PLDA using Gaussian Restricted Boltzmann Machines with application to Speaker Verification

Themos Stafylakis[1,2]*, Patrick Kenny*[1]*, Mohammed Senoussaoui*[1,2]*, Pierre Dumouchel*[1,2]

[1]Centre de Recherche Informatique de Montréal (CRIM), Quebec, Canada
[2]École de Technologie Superieure (ÉTS), Montréal, Quebec, Canada
{themos.stafylakis, patrick.kenny, mohammed.senoussaoui, pierre.dumouchel}@crim.ca

Abstract

A novel approach to supervised dimensionality reduction is introduced, based on Gaussian Restricted Boltzmann Machines. The proposed model should be considered as the analogue of the probabilistic LDA, using undirected graphical models. The training algorithm of the model is presented while its close relation to the cosine distance is underlined. For the problem of speaker verification, we applied it to i-vectors and attained a significant improvement compared to the Fisher's Discriminant LDA projection using less than half of the number of eigenvectors required by LDA.

Index Terms: Speaker Recognition, Restricted Boltzmann Machines

1. Introduction

During the last several years, speaker recognition technologies are being increasingly dominated by the so-called i-vector representation of speech utterances, [1]. This representation enables us to map utterances of variable duration into a fixed dimensional subspace $\Re^d$, with $d \in [600, 800]$ being the typical range. The i-vectors originate from the Joint-Factor Analysis (JFA) decomposition of supervectors into speaker and channel factors, where the supervector is defined as the concatenated 60-dimensional mean values of a 2048-component GMM, [2]. The way JFA-derived speaker factors differ from i-vectors is the labeling of the utterances during training. It was shown that by treating each utterance as a separate class (i-vectors) instead of the each speaker (JFA), one may obtain a representation of the same or better performance, by treating them effectively as the actual features for classification, verification and clustering algorithms.

However, the fact that i-vectors extractors are trained without speaker-level labeling, indicates that further transformations should apply in order to increase their speaker discriminative capacity. In particular, it was shown that by projecting i-vectors onto a Linear Discriminative Analysis (LDA) basis, trained using representative enrollment data and speaker-labels to defined classes, the performance can be improved significantly. Further projections, like the Within-Class Covariance Normalization (WCCN), proved capable of reducing the Equal-Error Rate (EER), while the extended use of the so-called cosine distance (instead of the Euclidean or the unnormalized correlation) and length-normalization showed that the speaker discriminative strength of the i-vector representation is conveyed in their direction in $\Re^d$, rather than in their magnitude, [1], [3]. Finally, Gaussian and heavy-tailed probabilistic LDA (PLDA), performed either on i-vectors directly or on the LDA-projected length-normalized i-vectors, yield state-of-the-art speaker verification results, [2].

In this paper, we examine an alternative way to apply PLDA, that is based on the powerful and fledgeling framework of Restricted Boltzmann Machines (RBMs). An RBM is an undirected graphical model, with a specific type of sparse connectivity and is used either as a stand-alone model, or as the primary building block for deeper architectures, such as Deep Boltzmann Machines and Deep Belief Networks, [4]. It turn, like its counterpart in directed graphical model, RBM-based PLDA can either be used as a fully-probabilistic back-end model for threshold-free verification, or as a feature extractor that enhances the speaker-discriminative capacity, while reduces their dimensionality.

The rest of the paper is organized as follows. In Sect. 2, the basic theory of Gaussian RBMs is covered, along with its relation to Gaussian Markov Random Fields (GMRFs). In Sect. 3, the topology of the proposed PLDA model is introduced, together with the way to evaluate it for speaker recognition. In Sect. 4, the training algorithm for the proposed model is described in details. In Sect. 5, experimental results on NIST-2010 data are demonstrated and compared to other state-of-the-art methods, while conclusions and future work directions are discussed in Sect. 6.

