DISCRIMINATIVE ALGORITHM FOR COMPACTING MIXTURE MODELS WITH APPLICATION TO LANGUAGE RECOGNITION

Yossi Bar-Yosef [1,2] and *Yuval Bistritz* [1]

[1]School of Electrical Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel
[2]Nice Systems, Ra'anana, Israel

yossibaryosef@gmail.com, bistritz@eng.tau.ac.il

ABSTRACT

In this paper we explore a discriminative algorithm for compacting large order mixture models. Several studies investigated efficient algorithms for finding a reduced-order model that best approximates a referenced model using only the original mixture parameters. Recently, a discriminative approach named *maximum correct association* (MCA) was introduced to efficiently construct a set of compact models for improved classification. In this paper we suggest a two stage procedure that applies the MCA algorithm after initially obtaining a compact model through the variational-EM method (which is a non-discriminative algorithm). The proposed method is validated in a language recognition task where large order mixture models are compacted into low order models. Experiments showed that the MCA-refined models performed consistently better than reduced models derived with the non-discriminative methods including boosting performance over the standard maximum-likelihood trained from the original data.

Index Terms: Gaussian mixture models, hierarchical clustering, discriminative learning, language recognition

1. INTRODUCTION

The Gaussian Mixture Model (GMM) is a very powerful framework, frequently used in statistical learning. Many applications that use Gaussian mixture models require large order models to achieve adequate data representation. On the other hand, using large-order models is expensive in terms of computational complexity and therefor there is a constant need to decrease model complexity. Reducing the computational requirements can be helpful not only in the testing phase, but also in the model training process. Given an adequate model that is too large, a reasonable approach to decrease its complexity is to efficiently reduce the number of components of the mixture, using only the model's parameters, without returning to the original samples and avoiding Monte-Carlo re-sampling of data-points. Recent studies proposed hierarchical clustering algorithms for generating a simplified model while trying to preserve the similarity to the original mixture model [1]-[6].

The most natural and commonly used measure of distance between two probability densities is the Kullback-Leibler (KL) divergence, also known as the relative entropy. Since, there is no closed-form expression for the KL measure between two mixture models, most approaches use an analytical approximation. Goldberger and Rowies [1] introduced a component grouping algorithm that minimizes an approximation of the KL-divergence which is based on Gaussian matching. Later, a soft version of the Gaussian-matching clustering algorithm appeared in [2], based on maximizing the cross-entropy approximation between the two models. Dognin et al. [3], followed the variational KL approximation introduced by Hershey et al. [7], to suggest the *variational expectation-minimization* (varEM) algorithm for Gaussian component clustering in the framework of automatic speech recognition. Other studies used several different approximations for the KL divergence. Goldberger et al., in [2], introduced the unscented-transform approximation and derived an EM-based algorithm to learn the reduced representation. In [4], a fast algorithm, based on the Bregman k-means clustering was suggested. Garcia et al. introduced a clustering algorithm based on Bregman divergence [5]. In [6], Zhang et al. perform the simplification by minimizing an upper bound of the approximation error using the L_2 distance measure.

In the previously mentioned approaches, the simplification is based on optimizing a parametric distance or similarity measure related to the referenced model, while the discrimination characteristics between models of different classes are discarded. Recently, we introduced the *maximum correct association* (MCA) algorithm based on optimizing the probability of associating the components of the original models to their correct class in the reduced representation [8]. The MCA algorithm, also based on the variational approximation of the KL divergence, provided superior results, on pure phone classification task, reducing 128-order mixture models to small-order models. In the current study we introduce a two-stage procedure for discriminative simplification of mixture models. In the first stage we apply the varEM algorithm (as in [3]) to produce the initial reduced model set. Then, the MCA algorithm is applied to refine the reduced models in the direc-

tion of maximizing the *correct association* criteria.

In the next section, the MCA algorithm for simplifying mixture models is described. In section 3 we evaluate the suggested procedure by deriving reduced order GMMs in a language recognition task. Finally, section 4 concludes the paper.

