Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis

Pierre-Michel Bousquet [1], *Anthony Larcher* [2],
Driss Matrouf[1], *Jean-François Bonastre*[1], *Oldřich Plchot* [3]

(1) University of Avignon - LIA, France
(2) Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
(3) Brno University of Technology, Speech@FIT, Brno, Czech Republic,
{pierre-michel.bousquet, driss.matrouf, jean-francois.bonastre}@univ-avignon.fr
alarcher@i2r.a-star.edu.sg, iplchot@fit.vutbr.cz

Abstract

I-vector extraction and Probabilistic Linear Discriminant Analysis (PLDA) has become the state-of-the-art configuration for speaker verification. Recently, Gaussian-PLDA has been improved by a preliminary length normalization of *i*-vectors. This normalization, known to increase the Gaussianity of the *i*-vector distribution, also improves performance of systems based on standard Linear Discriminant Analysis (LDA) and "two-covariance model" scoring. But this technique follows a standardization of the i-vectors (centering and whitening i-vectors based on the first and second order moments of the development data). We propose in this paper two techniques of normalization based on total, between- and within-speaker variance spectra [1]. These "spectral" techniques both normalize the *i*-vectors length for Gaussianity, but the first adapts the *i*-vectors representation to a speaker recognition system based on LDA and two-covariance scoring when the second adapts it to a Gaussian-PLDA model. Significant performance improvements are demonstrated on the male and female telephone portion of NIST SRE 2010.

Index Terms: i-vectors, probabilistic linear discriminant analysis, speaker recognition.

1. Introduction

Based on Factor Analysis (FA), the Total Variability space framework has become a new standard for speaker verification systems. Total Variability space provides a compact representation of speech sessions, so called *i*-vectors, that carries out the classification task into a low dimensional factor space rather than in the GMM-super-vector space. By nature, the Total Variability space contains the overall between-utterance variability and it is ultimately the role of intersession compensation methods to define how speakers are discriminated from one another in this subspace.

The initial *i*-vector speaker recognition system developed by Dehak et al.[1, 2] enhanced discrimination in Total Variability space by reducing the dimension of *i*-vectors using standard Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN). Classification was performed by using a cosine kernel. Later, normalized and weighted versions of LDA have been proposed [3] to address the issue of unbalanced development data.

Since generative models have been introduced, *i*-vectors are seen as an observation from a probabilistic model, ignoring their extraction mechanism. The two-covariance model introduced in the speaker recognition field by N. Brummer et al. [4] proposes an efficient alternative to existing scoring. This generative model is a particular case of Probabilistic Linear Discriminant Analysis (PLDA), itself a special case of Factor Analysis (FA) [5] as it considers a single component. Gaussian-PLDA (G-PLDA) assumes Gaussian priors of both channel and speaker factors. Introduced in [6], Heavy-tailed version of PLDA (HT-PLDA) replaces Gaussian distributions with Student-t distributions and was shown to substantially improve the performance.

Recently it has been shown [7, 8] that preliminary standardization and length normalization of *i*-vectors (normalizing centered and scaled *i*-vectors by their magnitude) significantly improves performance of Gaussian-based systems like two-covariance scoring [9], Mahalanobis Scoring [8] or Gaussian PLDA [7]. In particular, Gaussian-PLDA preceded by a standardization and length normalization compares very favorably with the Heavy Tailed version of PLDA [7]. This work deals with the length normalization step.

Once norms of all the *i*-vectors are equal, the representation space becomes a spherical surface of finite volume. In this context, we propose two normalization techniques based on total, between- and within-speaker distributions of development data. The first one is adapted to a speaker recognition system based on LDA and two-covariance scoring. This normalization aims at enhancing the optimization criterion of LDA. The second one is adapted to a Gaussian-PLDA model. It moves data towards an appropriate starting point in the optimization landscape of Gaussian-PLDA.

Next two sections give first a description of the *i*-vector paradigm then describe LDA, PLDA and two-covariance models. Section 4 describes the two "variance-spectra based normalization" techniques after presenting a simple visualization tool dedicated to the analysis of development dataset before and after transform. Section 5 analyses the behaviour of these techniques on the telephone portion of NIST SRE 2010.

[1] We speak of "variance-spectra" by analogy with the spectrum of a matrix, i.e. the set of its eigenvalues.

2. I-vector paradigm

The *i*-vector approach has become state of the art in the field of speaker verification [1, 2]. In this approach, an *i*-vector extractor converts a sequence of acoustic vectors into a single low-dimensional vector representing the whole speech utterance. The speaker- and session-dependent super-vector $\mathbf{s}$ of concatenated Gaussian Mixture Model (GMM) means is assumed to obey a linear model (Factor Analysis) of the form:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where $\mathbf{m}$ is the mean super-vector of the Universal Background Model (UBM), $\mathbf{T}$ is the low-rank variability matrix obtained from a large dataset by MAP estimation [10] and the standard-normally distributed latent variable $\mathbf{w}$ is the resulting *i*-vector.

3. I-vectors recognition systems

To improve comparison of *i*-vectors in a speaker verification trial, dimensionality reduction techniques, like Linear Discriminant Analysis (LDA) [1, 2, 9], can be applied. I-vectors can also be seen as observations from a probabilistic generative model (two-covariance model [4], Probabilistic Linear Discriminant Analysis [5, 6]). In the following, these dimension reduction techniques and generative models used in our experiments are described.