2. Modeling via Gaussian Restricted Boltzmann Machines

2.1. General properties of RBMs

One may consider GRBMs are Gaussian MRFs having both visible and hidden nodes, and of a specific (or restricted) bipartite structure that allows no connections between nodes of the same layer (see 1(a)). The visible layer is the one which holds the observations $\mathbf{v} = \{v_i\}_{i=1,\dots,d_\mathbf{v}}$ while the hidden layer holds the latent variables, or missing data $\mathbf{h} = \{h_j\}_{j=1,\dots,d_\mathbf{h}}$. The two layers are connected though the connectivity matrix $W = \{w_{i,j}\}$ and each node represents a univariate Gaussian, with mean and standard deviation denoted by μ and σ respectively. Let the model parameters be denoted by $\Theta = \{W, \{\mu_i\}, \{\mu_j\}, \{\sigma_i\}, \{\sigma_j\}\}$.

Viewing GRBMs as generative models, what differentiates them from multivariate normal (MVN) distributions is the indirect way they capture correlations between nodes in $\mathbf{v}$. They do so not by placing edges between these nodes - like the MVN distributions does - but rather by integrating out a set of the hidden variables $\mathbf{h}$. The resulting marginal distribution $P(\mathbf{v};\Theta) = \int_{\mathbf{h}\in\Re^{d_h}} P(\mathbf{v},\mathbf{h};\Theta)d\mathbf{h}$ is a multivariate normal distribution with non-zero correlation. We emphasize though that the two models, despite of having the same functional form for

the pdf of $\mathbf{v}$, they express different underlying models.
Furthermore, the absence of connections between nodes of the same layer makes the estimation of Θ very efficient, because it allows blocked-Gibbs sampling to be applied, i.e. variables of the same layer can be sampled simultaneously. A variant of the Gibbs sampler is the contrastive divergence algorithm that is commonly used to train RBMs, and is the one we use in this paper as well, [5].
The joint pdf of $(\mathbf{v}, \mathbf{h})$ given Θ is a $(d_{\mathbf{v}} + d_{\mathbf{h}})$-dimensional Gaussian, $P(\mathbf{v}, \mathbf{h}; \Theta) = Z(\Theta)^{-1} \exp(-E(\mathbf{v}, \mathbf{h}))$, where $Z(\Theta)$ the partition function and $E(\mathbf{v}, \mathbf{h})$ the energy function given below

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i \in vis} \frac{v_i^2}{2\sigma_i^2} + \sum_{j \in hid} \frac{h_j^2}{2\sigma_j^2} - \sum_{i,j} \frac{v_i}{\sigma_i} \frac{h_j}{\sigma_j} w_{ij} \quad (1)$$

where zero means are assumed.The precision (i.e. inverse covariance) matrix $\mathbf{P}$ is as follows

$$\mathbf{P} = \Sigma_{diag}^{-1/2} (I - \mathbf{W}) \Sigma_{diag}^{-1/2} \quad (2)$$

where Σ_{diag} the diagonal covariance matrix of $(\mathbf{v}, \mathbf{h})$ with $\{\sigma_i^2\}$ and $\{\sigma_j^2\}$ as entries, while

$$\mathbf{W} = \begin{pmatrix} \mathbf{0} & W \\ W^T & \mathbf{0} \end{pmatrix}$$

the connectivity matrix into its joint form. Note that the zero entries of the diagonal blocks are due to the absence of connections between nodes of the same layer.

2.2. Parameter estimation

The gradient of logarithmic marginal likelihood of $\{\mathbf{v}^{(k)}\}_{k=1}^{n_s}$ given W is as follows

$$\frac{\partial}{\partial W} \log P\left(\{\mathbf{v}^{(k)}\}_{k=1}^{n_s}; W\right) = \mathbb{E}_{P_D}[\mathbf{v}\mathbf{h}^T] - \mathbb{E}_{P_M}[\mathbf{v}\mathbf{h}^T] \quad (3)$$

