$ and a *variational*-ML (VML) for $h = 0$ (which can be optimized by varEM for each class separately). In the remaining of this paper we consider algorithms for the optimization of the above objective function (20) and their evaluation in reducing GMMs for phoneme and language recognition. We shall also examine the impact of the interpolation parameter h on classification accuracy.

IV. DIRECT OPTIMIZATION OF THE VMMI

There are several gradient-based approaches for maximizing a non-concave objective function like the VMMI. Best known are the steepest gradient and the conjugate gradient, and more optimal in the sense of steepest direction (in the Riemannian parameter space) is the natural gradient [23]. More advanced techniques also involve calculations or predictions of second order derivatives. In the reported experiments we have used the Generalized Probabilistic Descent (GPD) [16] technique, which is based on the probability descent theorem [24], and since its introduction was reported to perform well in several cases of discriminative training of GMMs from raw data. Its general idea is as follows. Let $\nabla_{\theta_t} \mathcal{J}$ denote the gradient vector of the objective function at iteration t . Then, with a properly designed transformation matrix \mathbf{U}_t , the parameters are updated by $\epsilon_t \mathbf{U}_t (\nabla_{\theta_t} \mathcal{J})$, where ϵ_t is the learning rate parameter and the transformation \mathbf{U}_t is a positive definite matrix that transforms the parameter space to be updated in the steepest direction.

The GPD is an unconstrained optimization scheme. Therefore, in order to meet some constraints related to GMMs' parameters certain parameter transformations are used. To ensure that the weights fulfill the constraints $\beta_{cj} > 0$ and $\sum_{j=1}^{R_c} \beta_{cj} = 1$, they are replaced by parameters ω_{cj} which maintain

$$\beta_{cj} = \frac{e^{\omega_{cj}}}{\sum_{j'=1}^{R_c} e^{\omega_{cj'}}}. \quad (26)$$

Similarly, positive definiteness of the covariance matrix, Σ_{cj} , is attained by maximizing it via $\tilde{\Sigma}_{cj}$ which satisfies

$$\Sigma_{cj} = \exp \left(\tilde{\Sigma}_{cj} \right). \quad (27)$$

This way, updates of ω_{cj} and $\tilde{\Sigma}_{cj}$ define corresponding updates for β_{cj} and Σ_{cj} that meet their required constraints. In addition, we also normalize the means by $\Sigma_{cj}^{-1/2}$ and use

$$\tilde{\mu}_{cj} = \Sigma_{cj}^{-\frac{1}{2}} \mu_{cj}. \quad (28)$$

This normalization of the means fixes the parameter space curvature such that \mathbf{U}_t can now be taken as the identity matrix. In fact, it can be shown that this normalization is identical to taking the natural gradient with respect to the means. Carrying out the differentiation of \mathcal{J}_h w.r.t. ω_{cj} , $\tilde{\mu}_{cj}$, and $\tilde{\Sigma}_{cj}$ gives for the transformed weights

$$\frac{\partial \mathcal{J}}{\partial \omega_{cj}} = \left(p_c \sum_{i=1}^{M_c} \alpha_{ci} \hat{q}_{c(j|i)} - h \sum_{k=1}^N p_k \sum_{i=1}^{M_k} \alpha_{ki} \hat{w}_{(cj|ki)} \right) (1 - \beta_{cj}), \quad (29)$$

for the means

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \tilde{\mu}_{cj}} &= p_c \sum_{i=1}^{M_c} \alpha_{ci} \hat{q}_{c(j|i)} \Sigma_{cj}^{-\frac{1}{2}} (m_{ci} - \mu_{cj}) \\ &\quad - h \sum_{k=1}^N p_k \sum_{i=1}^{M_k} \alpha_{ki} \hat{w}_{(cj|ki)} \Sigma_{cj}^{-\frac{1}{2}} (m_{ki} - \mu_{cj}), \end{aligned} \quad (30)$$

and for the transformed covariance matrices,

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \tilde{\Sigma}_{cj}} &= \frac{1}{2} p_c \sum_{i=1}^{M_c} \alpha_{ci} \hat{q}_{c(j|i)} \\ &\quad \left\{ \Sigma_{cj}^{-1} \left[V_{ci} + (m_{ci} - \mu_{cj})(m_{ci} - \mu_{cj})^T \right] - I \right\} \\ &\quad - h \frac{1}{2} \sum_{k=1}^N p_k \sum_{i=1}^{M_k} \alpha_{ki} \hat{w}_{(cj|ki)} \\ &\quad \left\{ \Sigma_{cj}^{-1} \left[V_{ki} + (m_{ki} - \mu_{cj})(m_{ki} - \mu_{cj})^T \right] - I \right\}. \end{aligned} \quad (31)$$

The expressions (29)–(31) provide the key ingredients for applying a gradient-based algorithm to obtain the parameters of the reduced GMMs that optimize \mathcal{J}_h . The optimization of \mathcal{J}_h with this implementation of the GPD algorithm achieved good results in our experiments.

V. OPTIMIZATION VIA CONCAVE ASSOCIATED FUNCTIONS

The efficiency and robustness of a learning procedure for a non-convex problem is highly dependent on the characteristics of the observed random process and on model structure and complexity. The Extended Baum-Welch (EBW) algorithm is probably the most widely used training scheme in speech and language processing for GMMs under the MMI criterion [25]–[27]. The popularity of EBW comes from its simplicity in casting the optimization into EM-like iterations, that has been shown to be robust and efficient for large-scale speech recognition [21], [27]. Extensions and improvements for EBW continued to appear over the years, such as I-smoothing [28], Boosted-MMI (BMMI) [29], and ‘‘Generalized-BW’’ (GBW)

[30], alongside with studies on how to control the EBW's update steps [31], [32].

Recently, the EBW algorithm has been generalized into an optimization technique called *Line Search A-functions* (LSAF) [17]. This technique uses only first order derivatives but provides effective update equations that empirically give good step size guesses. The LSAF approach consists of two basic parts: the first involves the selection of some "associated function", referred to as "A-function", that satisfies certain conditions that simplify the optimization, and the second involves a line search that forces the objective function to increase (or decrease in minimization). This section is devoted to outline the derivation of an LSAF-based optimization algorithm for the generalized VMMI function (25) (for clarity, some of its derivation details and proofs are given in the Appendix). The LSAF-based optimization of the VMMI function achieved the best results in our GMM reduction experiments.

A. Line Search A-Functions (LSAF)

Consider a function $\mathcal{F}(\theta) : \mathbb{R}^n \rightarrow \mathbb{R}$, differentiable for all θ in an open set $\mathcal{U} \subset \mathbb{R}^n$. Following [17], we call a function $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0)$ that is differentiable in θ and θ_0 in \mathcal{U} an associated function, or an **A-function**, for \mathcal{F} if the following properties hold:

- 1) $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0)$ is a concave (or convex) function of $\theta \in \mathcal{U}$ for all $\theta_0 \in \mathcal{U}$;
- 2) $\nabla_{\theta} \mathbf{A}_{\mathcal{F}}(\theta, \theta_0)|_{\theta=\theta_0} = \nabla_{\theta} \mathcal{F}(\theta_0)$. Namely, the first partial derivatives of $\mathbf{A}_{\mathcal{F}}(\cdot, \theta_0)$ and $\mathcal{F}(\theta)$ are identical for all $\theta_0 \in \mathcal{U}$.