3.1. Linear Discriminant Analysis

Standard Linear Discriminant Analysis (LDA) is a technique for dimensionality reduction that projects the data onto a subspace which satisfies the requirement of maximizing between-class variance and minimizing within class variance. The LDA optimization problem (finding a basis of this subspace) can be defined according to the following ratio:

$$J(v) = \frac{v^t \mathbf{B} v}{v^t \mathbf{W} v} \tag{2}$$

where $\mathbf{B}$ is the between-speaker covariance matrix and $\mathbf{W}$ the within-speaker covariance matrix defined by:

$$\mathbf{B} = \sum_{s=1}^{S} \frac{n_s}{n} (\mathbf{y}_s - \boldsymbol{\mu})(\mathbf{y}_s - \boldsymbol{\mu})^t \tag{3}$$

$$\mathbf{W} = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \mathbf{y}_s)(\mathbf{w}_i^s - \mathbf{y}_s)^t \tag{4}$$

where n_s is the number of utterances for speaker s, n is the total number of utterances, $\mathbf{w}_i^s$ are the *i*-vectors of sessions of speaker s, $\mathbf{y}_s$ is the mean of all the i-vectors of speaker s and μ represents the overall mean of the training dataset.

This ratio of Equation 2 is often referred to as the Rayleigh coefficient for space direction v. The optimal subspace is comprised by the first eigenvectors (those with highest eigenvalues) of $\mathbf{W}^{-1}\mathbf{B}$.

In the speaker recognition domain, it turned out that better performance were achieved [1, 2] by replacing $\mathbf{B}$ and $\mathbf{W}$ with the scatter-matrices:

$$\mathbf{S_B} = \sum_{s=1}^{S} (\mathbf{y}_s - \boldsymbol{\mu})(\mathbf{y}_s - \boldsymbol{\mu})^t \tag{5}$$

$$\mathbf{S_W} = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \mathbf{y}_s)(\mathbf{w}_i^s - \mathbf{y}_s)^t \tag{6}$$

Unlike $\mathbf{B}$ and $\mathbf{W}$, these matrices do not take into account prior of each speaker, i.e. the amount of utterances per speaker training sample. In the following, we refer to $\text{LDA}_{\mathbf{B},\mathbf{W}}$ (resp. $\text{LDA}_{\mathbf{S_B},\mathbf{S_W}}$) for linear discriminant analysis based on the Rayleigh coefficient computed with $\mathbf{B}$ and $\mathbf{W}$ (resp. $\mathbf{S_B}$ and $\mathbf{S_W}$).

3.2. Probabilistic Linear Discriminant Analysis

3.2.1. PLDA Model

Introduced in [5], the Gaussian Probabilistic Linear Discriminant Analysis (PLDA) is a generative *i*-vector model which assumes that each p-dimensional *i*-vector $\mathbf{w}$ of a speaker s can be decomposed as:

$$\mathbf{w} = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{y}_s + \boldsymbol{\Gamma}\mathbf{z} + \boldsymbol{\varepsilon} \tag{7}$$

This model comprises two parts: (i) the component $\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{y}_s$ which depends only on the speaker, (ii) the noise component $\boldsymbol{\Gamma}\mathbf{z} + \boldsymbol{\varepsilon}$ which is different for every session and represents within-speaker noise. Matrix $\boldsymbol{\Phi}$ is rectangular, with r_{voices} columns ($r_{voices} < p$) providing a basis for a speaker subspace, usually called "eigenvoices". Likewise, $\boldsymbol{\Gamma}$ is rectangular, with $r_{channels}$ columns providing a basis for a channel subspace, usually called "eigenchannels". Standard normal priors are assumed for $\mathbf{y}_s$ and $\mathbf{z}$. Lastly the residual term $\boldsymbol{\varepsilon}$ is assumed to be Gaussian with zero mean and diagonal covariance $\boldsymbol{\Sigma}$.

The particular case of $r_{channels} = p$ (full-dimensional matrix $\boldsymbol{\Gamma}$) is equivalent to the version of the PLDA proposed in [6]: the eigenchannels are removed from Equation 7 and the residual noise is assumed to have a full covariance matrix. The PLDA model becomes:

$$\mathbf{w} = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{y}_s + \boldsymbol{\varepsilon} \tag{8}$$

where $\boldsymbol{\Phi}$ is a $p \times r$ matrix ($r < p$) and $\boldsymbol{\varepsilon}$ a p-dimensional vector with a full covariance matrix.

3.2.2. PLDA scoring

After estimation of the PLDA meta-parameters, the speaker verification score given two *i*-vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ is the likelihood-ratio described by Equation 9, where the hypothesis θ_{tar} states that inputs $\mathbf{w}_1$ and $\mathbf{w}_2$ are from the same speaker and the hypothesis θ_{non} states they are from different speakers.

$$score = \log \frac{P(\mathbf{w}_1, \mathbf{w}_2 | \theta_{tar})}{P(\mathbf{w}_1, \mathbf{w}_2 | \theta_{non})} \tag{9}$$