The term $\mathbb{E}_{P_D}[\mathbf{v}\mathbf{h}^T]$ denotes the expectation w.r.t. the data distribution, i.e. $P_D(\mathbf{v}, \mathbf{h}; W) = P(\mathbf{h}; \mathbf{v}, W) P_D(\mathbf{v})$ and $P_D(\mathbf{v}) = \frac{1}{n} \sum_{k=1}^{n} \delta(\mathbf{v} - \mathbf{v}^{(k)})$ the empirical distribution of $\mathbf{v}$. In RBM, this term is calculated exactly, in our case by the positive phase of CD-1, but in general BM architectures (e.g. DBM), mean-field approximation is usually applied. The term $\mathbb{E}_{P_M}[\mathbf{v}\mathbf{h}^T]$, that is derived by differentiating the log-partition function w.r.t. W, is approximated by the negative phase of CD-1. It should be seen as a transformation procedure, where the current estimate of the model parameters is transformed from the *natural* parametrization (precision matrix) to the *expectation* parametrization (covariance matrix), without having to apply matrix inversion.
Let W^t be the current estimate of W. Due to the stochastic nature of the optimization algorithm, a momentum term is added, with $0 \leq c_m < 1$ coefficient, so that the trajectory of the sequence $\{W^t\}_{t=0,1,\ldots}$ becomes smoother. Finally, an L_2 regularization term is also added to prevent W from overfitting the data. Adding this term is crucial, because the matrix $I - WW^T$ should be invertible in order to have positive definite covariance matrix. Therefore, the implied objective function becomes the log-marginal likelihood of $\{\mathbf{v}^{(k)}\}_{k=1}^{n}$ minus $c_{sm} tr\left(WW^T\right)$, where $tr(\cdot)$ denotes trace, while $c_{L_2} \geq 0$ the L_2 smoothing coefficient. By including those terms, the overall updating rule becomes as follows

$$\Delta W^{t+1} = c_m \Delta W^t + c_u \left(\delta W^{t+1} - c_{L_2} W^t\right) \quad (4)$$

where $\Delta W^{t+1} = W^{t+1} - W^t$ and δW^{t+1} is given in (3).

3. Proposed model for PLDA using Gaussian RBMs

In this section, the proposed model is demonstrated, together with its log-likelihood ratio that will be used for verification. The topology and its motivation will be analyzed, while the training algorithm is presented of a separate section.

3.1. The topology for Gaussian PLDA

We now consider how to make this model suited to apply PLDA, [2], [6]. First, we should distinguish the hidden layer into speaker ($\mathbf{s}$) and channel factors ($\mathbf{c}$), i.e. $\mathbf{h}^T = [\mathbf{s}^T, \mathbf{c}^T]$, and partition the connectivity matrix accordingly, i.e. $W = [W_{\mathbf{s}}, W_{\mathbf{c}}]$. Note that we will be using s and $\mathbf{s}$ to refer to variables or model parameters that are dependent and independent of the training or testing speaker, respectively. By fixing means and standard deviations equal to zero and one respectively, the aim of the algorithm is to train only the connectivity matrix W. During training, the proposed model has a variable number of nodes, depending of the number of i-vectors available for each speaker, $n_s, s = 1, 2, \ldots$. The proposed model is illustrated in Fig. 1(b). From now on, we will be using the term sublayer so that we refer to sets of nodes that lie on the same layer and form a distinct entity. Therefore, the i-vectors are the sublayers of the visible layer, while the layers are two, i.e. visible and hidden.

As in the PLDA with directed graphical models, each i-

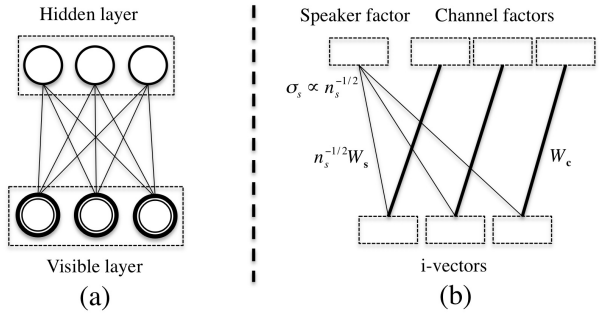

Figure 1: (a) *General structure of a Restricted Boltzmann Machine.* (b) *The proposed PLDA model, for a speaker with $n_s = 3$ i-vectors. For illustration purposes, sublayers are depicted using rectangles and W with single lines. Note that the standard deviation at the speaker factor sublayer is $\propto n_s^{-1/2}$.*