2. THE MCA ALGORITHM

Assume a classification problem involves a decision among N classes, where each class has a GMM representation

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

where $f_{ci}(x) \sim \mathcal{N}(m_{ci}, V_{ci})$, and N_c is the model order of class c. The objective is to use the given models $\{F_c\}$ to derive a new set of simplified mixture models of order $M_c < N_c$,

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

where $g_{cj}(x) \sim \mathcal{N}(\mu_{cj}, \mathbf{\Sigma}_{cj})$, such that the reduced set obtains a good discrimination for the classification problem.

2.1. The correct association criterion

A commonly used and most natural measure for the distance between two distributions, $p(x)$ and $q(x)$, is the relative entropy also named the Kullback-Leibler (KL) divergence, given by:

$$D(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx. \qquad (3)$$

We use the KL-divergence to define the following discriminative criterion:

$$\Lambda = \sum_{c=1}^{N} [D(F_c\|G_w) - D(F_c\|G_c)] \qquad (4)$$

where

$$G_w(x) = \frac{1}{N} \sum_{c=1}^{N} G_c(x). \qquad (5)$$

G_w is regarded as the *world* model since it is defined as a combination of all the simplified models to represent the background distribution. Maximizing Λ with respect to the parameters of $\{G_c\}$ means that, in average, we wish to minimize the KL-divergence of the reduced model G_c from its original respective (i.e. same-class) model F_c, while maximizing the KL-divergence between F_c and the world model G_w. The KL-divergence between GMMs cannot be computed analytically, and therefore we need to use an analytical approximation. The authors of [7] introduced the variational approximation and showed that it was superior to other methods, at measuring divergence among phonetic models.

2.2. The variational approximation of relative entropy

We consider two Gaussian mixture models: a source model, $F(x) = \sum_i \alpha_i f_i(x)$, where $f_i(x) \sim N(m_i, V_i)$, and a target model $G(x) = \sum_j \beta_j g_j(x)$, where $g_j(x) \sim N(\mu_j, \mathbf{\Sigma}_j)$. The cross-entropy between the source and target models is defined by

$$\begin{aligned} L(F\|G) &\equiv \int F(x) \log G(x) dx \\ &= \sum_i \alpha_i \int f_i(x) \log \sum_j \beta_j g_j(x) dx. \end{aligned} \qquad (6)$$

We can find a variational lower bound for the cross-entropy by introducing variational distribution parameters, $q_{(j|i)} > 0$, such that $\sum_j q_{(j|i)} = 1$. The variational parameters $q_{(j|i)}$ can be regarded as the association probability of component i from model F, to component j of model G. By Jensen's inequality we obtain a lower bound for (6) as follows

$$\begin{aligned} L(F\|G) &= \sum_i \alpha_i \int f_i(x) \log \sum_j q_{(j|i)} \frac{\beta_j g_j(x)}{q_{(j|i)}} dx \\ &\geq \sum_i \alpha_i \int f_i(x) \sum_j q_{(j|i)} \log \frac{\beta_j g_j(x)}{q_{(j|i)}} dx \\ &= \sum_i \alpha_i \sum_j q_{(j|i)} \left[\log \frac{\beta_j}{q_{(j|i)}} + \int f_i(x) \log g_j(x) dx \right] \\ &= \sum_i \alpha_i \sum_j q_{(j|i)} \left[\log \frac{\beta_j}{q_{(j|i)}} - D(f_i\|g_j) + H(f_i) \right] \\ &:= L_q(F\|G), \end{aligned} \qquad (7)$$

where $H(f_i)$ is the entropy of mixture component f_i. The lower bound $L_q(F\|G)$ can be maximized with respect to the variational parameters $q_{(j|i)}$, yielding the closest bound by

$$\hat{q}_{(j|i)} = \frac{\beta_j e^{-D(f_i\|g_j)}}{\sum_{j'} \beta_{j'} e^{-D(f_i\|g_{j'})}}. \qquad (8)$$

Replacing $\hat{q}_{(j|i)}$ in (7) yields the maximum of the lower bound

$$L_{\hat{q}}(F\|G) = \sum_i \alpha_i \log \sum_j \beta_j e^{-D(f_i\|g_j) + H(f_i)}. \qquad (9)$$