As proposed by S.Prince [5, 11], the generative Equation 7 under assumption θ_{tar} looks like:

$$\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Gamma} & 0 \\ \boldsymbol{\Phi} & 0 & \boldsymbol{\Gamma} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \tag{10}$$

and under assumption θ_{non}:

$$\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Phi} & 0 & \boldsymbol{\Gamma} & 0 \\ 0 & \boldsymbol{\Phi} & 0 & \boldsymbol{\Gamma} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \tag{11}$$

For the Gaussian-PLDA case, all the marginal likelihoods are Gaussian and the score 9 can be evaluated analytically [11]. The final expressions do not include the hidden variables. Numerator of Equation 9 is equal to:

$$\mathcal{N}\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Phi}\boldsymbol{\Phi}^t+\boldsymbol{\Gamma}\boldsymbol{\Gamma}^t+\boldsymbol{\Sigma} & \boldsymbol{\Phi}\boldsymbol{\Phi}^t\\ \boldsymbol{\Phi}\boldsymbol{\Phi}^t & \boldsymbol{\Phi}\boldsymbol{\Phi}^t+\boldsymbol{\Gamma}\boldsymbol{\Gamma}^t+\boldsymbol{\Sigma}\end{bmatrix}\right) \tag{12}$$

and denominator to:

$$\mathcal{N}\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Phi}\boldsymbol{\Phi}^t+\boldsymbol{\Gamma}\boldsymbol{\Gamma}^t+\boldsymbol{\Sigma} & 0\\ 0 & \boldsymbol{\Phi}\boldsymbol{\Phi}^t+\boldsymbol{\Gamma}\boldsymbol{\Gamma}^t+\boldsymbol{\Sigma}\end{bmatrix}\right) \tag{13}$$

3.3. The two-covariance model

Described in [4, 12], the two-covariance model can be seen as a particular case of the Probabilistic Linear Discriminant Analysis [13]. It consists of a simple linear-Gaussian generative model in which an *i*-vector $\mathbf{w}$ of a speaker s can be decomposed in:

$$\mathbf{w} = \mathbf{y}_s + \boldsymbol{\varepsilon} \tag{14}$$

where the speaker model $\mathbf{y}_s$ is a vector of the same dimensionality as an *i*-vector, $\boldsymbol{\varepsilon}$ is Gaussian noise and

$$P(\mathbf{y}_s) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{B}), \tag{15}$$

$$P(\mathbf{w}|\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s, \mathbf{W}). \tag{16}$$

$\mathcal{N}$ denotes the normal distribution, $\boldsymbol{\mu}$ represents the overall mean of the training dataset, $\mathbf{B}$ and $\mathbf{W}$ are the between- and within-speaker covariance matrices defined in Equations 3, 4.

Under assumptions (15)(16), the score from Equation 9 can be expressed as:

$$s = \frac{\int \mathcal{N}(\mathbf{w}_1|y,\mathbf{W})\,\mathcal{N}(\mathbf{w}_2|y,\mathbf{W})\,\mathcal{N}(y|\boldsymbol{\mu},B)\,dy}{\prod_{i=1,2}\int \mathcal{N}(\mathbf{w}_i|y,\mathbf{W})\,\mathcal{N}(y|\boldsymbol{\mu},B)\,dy} \tag{17}$$

The explicit solution of (17) is given in [4].

4. I-vector normalizations

In the following, we call "normalization" any transformation that projects *i*-vectors onto the surface area of the p-dimensional sphere of radius 1, and "length normalization" the straight division of *i*-vectors by their Euclidean norm. In the last paragraphs of this section, we propose two variance-spectra based normalization techniques, one for LDA system and the other for PLDA system. First, we introduce a visualization tool intended to better analyze the spectral distributions of a dataset.

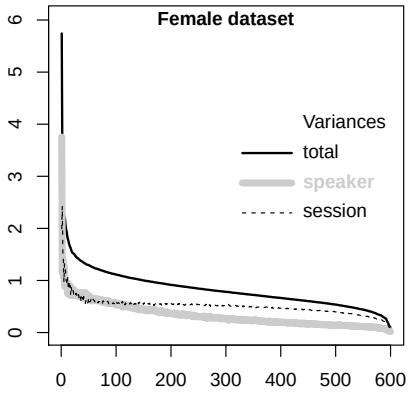
4.1. Spectral graph

Given an *i*-vector dataset expressed in a given basis, we call "spectral graph" a visualization of the total and related speaker, session variabilities across the space dimensions.
To proceed, matrices $\mathbf{B}$ and $\mathbf{W}$ of Equation 3 and 4 are computed. The total covariance matrix is equal to $\mathbf{B}+\mathbf{W}$. The spectral graph displays the three series $diag(\mathbf{B}+\mathbf{W})$, $diag(\mathbf{B})$ and $diag(\mathbf{W})$, which contain the proportion of total, speaker and session variance per dimension.

Figure 1 shows the spectral graph of female development dataset described in section 5, in the eigenvector basis of total covariance matrix [2]. Dimension of the *i*-vector space is $p = 600$. Note that the "total" curve is just the eigenvalue spectrum of total covariance matrix but the two others are not eigenvalue spectra of $\mathbf{B}$ and $\mathbf{W}$ (their eigenvectors basis are generally distinct). It is of interest to analyze this graph:

- the total covariance matrix is far from the identity matrix: its eigenvalue spectrum is not flat, though Factor Analysis-Total Variability assumes that *i*-vectors are standard-normally distributed;
- the speaker spectrum is highly correlated with the total spectrum. As FA-Total Variability ignores the speaker information, the latent variable seems to induce a bias into the training sample;
- the session spectrum is flatter, thus less correlated with other spectra.