Next, we discuss the line search rule for the maximization of the objective function (the discussion can be easily inverted to minimization). When an **A-function** is available for an objective function \mathcal{F} , it can assist the iterations for the maximization of \mathcal{F} as follows. Let θ_0 be some point in \mathcal{U} and $\hat{\theta} \in \mathcal{U}$ be a solution to the equation $\nabla_{\theta} \mathbf{A}_{\mathcal{F}}(\theta, \theta_0)|_{\theta=\hat{\theta}} = 0$. That is,

$$\hat{\theta} = \arg \max_{\theta \in \mathcal{U}} \mathbf{A}_{\mathcal{F}}(\theta, \theta_0).$$

Then a line search from the current point θ_0 can be performed toward $\hat{\theta}$, using a parameter $\alpha \in (0, 1]$ by

$$\theta^+ = \alpha \hat{\theta} + (1 - \alpha) \theta_0. \quad (32)$$

In general, an associated $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0)$ guarantees that for some small enough $\alpha \in (0, 1]$, we receive $\mathcal{F}(\theta^+) > \mathcal{F}(\theta_0)$ (except for a situation where the gradients at θ_0 equal 0), as illustrated in part (a) of Fig. 1. If the concave **A-function** fully underestimates the objective function (as illustrated in part (b) of Fig. 1), the associated function reduces to the commonly called auxiliary function for \mathcal{F} , that satisfies the Jensen's inequality: $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0) - \mathbf{A}_{\mathcal{F}}(\theta_0, \theta_0) \leq \mathcal{F}(\theta) - \mathcal{F}(\theta_0)$. In this case, the growth property in the objective function holds for all $\alpha \in (0, 1]$ (including $\alpha = 1$).

B. A-Function for Concave Log-sum Functions

As an auxiliary step toward the optimization of the VMMI objective function, consider the case of a function \mathcal{F} that is defined

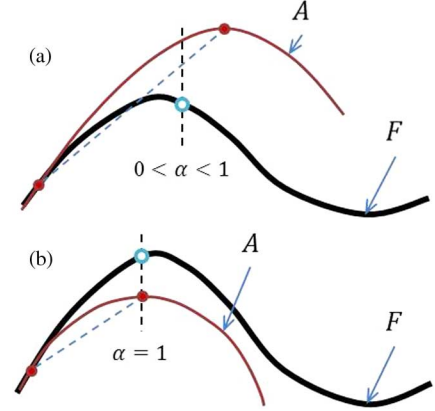


Fig. 1. LSAF optimization with concave associated functions: (a) LSAF with a general form of a concave **A-function**. (b) LSAF in the special case where **A** is an auxiliary function (as in EM).

by a logarithm of the sum of real functions $f_k(\theta)$ differentiable for all θ in an open set $\mathcal{U} \subset \mathbb{R}^n$,

$$\mathcal{F}(\theta) = \log \left(\sum_k f_k(\theta) \right). \quad (33)$$

Then, its gradient is given by

$$\nabla \mathcal{F}(\theta) = \sum_k \left(\frac{1}{\sum_{k'} f_{k'}(\theta)} \right) \nabla f_k(\theta). \quad (34)$$

We now show that, if the logarithm of each component $f_k(\theta)$ is concave, then the following function is an **A-function** for $\mathcal{F}(\theta)$ in (33):

$$\mathbf{A}_{\mathcal{F}}(\theta, \theta_0) = \sum_k \left(\frac{f_k(\theta_0)}{\sum_{k'} f_{k'}(\theta_0)} \right) \log f_k(\theta). \quad (35)$$

First, the concavity condition is satisfied since $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0)$ is a sum of concave functions, $\log(f_k(\theta))$, and therefore it is concave too. Second, the gradient of $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0)$ at θ_0 equals the gradient of $\mathcal{F}(\theta)$ at θ_0 :

$$\begin{aligned} \nabla \mathbf{A}_{\mathcal{F}}(\theta, \theta_0)|_{\theta=\theta_0} &= \sum_k \left(\frac{f_k(\theta_0)}{\sum_{k'} f_{k'}(\theta_0)} \right) \frac{1}{f_k(\theta_0)} \nabla f_k(\theta) \Big|_{\theta=\theta_0} \\ &= \sum_k \left(\frac{1}{\sum_{k'} f_{k'}(\theta_0)} \right) \nabla f_k(\theta) \Big|_{\theta=\theta_0} \\ &= \nabla \mathcal{F}(\theta) \Big|_{\theta=\theta_0}. \end{aligned}$$

In this concave log-sum case, the associated function $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0)$ obeys the Jensen's inequality:

$$\begin{aligned} \mathbf{A}_{\mathcal{F}}(\theta, \theta_0) - \mathbf{A}_{\mathcal{F}}(\theta_0, \theta_0) &= \sum_k \left(\frac{f_k(\theta_0)}{\sum_{k'} f_{k'}(\theta_0)} \right) \log \frac{f_k(\theta)}{f_k(\theta_0)} \\ &\leq \log \sum_k \left(\frac{f_k(\theta_0)}{\sum_{k'} f_{k'}(\theta_0)} \right) \frac{f_k(\theta)}{f_k(\theta_0)} \\ &= \log \left(\frac{\sum_k f_k(\theta)}{\sum_{k'} f_{k'}(\theta_0)} \right) \\ &= \mathcal{F}(\theta) - \mathcal{F}(\theta_0). \end{aligned}$$

In other words, if all $f_k(\theta)$ in (33) are concave, $\mathbf{A}_{\mathcal{F}}(\theta, \theta_0)$ in (35) is actually an auxiliary function for $\mathcal{F}(\theta)$.

C. A Concave \mathbf{A} -Function for Discriminative GMM Reduction

With the above preparations, we proceed to the problem at hand which is obtaining reduced models for a given set of large GMMs such that the VMMI criterion is maximized. It is possible to write the objective function \mathcal{J}_h in (25) as follows

$$\mathcal{J}_h(\theta) = Q(\theta) - hW(\theta), \quad (36)$$

where

$$Q(\theta) = \sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \log \left(p_c \sum_{j=1}^{R_c} \beta_{cj} e^{-KL(f_{ci} \| g_{cj})} \right) \quad (37)$$

and

$$W(\theta) = \sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \log \left(\sum_{k=1}^N p_k \sum_{j=1}^{R_k} \beta_{kj} e^{-KL(f_{ci} \| g_{kj})} \right). \quad (38)$$

