Discriminative Joint Vector and Component Reduction for Gaussian Mixture Models

Yossi Bar-Yosef and Yuval Bistritz
School of Electrical Engineering, Tel Aviv University, Tel Aviv 69978, Israel.
yossibaryosef@gmail.com, bistritz@tau.ac.il

***Abstract*—We introduce a discriminative parametric vector dimensionality reduction algorithm for Gaussian mixtures that is performed jointly with mixture component reduction. The reduction algorithm is based on the variational maximum mutual information (VMMI) method, which in contrast to other reduction algorithms, requires only the parameters of existing high order and high dimensional mixture models. The idea behind the proposed approach, called JVC-VMMI (for joint vector and component VMMI), differs significantly from traditional classification approaches that perform separately dimensionality reduction first, and then use the low-dimensional feature vector for training lower order models. The fact that the JVC-VMMI approach is relieved from using the original data samples admits an extremely efficient computation of the reduced models optimized for the classification task. We report experiments in vowel classification in which JVC-VMMI outperformed conventional Linear Discriminant Analysis (LDA) and Neighborhood Component Analysis (NCA) dimensionality reduction methods.**

***Index Terms*—Dimensionality reduction, Gaussian mixture models, Discriminative learning, Hierarchical clustering.**

I. Introduction

The Gaussian mixture model (GMM) is a very powerful parametric modelling tool that is often used to represent complex data distributions. In many learning processes, high-order models (models containing a large number of components) are trained in a high-dimensional feature vector space. The use of large mixture models often becomes computationally expensive for practical implementations. An effective approach to deal with such situations is to use reduction methods that can be applied to the parameters of the already known, but too complex, mixture models. Various component reduction algorithms (sometimes called hierarchical clustering) that generate a simplified model that maximizes its similarity to the original mixture model were proposed [1]-[4]. Recently, it has been demonstrated that for classification tasks, it is more effective to learn the set of reduced order models by a joint multi-class optimization, rather than learning separately a reduced model for each class. [5] [6] [7] The joint model reduction problem in [7] was posed as an information theoretic principle named variational maximum mutual information (VMMI).

Another aspect of model simplification relates to the dimensionality reduction of the feature vector space. Finding a lower-dimensional representation for the high-dimensional feature vectors is a key ingredient in machine learning. It is traditionally performed as part of the feature extraction phase before models are trained using the lower dimensional vectors as data sets. Dimension reduction is done to combat the "curse of dimensionality", and it carries several benefits: leads to computational cost reduction; finds representations on independent principal components that usually improve the trained model; provides insightful 2D and 3D data visualizations.

Many well known vector reduction (VR) methods essentially assume that the underlying structure of the data can be approximated by Gaussian clusters. Principal Component Analysis (PCA) [9] and probabilistic-PCA (PPCA) [10] assume a single Gaussian structure for all data; Linear Discriminant Analysis (LDA) [11] assumes a single Gaussian for each class with shared covariance; Heteroscedastic Discriminant Analysis (HDA) [12] assumes a single Gaussian for each class with unequal covariance. More complex methods, that emphasize the strength of Gaussian mixtures, include mixture of factor analyzers (MFA) [13], mixtures of probabilistic principal component analyzers [14], and mixture of probabilistic LDA (PLDA) suggested in [15]. Basically, these methods search for a linear transformation for the original high-dimensional feature vectors, without giving direct attention to the constraints and structure of the final modelling scheme. In addition, discriminative methods that directly process the data, like Neighborhood Component Analysis (NCA) [16], pose a demanding cost of $O(n^2)$ pairwise distance calculations per iteration, for processing n samples.

Some related studies attempt to alleviate the latter computational burden by explicitly using Gaussian mixtures. In [17], the authors proposed a method called DCA-GM (discriminative component analysis by Gaussian mixtures), whose computational complexity has only $O(n \cdot r)$ calculations per iteration, where r is the total number of Gaussians, and using $r \ll n$. It uses a reduced-order GMM per class to learn a linear discriminative transformation and it achieves comparable results to the performance of NCA in K-Nearest Neighbor (KNN) classification.