Figure 1: Spectral graph of female development data (before any transformation). For the 600 dimensions (x axis), the y axis shows the total, speaker and session parts of variance.



4.2. Length normalization

Proposed in [7, 8], the length normalization (scaling the *i*-vectors by their magnitude) is intended to make them more Gaussian. It can be shown[3] that, in a p high-dimensional space, the spherical surface of radius 1 is the maximal density shell of a normal distribution of covariance matrix $p^{-1}\mathbf{I}$. As a result, length normalization combats bad modelling assumptions, but also dataset shift between development and trial *i*-vectors.
However, two points caught our attention:

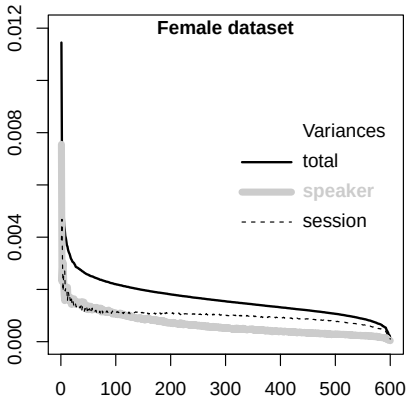
- Length normalization is preceded by a standardization (centering and whitening) based on he first and second order moments of the development data, It could be of interest to explore other normalizations based on total, speaker or noise covariance of the training dataset. We call "variance-spectra based normalization" such normalizations.
- In the current framework, normalization is applied previously to a discriminative technique, here LDA or PLDA. We propose to apply normalizations to transform the development and trial *i*-vector representations in such a way they better fit assumptions of the following discriminant techniques.

[2] Here and in the following, only female dataset spectral graphs are shown, since it turned out that all the male dataset graphs are almost identical.

[3] http://ontopo.wordpress.com/2009/03/10/reasoning-in-higher-dimensions-measure/

Figure 2 displays the spectral graph of female development dataset, in the eigenvector basis of total covariance matrix, after length normalization. We observe that this technique does not modify spectral distributions. We present below two variance-spectra based normalizations suitable for LDA and PLDA respectively.

Figure 2: Spectral graph of female development data after application of length normalization.



4.3. Variance-spectra based normalization for LDA

LDA assumes that it exists a subspace of significantly high speaker variability and low noise variability. Thus projecting the data onto this subspace would improve speaker discrimination.

There is usually no optimal solution to the LDA problem, i.e. finding axes simultaneously maximizing the speaker variance and minimizing the session variance. The aim of a variance-spectra based normalization before LDA is to increase the Rayleigh quotient on the surface of the sphere in such a way that the dataset moves towards the optimal solution. We presented in Interspeech 2011 [8] an algorithm intended to prepare *i*-vectors for scoring. Given a training dataset $\mathcal{T}$, the algorithm is:

Algorithm for training dataset

for $i = 1$ to $nb\_iterations$

Compute mean μ_i and total covariance matrix $\mathbf{\Sigma}_i$ of $\mathcal{T}$

For each $\mathbf{w}$ of $\mathcal{T}$: $\quad \mathbf{w} \leftarrow \dfrac{\mathbf{\Sigma}_i^{-\frac{1}{2}} (\mathbf{w} - \boldsymbol{\mu}_i)}{\left\| \mathbf{\Sigma}_i^{-\frac{1}{2}} (\mathbf{w} - \boldsymbol{\mu}_i) \right\|}$

$\mathbf{\Sigma}_i^{-\frac{1}{2}}$ can be computed using a singular value decomposition. For each test vector of an evaluation set the same algorithm is applied, but with the successive parameters $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ computed on the training dataset.

Algorithm for test vectors

Given a test vector $\mathbf{w}$,

for $i = 1$ to $nb\_iterations$: $\quad \mathbf{w} \leftarrow \dfrac{\mathbf{\Sigma}_i^{-\frac{1}{2}} (\mathbf{w} - \boldsymbol{\mu}_i)}{\left\| \mathbf{\Sigma}_i^{-\frac{1}{2}} (\mathbf{w} - \boldsymbol{\mu}_i) \right\|}$

In each iteration, *i*-vectors of the training dataset are standardized then length normalized. With only one iteration, the algorithm is just a standardization followed by normalization, already used in current i-vector based systems [7, 8].

We showed empirically in [8] that this algorithm makes the global mean $\boldsymbol{\mu}_i$ tending to 0 and the covariance matrix $\mathbf{\Sigma}_i$ to $p^{-1}\mathbf{I}$ (the identity matrix divided by the dimension p of the Total Variability space). Note that this algorithm is stationary after several iterations.

Figure 3: Spectral graph of female development data, in the eigenvector basis of $\mathbf{B}$, after 3 iterations of variance-spectra based normalization (Eigen Factor Radial).