The approximation of the KL-divergence can be obtained directly by the variational approximation of the cross-entropy in (9), and the equality $D(F\|G) = L(F\|F) - L(F\|G)$, yielding

$$\begin{aligned} \tilde{D}(F\|G) &= L_{\hat{q}}(F\|F) - L_{\hat{q}}(F\|G) \\ &= \sum_i \alpha_i \log \frac{\sum_{i'} \alpha_{i'} e^{-D(f_i\|f_{i'})}}{\sum_j \beta_j e^{-D(f_i\|g_j)}}. \end{aligned} \qquad (10)$$

The variational approximation of (10) is tractable, since the KL-divergence for two Gaussian distributions $f(x) \sim \mathcal{N}(m, V)$ and $g(x) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ has a closed-form expression

$$D(f\|g) = \frac{1}{2}\log\frac{|\mathbf{\Sigma}|}{|V|} + \frac{1}{2}\mathrm{Tr}(\mathbf{\Sigma}^{-1}V) - \frac{d}{2} + \frac{1}{2}(m-\mu)^T\mathbf{\Sigma}^{-1}(m-\mu)$$

where d is the dimension of x.

2.3. Algorithm Derivation

It can be easily shown that setting the approximation of (10) into expression (4), yields (after dropping the $\log N$ term that has no influence on the optimization) the following objective function as approximation for the correct association criterion Λ in (4):

$$\mathcal{J}=\sum_{c=1}^{N}\sum_{i=1}^{N_c}\alpha_{ci}\log\frac{\sum_{j=1}^{M_c}\beta_{cj}e^{-D(f_{ci}\|g_{cj})}}{\sum_{k=1}^{N}\sum_{j=1}^{M_k}\beta_{kj}e^{-D(f_{ci}\|g_{kj})}}. \quad (11)$$

For further description of the algorithm, we introduce the following definition

$$P_{kj|ci} \equiv \frac{\beta_{kj}e^{-D(f_{ci}\|g_{kj})}}{\sum_{k'=1}^{N}\sum_{j'=1}^{M_{k'}}\beta_{k'j'}e^{-D(f_{ci}\|g_{k'j'})}}. \quad (12)$$

$P_{kj|ci}$ has a meaningful stochastic interpretation. It can be shown to be the optimal probability of associating Gaussian component i of the source model c (noted as f_{ci}), with Gaussian component j of the reduced model k (noted as g_{kj}) that is obtained by maximizing the variational principle presented in rate-distortion theory. Next, we use (12) to compute the probability of correct association of f_{ci} by:

$$P_{c|ci} = \sum_{j=1}^{M_c}P_{cj|ci}. \quad (13)$$

Now, the optimization criterion (11) can be expressed in terms of the expected "log-likelihood" of correct association:

$$\mathcal{J} = \sum_{c=1}^{N}\sum_{i=1}^{N_c}\alpha_{ci}\log P_{c|ci}. \quad (14)$$

The choice of reduced models (2) that optimize $\mathcal{J}$ is regarded as the maximum correct association (MCA) solution to the GMM simplification problem.

Having defined the mean correct association functional (14), the next step is to maximize it w.r.t. the parameters of the reduced model set (2), β_{cj}, μ_{cj}, and $\mathbf{\Sigma}_{cj}$. Since $\mathcal{J}$ is not convex the maximization is performed using gradient based optimization. In order to ensure that the weights fulfill the conditions $\beta_{cj} > 0$ and $\sum_{j=1}^{M_c}\beta_{cj} = 1$, and that the estimated covariance matrices, $\mathbf{\Sigma}_{cj}$, are positive definite, we use the following transformations:

$$\beta_{cj} = \frac{e^{w_{cj}}}{\sum_{j'=1}^{M_c}e^{w_{cj'}}} \quad (15)$$

$$\mathbf{\Sigma}_{cj} = e^{V'_{cj}}, \quad (16)$$

such that updating $\{w_{cj}\}_{j=1}^{M_c}$ and V'_{cj} yields a direct derivation of $\{\beta_{cj}\}_{j=1}^{M_c}$ and $\mathbf{\Sigma}_{cj}$, respectively. Differentiating the objective function w.r.t. w_{cj}, μ_{cj}, and V'_{cj} yields the following gradient rules; For the transformed weights,