Pertinent to the approach in this paper are also some previous VR studies. In [18] the authors reduce the computational burden by describing the large data set by large GMMs, then a linear transformation optimization is performed over the parameters of the Gaussian components involving computational cost of $O(m^2)$ of pairwise divergence (KL-distance) calculations per iteration, where m is the number of Gaussian components of the model and typically $m \ll n$. This method can be thought of as a "hierarchical" neighborhood component analysis, where a linear transformation is learned through maximizing a mutual information criterion (Renyi's

978-9-0827-9703-9/19/$31.00

entropy). In [19], the authors proposed discriminative VR based on MMI, in which the transformed data vectors are scored against a set of GMMs that represent the data distribution. Thangavelu and Raich [20] presented linear VR with multi-class GMMs using Chernoff union bound. They pre-trained low-order GMMs to present the classes and optimized a linear transformation that minimizes the Chernoff union upper bound on the probability of the error after projection onto the reduced subspace. Differently from [19], they use a parametric optimization scheme to deliver the transformation matrix.

All the aforementioned approaches perform VR without taking into account the structure and capacity (e.g. the number of clusters) of the final parametric model that is designated to use the reduced size vectors. Whatever space dimensionality reduction approach is taken, next follows the question of adjusting the model's size/capacity for training. Complex methods, like NCA, that do not pose constraints on model capacity, can be optimal for high-capacity models, but may result in poor accuracy when low-capacity models were used. Recently, Yang et. al. [21] addressed this issue by introducing joint learning of unsupervised VR for training a GMM over raw data samples. Their joint optimization outperformed traditional methods that do VR first and GMM training only afterwards.

In this paper we propose to carry out reduction of the dimension of vectors within the discriminative hierarchical VMMI method [7]. We introduce an extension to the VMMI method that can perform joint component reduction and vector reduction, named JVC-VMMI for short. In this setting, a linear transformation matrix is embedded into the VMMI criterion, to be able to perform VR along with component reduction (CR), in a discriminative manner. The suggested method combines the following major advantages: it considers limitations on the final reduced models while reducing the vector dimension; it provides an extremely fast parametric learning algorithm based on the parameters of pre-trained high-order mixtures; it uses a criterion that approximates the mutual information among the original GMMs and the reduced GMMs classes, resulting in a discriminative criterion.

Section II brings background on the VMMI criterion for mixture models. In Section III we describe the VMMI algorithm with linear projection for joint GMM reduction. In Section IV we evaluate the JVC-VMMI procedure in vowel classification. Conclusion is brought in the final section.

II. Background - VMMI for Mixture Models

The VMMI method enables an extremely fast reduction of a given set of high-order GMMs into a new set of reduced-component models, with improved accuracy in classification tasks.

We consider a set of N class models $\{F_c(x)\}_{c=1}^{N}$, where each $F_c(x)$ is a GMM of high order M_c

$$F_c(x) = \sum_{i=1}^{M_c} \alpha_{ci} f_{ci}(x), \tag{1}$$

where M_c presents the number of Gaussian components $f_{ci}(x)$ in class c and α_{ci} are the mixing weights, $0 < \alpha_{ci} < 1$, $\sum_{i=1}^{M_c} \alpha_{ci} = 1$. Each $f_{ci}(x)$ is a Gaussian

$$f_{ci}(x) = \mathcal{N}(x|m_{ci}, V_{ci}), \tag{2}$$

with mean vector m_{ci} and covariance matrix V_{ci}. We define a new set of N GMMs of lower orders $R_c < M_c$,

$$G_c(x) = \sum_{j=1}^{R_c} \beta_{cj} g_{cj}(x) \quad , \quad c = 1, \ldots, N \quad , \tag{3}$$

with mixing weights β_{cj}, $\sum_{j=1}^{R_c} \beta_{cj} = 1$, and Gaussian components $g_{cj}(x)$ with means and covariance matrices denoted by

$$g_{cj}(x) = \mathcal{N}(x|\mu_{cj}, \mathbf{\Sigma}_{cj}). \tag{4}$$

In [7] it was shown that the realization of the mutual information between the observations $x \in X$, modeled by $G_c(x)$, and the class variable $c \in C$, given that $x|c$ is distributed according to its high-order representation $F_c(x)$, can be approximated by a variational Bayes approach as

$$\mathcal{J} = \sum_{c=1}^{N} p_c \sum_{i=1}^{M_c} \alpha_{ci} \log \left(\frac{p_c \sum_{j=1}^{R_c} \beta_{cj} e^{-KL(f_{ci} \| g_{cj})}}{\sum_{k=1}^{N} p_k \sum_{j=1}^{R_k} \beta_{kj} e^{-KL(f_{ci} \| g_{kj})}} \right), \tag{5}$$

where p_c is the a-priori probabilities for each class, $0 < p_c < 1$, $\sum_{c=1}^{N} p_c = 1$, and KL denotes the Kullback-Leibler (KL) divergence between the pdf components. Since the KL divergence for two multi-dimensional Gaussian distributions (e.g. $KL(f_{ci} \| g_{kj})$) has a well known analytical expression, $\mathcal{J}$ is differentiable and its maximization is analytically tractable.