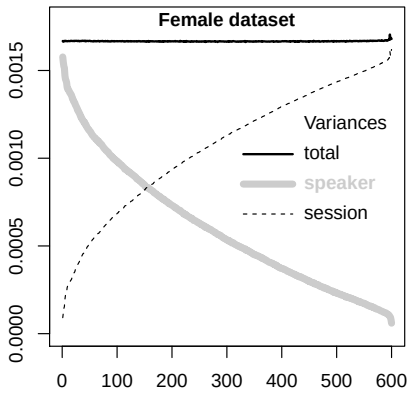


Figure 3 displays the spectral graph of female development data, in the eigenvector basis of $\mathbf{B}$, after 3 iterations of this variance-spectra based normalization. The convergence of $\mathbf{\Sigma}$ to $p^{-1}\mathbf{I}$ is almost totally achieved. After this normalization, speaker and session variances are complementary and therefore the first eigenvectors of $\mathbf{B}$ contain a major proportion of speaker variability and a minor proportion of session variability.
This fact can be demonstrated. As $\mathbf{\Sigma}$ is approximately equal to $p^{-1}\mathbf{I}$ and $\mathbf{\Sigma} = \mathbf{B} + \mathbf{W}$, each eigenvector v of $\mathbf{B}$ for an eigenvalue λ verifies almost exactly:

$$\mathbf{W}v = \left(p^{-1} - \lambda\right) v \tag{18}$$

Thus v is an eigenvector of $\mathbf{W}$ for the eigenvalue $\left(p^{-1} - \lambda\right)$. Accordingly, sum of $\mathbf{B}$ and $\mathbf{W}$ eigenvalues for each axis is equal to p^{-1}.

Therefore the first eigenvectors of $\mathbf{B}$ are the solution of the $\text{LDA}_{\mathbf{B},\mathbf{W}}$ optimization problem of Equation 2. As a same orthogonal basis contains speaker and session principal subspaces (often called *eigenvoices* and *eigenchannels*) and as *i*-vectors lie on a sphere, we call *EigenFactors Radial* (EFR) this variance-spectra based normalization technique.
It is worth noting that after EFR, retaining the r first dimensions

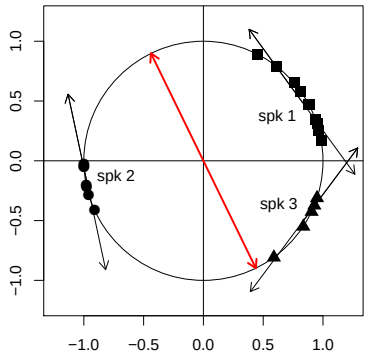
of the eigenvector basis of $\mathbf{B}$ becomes equivalent to removing the $(p-r)$ first dimensions of $\mathbf{W}$. This involves that LDA after normalization is equivalent to Nuisance Attribute Projection[4].

4.4. Variance-spectra based normalization for PLDA

4.4.1. Method

Considering a normalization of *i*-vectors that would be adapted to PLDA modeling, we raise the following issue: once all *i*-vectors lie on a spherical surface, it is difficult to estimate the within-class covariance matrix. To depict this fact, Figure 4 shows a two-dimensional example: *i*-vectors of three speakers lie on a circle, with a good level of isolation between classes. The arrows indicate the first principal axis (1D) of session variability for each speaker and the central red arrows indicate the first principal axis of $\mathbf{W}$. Estimating the principal axis of the

Figure 4: Two-dimensional example of session variability onto a spherical surface. I-vectors of three speakers are shown. All *i*-vectors are of norm 1. Arrows indicate the first principal session axis of each speaker. Central red arrows indicate the first principal axis of $\mathbf{W}$.



session variability over all speakers by the mean of the three axes gives a poorly accurate result. More generally, it can be questioned to base a generative model on an overall linear estimation of session covariances meanwhile data lie on a spherical surface. Facing this fact, the most appropriate alternative seems to apply a transformation which keeps data on the non-linear surface for Gaussianity, but without any principal directions of session variability nor without dependences between directions. Thus, to have a *spherical* within-class covariance matrix (a spherical matrix is a matrix $\sigma^2\mathbf{I}$).

Finding a way to conciliate length normalization and spherical noise matrix is greatly facilitated by the previous variance-spectra based normalization. Based on the fact that the EFR normalization makes $\mathbf{\Sigma}$ tending to $p^{-1}\mathbf{I}$, which is spherical, we propose to carry out a similar algorithm, by replacing $\mathbf{\Sigma}$ with $\mathbf{W}$. The algorithm becomes:

[4]Details on Nuisance Attribute Projection (NAP) can be found in [14].

Algorithm for training dataset

for $i = 1$ to $nb\_iterations$

Compute mean $\boldsymbol{\mu}_i$ and within-class cov. matrix $\mathbf{W}_i$ of $\mathcal{T}$

For each $\mathbf{w}$ of $\mathcal{T}$: $\quad \mathbf{w} \leftarrow \dfrac{\mathbf{W}_i^{-\frac{1}{2}}(\mathbf{w}-\boldsymbol{\mu}_i)}{\left\|\mathbf{W}_i^{-\frac{1}{2}}(\mathbf{w}-\boldsymbol{\mu}_i)\right\|}$

For each test vector of an evaluation, the same algorithm is applied, with the successive parameters $\boldsymbol{\mu}_i$ and $\mathbf{W}_i$ computed on the training dataset.

Algorithm for test vectors

Given a test vector $\mathbf{w}$,

for $i = 1$ to $nb\_iterations$: $\quad \mathbf{w} \leftarrow \dfrac{\mathbf{W}_i^{-\frac{1}{2}}(\mathbf{w}-\boldsymbol{\mu}_i)}{\left\|\mathbf{W}_i^{-\frac{1}{2}}(\mathbf{w}-\boldsymbol{\mu}_i)\right\|}$

We call Spherical Nuisance normalization this technique, because the combined session and residual nuisances have a spherical covariance matrix [5].