$$\frac{\partial\mathcal{J}}{\partial w_{cj}} = \left(\sum_{i=1}^{N_c}\alpha_{ci}\frac{P_{cj|ci}}{P_{c|ci}} - \sum_{k=1}^{N}\sum_{i=1}^{N_k}\alpha_{ki}P_{cj|ki}\right)(1-\beta_{cj}) \quad (17)$$

for the means,

$$\frac{\partial\mathcal{J}}{\partial\mu_{cj}} = \sum_{i=1}^{N_c}\alpha_{ci}\frac{P_{cj|ci}}{P_{c|ci}}{\mathbf{\Sigma}_{cj}}^{-1}(m_{ci}-\mu_{cj}) - \sum_{k=1}^{N}\sum_{i=1}^{N_k}\alpha_{ki}P_{cj|ki}{\mathbf{\Sigma}_{cj}}^{-1}(m_{ki}-\mu_{cj}), \quad (18)$$

and for the transformed covariance matrices,

$$\frac{\partial\mathcal{J}}{\partial V'_{cj}} = \frac{1}{2}\sum_{i=1}^{N_c}\alpha_{ci}\frac{P_{cj|ci}}{P_{c|ci}}\left[{\mathbf{\Sigma}_{cj}}^{-1}h(m_{ci},\mu_{cj},V_{ci}) - I\right] -\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{N_k}\alpha_{ki}P_{cj|ki}\left[{\mathbf{\Sigma}_{cj}}^{-1}h(m_{ki},\mu_{cj},V_{ki}) - I\right], \quad (19)$$

where

$$h(m_i,\mu_j,V_i) \equiv V_i + (m_i-\mu_j)(m_i-\mu_j)^T.$$

With this we obtain the analytical computation of the gradients of $\mathcal{J}$ w.r.t. the parameters of the reduced GMMs (equations (17)-(19)), using only the parameters of the original GMMs. The optimization of the reduced set can be performed by simple gradient ascent iterations, or by using other optimizers (e.g. conjugate gradient) to avoid local maxima traps.

2.4. Implementation issues

We used a simple gradient ascent optimization. There are methods to find the optimal step size in each iteration, but they require more computations. We applied a simpler approach where we used an adjustable step size: in each iteration, model parameters are updated and the new objective function $\mathcal{J}$ is computed. When getting an improvement, the step is increased by factor 1.1 and we continue to the next iteration. If the objective degrades the step is reduced by factor 2, model parameters are re-calculated and we check the new objective. Only after the objective is improved we continue to the next iteration. Another issue refers to ignoring extreme errors during the learning process. The value of the correct association probability greatly affects the update of the gradients such that the algorithm emphasizes "learning from errors". When $P_{c|ci} \to 1$, the update of the gradients regarding component f_{ci} tends toward zero. As the correct association probability becomes smaller, the updates become more substantial. Sometimes it is useful to ignore extreme errors in order to avoid over-fitting to extreme situations. In the MCA method, it is simple to define a threshold on the probability

$P_{c|ci}$, such that for values under a certain $P_{threshold}$ the gradients' update is skipped. In all the experiments reported in the next section the threshold $P_{threshold} = 0.02$ was used.