Note that both Q and W include a weighted summation of log-sum functions in the form considered in Section V-B. Therefore we can use the property shown there to choose a suitable associated function for each of them. Let θ present the parameter set of the reduced GMM, $\theta \sim \{\beta_{cj}, g_{cj}\}$, where g_{cj} is the Gaussian component $g_{cj} \sim N(\mu_{cj}, \Sigma_{cj})$, for all $c = 1, \dots, N$, and $j = 1, \dots, R_c$. Similarly, let $\theta_0 \sim \{\beta_{cj}^0, g_{cj}^0\}$, where $g_{cj}^0 \sim N(\mu_{cj}^0, \Sigma_{cj}^0)$, be the current point in the parameter space. Hence, according to (35),

$$\mathbf{A}_Q(\theta, \theta_0) = \sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \sum_{j=1}^{R_c} q_{c(j|i)} \log \left(\beta_{cj} e^{-KL(f_{ci} \| g_{cj})} \right) \quad (39)$$

provides an associated function for Q (37), where

$$q_{c(j|i)} = \frac{\beta_{cj}^0 e^{-KL(f_{ci} \| g_{cj}^0)}}{\sum_{j'=1}^{R_c} \beta_{cj'}^0 e^{-KL(f_{ci} \| g_{cj'}^0)}}, \quad (40)$$

and

$$\mathbf{A}_W(\theta, \theta_0) = \sum_{c=1}^N p_c \sum_{i=1}^{M_c} \alpha_{ci} \sum_{k=1}^N p_k \sum_{j=1}^{R_k} w_{(kj|ci)} \log \left(\beta_{kj} e^{-KL(f_{ci} \| g_{kj})} \right), \quad (41)$$

provides an associated function for W (38), where

$$w_{(kj|ci)} = \frac{p_k \beta_{kj}^0 e^{-KL(f_{ci} \| g_{kj}^0)}}{\sum_{k'=1}^N p_{k'} \sum_{j'=1}^{R_{k'}} \beta_{k'j'}^0 e^{-KL(f_{ci} \| g_{k'j'}^0)}}. \quad (42)$$

Notice that $q_{c(j|i)}$ and $w_{(kj|ci)}$ correspond to the variational parameters in (37) and (38), respectively, at the fixed point θ_0 . Still, the expression $\mathbf{A}_Q(\theta, \theta_0) - h\mathbf{A}_W(\theta, \theta_0)$ is not concave. To

overcome this difficulty, we can add a concave smoothing function $\mathbf{A}_S(\theta, \theta_0)$ such that its gradient at $\theta = \theta_0$ equals zero. This way, the local gradient will remain identical to the gradient of our VMMI function $J_h(\theta)$ (25) at θ_0 . This leads to the following casting of the associated function for the VMMI criterion:

$$\mathbf{A}_{\mathcal{J}}(\theta, \theta_0) = \mathbf{A}_Q(\theta, \theta_0) - h\mathbf{A}_W(\theta, \theta_0) + \mathbf{A}_S(\theta, \theta_0). \quad (43)$$

The smoothing function needs to ensure concavity for the optimization, but it should not lose the discrimination quality or slow down the process too much. We select it as a regularizer that penalizes new estimates that are distant from the current parameters. Since the parameters β and $g \sim \{\mu, \Sigma\}$ are separable in the term $\log(\beta e^{-KL(f \| g)})$, the smoothing function can be separated into two independent terms,

$$\mathbf{A}_S(\theta, \theta_0) = \mathbf{A}_{S_\beta}(\theta, \theta_0) + \mathbf{A}_{S_g}(\theta, \theta_0), \quad (44)$$

where \mathbf{A}_{S_g} is selected for the Gaussian components, and \mathbf{A}_{S_β} is designed for the mixture weights. For the Gaussian components we use the following smoothing function,

$$\mathbf{A}_{S_g}(\theta, \theta_0) = \sum_{k=1}^N \sum_{j=1}^{R_k} D_{kj} [-KL(g_{kj}^0 \| g_{kj})], \quad (45)$$

where D_{kj} are positive smoothing constants assigned for each Gaussian component of the reduced model set. Clearly, \mathbf{A}_{S_g} is concave, and has a maximum at $\theta = \theta^0$. For mixture weights we also use the KL divergence, which yields the following smoothing function:

$$\mathbf{A}_{S_\beta}(\theta, \theta_0) = \sum_{k=1}^N B_k \left(\sum_{j=1}^{R_k} \beta_{kj}^0 \log \frac{\beta_{kj}}{\beta_{kj}^0} \right), \quad (46)$$

where each term B_k is a positive constant assigned to a subset of weights that belong to the target GMM G_k . Clearly, this smoothing function is concave and has a zero gradient at $\theta = \theta_0$.

The optimization speed depends on the smoothing constants, where larger values of B_k and D_{kj} cause smaller steps in the process. Setting different smoothing constant, D_{kj} , per each Gaussian increases the flexibility to converge to a better discriminative solution.

To derive the solution we first assume sufficiently large B_k s that force concavity with respect to the weights, and sufficiently large D_{kj} s that force concavity with respect to the parameters of the Gaussian components. For convenience, we define for (39) and (41), the following ‘‘occupation’’ probabilities

$$O_{cij} = p_c \alpha_{ci} q_{c(j|i)}; \quad U_{cijk} = p_c \alpha_{ci} p_k w_{(kj|ci)}. \quad (47)$$

The maximization of the concave associated function $\mathbf{A}_{\mathcal{J}}(\theta, \theta_0)$ (43), can now be obtained by setting its derivatives to zero, as detailed in the Appendix, part A. As shown there, in the derivation of the following formula for the update of the weights, Lagrange multipliers are used to enforce the constraints $\sum_{j=1}^{R_c} \beta_{cj} = 1$. Denote the current parameters

by $\theta_0 = \{\beta_{cj}^0, \mu_{cj}^0, \Sigma_{cj}^0\}$. Then, the update formulas for the weights (see (59) in the Appendix) are obtained by

$$\beta_{cj}^+ = \frac{\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + B_c \beta_{cj}^0}{\sum_{j'=1}^{R_c} \left(\sum_{i=1}^{M_c} O_{cij'} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk'} \right) + B_c}; \quad (48)$$

The updates for the means are (see (62))

$$\mu_{cj}^+ = \frac{\sum_{i=1}^{M_c} O_{cij} m_{ci} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} m_{ki} + D_{cj} \mu_{cj}^0}{\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + D_{cj}}; \quad (49)$$

And the updates for the covariances (see (63)) are given by

$$\Sigma_{cj}^+ = \frac{\sum_{i=1}^{M_c} [V_{ci} + m_{ci} m_{ci}^T] - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} [V_{ki} + m_{ki} m_{ki}^T]}{\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + D_{cj}} + \frac{D_{cj} [\Sigma_{cj}^0 + \mu_{cj}^0 \mu_{cj}^{0T}]}{\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + D_{cj}} - \mu_{cj}^+ \mu_{cj}^{+T}. \quad (50)$$

D. Choosing the Smoothing Parameters

The setting of the smoothing constants in each iteration is crucial for the optimization. We demonstrate next that the smoothing parameters play two roles: i) they ensure the concavity of the objective function, ii) they control the line search.