vector has its corresponding channel factor $\{\mathbf{c}^{(k)}\}_{k=1}^{n_s}$, connected via $W_{\mathbf{c}}$. However, speaker factors should be treated in a slightly different way than the one in the directed-model graph. As we depict in Fig. 1(b) the model has a unique speaker factor sublayer with $\sigma_s = n_s^{-1/2} \sigma_{\mathbf{s}}$ and connectivity matrix $W_s = n_s^{-1/2} W_{\mathbf{s}}$ and moreover we set $\sigma_{\mathbf{s}} = 1$. The rationale for using this topology is the following. Ignoring channel factors, we simultaneously like (a) the conditional of $\mathbf{s}|\{\mathbf{v}^{(k)}\}_{k=1}^{n_s}$ to be centered at $n_s^{-1} W_{\mathbf{s}}^T \sum_{k=1}^{n_s} \mathbf{v}^{(k)}$, i.e. at the mean value of the sth speaker, on the speaker factor subspace spanned by the columns of $W_{\mathbf{s}}$, while (b) the conditionals of $\mathbf{v}^{(k)}|\mathbf{s}$ to be centered at $\mathbf{s}$. Applying the rule of conditioning, for (a) we obtain, $\mathbf{s} n_s^{1/2} \leftarrow {n_s}^{-1/2} W_{\mathbf{s}}^T \sum_{k=1}^{n_s} \mathbf{v}^{(k)}$, while for (b) we obtain $\mathbf{v}^{(k)} \leftarrow {n_s}^{-1/2} W_{\mathbf{s}} \mathbf{s} {n_s}^{1/2}$ and therefore both requirements are met. Moreover, the fact that the variance of $\mathbf{s}|\{\mathbf{v}^{(k)}\}_{k=1}^{n_s}$ is shrinking by a factor of n_s reflects the variable uncertainty in the estimate of the center of each speaker.

3.2. Evaluation phase

We now derive the log-likelihood ratio (LLR) between the two hypotheses that is useful for speaker verification, i.e. given two i-vectors $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$, the log-likelihood of the target H_1 minus the log-likelihood of non-target H_0 hypothesis. To evaluate the likelihood of $\{\mathbf{v}^{(k)}k\}_{k\in\{1,2\}}$ for each of the two models, one should apply the expectation operator to minus the energy function of the model w.r.t. $P(\mathbf{h};\mathbf{v},\Theta) = \prod_{j\in hid} P(h_j;\mathbf{v},\Theta)$, and subtract $\log Z(\Theta)$. The derived expression is partitioned into data-dependent and data-independent terms. In this paper, we will ignore the data-independent terms, since their difference can be absorbed by the user-defined threshold.

Let us consider the projections $\mathbf{y}^{(k)} = W_{\mathbf{s}}^T\mathbf{v}^{(k)}$, $k \in \{1,2\}$. Note that these projections are the expected values of the speaker factors, under H_0, while their mean corresponds to the speaker factor under H_1. Using this fact, the following results can be derived

$$LLR(\mathbf{v}^{(1)},\mathbf{v}^{(2)};\Theta) \equiv -\frac{1}{2}(\mathbf{y}^{(1)}-\mathbf{y}^{(2)})^T(\mathbf{y}^{(1)}-\mathbf{y}^{(2)}) \quad (5)$$

where "$\equiv$" denotes equality up to a fixed additive constant. Hence, the LLR is simply half the negative squared Euclidean distance between the estimated speaker factors under H_0, plus some terms that are independent of $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$. Note also that the channel factors are being canceled out when forming LLR, and therefore their only role is to increase the speaker discriminative capacity of $W_{\mathbf{s}}$. Finally, to underline the relation with the cosine distance, consider the case where $\mathbf{s}^{(k)}$ are constrained to lie on a $d_{\mathbf{s}}$-sphere. By making such an assumption, LLR becomes the dot product $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, where $\mathbf{y}^{(k)} = \frac{W_{\mathbf{s}}^T\mathbf{v}^{(k)}}{\|W_{\mathbf{s}}^T\mathbf{v}^{(k)}\|}$, and therefore equivalent to the negative cosine distance, [1].

