Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Constrained discriminative speaker verification specific to normalized i-vectors

Pierre-Michel Bousquet, Jean-François Bonastre

University of Avignon - LIA, France
{pierre-michel.bousquet, jean-francois.bonastre}@univ-avignon.fr

Abstract

This paper focuses on discriminative trainings (DT) applied to i-vectors after Gaussian probabilistic linear discriminant analysis (PLDA). If DT has been successfully used with non-normalized vectors, this technique struggles to improve speaker detection when i-vectors have been first normalized, whereas the latter option has proven to achieve best performance in speaker verification. We propose an additional normalization procedure which limits the amount of coefficient to discriminatively train, with a minimal loss of accuracy. Adaptations of logistic regression based-DT to this new configuration are proposed, then we introduce a discriminative classifier for speaker verification which is a novelty in the field.

1. Introduction

Once the i-vector paradigm of low rank total variability factor has been introduced in the field of speaker recognition [1], Gaussian modelings of i-vector have been worked out [2, 3], intended to estimate the latent speaker variable. In the case of single-cut enrollment trial, the formulation of the speaker verification log-likelihood ratio reduces to a quadratic function of the i-vector pair of a trial, that is, a second degree polynomial of i-vector components. This functional form led to attempts to optimize parameters of the model by using a discriminative approach. Discriminative trainings for speaker verification rely on minimization of non-linear objective functions [4, 5, 6].

First Gaussian modeling provided disappointing results. Application of normalization pre-procedures (within-class covariance and length [1, 7, 8]) to i-vectors allows Gaussian modeling to be competitive versus more complicated systems, as heavy-tailed PLDA [2], in terms of performance.

In [9], the same performance than a non-normalized and optimal PLDA system is achieved by using a SVM based-discriminative system, without the need to tune the rank of the PLDA speaker variability subspace. But DT struggles to improve accuracy of detection with normalized vectors. Normalization techniques are known to improve Gaussianity of i-vector distribution [7], and seem to limit the impact of an additional DT step. [10] remarks that DT may suffer from training data insufficiency. Thus, constrained DT systems have been proposed (of order the i-vector dimension, instead of its square), in order to limit the amount of parameters to optimize with respect to the total amount of training data [10, 11]. But it appeared that effectiveness of DT approach may also suffer from over-fitting to development data. Very low order discriminative trainings have been proposed, which only optimize parts of score [10], or matrix scalings of PLDA matrices [11]. Also, constraints on coefficients trained from DT do have to be taken into account, in order to respect properties of PLDA covariance matrices (definiteness, positivity/negativity). Finally, [10] points out that statistical independence between training trials used for DT is questionable, as the need of a large amount of data for DT training obliges to use same speaker and utterances in more than one training trial.

We propose to train a small order DT by taking advantage of normalized i-vector properties. A simple additional normalization procedure is presented, which does not modify distances between i-vectors and allows to train low order DT classifiers with a minimal loss of accuracy. We adapt two state-of-the-art DT based on logistic regression (LR) to this new configuration, the first is optimizing score coefficients, the second one PLDA parameters.

However, the ability of LR based-discriminative classifiers to improve Gaussian PLDA system by using cross entropy minimization, based itself on Gaussian assumptions, may be questioned, as normalization techniques are known to improve Gaussianity. In order to overcome this issue, a new specific discriminative classifier is proposed, based on covariance of target and non-target trials when they are represented by expanded vectors of score.

This paper is organized as follows. Sec. 2 summarizes the Gaussian generative model for i-vector. State-of-the-art discriminative trainings of PLDA hyperparameters are recalled in Sec. 3. In Sec. 4, the additional normalization procedure intended to simplify discriminative training is presented, and Sec. 5 proposes discriminative classifiers specific to normalized i-vectors. Experimental results are reported in Sec. 6, and conclusions are provided in Sec. 7.

2. Gaussian generative model for i-vector

The total variability factor approach for speaker recognition [1] provides a representation of speech segments by low dimensional feature vectors (commonly less than 600), independent of the length of the utterance, referred to as i-vector. This approach defers the problem of intersession variability to a second stage. To facilitate accurate comparison in a verification trial, i-vector modelings have been proposed, assuming additive decomposition of speaker and noise components. We focus here on the most commonly used Gaussian model in speaker verification [2]. Introduced in [12, 13], Gaussian PLDA (G-PLDA) is a generative model where latent vector $\mathbf{y}_s$ representing speaker s is assumed to be distributed according to standard normal prior and a d-dimensional vector $\mathbf{w}$ can be decomposed as follows:

$$\mathbf{w} = \mu + \mathbf{\Phi}\mathbf{y}_s + \varepsilon \qquad (1)$$

where $\mathbf{\Phi}\mathbf{y}_s$ and ε are assumed to be statistically independent and ε follows a centered Gaussian distribution with full covariance matrix $\mathbf{\Lambda}$. Speaker factor $\mathbf{y}_s$ can be a full-rank d-vector (this model is referred to as *two-covariance model* [3]) or con-

doi: 10.21437/Odyssey.2016-8

strained to lie in the r-linear range of the $d \times r$ matrix $\mathbf{\Phi}$, referred to as *eigenvoice subspace* [2].