3. EXPERIMENTS

Experiments were conducted on a basic language recognition task, in order to compare the performance of reduced order models obtained by varEM with and without the MCA algorithm. In the language recognition framework every speech frame was 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) as in [9]. RASTA (Relative Spectral Transform) filtering was applied to reduce channel effects [10]. Language GMMs, with diagonal covariance matrices, were trained with the conventional maximum-likelihood EM algorithm. Mixed-gender language models were trained from a subset of the CallFriend Corpus, and recognition was performed on the 30 sec segments from NIST language recognition evaluation (LRE) 2003 and from the evaluation part of NIST 1996 (noted as LRE 96e) [1]. There were 12 target languages in the evaluation. For performance evaluation, we used the standard Detection Error Trade-off (DET) curve and the Equal Error Rate (EER) measure. DET is a plot of false accept probability against miss probability dependent on the score threshold. The EER is the point were the false-accept probability and the miss probability become equal. A simple log likelihood ratio score was used for each test utterance X_i, per language model G_c, as follows:

$$S_c(X_i) = \log P(X_i|G_c) - \max_{k \neq c} \{\log P(X_i|G_k)\}. \quad (20)$$

To begin with, we examined the effect of different GMM orders on the recognition performance in standard maximum-likelihood (ML) training from the original samples. In Figure 1 we observe that performance keeps improving as model order increases.

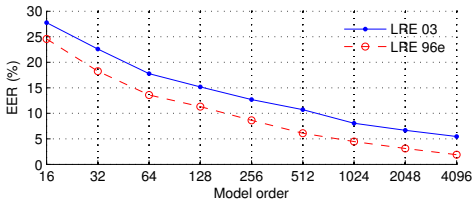

Fig. 1. *Language recognition results for different orders of GMMs, trained from the original data. Results were obtained using the 30 sec excerpts of LRE 03 and of LRE 96e.*

Figure 2 demonstrates the optimization process of the MCA. The optimization starts with initial models obtained by the varEM algorithm (appearing as iteration 0 in the figure).

[1] http://www.itl.nist.gov/iad/mig/tests/lre/

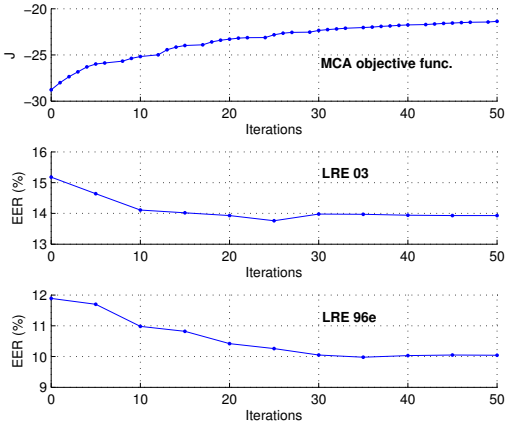

Fig. 2. *Demonstration of MCA iterations in reducing 4096-order models to 128-order models. The upper plot brings the values of the MCA objective function (10) in each iteration. The middle and lower plots present the performance obtained on LRE 03 and LRE 96e, respectively, expressed by EER evaluated every 5 iterations.*

It is seen that the objective function (14) increases with each iteration of the gradient ascent. Around the 30th iteration, the objective improvement slows down and becomes irrelevant to the performance on the test set. Our experiments studied the simplification of 4096-order GMMs into reduced models with orders in the range of 16 to 256 components. For comparison, reduced order models were generated using three methods: the traditional ML training from the data, the varEM algorithm, and the varEM followed by the MCA algorithm.

The results in Figure 3, clearly indicate that the MCA algorithm obtains improved performance when it is applied after the non-discriminative EM algorithm. The improvement grows from about 10% reduction in EER for higher order models (256, 128) up to 25% reduction in EER for low order models (16, 32). For a different view on the achieved improvement we bring in Figure 4 the DET plot for one instance where models of order 4096 were reduced to 256 components. Figure 4 shows that the MCA algorithm obtains a significant improvement at each working point of the graph. It is apparent that the MCA algorithm consistently improves the performance of the reduced models compared to the non-discriminative EM-based algorithm in all the language recognition tests. While the efficient hierarchical EM method (varEM) reaches, more or less, the performance of a similar order model trained directly from the data by conventional maximum-likelihood, the MCA algorithm consistently outperforms both methods. A most remarkable effect considering the fact that it relies only on the given high order model parameters without further accessing the original data.