Consider first the smoothing parameter D_{cj} for the Gaussian components. It can be separated into $D_{cj} = D_{cj}^{(0)} + D_{cj}^{(1)}$, where $D_{cj}^{(0)}$ is the nominal value that ensures a concave \mathbf{A} -function, and $D_{cj}^{(1)}$ controls the line search between the initial value and the estimation obtained with the smoothing of $D_{cj}^{(0)}$. The line search parameter $\alpha \in (0, 1]$ in (32) becomes

$$\alpha_{cj} = \frac{\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + D_{cj}^{(0)}}{\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + D_{cj}^{(0)} + D_{cj}^{(1)}}. \quad (51)$$

Namely, it can be shown that the update of the means (49) can be written with the above parameter by

$$\mu_{cj}^+ = \alpha_{cj} \hat{\mu}_{cj} + (1 - \alpha_{cj}) \mu_{cj}^0, \quad (52)$$

where $\hat{\mu}_{cj}$ is the estimate when the smoothing constant in (49) is set to $D_{cj} = D_{cj}^{(0)}$. Similarly, the update of the covariance matrices (50) can be written as the following line search rule

$$\Sigma_{cj}^+ + \mu_{cj}^+ \mu_{cj}^{+T} = \alpha_{cj} \left(\hat{\Sigma}_{cj} + \hat{\mu}_{cj} \hat{\mu}_{cj}^T \right) + (1 - \alpha_{cj}) \left(\Sigma_{cj}^0 + \mu_{cj}^0 \mu_{cj}^{0T} \right), \quad (53)$$

where $\hat{\Sigma}_{cj}$ is the covariance estimate with the smoothing value of $D_{cj}^{(0)}$. The Gaussian-specific constants are set empirically as follows. They are chosen to be the maximum of i) two

times of the value necessary to ensure positive variances, i.e., $2D_{cj}^{min}$, and ii) a global constant E (in our experiments we used $E = 1$) multiplied by the ‘‘denominator occupancy’’ (created by the denominator of the VMMI criterion (20)), which is $h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk}$. This empirical setting is adopted from [27] where it was shown to produce efficient training of GMMs for speech recognition, after we found it to be also effective for the parametric reduction of GMMs. Hence, the setting for D_{cj} can be written as

$$D_{cj} = \max \left\{ 2D_{cj}^{min}, E h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} \right\}. \quad (54)$$

In part B of the Appendix we prove that the above controlling scheme satisfies the concavity of $\mathbf{A}_{\mathcal{J}}$ (43) with respect to the Gaussian parameters in the LSAF technique.

A very similar line-search principle is applied for setting the smoothing parameters B_c to control the weights updates. We skip the detailed discussion on this issue since it does not bring additional insight to the problem. In practice, the setting for B_c is chosen as follows

$$B_c = E_{\beta} h \max_j \left\{ \frac{1}{\beta_{cj}} \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} \right\}, \quad (55)$$

where E_{β} is a globally fixed constant (we used $E_{\beta} = 2$). In part B of the Appendix we show that this update rule satisfies the concavity condition for the associated function $\mathbf{A}_{\mathcal{J}}$ (43) with respect to the weights. It should be noted that in our experiments, the update of mixture weights had relatively little influence on the performance.

It is important to stress that satisfying the necessary conditions for concavity is not enough to guarantee a robust optimization. For robust convergence, it is also important to stay away from covariance estimates that approach zero and from mixture weights that approach a single dominant weight. Therefore, the smoothing control scheme requires further empirical adjustment by experiments. In the context of our experiments, the smoothing rules suggested in (54) and (55) have worked quite effectively.

VI. EXPERIMENTS

To evaluate the quality of the VMMI objective function and its optimization by the GPD and the LSAF algorithms, experimental studies were conducted in two speech processing tasks: phone recognition and language authentication. In order to save computations, and in accordance with customary practice in most speech processing applications, models with diagonal covariance were used (even though the presented formulation admits full-covariance matrices). We also used equal probabilities $p_c = 1/N$, for all classes $c = 1, \dots, N$. To compare our discriminative approach with non-discriminative GMM reduction methods, we bring as reference the varEM algorithm in [6] which is a leading method among the recently proposed non-discriminative reduction methods (although the differences between the variety of proposed non-discriminative techniques have not been found to be very significant). The varEM has been reported and found also in our experiments to attain reduced models with classification performance close to those of low

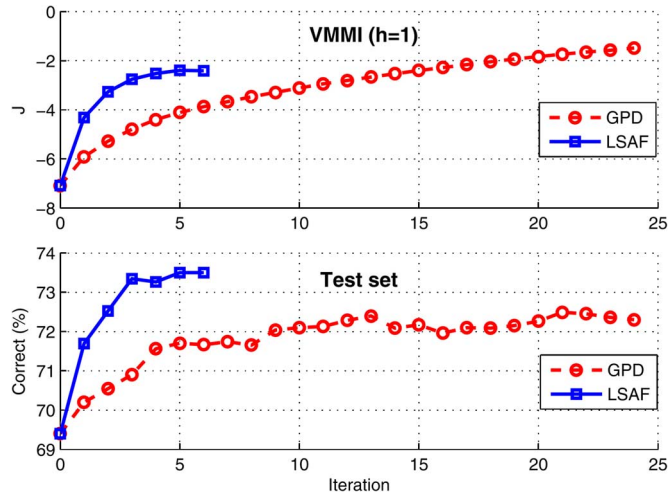


Fig. 2. VMMI iterations of learning 8-order GMMs from original 128-order GMMs. The higher plot presents the values of $\mathcal{J}(20)$ as function of the number of GPD and LSAF iterations. The lower plot presents the corresponding phone recognition accuracy.

order models trained directly from raw samples using standard EM. We also used reduced models obtained by the varEM algorithm as the initial parameters for the VMMI maximization with the GPD and LSAF algorithms.

A. Acoustic Phone Recognition

Phone recognition tests were performed on the TIMIT corpus [33] which includes accurate manual phonetic segmentation for evaluating acoustic modeling performance. The original data partition was used, 3696 training sentences and 1344 test sentences. As features, we used mel-cepstra (cepstra + Δ + $\Delta\Delta$) of dimension 38, with mean subtraction, followed by a PCA transformation. A model-set of 39 mono-phones was used, each model built of a 3-state HMM. Model size reduction of the middle-state GMMs was tested (the first and third states of each phone remained fixed to a 12-component GMM). Recognition was performed using a bi-gram phonetic language model learnt from the training set.

First we examined the evolution of the VMMI criterion (the resulting value of \mathcal{J} in (20)) and the accuracy of actual phone recognition along consecutive iterations of the optimization algorithms (briefly denoted by GPD and LSAF). Fig. 2 presents the results when using interpolation factor $h = 1$ (which relates to full VMMI). Iteration zero presents the performance of the non-discriminative varEM algorithm (which is identical to running the LSAF algorithm with $h = 0$ together with setting the smoothing parameters to zero). It will also serve as a baseline. For the GPD algorithm, it is seen that while the VMMI measure keeps increasing, the actual recognition performance becomes relatively steady after about 12 iterations, experiencing an over-fitting effect. In comparison, the LSAF algorithm is seen to converge to a good value of the VMMI function with fewer iterations while reaching a better steady state in phone recognition accuracy.