Using a Bayesian approach, the goal of evaluating hypotheses $\mathcal{H}_1$ that two i-vectors $\mathbf{w}_i$, $\mathbf{w}_j$ are produced by the same source and $\mathcal{H}_0$ that they are produced by different sources reduces to estimating the log-likelihood ratio score (LLR):

$$\begin{aligned} s_{i,j} &= \log \frac{P(\mathbf{w}_i, \mathbf{w}_j | \mathcal{H}_1)}{P(\mathbf{w}_i, \mathbf{w}_j | \mathcal{H}_0)} \\ &= \log \frac{\int_y \prod_{l=i,j} P(\mathbf{w}_l | \mu + \mathbf{\Phi} y, \mathbf{\Lambda}) P(y | \mathbf{0}, \mathbf{I})\, dy}{\prod_{l=i,j} \int_y P(\mathbf{w}_l | \mu + \mathbf{\Phi} y, \mathbf{\Lambda}) P(y | \mathbf{0}, \mathbf{I})\, dy} \end{aligned} \quad (2)$$

For the Gaussian case, the closed-form solution of (2) is the following second degree polynomial function of $\mathbf{w}_i$ and $\mathbf{w}_j$:

$$\begin{aligned} s_{i,j} &= \mathbf{w}_i^t \mathcal{P} \mathbf{w}_j + \frac{1}{2}\left(\mathbf{w}_i^t \mathcal{Q} \mathbf{w}_i + \mathbf{w}_j^t \mathcal{Q} \mathbf{w}_j\right) \\ &\quad - \mu^t (\mathcal{P} + \mathcal{Q})(\mathbf{w}_i + \mathbf{w}_j) \\ &\quad + \mu^t (\mathcal{P} + \mathcal{Q}) \mu + \frac{1}{2} \log |\mathbf{A}_t| - \log |\mathbf{A}_n| \end{aligned} \quad (3)$$

where

$$\begin{aligned} \mathcal{P} &= \mathbf{\Lambda}^{-1} \mathbf{\Phi} \left(2 \mathbf{\Phi}^t \mathbf{\Lambda}^{-1} \mathbf{\Phi} + \mathbf{I}_r\right)^{-1} \mathbf{\Phi}^t \mathbf{\Lambda}^{-1} \\ \mathcal{Q} &= \mathcal{P} - \mathbf{\Lambda}^{-1} \mathbf{\Phi} \left(\mathbf{\Phi}^t \mathbf{\Lambda}^{-1} \mathbf{\Phi} + \mathbf{I}_r\right)^{-1} \mathbf{\Phi}^t \mathbf{\Lambda}^{-1} \\ \mathbf{A}_t &= \left(2 \mathbf{\Phi}^t \mathbf{\Lambda}^{-1} \mathbf{\Phi} + \mathbf{I}_r\right)^{-1} \\ \mathbf{A}_n &= \left(\mathbf{\Phi}^t \mathbf{\Lambda}^{-1} \mathbf{\Phi} + \mathbf{I}_r\right)^{-1} \end{aligned} \quad (4)$$

and $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. It can be shown that this formulation is equivalent to those of [5, 6, 7] initially proposed in [13].

As remarked in [2], G-PLDA failed to produce accurate model for i-vectors. A normalization step, applied before any i-vector modeling, has been incorporated, including within-class covariance matrix $\mathbf{W}$ (centering and scaling) and length normalization [1, 7, 8]. These transformations involve some properties related to intersession compensation [14, 7, 8]. Performance of G-PLDA becomes competitive versus the heavy-tailed PLDA [2], when the latter shows a significant higher complexity.

3. Discriminative PLDA training

Generative Gaussian PLDA model provides the functional second degree polynomial of speaker verification score (3). Discriminative classifiers aim at enhancing the accuracy of speaker detection by optimizing parameters of this polynomial, according to an objective function to minimize. The total cross entropy (TCE) of a trial dataset is the log-probability of correctly classifying all trials of this dataset. When the trials are made up of all the pairs of a n-size training segment set, TCE can be written as [5]:

$$TCE = -\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{i,j} \log \sigma(t_{i,j} s_{i,j}) \quad (5)$$

where label $t_{i,j}$ is equal to 1 if $\mathbf{w}_i, \mathbf{w}_j$ are from the same speaker and -1 if they are from different speakers, σ denotes the sigmoid activation function and $\alpha_{i,j}$ is used to assign the weight $P(\mathcal{H}_1)/n_1$ (resp. $P(\mathcal{H}_0)/n_0$) to same-speaker (resp. different-speaker) trials, where $P(\mathcal{H}_1)$,$P(\mathcal{H}_0)$, n_1,n_0 are their prior and cardinality, respectively.

Stating TCE as a function $E(\omega)$ of a set of parameters ω, then minimizing this loss function, has led in [4, 5] to the case of logistic regression (LR) based DT for speaker verification. Using a *hinge* loss function has led to the SVM case [6, 5]. Minimization of the non-linear function $E(\omega)$ is done by gradient-descent. Parameters ω to optimize can be the coefficients of the LLR score polynomial, as proposed in [5], or the PLDA parameters $(\mu, \mathbf{\Phi}, \mathbf{\Lambda})$ of (1), as proposed in [11]. In the first case, the score can be written as a dot product of an *expanded vector* $\varphi_{i,j}$ stacking all the monomials of the second degree polynomial, and a vector stacking all its coefficients. The gradient of E is equal to:

$$\nabla E(\omega) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{i,j} (-t_{i,j}) \sigma(-t_{i,j} s_{i,j}) \frac{\partial s_{i,j}}{\partial \omega} \quad (6)$$

where, in the first case, $\frac{\partial s_{i,j}}{\partial \omega}$ just gives the expanded vector $\varphi_{i,j}$. DT uses the conjugate gradient trust region method [15] as implemented in [16]. Parameter ω is currently updated in the field by the following formula:

$$\omega^{(l+1)} = \omega^{(l)} - \frac{u^t . \nabla E\left(\omega^{(l)}\right)}{u^t \mathbf{H} u} \nabla E\left(\omega^{(l)}\right) \quad (7)$$

where $\mathbf{H}$ is the Hessian and u is estimated by using the Hestenes -Stiefel variant of line search along a direction.