Figure 5: Spectral graph of female development data, in the eigenvector basis of $\mathbf{B}$, after application of 3 iterations of Spherical Nuisance normalization.

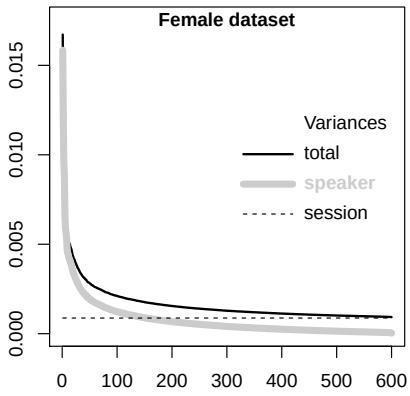


Figure 5 displays the spectral graph of female development data, in the eigenvector basis of $\mathbf{B}$, after 3 iterations of this algorithm. The $\mathbf{W}$ spectrum is almost exactly flat. As not shown in Figure 5, $\mathbf{W}$ is also close to be diagonal. Note that since $\mathbf{W}$ tends to a matrix $\sigma^2\mathbf{I}$ (σ scalar) this algorithm is stationary after some iterations.
The main point here is that the energy of the $\mathbf{B}$-spectrum has been maintained. Less than 200 axes contain a major proportion of the speaker variability. The part of speaker variance in the total variance even increased (from 41% to 50% for the male dataset and 40% to 47% for the female dataset).

4.4.2. PLDA initialization

PLDA meta-parameters are estimated through an iterative Expectation Maximization (EM) algorithm. In the absence of any

[5]The case of *isotropic* noise $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$ is referred to as "homoscedastic residuals model" [15].

other information, speaker and channel matrices $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ of Equation 7 are randomly initialized. After the above algorithm, it can be taken advantage of the new representation to initialize the PLDA matrices. To proceed, development and evaluation data are expressed in the eigenvector basis of $\mathbf{B}$ (each vector is multiplied by the $\mathbf{B}$ eigenvector matrix sorted by decreasing order of eigenvalues). In this basis, which is the one of Figure 5, the r first dimensions contain the principal speaker variability. Hence, the speaker matrix is initialized with:

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{I}_{r,r} \\ \mathbf{0}_{p-r,r} \end{bmatrix} \tag{19}$$

where $\mathbf{I}_{r,r}$ is the $r \times r$ identity matrix and $\mathbf{0}_{p-r,r}$ is the $(p-r) \times r$ null matrix.

$\mathbf{\Gamma}$ matrix is initialized by the Cholesky decomposition of $\mathbf{W}$, so that $\mathbf{W} = \mathbf{\Gamma}\mathbf{\Gamma}^t$. By this way, the estimation algorithm is initialized at an meaningful starting point in the optimization landscape of the PLDA EM-algorithm. PLDA will refine this model by distributing speaker variabilities in the subspace and normalizing anisotropic nuisances.

5. Experiments and results

5.1. Experimental setup

All our experiments were carried out on the extended-core telephone-telephone condition (i.e., condition 5) from the NIST SRE 2010 evaluation. For all our experiments, we have used the *i*-vectors provided by Brno University of Technology (BUT) [16].

5.2. Voice Activity Detection

Speech/silence segmentation is performed by BUT Hungarian phoneme recognizer [17], where all phoneme classes are linked to the *speech* class. More details on VAD are provided in [18].

5.3. Feature Extraction

We use MFCC 19 + energy augmented with their delta and double delta coefficients, making 60 dimensional feature vectors.

The analysis window has 20 ms with shift of 10 ms. First we remove silence frames according to VAD and after that we apply short-time cepstral mean and variance normalization which uses a window of 300 frames.

5.4. GMM UBM Training

One gender-independent UBM was represented as a full covariance 2048-component GMM. It was trained on the 7823 female and 7779 male segments balanced over the telephone/telmic/interview channels from NIST SRE 2004, 2005, 2006 and 2008 data. The variance flooring was used in each iteration of EM algorithm during the UBM training [16].

5.5. I-vector Extractor Training

Gender-dependent i-vector extractor was trained on the following telephone data: NIST SRE 2004, 2005, 2006, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2 giving 9627 female speakers in 45195 segments of speech, and 6989 male speakers in 33645 segments of speech (both after VAD). The results are reported with 600 dimensional i-vectors.

5.6. LDA and PLDA Training

The gender-dependent training lists for two-covariance, LDA and PLDA are from sessions of NIST SRE 2004, 2005, 2006, Switchboard II Phases 2 and 3 and Switchboard Cellular Parts 1 and 2, telephone only, nominal length higher than 180 seconds. No minimum amount of utterances per speaker has been stated. This gives 21475 sessions of 1575 speakers for male, 27155 sessions of 2012 speakers for female.

5.7. Results

Tables 1, 2, 3 give comparison result for the core extended condition 5 (telephone-telephone) from the NIST SRE 2010 evaluation, between systems based on three techniques: two-covariance scoring, LDA followed by two-covariance scoring and PLDA, in terms of Equal Error Rate (EER) and normalized minimum Decision Cost Functions (DCF) for the two operating points as defined by NIST for the SRE 2010 evaluations.