Fig. 3 brings the classification performance of reduced models of varying orders derived from a fixed 128-order GMM

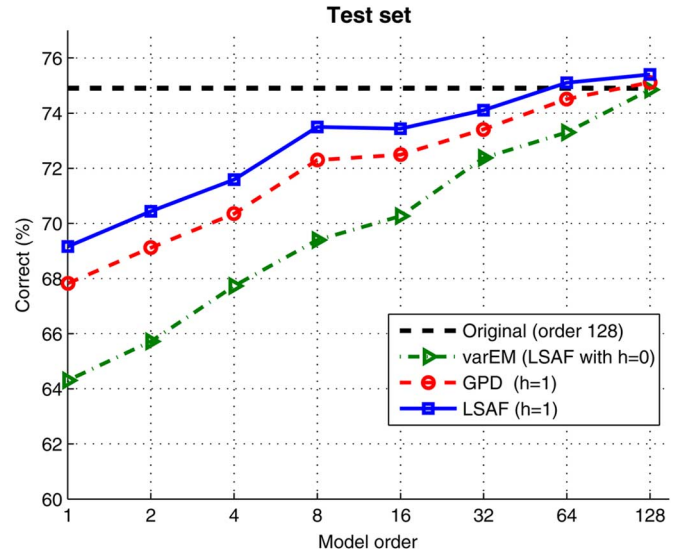


Fig. 3. Phone recognition accuracy of reduced models of different orders obtained from source GMMs of order 128.

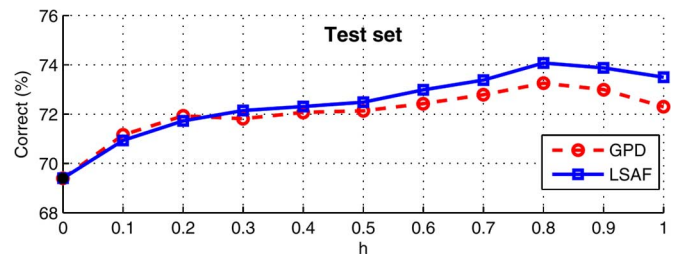


Fig. 4. Dependency of phone recognition accuracy (for reduced GMMs of order 8) on the interpolation factor h used in the generalized VMMI criterion.

set. The utmost horizontal line presents the performance of 128-order GMMs trained by the standard EM algorithm on the original samples. The figure shows that the accuracy of the varEM degrades sharply as the order of the reduced models decreases. The GPD and the LSAF algorithms degrade more gracefully for reduced orders. It is evident that the proposed discriminative algorithms exhibit an ability to boost the performance of the reduced models during the inevitable loss imposed by the size reduction. It also appears that the proposed LSAF optimization constantly provides better results than the GPD optimization.

To examine the impact of the interpolation factor h , used in $\mathcal{J}_h(25)$ to control the tradeoff between full-VMMI and non-discriminative VML (as presented in Section III-D), we examined the GPD and the LSAF algorithms in reducing the number of mixture components from 128 to 8 with variable h values. The results, presented in Fig. 4, show that a slight back-off from the full VMMI criterion to values of h in the range 0.8–0.9, is beneficial.

Even though phone recognition and language recognition are different tasks, the value of $h = 0.8$ shown above for the first task, was found to be effective also for the second task. In general, a value less than 1 seems to improve performance by relaxation of over-fitting effects. However, a fine-tuning of h over the development set for other application may be rewarding.

B. Language Authentication

Language authentication tests were performed on the 30 sec segments of the NIST language recognition evaluation (LRE) 2003 [34]. A subset of the CallFriend corpus [35], was used to train 12 language models, using 12 hours of audio per language (around 3 million vectors per model). Speech frames were mapped to a 56-dimensional feature vector composed from shifted delta cepstra (SDC) 7-1-3-7 coefficients and additional 7 MFCC coefficients (including C_0), with RASTA filtering (following [36]). Each language was initially modeled by a large GMM of order 4096 with diagonal covariance matrices, trained by a conventional EM algorithm on feature vectors pooled from all recordings of the corresponding language. The original language models were reduced to target models of order 256. A verification score was computed using Log-Likelihood Ratio Test (LLRT), for each test utterance X_i , per language model G_c , as follows:

$$S_c(X_i) = \log P(X_i|G_c) - \max_{k \neq c} \{\log P(X_i|G_k)\}. \quad (56)$$

Performance was evaluated by the standard Detection Error Trade-off (DET) curve. For comparison, benchmark reduced model sets were generated in two ways. One by ML training of 256-component models directly from the original feature vectors using EM, and a second set by reducing the original 4096-component GMMs to 256-component models using varEM. The latter models also served as the initial parameter set for both GPD and LSAF algorithms. We present results using the criterion \mathcal{J}_h with the setting $h = 0.8$ (based on the improvement observed for this value of the interpolation factor compared to full VMMI). In all our experiments, the LSAF algorithm typically required between 5 to 7 iterations while for GPD we used 30 to 50 iterations.

Fig. 5 brings the results of language authentication tests with reduced order models of 256 Gaussian components. As seen, the performance of models trained by varEM is similar to the performance of models trained directly from the data. Both discriminative VMMI algorithms achieved a significant improvement over the non-discriminative ML procedures (both traditional EM and parametric varEM), with an advantage to the LSAF algorithm over the GPD algorithm.

The effect of using source models of different orders was also examined. Fig. 6 brings the results for reduced models of order 256 obtained by the LSAF algorithm (optimizing VMMI with $h = 0.8$) from several higher order source model sets (that are well trained by conventional EM directly from the raw data). The results demonstrate that the performance of the reduced set improves as the order of the source model set, from which it is learned, increases. Clearly, the performance of lower order models can be greatly improved by learning it from a higher order source that provides a more accurate representation of the data. A less expected result is noticed when the LSAF process introduces a slight improvement even when it is applied without the intention of order reduction (i.e. the case of deriving target models of order 256 from source models of order 256).

VII. CONCLUDING REMARKS

We have introduced a new approach for the reduction of a set of high order Gaussian mixture models (GMMs) by a system

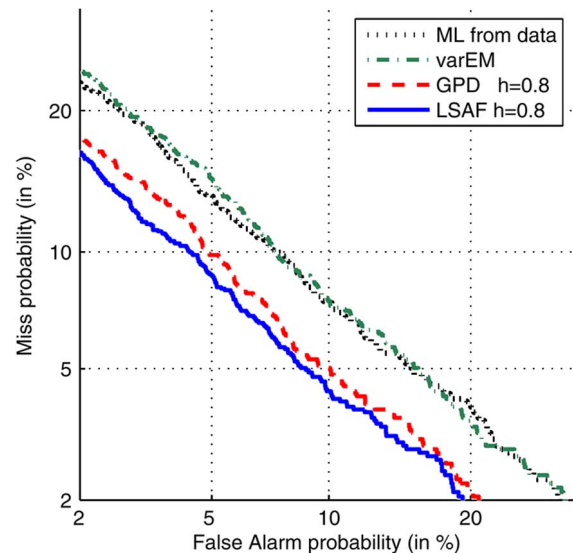


Fig. 5. Language authentication performance of GMMs of order 256 (trained from data versus reduced from order 4096 GMMs).