III. VMMI with dimensionality reduction

The VMMI method described in the previous section solves the problem of component-reduction from high-order mixtures to lower-order mixtures embedded in the same vector space. In this section we extend the VMMI setting to deal also with dimensionality reduction of the vector space.

We consider the set of GMMs $\{F_c(x)\}_{c=1}^{N}$ representing the distribution of N classes, as described in the previous section. Each model $F_c(x)$, represents the probability density function of vectors $x \in \mathbb{R}^D$, as in Equations (1) and (2).

Now we want to derive a simplified set of models $\{G_c(y)\}_{c=1}^{N}$ with both reduced-dimension vectors $y \in \mathbb{R}^d$, where $d < D$, and reduced-order mixtures with $R_c < M_c$. The reduced GMM $G_c(y)$ is defined by

$$G_c(y) = \sum_{j=1}^{R_c} \beta_{cj} g_{cj}(y), \tag{6}$$

with components

$$g_{cj}(y) = \mathcal{N}(y|\mu_{cj}, \mathbf{\Sigma}_{cj}). \tag{7}$$

We define a linear projection matrix $\mathbf{A}$ of size $(d \times D)$, that projects the high-dimensional vector x to a vector y in the

lower-dimensional space by $y = \mathbf{A}x$. Consequently, the linear projection of each original Gaussian, $f_{ci}(x) = \mathcal{N}(x|m_{ci}, V_{ci})$, onto the lower-dimensional space, can be written by a Gaussian of correspondingly transformed mean and covariance

$$f_{ci}^{\mathbf{A}}(y) = f_{ci}^{\mathbf{A}}(\mathbf{A}x) = \mathcal{N}(y|\mathbf{A}m_{ci}, \mathbf{A}V_{ci}\mathbf{A}^T) \ . \quad (8)$$

Following (8), the corresponding projection of the entire mixture model is denoted by

$$F_c^{\mathbf{A}}(\mathbf{A}x) = \sum_{i=1}^{M_c} \alpha_{ci} f_{ci}^{\mathbf{A}}(\mathbf{A}x) \quad , \quad c = 1, \ldots, N \ . \quad (9)$$

With these definitions, a VMMI objective function similar to $\mathcal{J}$ in (5) is obtained with $\{F_c^{\mathbf{A}}(y)\}_{c=1}^N$ and $\{G_c(y)\}_{c=1}^N$. Namely, the original high-dimensional parameters $\{\alpha_{ci}, m_{ci}, V_{ci}\}$ are simply replaced by the parameters of the reduced set $\{\alpha_{ci}, \mathbf{A}m_{ci}, \mathbf{A}V_{ci}\mathbf{A}^T\}$. Then, for fixed values of $\{\beta_{cj}, \mu_{cj}, \mathbf{\Sigma}_{cj}\}_{c=1}^N$ each increment of $\mathcal{J}$ w.r.t. the matrix $\mathbf{A}$ will deliver a better projected set $F_c^{\mathbf{A}}(\mathbf{A}x)$ in the sense of the approximated mutual information. The goal of joint reduction of both dimension and order of the models amounts to a VMMI simultaneous optimization of the projection matrix $\mathbf{A}$ and the parameters of $\{G_c\}_{c=1}^N$.

The VMMI objective function $\mathcal{J}$ (5) can be maximized with respect to the projection matrix $\mathbf{A}$ and the parameter set of the reduced models $\theta \sim \{\beta_{cj}, \mu_{cj}, \mathbf{\Sigma}_{cj}\}_{c=1}^N$ by a gradient-based optimization since all the required gradients have closed-form expressions. Evidently, the objective function is not convex and therefore its optimization should be done with some care to avoid poor local maxima. In the following we derive closed forms expressions for the optimization of the JVC-VMMI objective function by the Generalized Probabilistic Descent (GPD) algorithm that was shown to work well for Gaussian mixtures [7].