4. Algorithmic description

In this section, we explain the algorithm we used in order to train the model. The algorithm is an extension of the single-step Contrastive Divergence (CD-1) to fully Gaussian RBMs with tied nodes. The notation used refers to the model in Fig. 1(a), i.e. the one with a single sublayer and $\sigma_s = n_s^{-1/2}\sigma_{\mathbf{s}}$. Recall that in our case, $\sigma_{\mathbf{v}} = \sigma_{\mathbf{s}} = \sigma_{\mathbf{c}} = 1$.

4.1. Sampling phase

Assume we have a labeled training set of N_s speakers, denoted by $\mathbf{X} = \{X_s\}_{s=1}^{N_s}$. Let $X_s = \{\mathbf{v}^{(k)}\}_{k=1}^{n_s}$ be the set of n_s d-dimensional *prewhitened* i-vectors that belong to the sth speaker. The biases are fixed to zero and are being omitted from the expressions while the standard deviations are set equal to one.

The first step is the calculation of the conditional of the hidden layers, given the n_s visible layers to the hidden layer. Therefore, the expected values of $\{\mathbf{s},\mathbf{c}^{(k)}\}_{k=1}^{n_s}$ are obtained as follows

$$\mathbf{s}_p \leftarrow \sigma_{\mathbf{s}} W_{\mathbf{s}}^T\frac{\bar{\mathbf{v}}_p}{\sigma_{\mathbf{v}}}, \qquad \mathbf{c}_p^{(k)} \leftarrow \sigma_{\mathbf{c}} W_{\mathbf{c}}^T\frac{\mathbf{v}_p^{(k)}}{\sigma_{\mathbf{v}}} \quad (6)$$

In the above notation, $\bar{\mathbf{v}}_p = n_s^{-1}\sum_{k=1}^{n_s}\mathbf{v}_p^{(k)}$, while the subscript p is used to denote quantities of the positive phase of CD-1, during which the visible layer holds the actual i-vectors. Note that $\mathbf{s}_p$ defines the expectation of n_s visible vectors projected onto $W_{\mathbf{s}}$. This means that the variance in the estimate of speaker factors is n_s times smaller than the variance of each of the n_s channel factors.

The negative phase of CD-1 begins by sampling a d_s-dimensional uncorrelated Gaussian distribution. Recall that the Gaussian retains its diagonal-covariance form, due to the conditioning on $\mathbf{v}_p$ and the RBM structure of the model, which makes blocked-Gibbs sampling feasible. We sample the Gaussians as follows

$$\mathbf{s}_{st} \sim \mathcal{N}\left(\mathbf{s}_p, n_s^{-1}{\sigma_{\mathbf{s}}}^2 I\right),\ \mathbf{c}_{st}^{(k)} \sim \mathcal{N}\left(\mathbf{c}_p^{(k)}, {\sigma_{\mathbf{c}}}^2 I\right) \quad (7)$$

where the subscript denotes *state*. The values of the factors are then propagated to the network, so that they define the expected value of the visible layer, i.e.

$$\mathbf{v}_n^{(k)} \leftarrow \sigma_{\mathbf{v}}\left[W_{\mathbf{c}}\frac{\mathbf{c}_{st}^{(k)}}{\sigma_{\mathbf{c}}} + W_{\mathbf{s}}\frac{\mathbf{s}_{st}}{\sigma_{\mathbf{s}}}\right] \quad (8)$$