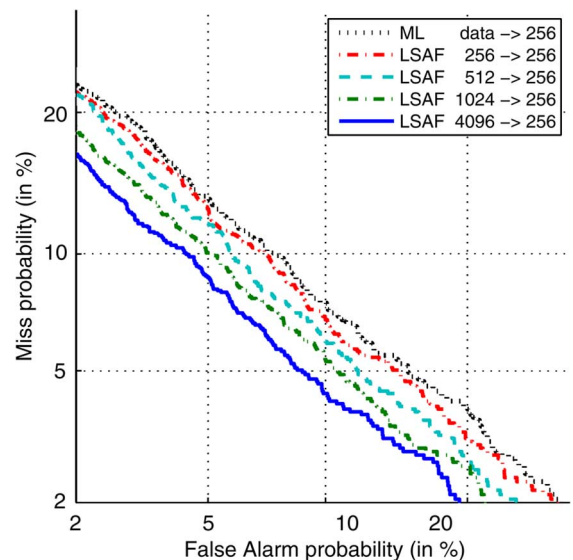


Fig. 6. Language authentication performance GMMs of order 256 optimized by VMMI ($h = 0.8$, using LSAF) from different higher order source models (and directly by EM from raw data).

of GMMs with a lower number of components, based on optimizing a new criterion that characterizes the classification capacity of the reduced models. The proposed method involves a parametric realization of the maximum mutual information (MMI) criterion and its variational approximation by a closed-form objective function called variational-MMI (VMMI). As a result, the maximization of the VMMI objective function can be carried out by analytically tractable algorithms that require only the parameters of the high order models.

Experiments, held with two different speech classification tasks (phoneme classification and language authentication), showed that reduced models obtained by the optimization of the VMMI criterion, significantly outperform previous methods that optimize a non-discriminative criterion as well as a system that derives the lower sized models directly from the original samples by ML training. The discrimination ability of the

reduced system depends on the availability of well trained high order models. This was demonstrated in our experiments by showing that the classification accuracy of a reduced system of a certain order increases with the order of the source system from which it was deduced. We have also proposed a generalized VMMI criterion, one that uses an interpolation parameter h to control a tradeoff between full VMMI ($h = 1$) and non-discriminative *variational*-ML ($h = 0$), and showed that it can enhance the performance when used with a value slightly below full VMMI.

Still, the VMMI objective function poses a non-concave maximization problem, and its optimization involves the challenge of avoiding poor local maxima traps. Two optimization algorithms were studied, the gradient-based GPD algorithm, and the LSAF algorithm that uses a concave associated function. Our results demonstrate that the VMMI optimization, carried out by the LSAF algorithm, converges after a few cycles of iterations and yields better results than the GPD algorithm. The more efficient LSAF algorithm is also more attractive in offering a simple EM-like optimization procedure.

Additional research may focus on finding improved techniques to optimize the VMMI criterion. Within gradient algorithms, it would be interesting to examine a natural gradient scheme, which performs the parameter updates in an optimal direction that depends on the nature of the manifold [23]. It is noted that the update of the mixture's means in the GPD scheme (28) acts already as natural gradient, however it could be interesting to have it compared with a full gradient scheme that updates all the parameters along their natural gradients (subjected to appropriate constraints on the weights and on the covariance matrices). The better performing LSAF algorithm uses concave associated function, but since it is not a true auxiliary function it still requires a smart line search guess. While, it is not expected that a highly complex objective function like VMMI could be treated by true convex auxiliary functions, future research may investigate other ways to relax the VMMI criterion in order to admit the use of true auxiliary functions in a convex optimization structure.

Although the VMMI criterion was evaluated in two speech processing classification tasks, it is expected to be effective also for deriving reduced Gaussian mixture models in a wider range of pattern classification tasks. More generally, the idea of discriminative model reduction for classification purposes, that underlies the VMMI criterion (and the MCA criterion in [12], [13]), is not restricted to models with Gaussian components. It may also be adapted to other distributions (e.g. to other exponential family distributions). In principle, any mixture of probability distributions for which the KL divergence between pairs of mixture components is available in closed-form and is twice-differentiable can lead to analytically tractable VMMI optimization algorithms to produce reduced mixture models with enhanced classification accuracy.

APPENDIX

This Appendix brings some derivation details and proofs for VMMI optimization by the LSAF algorithm in Section V.

A. Derivation of the Update Equations (48)–(50)

For the following proofs we shall need in the partial derivatives of $KL(f||g)$ (4) with respect to the parameters of $g(x)$. They are given by

$$\frac{\partial KL(f||g)}{\partial \mu} = -\Sigma^{-1}(m - \mu) \quad (57)$$

and

$$\frac{\partial KL(f||g)}{\partial \Sigma} = \frac{1}{2}\Sigma^{-1} \left\{ I - [V + (m - \mu)(m - \mu)^T] \Sigma^{-1} \right\}. \quad (58)$$

The maximization of the concave associated function $\mathbf{A}_{\mathcal{J}}(\theta, \theta_0)$ (43) with respect to the optimized parameters is achieved in principle by setting its corresponding derivatives to zero. Specifically, we derive the update formula for the weights (48), by first adding a set of Lagrange multipliers L_c to enforce the constraints $\sum_{j=1}^{R_c} \beta_{cj} = 1$. Thus, we differentiate $\mathbf{A}_{\mathcal{J}}(\theta, \theta_0) + L_c(\sum_{j=1}^{R_c} \beta_{cj} - 1)$, with respect to β_{cj} and solve the equation

$$\begin{aligned} 0 &= \frac{\partial \mathbf{A}_{\mathcal{J}}(\theta, \theta_0)}{\partial \beta_{cj}} - \frac{\partial}{\partial \beta_{cj}} \left[L_c \left(\sum_{j=1}^{R_c} \beta_{cj} - 1 \right) \right] \\ &= \frac{1}{\beta_{cj}} \left(\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + B_c \beta_{cj}^0 \right) - L_c. \end{aligned} \quad (59)$$

to obtain

$$\beta_{cj} = \frac{1}{L_c} \left(\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + B_c \beta_{cj}^0 \right). \quad (60)$$

Summing β_{cj} over all j and using the constraint $\sum_{j=1}^{R_c} \beta_{cj} = 1$ gives

$$\begin{aligned} L_c &= \sum_{j=1}^{R_c} \left[\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + B_c \beta_{cj}^0 \right] \\ &= \sum_{j=1}^{R_c} \left[\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} \right] + B_c \sum_{j=1}^{R_c} \beta_{cj}^0. \end{aligned} \quad (61)$$

Using $\sum_{j=1}^{R_c} \beta_{cj}^0 = 1$ and inserting (61) into (60) verifies the update formula (48). Next, differentiating $\mathbf{A}_{\mathcal{J}}$ with respect to μ_{cj} and using (57) gives

$$\begin{aligned} \frac{\partial \mathbf{A}_{\mathcal{J}}(\theta, \theta_0)}{\partial \mu_{cj}} &= \Sigma_{cj}^{-1} \left[\sum_{i=1}^{M_c} O_{cij}(m_{ci} - \mu_{cj}) \right. \\ &\quad \left. - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk}(m_{ki} - \mu_{cj}) + D_{cj}(\mu_{cj}^0 - \mu_{cj}) \right] \\ &= \Sigma_{cj}^{-1} \left(\sum_{i=1}^{M_c} O_{cij} m_{ci} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} m_{ki} + D_{cj} \mu_{cj}^0 \right) \\ &\quad - \Sigma_{cj}^{-1} \mu_{cj} \left(\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cijk} + D_{cj} \right). \end{aligned} \quad (62)$$