To ensure non-negativity conditions and achieve better convergence, the GPD technique defines the following transformations:

$$\beta_{cj} = \frac{e^{w_{cj}}}{\sum_{j'=1}^{R_c} e^{w_{cj'}}}; \quad \mathbf{\Sigma}_{cj} = \exp\left(\tilde{\mathbf{\Sigma}}_{cj}\right); \quad \tilde{\mu}_{cj} = \mathbf{\Sigma}_{cj}^{-\frac{1}{2}} \mu_{cj}.$$

The advantage in GPD is that the optimization of $\{\beta_{cj}, \mu_{cj}, \mathbf{\Sigma}_{cj}\}_{c=1}^N$ can be performed through simple gradient ascent optimization of $\{w_{cj}, \tilde{\mu}_{cj}, \tilde{\mathbf{\Sigma}}_{cj}\}_{c=1}^N$ without setting additional external non-negativity constraints. Let us first mention two intermediate calculations (named "association probabilities") that will be used next in the derivatives' equations,

$$\hat{q}_{c(j|i)} = \frac{\beta_{cj} e^{-KL(f_{ci}\|g_{cj})}}{\sum_{j'=1}^{R_c} \beta_{cj'} e^{-KL(f_{ci}\|g_{cj'})}} \ , \quad (10)$$

$$\hat{w}_{(kj|ci)} = \frac{p_k \beta_{kj} e^{-KL(f_{ci}\|g_{kj})}}{\sum_{k'=1}^{N} p_{k'} \sum_{j'=1}^{R_{k'}} \beta_{k'j'} e^{-KL(f_{ci}\|g_{k'j'})}}. \quad (11)$$

The following derivatives are required. Their derivation was carried out in [8] where some related derivation can be found also in [7]. The gradient for the optimization of the transformed weights is

$$\frac{\partial \mathcal{J}}{\partial w_{cj}} = \left(p_c \sum_{i=1}^{M_c} \alpha_{ci} \, \hat{q}_{c(j|i)} - \sum_{k=1}^{N} p_k \sum_{i=1}^{M_k} \alpha_{ki} \, \hat{w}_{(cj|ki)} \right) \left(1 - \beta_{cj}\right) , \quad (12)$$

for the means it is

$$\frac{\partial \mathcal{J}}{\partial \tilde{\mu}_{cj}} = p_c \sum_{i=1}^{M_c} \alpha_{ci} \, \hat{q}_{c(j|i)} \mathbf{\Sigma}_{cj}^{-\frac{1}{2}} \left(\mathbf{A}m_{ci} - \mu_{cj}\right) - \sum_{k=1}^{N} p_k \sum_{i=1}^{M_k} \alpha_{ki} \, \hat{w}_{(cj|ki)} \mathbf{\Sigma}_{cj}^{-\frac{1}{2}} \left(\mathbf{A}m_{ki} - \mu_{cj}\right) , \quad (13)$$

and the gradient for the transformed covariance matrices, $\frac{\partial \mathcal{J}}{\partial \tilde{\mathbf{\Sigma}}_{cj}}$, are given below in Eq. (14) assuming the new means $\hat{\mu}_{cj}$ are already obtained in the gradient step after using (13). Finally, the gradient expression with respect to the linear projection matrix $\mathbf{A}$ is obtained in Eq. (15). Equations (12) - (15) provide analytical expressions for all the gradients required for the optimization of $\mathcal{J}$. Depending on the application, gradient-based updates can be jointly done for all the parameters along iterations, or if desired, only the linear projection matrix can be learned, through (15), subjected to a fixed embedding of mixture components.

Notice that the update of $\mathbf{A}$ should be treated with caution. In our experiments, we were required at times to apply some preconditioning to the system in order to prevent the matrix $(\mathbf{A}V_{ci}\mathbf{A}^T)$, in Eq. (15), from approaching singularity. In these situations some rescaling of the given space can be performed to ensure that $|\mathbf{A}V_{ci}\mathbf{A}^T|$ remains above some minimum value. In addition, a proper regularization on the scale of $\mathbf{A}$ can also be used.