Thus, DT for speaker verification reduces to minimizing an $O(d^2)$ parameter function. It is remarked in [10] that these methods can suffer from data insufficiency. Moreover, the large amount of parameters to estimate may cause over-fitting to development data, recalling limits of the approach. Therefore, constrained DT have been proposed. That is, DT training only a small amount of parameters, of order $O(d)$ or even $O(1)$. These approaches have provided interesting performance. In [10], a single coefficient is optimized for each dimension of the i-vector ($O(d)$ DT) or even the four *feature kinds* that make up score of (3) ($O(4)$ DT). In [11], mean vector μ and only eigenvalues of PLDA matrices $\mathbf{\Phi}\mathbf{\Phi}^t$ and $\mathbf{\Lambda}$ are trained by DT and, even, their scaling factors only.

On the other hand, parameters to estimate derive from covariance matrices of a probabilistic model. Matrices $\mathbf{\Phi}\mathbf{\Phi}^t$ and $\mathbf{\Lambda}$ are positive-(semi)definite, and it has been noticed [7] that matrices $\mathcal{P}$ and $\mathcal{Q}$ of (4) are positive-(semi)definite and negative-(semi)definite. These properties impose constraints during the DT-minimization phase, keeping in mind that objective functions as TCE rely on Gaussian modeling. Having to handle $\mathbf{\Phi}\mathbf{\Phi}^t$ and $\mathbf{\Lambda}$ eigenvalues, [11] includes a flooring of parameters. In [10], working with singular value decomposition of these matrices attempts to respect these conditions.

4. Additional normalization procedure

Once i-vectors have been normalized, their length is equal to 1, but it is also worth noting that their within-class covariance matrix $\mathbf{W}$ is almost exactly isotropic [14], i.e. $\mathbf{W} \approx \sigma \mathbf{I}$ where σ is a positive real and $\mathbf{I}$ is the $d \times d$ identity matrix. This entails properties which could be taken advantage of for discriminative classification.

We propose to add a supplementary step to the normalization procedure. I-vectors of training and test are rotated by the eigenvector basis of between-class covariance matrix $\mathbf{B}$ of the training dataset. By this way, the new between-class covariance matrix is diagonal, with its diagonal equal to its eigenvalue spectrum. Moreover, the new within-class covariance matrix remains almost isotropic (and therefore diagonal), as the eigenvector basis is orthogonal. It stems from these remarks that the new total covariance matrix is almost diagonal. We consider that metaparameters $\mathbf{\Phi\Phi}^t$, $\mathbf{\Lambda}$ of PLDA also become almost diagonal, and even isotropic for $\mathbf{\Lambda}$. As a consequence, $\mathcal{P}$ and $\mathcal{Q}$ of score (3) are almost diagonal. PLDA score of (3) can be rewritten as:

$$s_{i,j} = \sum_{k=1}^{d} \left\{ \begin{array}{l} p_k \mathbf{w}_{i,k}\mathbf{w}_{j,k} + \frac{1}{2}q_k\left(\mathbf{w}_{i,k}^2 + \mathbf{w}_{j,k}^2\right) \\ -\left(p_k + q_k\right)\mu_k\left(\mathbf{w}_{i,k} + \mathbf{w}_{j,k}\right) \end{array} \right\} + res_{i,j} \tag{8}$$

where $p \in \mathbb{R}^d$ (resp. q) denotes the diagonal of $\mathcal{P}$ (resp. $\mathcal{Q}$) and the residual term $res_{i,j}$ sums all the *off-diagonal* terms and constant offsets. Thus, we assume that the major proportion of variability in this score is contained into these d diagonal terms.

Moreover, if the eigenvector basis of $\mathbf{B}$ used for rotation is sorted in decreasing order of eigenvalues, the first r terms are approximately those of the r-dimensional eigenvoice subspace of PLDA. This assumption stems from the fact that these first r dimensions are approximately those of the r-rank subspace of deterministic LDA, since the LDA solution is the r-range of the first eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$, which is here approximately equal to $\mathbf{B}$. The score can be rewritten as:

$$s_{i,j} = \sum_{k=1}^{r} \left\{ \begin{array}{l} p_k \mathbf{w}_{i,k}\mathbf{w}_{j,k} + \frac{1}{2}q_k\left(\mathbf{w}_{i,k}^2 + \mathbf{w}_{j,k}^2\right) \\ -\left(p_k + q_k\right)\mu_k\left(\mathbf{w}_{i,k} + \mathbf{w}_{j,k}\right) \end{array} \right\} + res_{i,j}^{'} \tag{9}$$

where $res_{i,j}^{'}$ sums all the diagonal terms beyond the r^{th} dimension, all the off-diagonal terms and offsets.

Consequently, we assume that the major proportion of score variability is contained into the first r diagonal terms of (8). By this way, the $O(d^2)$ initial discriminative classifier can be replaced by a constrained $O(r+1)$ discriminative classifier with a minimal loss of accuracy. This classifier relies on the complete PLDA score (including all its terms) but focuses on its principal *components*. Section 6 includes an empirical analysis of PLDA parameters diagonality and isotropy, after the additional normalization procedure.

5. Discriminative classifiers specific to normalized i-vectors

5.1. With logistic regression

Both DT approaches for speaker verification recalled in Sec. 3 can be adapted to the previous score decomposition. First, let define the $\mathbb{R}^{3r+1}$ expanded vector $\varphi_{i,j}$ equal to:

$$\varphi_{i,j} = \begin{bmatrix} \mathbf{w}_i^{(r)} \circ \mathbf{w}_j^{(r)} \\ \mathbf{w}_i^{(r)} \circ \mathbf{w}_i^{(r)} + \mathbf{w}_j^{(r)} \circ \mathbf{w}_j^{(r)} \\ \mathbf{w}_i^{(r)} + \mathbf{w}_j^{(r)} \\ res_{i,j}^{'} \end{bmatrix} \tag{10}$$

where the superscript $^{(r)}$ indicates the first r components of a vector and the symbol $\circ$ denotes the element wise product. The score of (9) becomes:

$$s_{i,j} = \varphi_{i,j}^t . \begin{bmatrix} p^{(r)} \\ \frac{1}{2}q^{(r)} \\ -\left(p^{(r)} + q^{(r)}\right) \circ \mu^{(r)} \\ 1 \end{bmatrix} \tag{11}$$

Logistic regression based-DT can be performed by optimizing $\omega = \begin{bmatrix} p^{(r)} & \frac{1}{2}q^{(r)} & -\left(p^{(r)} + q^{(r)}\right) \circ \mu^{(r)} & 1 \end{bmatrix}^t$.