Based on the arguments discussed in [7], we do not resample the visible layer, but we set its variables equal to their expected values given in (8). The final step required is the calculation of the conditional expectation of $\{\mathbf{s},\mathbf{c}^{(k)}\}_{k=1}^{n_s}$ given $\{\mathbf{v}_n^{(k)}\}_{k=1}^{n_s}$, using the following formula

$$\mathbf{s}_n \leftarrow \sigma_{\mathbf{s}} W_{\mathbf{s}}^T\frac{\bar{\mathbf{v}}_n}{\sigma_{\mathbf{v}}}, \qquad \mathbf{c}_n^{(k)} \leftarrow \sigma_{\mathbf{c}} W_{\mathbf{c}}^T\frac{\mathbf{v}_n^{(k)}}{\sigma_{\mathbf{v}}} \quad (9)$$

which completes CD-1 sampling phase. Note that in order to check convergence, the reconstruction squared error for each epoch, i.e. $\sum_{s=1}^{N_s}\sum_{k=1}^{n_s}\left(\mathbf{v}_p^{s,k}-\mathbf{v}_n^{s,k}\right)^T\left(\mathbf{v}_p^{s,k}-\mathbf{v}_n^{s,k}\right)$ is the most commonly used diagnostic, [7].

4.2. Parameter update phase

Once the sth minibatch is completed, the connectivity matrix $W = [W_{\mathbf{s}}, W_{\mathbf{c}}]$ is updated. The updating rules can be derived using the general RBM formula in (3) and (4). Excluding momentum and regularization terms, the updating formulas are as follows

$$\delta W_{\mathbf{c}} = \frac{1}{\sigma_{\mathbf{v}}\sigma_{\mathbf{c}}}\sum_{k=1}^{n_s}\mathbf{v}_p^{(k)}{\mathbf{c}_p^{(k)}}^T - \mathbf{v}_n^{(k)}{\mathbf{c}_n^{(k)}}^T \quad (10)$$

and

$$\delta W_{\mathbf{s}} = \frac{n_s}{\sigma_{\mathbf{v}}\sigma_{\mathbf{s}}}\left(\bar{\mathbf{v}}_p\bar{\mathbf{s}}_p^T - \bar{\mathbf{v}}_n\bar{\mathbf{s}}_n^T\right) \quad (11)$$

Note the linear increase in the average magnitude of both $\delta W_{\mathbf{c}}$ and $\delta W_{\mathbf{s}}$ with n_s. This property is analogous to the weighting by n_s of each class-covariance when estimating the within-class covariance matrix in standard LDA. The complete updating expressions can be reached by adding the momentum and smoothing terms given in (4).

5. Experimental results

We performed experiments on the *coreext-coreext* condition of the telephone speech NIST extended list. We focus on female data only, where the state-of-the-art performance is worst than the one on male data. We use the Equal Error Rate (EER) and the (new and old) normalized minimum Detection Cost Function 2 (DCF) of NIST as metrics.

5.1. Enrollment data, i-vectors and tuning parameters

5.1.1. Universal Background Model

We use a gender-independent GMM UBM containing 2048 Gaussians. This UBM is trained with the LDC releases of

Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and NIST 2004 and 2005 SRE. Speech parameters are represented by a 60-dimensional vector of Mel Frequency Cepstral Coefficients (MFCC) i.e. static MFCC, first and second derivative of MFCC.

5.1.2. i-vector extractor

We use a gender independent i-vector extractor of dimension 800. Its parameters are estimated on the following data: LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; Fisher data and NIST 2004 and 2005 SRE (i.e. telephone speech) and all NIST microphone data (i.e. NIST 05, 06 and 08 interview development microphone data).

5.1.3. RBM-PLDA training

We used the female portion of the above dataset to train our model, namely $N_s = 1682$ speakers and $n = 20003$ i-vectors. We set momentum, L_2 regularization and learning rate equal to $c_m = 0.5$, $c_{L_2} = 0.10$ and $c_u = 10^{-4}$, respectively. Based on the reconstruction error, the maximum number of epochs was found experimentally to be around $N_e = 200$ (denoted by RBM_2). However, in order to avoid overfitting, we tried the early stopping technique, in which case we stopped training after $N_e = 80$ epochs (denoted by RBM_1). Finally, no score normalization technique is applied to any of the systems.

5.2. Results on NIST-2010 data

We compare our algorithm with the traditional LDA, as well as with LDA followed by WCCN, with 200 eigenvectors. The scoring method is the cosine distance, which as explained in Sect. 3, is the LLR in case where the speaker factors under H_0 are forced to lie on the unit $d_{\mathbf{s}}$-sphere. The results and the figure are generated using the BOSARIS toolkit, [8].