The update formula for the means (49) follows from equating the above expression to zero. Finally, setting to zero the derivative of $\mathbf{A}_{\mathcal{J}}$ with respect to the covariance matrices and using (58)

$$\begin{aligned} \frac{\partial \mathbf{A}_{\mathcal{J}}(\theta, \theta_o)}{\partial \Sigma_{c_j}} &= \frac{1}{2} \Sigma_{c_j}^{-1} \left\{ \sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} + D_{c_j} \right. \\ &\quad - \sum_{i=1}^{M_c} O_{cij} [V_{ci} + (m_{ci} - \mu_{c_j}^+)(m_{ci} - \mu_{c_j}^+)^T] \Sigma_{c_j}^{-1} \quad (63) \\ &\quad + h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} [V_{ki} + (m_{ki} - \mu_{c_j}^+)(m_{ki} - \mu_{c_j}^+)^T] \Sigma_{c_j}^{-1} \\ &\quad \left. - D_{c_j} [\Sigma_{c_j}^0 + (\mu_{c_j}^0 - \mu_{c_j}^+)(\mu_{c_j}^0 - \mu_{c_j}^+)^T] \Sigma_{c_j}^{-1} \right\} = 0. \end{aligned}$$

gives, after a little manipulation, the expression (50).

B. Concavity Conditions of the Associated Function

We show that the controls posed on the smoothing parameters of the LSAF algorithm in (54) and (55) are sufficient to ensure the concavity of the associated function $\mathbf{A}_{\mathcal{J}}(\theta, \theta_o)$ with respect to each of the estimated parameters, the weights, means, and covariances. Concavity with respect to each set of parameters provides only necessary conditions for a common maximum. For strict concavity one needs in addition the negativity of the Hessian matrix whose entries are all second order partial derivatives of the function. However, using the indicated conditions seem to provide (according to our experiments with the LSAF algorithm) a common concavity condition in practice.

Claim 1: For B_c chosen as in (55), the value of $\beta_{c_j}^+$ in (48) (that solves $\partial \mathbf{A}_{\mathcal{J}}(\theta, \theta_o) / \partial \beta_{c_j} = 0$) is a maximum of $\mathbf{A}_{\mathcal{J}}$ with respect to β_{c_j} .

Proof: Obtain the second derivative of $\mathbf{A}_{\mathcal{J}}(\theta, \theta_o)$ with respect to β_{c_j} by differentiating again (59)

$$\frac{\partial^2 \mathbf{A}_{\mathcal{J}}(\theta, \theta_o)}{\partial \beta_{c_j}^2} = -\frac{1}{\beta_{c_j}^2} \left(\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} + B_c \beta_{c_j}^0 \right).$$

The term in the brackets is positive if

$$B_c > \frac{1}{\beta_{c_j}^0} \left(-\sum_{i=1}^{M_c} O_{cij} + h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} \right) \quad \forall j = 1, \dots, R_c.$$

Hence, the second derivative is negative for (55) with $E_{\beta} \geq 1$. ■

Taking into consideration that $\mathbf{A}_{\mathcal{J}}(\theta, \theta_o)$ is strictly separable in β_{c_j} expands the implication of claim 1 to obtaining a concave curvature of $\mathbf{A}_{\mathcal{J}}$ with respect to the weights independently of the other parameters.

Claim 2: For D_{c_j} chosen as in (54), the value of $\mu_{c_j}^+$ (49) (that solves $\partial \mathbf{A}_{\mathcal{J}}(\theta, \theta_o) / \partial \mu_{c_j} = 0$) is a maximum of $\mathbf{A}_{\mathcal{J}}$ with respect to μ_{c_j} .

Proof: Differentiating again (62) with respect to μ_{c_j} gives

$$\frac{\partial^2 \mathbf{A}_{\mathcal{J}}(\theta, \theta_o)}{\partial \mu_{c_j}^2} = -\Sigma_{c_j}^{-1} \left(\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} + D_{c_j} \right) < 0.$$

Since the covariance inverse $\Sigma_{c_j}^{-1}$ is positive-definite, the expression is negative definite if the term in the brackets is positive, i.e.

$$D_{c_j} \geq -\sum_{i=1}^{M_c} O_{cij} + h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj}. \quad (64)$$

Hence, the condition (54) satisfies the negativity of the second derivative if $E \geq 1$. ■

Claim 3: For D_{c_j} chosen as in (54), the value $\Sigma_{c_j}^+$ in (50) (that solves $\partial \mathbf{A}_{\mathcal{J}}(\theta, \theta_o) / \partial \Sigma_{c_j} = 0$) is a maximum of $\mathbf{A}_{\mathcal{J}}$ with respect to Σ_{c_j} .

Proof: The second derivative of $\mathbf{A}_{\mathcal{J}}(\theta, \theta_o)$ with respect to the covariance matrix, Σ_{c_j} , is obtained by differentiating again (63),

$$\begin{aligned} \frac{\partial^2 \mathbf{A}_{\mathcal{J}}(\theta, \theta_o)}{\partial \Sigma_{c_j}^2} &= -\frac{1}{2} \Sigma_{c_j}^{-2} \left(\sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} + D_{c_j} \right) \\ &\quad - \left\{ \sum_{i=1}^{M_c} O_{cij} [V_{ci} + (m_{ci} - \mu_{c_j}^+)(m_{ci} - \mu_{c_j}^+)^T] \right. \\ &\quad \left. - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} [V_{ki} + (m_{ki} - \mu_{c_j}^+)(m_{ki} - \mu_{c_j}^+)^T] \right. \\ &\quad \left. + D_{c_j} [\Sigma_{c_j}^0 + (\mu_{c_j}^0 - \mu_{c_j}^+)(\mu_{c_j}^0 - \mu_{c_j}^+)^T] \right\} \Sigma_{c_j}^{-3}. \quad (65) \end{aligned}$$

Let us write the above equation as follows

$$\frac{\partial^2 \mathbf{A}_{\mathcal{J}}(\theta, \theta_o)}{\partial \Sigma_{c_j}^2} = -\frac{1}{2} \Sigma_{c_j}^{-2} C - [L_O - hL_U + D_{c_j}L_S] \Sigma_{c_j}^{-3},$$

where

$$C = \sum_{i=1}^{M_c} O_{cij} - h \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} + D_{c_j}, \quad (66)$$

and L_O , L_U , and L_S are positive definite matrices given by

$$\begin{aligned} L_O &= \sum_{i=1}^{M_c} O_{cij} [V_{ci} + (m_{ci} - \mu_{c_j})(m_{ci} - \mu_{c_j})^T] \\ L_U &= \sum_{k=1}^N \sum_{i=1}^{M_k} U_{cikj} [V_{ki} + (m_{ki} - \mu_{c_j})(m_{ki} - \mu_{c_j})^T], \\ L_S &= D_{c_j} [\Sigma_{c_j}^0 + (\mu_{c_j}^0 - \mu_{c_j}^+)(\mu_{c_j}^0 - \mu_{c_j}^+)^T]. \end{aligned}$$