Second, a $(\mu, \mathbf{\Phi}, \mathbf{\Lambda})$ PLDA parameter based-DT can be performed. After $\mathbf{B}$-rotation, $\mathbf{\Phi\Phi}^t$ and $\mathbf{\Lambda}$ are close to be diagonal. We propose to approximate $\mathbf{\Phi\Phi}^t$ and $\mathbf{\Lambda}$ by their diagonal version. Let denote by δ, λ the diagonal of the $r \times r$-upper-left block of $\mathbf{\Phi\Phi}^t$, $\mathbf{\Lambda}$, respectively. In the score of (3) simplified in (9), p,q and offset become, respectively:

$$\begin{aligned} p_k &= \frac{\delta_k}{2\delta_k\lambda_k + \lambda_k^2} \\ q_k &= p_k - \frac{\delta_k}{\delta_k\lambda_k + \lambda_k^2} \\ \log \frac{\left|\mathbf{\Phi}^t\mathbf{\Lambda}^{-1}\mathbf{\Phi} + \mathbf{I}_r\right|}{\left|2\mathbf{\Phi}^t\mathbf{\Lambda}^{-1}\mathbf{\Phi} + \mathbf{I}_r\right|^{\frac{1}{2}}} &= \log\left(\frac{\delta_k}{\lambda_k} + 1\right) - \frac{1}{2}\log\left(\frac{2\delta_k}{\lambda_k} + 1\right) \end{aligned} \tag{12}$$

DT is performed by training $\omega = [\delta, \lambda, \mu]^t$. In order to preserve isotropy of the channel component, $\mathbf{\Lambda}$ is further constrained by only updating matrix scaling (i.e. $\lambda \rightarrow \omega_\lambda \lambda$, ω_λ real), so that $\omega = \begin{bmatrix} \delta & \omega_\lambda & \mu \end{bmatrix}^t \in \mathbb{R}^{2d+1}$. As $s_{i,j}$ is no longer a linear function of ω, computation of the Hessian $\mathbf{H}$ is more complicated. The second order derivative $\frac{\partial^2 E(\omega)}{\partial\omega_k \partial\omega_{k'}}$ of $E(\omega)$, for $1 < k, k' < 2d+1$, is equal to:

$$\mathbf{H}_{k,k'} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{i,j}\frac{\partial s_{i,j}}{\partial \omega_k} \times \frac{\partial s_{i,j}}{\partial \omega_{k'}} + b_{i,j}\frac{\partial}{\partial \omega_k}\left(\frac{\partial s_{i,j}}{\partial \omega_{k'}}\right)\right) \tag{13}$$

where $a_{i,j} = \alpha_{i,j}\sigma\left(-t_{i,j}s_{i,j}\right)\left(1 - \sigma\left(-t_{i,j}s_{i,j}\right)\right)$ and $b_{i,j} = \alpha_{i,j}\left(-t_{i,j}\right)\sigma\left(-t_{i,j}s_{i,j}\right)$.

As $\mathbf{\Phi\Phi}^t$ and $\mathbf{\Lambda}$ are close to be diagonal, it can be noticed that δ and λ are close to their eigenvalues spectra. Thus, this DT method is those proposed in [11] applied to a simplified modeling.

5.2. Orthonormal discriminative classifier

Defining the expanded $\mathbb{R}^{r+1}$ vector $\varphi_{i,j}$ of a trial $(\mathbf{w}_i, \mathbf{w}_j)$ by:

$$\varphi_{i,j} = \begin{bmatrix} \left\{ \begin{array}{c} p^{(r)} \circ \mathbf{w}_i^{(r)} \circ \mathbf{w}_j^{(r)} \\ +\frac{1}{2}q^{(r)} \circ \left(\mathbf{w}_i^{(r)} \circ \mathbf{w}_i^{(r)} + \mathbf{w}_j^{(r)} \circ \mathbf{w}_j^{(r)}\right) \\ -\mu^{(r)} \circ \left(p^{(r)} + q^{(r)}\right) \circ \left(\mathbf{w}_i^{(r)} + \mathbf{w}_j^{(r)}\right) \end{array} \right\} \\ res_{i,j}^{'} \end{bmatrix} \tag{14}$$

score of (9) can be written as $s_{i,j} = \varphi_{i,j}^t.\mathbf{1}_{r+1}$, where $\mathbf{1}_{r+1}$ is the $\mathbb{R}^{r+1}$ vector of ones. As formulated, PLDA score is geometrically the projection of an expanded vector $\varphi_{i,j}$ onto the axis $\mathbf{1}_{r+1}$, i.e. onto the normal vectorial line of a separation hyperplane. Considering this formulation, speaker detection measurement of PLDA score can be reinterpreted. PLDA claims that vector $\mathbf{1}_{r+1}$ is, statistically, the best speaker discriminant axis (the more discriminant and the more generalizable) in this expanded vector space. That is, that all other axes of this space are not speaker-discriminant, or too correlated to the latter to improve accuracy of detection. We assume that useful information could be extracted from the complementary of the normal vector $\mathbf{1}_{r+1}$.