5.7.1. Two-covariance scoring

Table 1 compares performance of two-covariance scoring (Equation 17) applied without any transformation of the initial data and after a simple length normalization (without any standardization or other spectral technique) of both development and evaluation data. Length normalization involves a great enhancement of performance for both genders. Data become more suitable for a scoring based on the two-covariance model.

Table 1: Comparison of two-covariance scoring without and with length normalization of *i*-vectors for the core extended condition 5 (tel-tel) from the NIST SRE 2010 evaluation, in terms of EER and minDCF.

	male		female	
	EER	DCF	EER	DCF
2cov	4.47	0.44	4.75	0.48
length-norm + 2cov	**1.38**	**0.34**	**2.34**	**0.44**

5.7.2. LDA and two-covariance scoring

Table 2 shows the performance of LDA systems, all followed by two-covariance scoring. LDA based on $\mathbf{S}_b$ and $\mathbf{S}_w$ ($\text{LDA}_{\mathbf{S_B},\mathbf{S_W}}$) or based on $\mathbf{B}$ and $\mathbf{W}$ ($\text{LDA}_{\mathbf{B},\mathbf{W}}$), are applied after either length normalization or n iterations of variance-spectra based normalization (EFRnorm n iter). For all systems, the best LDA dimension reduction is dim = 80.

Compared to Table 1, both LDA techniques improve performance in terms of EER and minDCF. Standardization before normalization (EFRnorm 1 iter.) enhances performance for female evaluation but not for male evaluation. Two iterations of standardization and normalization (EFRnorm 2 iter.) achieve the best performance for male and female evaluation, but only if LDA based on $\mathbf{B}$ and $\mathbf{W}$ is carried out. Beyond two iterations, performance are stationary.
These experiments show that standardization and *i*-vector representation after two iterations, which is adapted to LDA based on $\mathbf{B}$ and $\mathbf{W}$, is the most robust of these LDA-based systems.

Table 2: Comparison of five systems based on LDA for the core extended condition 5 (tel-tel) from the NIST SRE 2010 evaluation, in terms of EER and minDCF. All systems use two-covariance scoring.

		male		female	
		EER	DCF	EER	DCF
length-norm	$\text{LDA}_{S_{\mathbf{B}},S_{\mathbf{W}}}$	**1.27**	**0.31**	2.27	0.38
EFRnorm 1 iter.	$\text{LDA}_{S_{\mathbf{B}},S_{\mathbf{W}}}$	1.36	0.33	2.29	0.39
EFRnorm 1 iter.	$\text{LDA}_{\mathbf{B},\mathbf{W}}$	1.36	0.30	**1.89**	**0.35**
EFRnorm 2 iter.	$\text{LDA}_{S_{\mathbf{B}},S_{\mathbf{W}}}$	1.30	0.32	2.30	0.39
EFRnorm 2 iter.	$\text{LDA}_{\mathbf{B},\mathbf{W}}$	**1.27**	**0.31**	**1.89**	**0.35**

5.7.3. PLDA

First row of Table 3 presents the results of the baseline Gaussian-PLDA model (according to Equation 7) following a length normalization. The rank of the speaker-matrix $\mathbf{\Phi}$ is 80 and the rank of the channel-matrix $\mathbf{\Gamma}$ is 600 (full-dimension). For the male evaluation, 100 iterations of the EM algorithm have been required to achieve optimum performance. For the female evaluation, this value increased to 300 iterations. As the speaker and channel matrices of the PLDA are randomly initialized, performance of the system are subject to a variability. Therefore, the displayed values are the mean performance yielded by repeating 10 times the same experiment. For male evaluation, EERs vary in the range of 1.15 to 1.33 and minDCFs from 0.31 to 0.34. For female evaluation, EERs vary in the range of 1.78 to 1.86 and minDCFs from 0.34 to 0.35.

Table 3: Comparison of three systems based on PLDA for the core extended condition 5 (tel-tel) from the NIST SRE 2010 evaluation, in terms of EER and minDCF.

	male		female	
	EER	DCF	EER	DCF
length-norm + G-PLDA	1.22	0.32	1.81	0.34
EFRnorm 2 iter. + G-PLDA	1.27	0.35	1.94	0.35
SphNnorm + G-PLDA	1.08	0.31	1.77	0.34
SphNnorm + G-PLDA init.	**1.04**	**0.29**	**1.73**	**0.33**

Second row of Table 3 gives the mean result of the best Gaussian PLDA system following standardization and length-normalization (EFR): two iterations of these normalization have been required and the optimal ranks of $\mathbf{\Phi}$ and $\mathbf{\Gamma}$ are 80 and 600. Compared to first row (a single length-normalization), no improvement of performance is observed.
Third row of Table 3 shows results of the same Gaussian PLDA system following two iterations of Spherical Nuisance normalization for both development and evaluation data. As before, the rank of speaker-matrix $\mathbf{\Phi}$ is 80 and the rank of channel-matrix $\mathbf{\Gamma}$ is 600. For both genders, 100 iterations of the PLDA algorithm have been required to achieve optimum performance. Speaker and channel matrices are randomly initialized and the displayed values are the mean performance yielded by repeating 10 times the same experiment. For male evaluation, EERs vary in the range of 1.04 to 1.13 and minDCFs from 0.29 to 0.32. For female evaluation, EERs vary in the range of 1.73 to 1.84 and minDCFs from 0.33 to 0.34. The preliminary Spherical Nuisance normalization improves mean performance for both genders, in terms of EER and minDCF.