IV. Experiments in Vowel Classification

JVC-VMMI was tested in a vowel classification experiments on samples from the TIMIT database. To neutralize effects that are not related to pure GMM modeling, GMM models where trained using the middle part of a phone segments. For each phonetic class, the feature vectors from the middle part of each phone occurrence (75% of the manually segmented part) were pooled together and used to train a GMM. Testing was performed on phonetic segments, where a feature vector sequence $\{x_1, \ldots, x_T\}$, taken from the middle part of the vowel, was first transformed using $y_t = \mathbf{A}x_t$, and then scored against the reduced model set. The score for each class c is computed by

$$S_c = \frac{1}{T} \sum_{t=1}^{T} \log G_c(y_t),$$

and the sample was classified to the winner.

As references, the classic Linear Discriminant Analysis (LDA) [11], and Neighborhood Components Analysis (NCA) [16] were used. LDA and NCA follow significantly different ideas. LDA naively assumes that all class distributions are Gaussian with a single shared covariance. NCA, in difference, does not make any assumptions on the number and size of

$$\frac{\partial \mathcal{J}}{\partial \tilde{\boldsymbol{\Sigma}}_{cj}} = \frac{1}{2}\, p_c \sum_{i=1}^{M_c} \alpha_{ci}\, \hat{q}_{c(j|i)} \left\{ \boldsymbol{\Sigma}_{cj}^{-1} \left[\mathbf{A} V_{ci} \mathbf{A}^T + (\mathbf{A} m_{ci} - \hat{\mu}_{cj}) (\mathbf{A} m_{ci} - \hat{\mu}_{cj})^T \right] - I \right\}$$
$$- \frac{1}{2} \sum_{k=1}^{N} p_k \sum_{i=1}^{M_k} \alpha_{ki}\, \hat{w}_{(cj|ki)} \left\{ \boldsymbol{\Sigma}_{cj}^{-1} \left[\mathbf{A} V_{ki} \mathbf{A}^T + (\mathbf{A} m_{ki} - \hat{\mu}_{cj}) (\mathbf{A} m_{ki} - \hat{\mu}_{cj})^T \right] - I \right\} \tag{14}$$

$$\frac{\partial \mathcal{J}}{\partial \mathbf{A}} = \sum_{c=1}^{N} \sum_{i=1}^{M_c} \sum_{j=1}^{R_c} p_c\, \alpha_{ci}\, \hat{q}_{c(j|i)} \left\{ \left(\mathbf{A} V_{ci} \mathbf{A}^T \right)^{-1} \mathbf{A} V_{ci} - \boldsymbol{\Sigma}_{cj}^{-1} \left[\mathbf{A} V_{ci} + (\mathbf{A} m_{ci} - \mu_{cj}) m_{ci}^T \right] \right\}$$
$$- \sum_{c=1}^{N} \sum_{i=1}^{M_c} \sum_{k=1}^{N} \sum_{j=1}^{R_k} p_c\, \alpha_{ci}\, \hat{w}_{(kj|ci)} \left\{ \left(\mathbf{A} V_{ci} \mathbf{A}^T \right)^{-1} \mathbf{A} V_{ci} - \boldsymbol{\Sigma}_{kj}^{-1} \left[\mathbf{A} V_{ci} + (\mathbf{A} m_{ci} - \mu_{kj}) m_{ci}^T \right] \right\} \tag{15}$$

data clusters. It maximizes a stochastic variant of a K-Nearest Neighbor (KNN) score (using Leave-One-Out (LOO)), and can learn highly complex data structures. Often NCA outperforms the classic LDA in a KNN classification framework. However, in situations of limited computational resources, NCA presents a serious computational difficulty requiring, for n samples, a magnitude of $O(n^2)$ Euclidian distance calculations per iteration. Another problem with NCA is that, although it works well in a KNN scheme, it can fail with small parametric models, since in training it does not consider the limited capacity of the final model. In our experiments, NCA was trained with about 15,000 vectors per class. The reduced model set is obtained with two independent steps. First the linear transformation matrix $\mathbf{A}$ is trained using LDA or NCA and afterwards the sought low-order GMMs are trained by standard EM using the transformed low-dimensional vectors. These two techniques are referenced next as **LDA+EM** and **NCA+EM**, respectively.

We used the original data partition of 3696 training sentences from which the vowel segments were extracted, for model training. Testing was performed on vowels of the remaining 1344 sentences. A 38-dimensional feature vector of MFCC $+\Delta + \Delta\Delta$ was used. Since, for the original 38-dimensional feature vector a phonetic models of around 128 components achieved best results, we focused on simplifying 128-component GMMs into reduced-order models with lower-dimensional feature vectors. All GMMs used diagonal covariance matrices.