Let denote by $(\alpha_t, g_t, \mathcal{W}_t)$ and $(\alpha_n, g_n, \mathcal{W}_n)$ the order 0, 1 and 2 statistics (prior, mean and covariance) of a target and non-target trial expanded vector dataset, respectively. Mean value of this set is equal to $\alpha_t g_t + \alpha_n g_n$, within-class covariance matrix is equal to $\mathcal{W} = \alpha_t \mathcal{W}_t + \alpha_n \mathcal{W}_n$ and a straightforward computation shows that its between-class covariance matrix is equal to $\mathcal{B} = \alpha_t \alpha_n (g_t - g_n)(g_t - g_n)^t$. In the case of a two-class classifier, Fisher's linear discriminant [17] extracts a discriminant axis u by maximizing the following Fisher criterion:

$$\frac{u^t \mathcal{B} u}{u^t \mathcal{W} u} \tag{15}$$

This maximization problem has a closed-form solution, given as the eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ corresponding to the largest eigenvalue. It can be shown that the eigenvector u and eigenvalue λ provided by this method are equal to:

$$\begin{aligned} u &= \mathcal{W}^{-1}(g_t - g_n) \\ \lambda &= \frac{u^t \mathcal{B} u}{u^t \mathcal{W} u} = \alpha_t \alpha_n (g_t - g_n)^t \mathcal{W}^{-1}(g_t - g_n) \end{aligned} \tag{16}$$

As noticed in [17], the projected data onto the vectorial line of u can be used to construct a discriminant score.

As the rank of $\mathcal{W}^{-1}\mathcal{B}$ is equal to 1, only one axis can be found out by this method. The resulting score $\varphi_{i,j}^t.u$ would yield a poor performance. In [18], a method is proposed to extract more axes than classes, whilst using the Fisher criterion. We refer to this method as "Orthonormal Discriminative (OD) classifier". Once the unique eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ (i.e. the only one corresponding to a non-null eigenvalue) has been extracted, data are projected onto its orthogonal subspace and the Fisher criterion based-extractor is reiterated. Given a training corpus $\mathcal{T}$ of target and non-target trial expanded vectors, the following algorithm describes this method:

```
for $k = 1$ to $K$
    Compute means $g_t^{(k)}, g_n^{(k)}$ and covariance
    matrices $\mathcal{W}_t^{(k)}, \mathcal{W}_n^{(k)}$ of $\mathcal{T}$.
    Extract vector $u^{(k)}$ so that:
    $u^{(k)} = \left(\alpha_t \mathcal{W}_t^{(k)} + \alpha_n \mathcal{W}_n^{(k)}\right)^{-1} \left(g_t^{(k)} - g_n^{(k)}\right)$
    Project $\mathcal{T}$ onto the orthogonal subspace of $u^{(k)}$.
```

This method has been successfully used in fields such as face recognition [19, 20]. It has been noticed that Fisher's linear discriminant is equivalent to the maximum-likelihood parameter estimates of a Gaussian model, under some assumptions. But expanded vectors do not follow Gaussian distribution. It can be shown that the first r components of $\varphi_{i,j}$ given an i-vector $\mathbf{w}_i$ and i-vector normal prior follow independent non-central χ^2 distributions with 1 degree of freedom and distinct non-central parameters for target and non-target trials. However, considering the original Fisher *geometrical* approach, this method may succeed in improving speaker detection accuracy without assumptions of Gaussianity for the expanded vector distribution.

OD extracts a set of K discriminant axes $u^{(1)}, ..., u^{(K)}$. The main issue that needs to be addressed is to combine this set in a unique vector u for application in speaker verification. Weights $\{\omega_k\}_{k=1}^K$ have to be estimated, so that the score becomes:

$$\begin{aligned} s_{i,j} &= \varphi_{i,j}^t . \sum_{k=1}^K \omega_k u^{(k)} \\ &= \sum_{k=1}^K \omega_k \left(\varphi_{i,j}^t . u^{(k)}\right) \end{aligned} \tag{17}$$

Instead of a development step intended to estimate these weights, whose robustness could be questioned, we propose the following reasoning: given the random vector φ of target and non-target trials expanded vectors, let denote by $u / \|u\|$ its length-1 resulting vector as defined in (16), and by $\mathcal{E}$ its total covariance matrix $\mathcal{B} + \mathcal{W}$. The variance of scores $\varphi^t.u / \|u\|$ is equal to:

$$\begin{aligned} var\left(\varphi^t . \frac{u}{\|u\|}\right) &= \frac{u^t \mathcal{E} u}{\|u\|^2} = \frac{u^t \mathcal{B} u + u^t \mathcal{W} u}{\|u\|^2} \\ &= \frac{(\lambda + 1)}{\|u\|^2} u^t \mathcal{W} u \\ &= \frac{(\lambda + 1)}{\|u\|^2} (g_t - g_n)^t \mathcal{W}^{-1} (g_t - g_n) \\ &= \frac{\lambda(\lambda + 1)}{\alpha_t \alpha_n \|u\|^2} \end{aligned} \tag{18}$$

This shows that dispersion of OD-scores is a function of the Fisher criterion λ. As this criterion measures the two classes separation and has been shown to decrease along iterations [19], we propose to adjust score variance to the value $\lambda(\lambda + 1) / (\alpha_t \alpha_n)$, thus to assign the value $\mathcal{W}^{-1}(g_t - g_n)$ to u. The weighted sum of extracted vectors provides the following normal vector:

$$u = \sum_{k=1}^K \left(\alpha_t \mathcal{W}_t^{(k)} + \alpha_n \mathcal{W}_n^{(k)}\right)^{-1} \left(g_t^{(k)} - g_n^{(k)}\right) \tag{19}$$

where the decreasing series of score variance is equal to $\lambda^{(k)}(\lambda^{(k)} + 1)/(\alpha_t \alpha_n)$. The optimal number of kept axes K is determined on a development set.