Fourth row of Table 3 shows results of the same Gaussian PLDA system following two iterations of Spherical Nuisance normalization. Here, speaker and channel matrices have been initialized with the procedure described in paragraph 4.4.2. As before, the rank of speaker-matrix $\mathbf{\Phi}$ is 80 and the rank of channel-matrix $\mathbf{\Gamma}$ is 600. For the male evaluation, only 10 iterations of the EM algorithm have been required to achieve optimum performance, and only 2 iterations for the female evaluation. Beyond, performances are almost stationary for both genders. However, when performing only one iteration of EM, the system yields slightly worse EERs: 1.15% for male evaluation and 1.75% for female evaluation. This result shows that the PLDA modelling remains necessary to optimize *i*-vector based speaker recognition engines.

The Spherical Nuisance normalization preliminary applied makes the PLDA algorithm quickly converge to an optimization point which yields significant gain of performance, compared to baseline. In rows two and three of Table 3, two iterations of Spherical Nuisance normalization are performed. When performing only one iteration, performance are slightly worse (EER of 1.10% for male evaluation, 1.81% for female evaluation) but increasing the number of iterations doesn't improve much as it seems that this algorithm converges after 2 or 3 iterations. It achieves a relative improvement of the PLDA system of 14.8%-point on the EER for the male evaluation, and of 4.76%-point for the female evaluation. Note that with Lnorm+G-PLDA system (first row of Table 3), an equivalent initialization of speaker and channel matrices does not improve the performance of this system.

6. Conclusion

This paper presented two *i*-vector normalization techniques based on variance-spectra of training dataset. Those normalizations aim at adapting the *i*-vector representation to a speaker discriminative system.

In the case of Linear Discriminant Analysis followed by two-covariance scoring, we show that the preliminary algorithm of standardization and length-normalization can be iterated to enhance the optimization criterion of LDA and makes the *i*-vector distribution more Gaussian. We also suggest the best within- and between-covariance matrices to use.
In the case of Probabilistic Linear Discriminant Analysis, the lower-bound maximization EM algorithm is dependent of the initial representation of data, i.e. algorithm converges to a local maximum-likelihood point in the optimization landscape of the PLDA. In this context, Spherical Nuisance normalization, also an iterative algorithm, moves data from a local optimum to a better one in a different part of the parameter space.

We show that after performing two iterations of Spherical Nuisance normalization and initializing the PLDA matrices according to the new *i*-vector representation, our system achieves optimum performance after only a reduced number of EM iterations (between 2 and 10). Nevertheless, EM estimation of the PLDA parameters remains necessary as the system yields slightly worse performance when running only one iteration of EM (1.15% EER for male evaluation and 1.75% EER for female evaluation).

While conducting this work, we observed that, after applying two iterations of Spherical Nuisance normalization, removing the channel matrix from the PLDA model (channel rank of $\mathbf{\Gamma} = 0$) and keeping only a diagonal noise covariance matrix

does not degrade the best performance reported in this paper. After Spherical Nuisance normalization, the simplest speaker-factor analysis model (low rank speaker subspace and diagonal noise covariance matrix) is able to optimize the performance of our *i*-vector-based speaker recognition engine. This needs to be analyzed and clarified by future investigations.

The spectral graph of initial development data is a simple and useful visualization tool for speaker recognition. We observed that the spectral distributions in regard to the latent speaker variable may differ from an *i*-vector extraction configuration to another so as the performance gap between speaker recognition systems (LDA, PLDA, ...). Analysing this graph can help to assess the adequacy of an *i*-vector extraction configuration to these speaker recognition systems.

7. Acknowledgments

We thank Pavel Matějka for providing the *i*-vectors of BUT. Special thanks to Niko Brümmer for the matlab PLDA system and for helpful discussions throughout the NIST-SRE2011 Workshop in Atlanta.

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] Najim Dehak, Read Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," in *International Conference on Speech Communication and Technology*, 2009, pp. 1559–1562.

[3] Mitchell McLaren and David A. Van Leeuwen, "Source-normalised and weighted LDA for robust speaker recognition using I-Vectors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2011, pp. 5456–5459.

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

[5] 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.

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

[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-Francois 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] Niko Brummer, Jesus Villalba, and Eduardo Lleida, "Fully Bayesian likelihood ratio vs i-vector length normalization in speaker recognition systems," in *NIST SRE Analysis Workshop*, 2011.

[10] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep., CRIM, 2005.

[11] Simon J.D. Prince, *Computer Vision: Models Learning and Inference*, Cambridge University Press, 2012, In press.

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

[13] 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.

[14] William M. Campbell, Douglas E. Sturim, Douglas A. Reynolds, and Alex Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, vol. 1, pp. 97–100.

[15] Michael E. Tipping and Christopher M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[16] 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.

[17] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Toulouse, France, May 2006, pp. 325–328.

[18] N. Brummer, L. Burget, P. Kenny, P. Matejka, E. Villiers de, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 2010," in *Proc. NIST 2010 Speaker Recognition Evaluation*. 2010, pp. 1–20, Brno University of Technology.