If the condition in (64) is satisfied, we get according to (66), $C > 0$. Recalling that the inverse of a positive definite matrix remains positive definite we get $\Sigma_{c_j}^{-2} C > 0$, and now we only require

$$L_O - hL_U + D_{c_j}L_S > 0. \quad (67)$$

Since L_S is positive definite, we can choose a sufficiently large D_{c_j} to guarantee (67). Let $D_{c_j}^{min}$ be the minimal value of D_{c_j}

that ensures a positive definiteness in (67), then the condition $D_{c_j} \geq 2 \cdot D_{c_j}^{min}$ (posed in (54)) suffices. ■

REFERENCES

- [1] N. Vasconcelos and A. Lippman, "Learning mixture hierarchies," presented at the Neural Inf. Process. Syst. (NIPS), Denver, CO, USA, 1998.
- [2] J. Goldberger and S. Roweis, "Hierarchical clustering of a mixture model," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2004, pp. 505–512.
- [3] N. Petrovic, A. Ivanovic, N. Jovic, S. Basu, and T. Huang, "Recursive estimation of generative models of video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, 2006, pp. 79–86.
- [4] J. Goldberger, H. K. Greenspan, and J. Dreyfuss, "Simplifying mixture models using the unscented transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 1496–1502, 2008.
- [5] P. Bruneau, M. Gelgon, and F. Picarougne, "Parameter-based reduction of Gaussian mixture models with a variational-Bayes approach," presented at the 19th Int. Conf. Pattern Recognit. (ICPR), Tampa, FL, USA, 2008.
- [6] P. L. Dognin, J. R. Hershey, V. Goel, and P. A. Olsen, "Refactoring acoustic models using variational Expectation-Maximization," presented at the 10th Ann. Conf. ISCA (INTERSPEECH), Brighton, U.K., 2009.
- [7] F. Nielsen, V. Garcia, and R. Nock, "Simplifying Gaussian mixture models via entropic quantization," presented at the 17th Eur. Conf. Signal Process. (EUSIPCO), Glasgow, Scotland, U.K., 2009.
- [8] V. Garcia, F. Nielsen, and R. Nock, "Simplification and hierarchical representations of mixtures of exponential families," *Signal Process.*, vol. 90, pp. 3197–3212, 2010.
- [9] K. Zhang and J. T. Kwok, "Simplifying mixture models through function approximation," *Trans. Neural Netw.*, vol. 21, no. 4, pp. 644–658, 2010.
- [10] B. Han, D. Comaniciu, Y. Zhu, and L. Davis, "Incremental density approximation and kernel-based Bayesian filtering for object tracking," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2004, pp. 638–644.
- [11] B. Kurkoski and J. Dauwels, "Message-passing decoding of lattices using Gaussian mixtures," in *Proc. 30th Symp. Inf. Theory Appl. (SITA)*, 2007, pp. 877–882.
- [12] Y. Bar-Yosef and Y. Bistriz, "Discriminative simplification of mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 2240–2243.
- [13] Y. Bar-Yosef and Y. Bistriz, "Discriminative algorithm for compacting mixture models with application to language recognition," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, 2012, pp. 2203–2207.
- [14] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1986, pp. 49–52.
- [15] J. R. Hershey and P. A. Olsen, "Approximating the Kullback-Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, pp. 317–320.
- [16] B. H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.
- [17] D. Kanevsky, D. Nahamoo, T. N. Sainath, B. Ramabhadran, and P. A. Olsen, "A-Functions: A generalization of extended Baum-Welch transformations to convex optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 5164–5167.
- [18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [19] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo, and M. A. Picheny, "Decoder selection based on cross-entropies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1988, pp. 20–23.
- [20] J. Zheng, J. Butzberger, H. Franco, and A. Stolck, "Improved maximum mutual information estimation training of continuous density HMMs," in *Proc. Eurospeech*, 2001, pp. 679–682.
- [21] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Comput. Speech Lang.*, pp. 25–47, 2002.

- [22] A. Ben-Yishai and D. Burshtein, "A discriminative training algorithm for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 204–217, 2004.
- [23] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.
- [24] S. Amari, "Theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, vol. EC-16, no. 3, pp. 299–307, 1967.
- [25] S. J. Wright, D. Kanevsky, L. Deng, H. Xiaodong, G. Heigold, and L. Haizhou, "Optimization algorithms and applications for speech and language processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2231–2243, Nov. 2013.
- [26] Y. Normandin and S. D. Morgera, "An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1991, pp. 537–540.
- [27] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Eng. Dept., Cambridge Univ., Cambridge, U.K., 2003.
- [28] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, pp. 105–108.
- [29] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "BoostedMMI for model and feature-space discriminative training," presented at the IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Las Vegas, NV, USA, 2008.
- [30] R. Hsiao and T. Schultz, "Generalized Baum-Welch algorithm and its application to new extended Baum-Welch algorithm," in *Proc. Interspeech*, 2011.
- [31] S. Axelrod, V. Goel, R. A. Gopinath, P. A. Olsen, and K. Visweswariah, "Discriminative estimation of subspace constrained Gaussian mixture models for speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 172–189, Jan. 2007.
- [32] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition a unifying review for optimization-oriented speech recognition," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, Sep. 2008.
- [33] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "TIMIT Acoustic-phonetic continuous speech corpus," presented at the Linguistic Data Consortium (LDC), Philadelphia, PA, USA, 1993.
- [34] A. Martin and M. Przybocki, "NIST 2003 language recognition evaluation," in *Proc. Eurospeech*, 2003, pp. 1341–1344.
- [35] "CallFriend corpus, telephone speech of 15 different languages or dialects," [Online]. Available: www ldc.upenn.edu/Catalog
- [36] P. Matejka, "Phonotactic and acoustic language recognition," Ph.D. dissertation, Brno Univ. of Technol., Brno, Czech Rep., 2008.



Yossi Bar-Yosef received his B.Sc. and M.Sc. degrees in electrical engineering in 1993 and 2003, respectively, both from Tel-Aviv University. He is currently pursuing the Ph.D. degree in electrical engineering at Tel-Aviv University. His research interests include information theory, machine learning, and signal processing.



Yuval Bistriz (F'03) received the B.Sc. degree in physics and the M.Sc. and Ph.D. degrees in electrical engineering, in 1973, 1978, and 1983, respectively, all from Tel Aviv University. He is currently a professor there in the School of Electrical Engineering.

From 1979 to 1984, he held various assistant and teaching position in the department of Electrical Engineering, Tel Aviv University, and in 1987 he joined this department with tenure. From 1984 to 1986, he was a research scholar in the Information System Laboratory, Stanford University, Stanford, CA. From 1986 to 1987, he was with AT&T Bell Laboratories, Murray Hill NJ, and from 1994 to 1996 with DSP Group, Santa Clara CA, doing research in speech processing. His research interests are in signal processing and system theory.