For fast training, OD-extractor can be parallelized. Given a training dataset $\mathcal{T}$ and a partitioning $\{\mathcal{T}_q\}_q$ of $\mathcal{T}$, the mean vector and covariance matrix $(\mu, \mathbf{\Sigma})$ of $\mathcal{T}$ can be expressed as a combination of means and covariance matrices $(\mu_q, \mathbf{\Sigma}_q)$ of $\mathcal{T}_q$ subsets. Moreover, extracting these discriminant axes does not require to project data onto the orthogonal subspace of the latest axis and to compute their statistics, at each iteration. To our knowledge, this observation has never been made, thus we detail below the fast algorithm used for OD axis extraction:

```
$\mathbf{M} = \mathbf{I}_r$
for $k = 1$ to $K$
    $v = (\alpha_t \mathcal{W}_t + \alpha_n \mathcal{W}_n)^{-1}(g_t - g_n)$
    $u^{(k)} = \mathbf{M}v$
    Compute matrix $\mathbf{V}$ of last $(r - k)$ eigenvectors
    of $\frac{v}{\|v\|}\left(\frac{v}{\|v\|}\right)^t$.
    $g_t = \mathbf{V}^t g_t$ ; $g_n = \mathbf{V}^t g_n$
    $\mathcal{W}_t = \mathbf{V}^t \mathcal{W}_t \mathbf{V}$ ; $\mathcal{W}_n = \mathbf{V}^t \mathcal{W}_n \mathbf{V}$
    $\mathbf{M} = \mathbf{M}\mathbf{V}$
```

Table 1: Analysis of PLDA parameters before and after the $\mathbf{B}$-rotation additional normalization procedure.

	before		after	
	male	female	male	female
Diagonality of:				
$\mathbf{\Phi\Phi}^t$	0.23	0.15	0.95	0.97
$\mathcal{P}$	0.48	0.25	0.98	0.96
$\mathcal{Q}$	0.41	0.23	0.96	0.97
Isotropy of $\mathbf{\Lambda}$	0.98	0.96	0.99	0.97
Residual variance	0.29	0.42	0.004	0.004

We assume that this method could improve accuracy of probabilistic model for speaker verification, but only if each component of the expanded vector has some *discriminant power*. That is, if its value for target trials is likely higher or lower than for non-target trials. Thus, we do not split the two terms (p_k...,q_k...) of the k^{th} dimension, so that the expanded vector of (14) is left intact, except the $(r+1)^{th}$ residual term which is considered as non-discriminant and is dropped. The score becomes a dot product between $\mathbb{R}^r$ expanded vectors and the vector u estimated by using OD-training.

After OD-extraction, the resulting score can be summarized by the sum of two terms $s^{[1]}_{i,j} + s^{[2]}_{i,j}$ where:

$$s^{[1]}_{i,j} = \sum_{k=1}^{r} u_k p_k \left(\mathbf{w}_{i,k} - \mu_k\right)\left(\mathbf{w}_{j,k} - \mu_k\right)$$

$$s^{[2]}_{i,j} = \frac{1}{2}\sum_{k=1}^{r} u_k q_k \left(\left(\mathbf{w}_{i,k} - \mu_k\right)^2 + \left(\mathbf{w}_{j,k} - \mu_k\right)^2\right) \quad (20)$$

As written above, we consider that OD would fail to optimize weights of these two terms, as the $\mathbb{R}^2$ expanded vector components do not have discriminant power. Thus, in order to optimize these weights, we perform a fusion by logistic regression [16], estimating (ω_1, ω_2) so that the final score becomes:

$$s_{i,j} = \omega_1 s^{[1]}_{i,j} + \omega_2 s^{[2]}_{i,j} \quad (21)$$

6. Experiments

Our experiments operate on 19 LFCC parameters augmented with 19 first (Δ) and 11 second ($\Delta\Delta$) derivatives. A normalization process is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance for a given utterance. The low-energy frames (corresponding mainly to silence) are removed. Gender-dependent 512 diagonal component UBM and total variability matrix of low rank 400 are trained on NIST SRE 2004, 2005, 2006 and Switchboard data. PLDA systems and discriminative classifiers are trained using i-vectors extracted from 24100 utterances from 2012 female speakers and 15660 utterances from 1147 male speakers, from the same corpus. Results are presented for the extended condition 5 (tel-tel) from NIST SRE 2010 evaluation. The reported numbers are equal error rate (EER), the two minimum decision cost function (DCF) scores corresponding to the two operating points as defined by NIST for the SRE 2008 (minDCF08) and SRE 2010 (minDCF10) evaluations [21], and $C_{llr}^{\min}$ [22]. The minimum DCF score for 2010 evaluation is normalized by 10^3, whereas the minimum DCF score for 2008 is normalized by 10. Separate results are provided for male and female speakers.

Table 2: Speaker recognition results obtained by the different generative and discriminative systems.

Male set

Method	EER	minDCF10	minDCF08	$C_{llr}^{\min}$
PLDA	2.15	0.473	0.125	0.087
simpl.	2.15	0.476	0.127	0.087
LR $O(3r+1)$	2.40	0.458	0.140	0.093
LR param	2.17	0.498	0.127	0.086
OD	2.10	0.467	0.121	0.085
OD+LR $O(2)$	2.12	0.480	0.122	0.083

Female set

Method	EER	minDCF10	minDCF08	$C_{llr}^{\min}$
PLDA	3.01	0.585	0.167	0.112
simpl.	3.07	0.597	0.166	0.113
LR $O(3r+1)$	3.18	0.611	0.180	0.120
LR param	3.05	0.592	0.165	0.113
OD	2.88	0.588	0.162	0.108
OD+LR $O(2)$	2.88	0.589	0.162	0.108