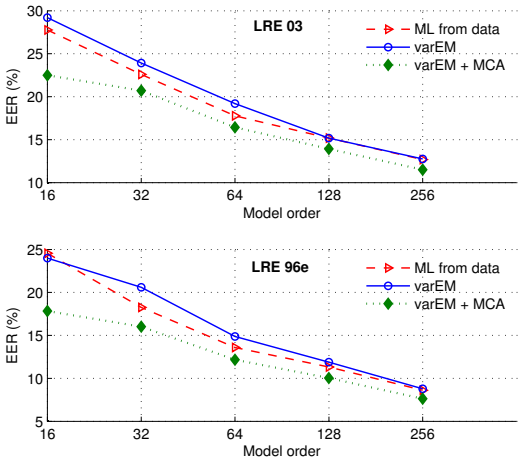

Fig. 3. *Recognition results for several orders of reduced models. The conventional EM training from the original data indicated as ML from data, serves as benchmark. The varEM and MCA models were derived from original models of order 4096.*

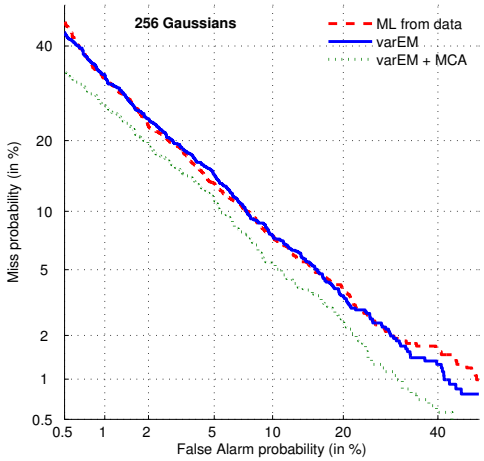

Fig. 4. *DET curve for 256 order reduced models on LRE 96e database. ML from data serves as a benchmark. The hierarchical algorithms derived the reduced set from GMMs of order 4096. The equal error rates are 8.6% for the benchmark, 8.8% for varEM and 7.6% for varEM + MCA.*

4. CONCLUSION

We proposed an efficient procedure for simplifying large order mixture models through hierarchical discriminative learning. The procedure has two stages. First, a reduced order model is created by the variational-EM algorithm. Then, the parameters of the model are refined by a *maximum correct association* (MCA) algorithm. The MCA algorithm is a hierarchical discriminative algorithm that iteratively refines the parameters of the compact GMMs in order to increase the probability of associating original Gaussian components to their correct class in the compact set. We applied the method to a basic language recognition task. The experimental results indicated a significant improvement in performance after the MCA procedure was performed over the reduced model set. The proposed algorithm is applicable to other parameter estimation problems that use large order Gaussian mixture models.

5. REFERENCES

[1] J. Goldberger and S. Roweis, "Hierarchical clustering of a mixture model," in *Neural Information Processing Systems (NIPS)*, 2005.

[2] J. Goldberger, H. K. Greenspan, and J. Dreyfuss, "Simplifying mixture models using the unscented transform," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1496–1502, 2008.

[3] P. L. Dognin, J. R. Hershey, V. Goel, and P. A. Olsen, "Refactoring acoustic models using variational Expectation-Maximization," in *10th Ann. Conf. of ISCA* INTERSPEECH 2009.

[4] F. Nielsen, V. Garcia, and R. Nock, "Simplifying gaussian mixture models via entropic quantization," in *17th European Conf. on Signal Processing* EUSIPCO 2009.

[5] V. Garcia, F. Nielsen, and R. Nock, "Hierarchical Gaussian mixture model," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, ICASSP 2010.

[6] K. Zhang and J. T. Kwok, "Simplifying mixture models through function approximation," *Trans. Neural Network*, vol. 21, no. 4, pp. 644-658, 2010.

[7] J. R. Hershey and P. A. Olsen, "Approximating the Kullback-Leibler divergence between Gaussian mixture models," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, ICASSP 2007.

[8] Y. Bar-Yosef and Y. Bistritz, "Discriminative simplification of mixture models," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* , ICASSP 2011.

[9] P. Matejka, *Phonotactic and Acoustic Language Recognition*, Phd. Thesis, BRNO university, 2008.

[10] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31-44, 1996.