Figure 1 brings comparative results of vowel classification tests, that included 7 English vowels taken from TIMIT: /aa/, /ah/, /ae/, /ao/, /eh/, /ih/, and /iy/. The four plots relate to different orders (number of components) of the reduced models (1, 2, 8, and 16), while the x-axis of each plot indicates the dimension of the new vector space (2, 4, 8, 16, 32, and the full dimension of 38). An interesting, although not surprising, observation is that LDA+EM outperforms NCA+EM in very low-order models. In fact, LDA performance remains quite stable and seems to be insensitive to model order increments, whereas the NCA outperforms it as model order increases. This outcome is in accordance with the conceptual differences between the two methods. The experiments demonstrate that

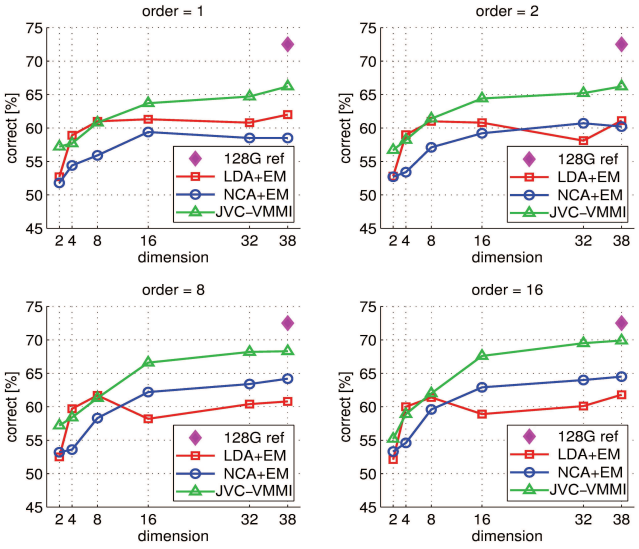

Fig. 1. Vowel classification accuracy. In the JVC-VMMI method, the original models, of order 128 with vectors of dimension 38, is the source. The accuracy of the original model is marked by "128G ref".

the joint vector reduction and component reduction using JVC-VMMI exhibits significantly better classification performance in comparison to LDA+EM and NCA+EM. Note also that the performance of JVC-VMMI keeps improving as long as the target models' order and space dimension keep increasing.

The superiority of the proposed method can be explained by the fact that the linear transformation, used to reduce the size of the feature vectors, is learned in conjunction with the model in which it is used afterwards for the classification task. Thus, the linear projection is optimized with the additional knowledge related to model capacity (in other words, the number of "dominant" clusters), a factor which seems to be a significant differentiator between JVC-VMMI and other methods that do not consider information about the final model.

The JVC-VMMI offers a highly efficient way to obtain discriminative dimensionality reduction without using the original data samples. Its starts by "compressing" the data into clusters

with a set of high-order models with a total number of components $M = \sum_{c=1}^{N} M_c$. After "compression" (having a set of GMMs trained by maximum-likelihood), in each iteration of the discriminative learning, these M components are scored against a reduced set of a total $R = \sum_{c=1}^{N} R_c$ components. Therefore, optimization complexity is proportional to $M \times R$. With this method the optimization complexity does not depend on the number of samples.

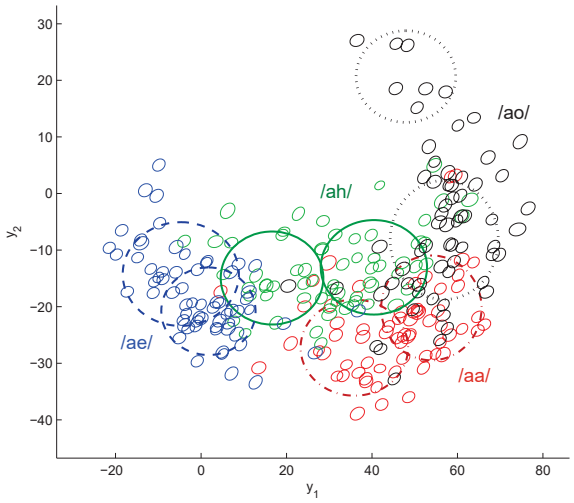

Fig. 2. Visualization of GMM reduction using JVC-VMMI. Linear projection from dimension $D = 38$ to $d = 2$ of 64-order GMMs to 2-order GMMs of 4 closely-pronounced English vowels: /aa/, /ah/, /ae/, and /ao/.