Table 1 shows results of an analysis of PLDA matrix parameters, before and after the additional normalization procedure of $\mathbf{B}$ based-rotation presented in Sec. 4, computed with our development data. Here and in all the following, the value of r is fixed to the optimal PLDA eigenvoice rank. Diagonality of $\mathbf{\Phi\Phi}^t$, $\mathcal{P}$ and $\mathcal{Q}$ of equations(1) and (3) is estimated, also isotropy of $\mathbf{\Lambda}$. Diagonality of a symmetric matrix $\mathbf{A}$ can be measured by the ratio:

$$\frac{Tr\left(diag\left(\mathbf{A}\right)^2\right)}{Tr\left(\mathbf{A}^2\right)} \quad (22)$$

where $Tr\left(\right)$ is the trace operator and $diag\left(.\right)$ denotes the diagonal "version" of $\mathbf{A}$ (all its off-diagonal values equal zero). This measure is the square-length ratio between $\mathbf{A}$ and its orthogonal projection onto the $d \times d$ diagonal matrix subspace (which can be identified as $\mathbb{R}^d$). The maximal value of 1 indicates that $\mathbf{A}$ is exactly diagonal. Table 1 confirms the assumptions of Sec. 4. All these matrices are very close to be diagonal after the additional normalization procedure, for both genders. To measure isotropy of $\mathbf{\Lambda}$, the ratio:

$$\frac{m_{\mathbf{\Lambda}}^2}{d \times Tr\left(\mathbf{\Lambda}^2\right)} \quad (23)$$

is computed, where $m_{\mathbf{\Lambda}}$ denotes the mean value of $\mathbf{\Lambda}$-diagonal. Similarly, this measure is the square-length ratio between $\mathbf{\Lambda}$ and its orthogonal projection onto the 1×1 matrix subspace (which can be identified as $\mathbb{R}$). Also, Table 1 confirms that $\mathbf{\Lambda}$ (which is almost isotropic before rotation, as data have been $\mathbf{W}$-normalized) remains close to isotropy after rotation. Last row of Table 1 shows the ratio of variance between the residual term $res^{'}_{i,j}$ and the whole score of (9). Ratios are close to 0 for both gender sets after rotation, confirming that PLDA score can be limited to this $O(r)$ sum of terms with a minimal loss of accuracy.

Table 2 provides speaker recognition results for the different systems presented above. 1st row of the Table reports the best performance yielded by PLDA system (the optimal eigenvoice rank is equal to 100 for female set, 200 for male set). 2nd row shows performance of a simplified score drawn from (9), in which the residual term has been dropped. Results are equivalent to those of PLDA system, confirming the assumptions of Sec. 4 in terms of performance.

3rd and 4th row show results of the two logistic regression based-DT proposed in Sec. 5, based on coefficients or PLDA parameters training. Both DT fail to optimizing PLDA parameters. Performance of the second system, which follows [11], are equivalent to those of PLDA, which raises questions. The first system, based on LLR score coefficients, even degrades performance. It seems that this method over-fits to training data, even if the number of parameters to be estimated has been constrained.

The fact that LR-DT are based on Gaussian assumptions may limit effectiveness of these methods, since Gaussianity of vectors has been significantly improved by pre-normalization, thus brought closer to an optimal Gaussian modeling. 5th row of Table 2 presents results of the OD-classifier, and 6th row of the same system than the 5th followed by the $O(2)$ LR-classifier of equation (21). The number of kept axes, estimated on a development set, is equal to 7. In order to take into account eventual distortions of the non-target expanded vector distribution in regions of false alarms, OD model is trained using only the non-target expanded vector subset providing the 10% highest PLDA scores. 5th row shows that this method succeeds in improving PLDA parameters, in terms of speaker detection. The gain is slight for male set, more significant for female set. The additional step of $O(2)$ LR-DT (6th row) turns out to be useless. It can be noticed that results of more sophisticated variants of OD (ULDA [23], Foley-Sammon LDA [24]) are not reported here, as they did not yield satisfying performance.

The gap of OD-classifier performance between gender sets raises an issue, about dependency of the method to the configuration. In order to better assess the ability of OD-classifier to improve PLDA models for speaker verification, we carried out the same experiments with i-vectors 2011 provided by Brno University of Technology (BUT). Detailed description of their configuration can be found in [25]. The i-vector size is equal to 600 and the optimal PLDA eigenvoice rank to 80 for both gender sets. PLDA training uses 21475 sessions of 1575 speakers for male, 27155 sessions of 2012 speakers for female. Table 3 presents results for this configuration. The same observations than for Table 2 apply to the first four systems. For OD-classifier systems, the number of kept axes is equal to 3. This method provides significant improvements of performance, in terms of all the detection measures, in particular of EER. The slight gain observed for the previous male evaluation of Table 2 may be explained by a lack of training vectors (about 15000 instead of more than 20000 for other sets). 6th row shows that, again, the additional $O(2)$ LR-DT does not improve performance once OD-classifier has been carried out.

Furthermore, the training of OD system matrices $\mathcal{W}_t$ and $\mathcal{W}_n$ presented in Sec. 5 was parallelized, as proposed in this section. Splitting the task in 20 processes, less than 30 minutes was needed to train gender system matrices on an usual configuration. Following extraction of OD-axes required less than 5 minutes with the fast algorithm of Sec. 5.

Table 3: Speaker recognition results with BUT i-vectors 2011.