Table 1: *Results on NIST-10 female tel. data (core condition).*

metric	*normalization*	LDA	RBM_1	RBM_2
ERR(%)	*none*	5.29	**3.38**	3.58
”	WCCN	3.45	**2.75**	2.85
micDCF_{old}	*none*	0.43	0.31	**0.30**
”	WCCN	0.33	0.28	**0.27**
micDCF_{new}	*none*	0.64	0.46	**0.43**
”	WCCN	0.49	0.46	**0.43**

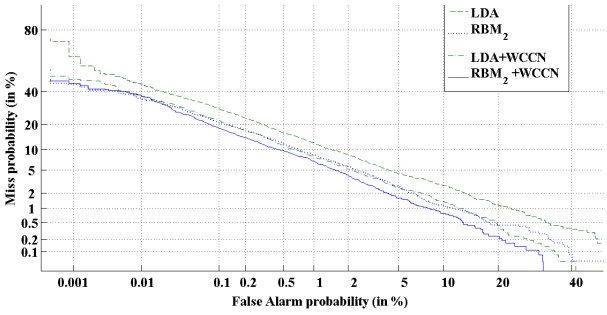

Figure 2: *Comparison of the proposed system to standard LDA using DET curves*

As the results demonstrate in Table 1 and Fig. 2, in all metrics the gain for using RBM-based PLDA is significant. In terms of EER, the results show that the proposed method performs better than the standard LDA+WCCN, even without applying WCCN. Moreover, early stopping is useful in terms of ERR, with a slight degradation in the minDCF metrics. In any case, the RBM-based PLDA seems to have a clear advantage over its traditional counterpart, which becomes more clear when considering their difference in dimensionality reduction (80 for the proposed vs. 200 for the traditional one).
Regarding the familiar PLDA, a direct comparison is not possible, since the reported results are based on a severely different feature space, namely on 200-dimensional, LDA-projected and length-normalized i-vectors, [8]. However, the ERR$= 2.47\%$ it attains is comparable to our result, which has been attained with using only WCCN. In any case, we are planning the examine our model using their configuration in order to make this comparison feasible.

6. Conclusions and future work

In this paper, we utilized Gaussian RBMs in order to perform PLDA on i-vectors. We proposed a topology that is capable of learning speaker and channel factor subspaces and yielding results that are comparable to state-of-the-art methods.
As feature work directions, we suggest the use of directional statistics in order to model the distribution of speaker factors, with the von Mises distribution being the most profound candidate, due to its exponential family membership. Moreover, we should examine the capacity of the model in making threshold-free decisions, which makes it an alternative back-end module for verification systems. The model, though, can still be used as a feature extractor, and be combined with several other proposed preprocessing operators, in a step-by-step threshold-based verification system.

7. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, ”Front-End Factor Analysis for Speaker Verification”, in *IEEE Transactions on Audio, Speech & Language Processing*, 2011.

[2] P. Kenny, Bayesian Speaker Verification with Heavy-Tailed Priors in *Proc. Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[3] D. Garcia-Romero and C. Y. Espy-Wilso, Analysis of i-vector length normalization in speaker recognition systems, in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.

[4] R. Salakhutdinov, ”Learning Deep Generative Models”, PhD thesis, University of Toronto, 2009.

[5] G. E. Hinton. ”Training products of experts by minimizing contrastive divergence”. in *Neural Computation*, 14(8):17111800, 2002.

[6] S. J. D. Prince and J. H. Elder, Probabilistic linear discriminant analysis for inferences about identity”, in *Proc. ICCV 2007*, Rio de Janeiro, Brazil, Oct. 2007.

[7] G. E. Hinton, ”A Practical Guide to Training Restricted Boltzmann Machines”, *Technical Report 2010003*, Department of Computer Science, University of Toronto, 2010.

[8] E. de Villiers and N. Brummer, BOSARIS toolkit, https://sites.google.com/site/bosaristoolkit/.

[8] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, P. Dumouchel: Mixture of PLDA models in I-vector space for gender independent speaker recognition, In *Proc. of Interspeech 2011*, Florence, Italy.