An attractive feature of JVC-VMMI is its possible use as a visualization tool. Minding that the JVC-VMMI procedure tries to discriminate between the components of different classes, JVC-VMMI admits insightful visualizations, having the flexibility of changing the number of underlined components. Figure 2 presents an example for such visualization. The original Gaussian components, that were projected onto a 2-dimensional space are drawn, and then the GMMs of reduced-order to 2 Gaussians each, are drawn on top of them. We used diagonal covariances, both for the original models and for the reduced-order models. Hence, the resulting rotation of the transformed Gaussian components is caused by the projection matrix $\mathbf{A}$, that was optimized to discriminate classes in a constrained 2-order mixture model structure. More insights can be gained by exploring more clusters in the low-dimension space, or getting visualizations at different stages along the steps of the JVC-VMMI optimization.

V. Conclusion

In this paper we have extended the framework of variational maximum mutual information (VMMI) to support linear dimensionality reduction jointly with component reduction. The method, dubbed JVC-VMMI, discriminatively simplifies models for classification tasks. The presented experiments show that the joint optimization of the projection matrix, together with the embedding of model parameters, is significantly more powerful than methods that do the vectors reduction before and irrespectively from the planned low order of the models in the classification task. While the JVC-VMMI was designed also to compact the number of components of the mixture models, this framework may be useful also solely for dimensionality-reduction. All over, the framework offers new and more flexible tradeoffs of computational complexity, model capacity, and classification performance.

References

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

[2] 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.

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

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

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

[6] Y. Bar-Yosef and Y. Bistritz, "Discriminative algorithm for compacting mixture models with application to language recognition," in *Proc. of the 20th Europian Signal Processing Conference, EUSIPCO,* 2012.

[7] Y. Bar-Yosef and Y. Bistritz, "Gaussian mixture models reduction by variational maximum mutual information," *IEEE Trans. Signal Processing*, vol. 63(6) pp. 1557–1569, 2015.

[8] Y. Bar-Yosef, "Component Analysis of Mixture Models - Theory and Applications" Ph.D. dissertation, Tel Aviv University, Tel Aviv, Israel, 2018.

[9] I. Jolliffe, *Principal component analysis*. Springer New York, 1986.

[10] M. E. Tipping, and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61(3) pp. 611–622, 1999.

[11] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol 7, pp. 179–188, 1936.

[12] N. Kumar, N. and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech Communication*, vol. 26(4), pp 283–297, 1998.

[13] Z. Ghahramani, G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," *Technical Report CRG-TR-96-1* , University of Toronto, 1996.

[14] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers", *Neural computation*, vol. 11(2), pp. 443–482, 1999.

[15] S. J. Prince, and J. H. Elder "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision, ICCV*, 2007.

[16] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems*, vol. 17, pp. 513–520. MIT Press, MA, 2005.

[17] J. Peltonen, J. Goldberger, and S. Kaski, "Fast discriminative component analysis for comparing examples," in *Neural Information Processing Systems, NIPS* , 2006.

[18] K. Torkkola, "Learning discriminative feature transforms to low dimensions in low dimensions," in proc. *Neural Information Processing Systems, NIPS* , 2001.

[19] J. M. Leiva-Murillo , and A. Artes-Rodriguez. "Linear dimensionality reduction with Gaussian mixture models," in *IAPR Workshop on Cognitive Information Processing*, Santorini, Greece, 2008.

[20] M. Thangavelu and R. Raich. "On linear dimension reduction for multiclass classification of Gaussian mixtures," in *the 2009 IEEE international workshop on Machine Learning for Signal Processing, MLSP*, 2009.

[21] X. Yang, K. Huang, J. Y. Goulermas, and R. Zhang, "Joint Learning of Unsupervised Dimensionality Reduction and Gaussian Mixture Model," in *Neural Processing Letters* , vol 45(3), pp. 791–806, 2017.

[22] Dognin, P. L., Hershey, J. R., Goel, V., and Olsen, P. A. (2009). "Refactoring acoustic models using variational expectation-maximization," in *Proceedings of INTERSPEECH* ISCA, 2009.