Male set

Method	EER	minDCF10	minDCF08	$C_{llr}^{\min}$
PLDA	1.03	0.309	0.061	0.040
simpl.	1.05	0.291	0.064	0.040
LR $O(3r+1)$	1.24	0.342	0.076	0.047
LR param	1.06	0.294	0.062	0.040
OD	0.95	0.282	0.060	0.038
OD+LR $O(2)$	0.96	0.281	0.059	0.038

Female set

Method	EER	minDCF10	minDCF08	$C_{llr}^{\min}$
PLDA	1.79	0.331	0.102	0.063
simpl.	1.77	0.326	0.099	0.061
LR $O(3r+1)$	1.78	0.331	0.101	0.064
LR param	1.72	0.336	0.101	0.061
OD	1.56	0.326	0.095	0.058
OD+LR $O(2)$	1.56	0.323	0.095	0.059

7. Conclusions

Discriminative training (DT) techniques have proven to be efficient for optimizing parameters of the initial speaker verification Gaussian modelings based on i-vector (two-covariance model, PLDA). But their benefits become less significant when i-vectors are initially normalized (the most currently used procedures being comprised of within-class covariance and length normalization), while these procedures allow Gaussian systems to achieve best performance. Several work have pointed out that discriminative approaches can suffer from various limitations, as data insufficiency, over-fitting on development data or metaparameters conditions, leading to constrained versions.

The additional normalization procedure that we propose (a simple rotation by between-class covariance matrix, which does not modify distances between i-vectors) leads to a functional form for verification scores derived from PLDA, which allows us to control the number of trainable parameters. The original discriminative classifier, of order the square of the i-vector size, can be replaced by constrained discriminative classifiers of low order with a minimal loss of accuracy.

Experiments carried out on current NIST evaluations sets show that logistic regression (LR) based-discriminative trainings, which we adapt to the new normalized distribution of i-vectors, continue to suffer from data insufficiency or over-fitting on development data. Giving up the LR approach, a new method in the field is proposed to extract discriminative axes by using the Fisher criterion (OD). Unlike the usual discriminative classifiers, which attempt to find out a unique normal vector of a separation hyperplane, the proposed method extracts a discriminant subspace (by decreasing variance, in a way similar to singular value decomposition), then combine its basis to find out the unique normal vector needed by speaker detection. This combination is done without the need to tune the weights of the discriminant axes set. Experiments show that this method is able to significantly improve performance of Gaussian PLDA systems, even when i-vectors are normalized. As far as training complexity is concerned, OD training also has the advantage of not being demanding in terms of time and memory requirements.

Future work should test OD method on specific conditions, as short duration or noisy utterances. Accurate estimation of the speaker variability is more difficult with these conditions, and Gaussian PLDA modeling could benefit from this additional discriminative training. Also, it has been shown in [26] that the normalization and PLDA framework can be successfully applied in speaker diarization to low rank total variability factors provided by a deep neural network. Testing OD method on i-vector-like representations would be of interest.

Acknowledgements

Many thanks to Honza Černocký and Pavel Matějka from Brno University of Technology, for letting us use the BUT i-vectors 2011.

8. References

[1] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[3] Niko Brummer and Edward de Villiers, "The speaker partitioning problem," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[4] Niko Brümmer, Lukas Burget, Jan Honza Cernocky, Ondrej Glembek, Frantisek Grezl, Martin Karafiat, David A. Van Leeuwen, Pavel Matejka, Petr Schwarz, and Albert Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[5] Lukas Burget, Oldrich Plchot, Sandro Cumani, Ondrej Glembek, Pavel Matejka, and Niko Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2011, pp. 4832–4835.

[6] Sandro Cumani, Niko Brummer, Lukas Burget, and Pietro Laface, "Fast discriminative speaker verification in the i-vector space," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2011, pp. 4852–4855.

[7] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *International Conference on Speech Communication and Technology*, 2011, pp. 249–252.

[8] Pierre-Michel Bousquet, Driss Matrouf, and Jean-François Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *International Conference on Speech Communication and Technology*, 2011, pp. 485–488.

[9] Sandro Cumani, Niko Brummer, Lukás Burget, Pietro Laface, Oldrich Plchot, and Vasileios Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[10] J. Rohdin, , S. Biswas, and K. Shinoda, "Robust discriminative training against data insufficiency in PLDA-based speaker verification," *Computer Speech and Language*, vol. 35, pp. 32–57, 2016.

[11] B. Börgstrom and A. Ahn Mac Cree, "Discriminatively trained bayesian speaker comparison of i-vectors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. IEEE; 1999, 2013, pp. 7659–7662.

[12] Sergey Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision*, pp. 531–542, 2006.

[13] Simon J.D. Prince and James H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[14] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre, and Oldřich Plchot, "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2012.

[15] Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, Springer, August 2000.

[16] Niko Brümmer. and Edward de Villiers, "The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing," 2011.

[17] Christopher M. Bishop, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.

[18] T. Okada and S. Tomita, "An optimal orthonormal system for discriminant analysis," *Pattern Recognition*, vol. 18, no. 2, pp. 139–144, 1985.

[19] Jinghua Wang, Yong Xu, David Zhang, and Jane You, "An efficient method for computing orthogonal discriminant vectors," vol. 73, no. 10-12, pp. 2168 – 2176, 2010.

[20] Wen Yi Zhao, "Discriminant component analysis for face recognition," *Pattern Recognition*, vol. 2, pp. 818–821, 2000.

[21] National Institute of Standards and Technology, "Speaker recognition evaluation plans," http://www.itl.nist.gov/iad/mig/tests/sre/.

[22] Niko Brümmer and Johan du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[23] Zhong Jin, Jing-Yu Yang, Zhong-Shan Hu, and Zhen Lou", "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognition*, vol. 34, no. 7, pp. 1405–1416, 2001.

[24] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors," *IEEE Transactions on Computers*, vol. 24, no. 3, pp. 281–289, March 1975.

[25] Pavel Matejka, Ondrej Glembeck, Fabio Castaldo, M.J. Alam, Oldrich Plchot, Patrick Kenny, Lukas Burget, and Jan Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *International Conference on Speech Communication and Technology*, 2011, pp. 4828–4831.

[26] M. Rouvier, P.M. Bousquet, and B. Favre, "Speaker diarization through speaker embeddings," in *European Signal and Image Processing Conference (EUSIPCO)*, 2015.