Exploring some limits of Gaussian PLDA modeling for i-vector distributions

Pierre-Michel Bousquet, Jean-François Bonastre, Driss Matrouf

Exploring some limits of Gaussian PLDA modeling for i-vector distributions

Pierre-Michel Bousquet, Jean-François Bonastre, Driss Matrouf

University of Avignon - LIA, France

{pierre-michel.bousquet, driss.matrouf, jean-francois.bonastre}@univ-avignon.fr

Abstract

Gaussian-PLDA (G-PLDA) modeling for i-vector based speaker verification has proven to be competitive versus heavy-tailed PLDA (HT-PLDA) based on Student's t-distribution, when the latter is much more computationally expensive. However, its results are achieved using a length-normalization, which projects i-vectors on the non-linear and finite surface of a hypersphere. This paper investigates the limits of linear and Gaussian G-PLDA modeling when distribution of data is spherical. In particular, assumptions of homoscedasticity are questionable: the model assumes that the within-speaker variability can be estimated by a unique and linear parameter. A non-probabilistic approach is proposed, competitive with state-of-the-art, which reveals some limits of the Gaussian modeling in terms of goodness of fit. We carry out an analysis of residue, which finds out a relation between the dispersion of a speaker-class and its location and, thus, shows that homoscedasticity assumptions are not fulfilled.

1. Introduction

Introduced in [1], the i-vector representation of speech utterances provides a feature vector of low dimension (less than 600), independent of the length of the utterance. A speaker verification system using these features and a simple classifier outperforms the previous approaches, like Joint Factor Analysis (JFA) [2, 3]. The Bayesian generative model designed to provide a consistent probabilistic framework for i-vectors is the PLDA, introduced in [4] for face recognition and adapted for speaker verification in [5, 6]. The first approach, Gaussian PLDA (G-PLDA), assumed that speaker and residual components have Gaussian distributions. To deal with severe within-class distortions and increase the robustness to outliers, a specific PLDA approach for modeling i-vector distributions was introduced in [5], based on Student's t-distribution and referred to as heavy-tailed PLDA. Student's t-distribution has heavier tails compared to the exponentially-decaying tails of a Gaussian, providing a better representation of the speaker and residual subspaces, including the outliers [7].

More recently, a pre-conditioning before any i-vector Gaussian modeling has been introduced [8, 9]. I-vectors are whitened and length-normalized, projecting data onto a spherical surface. It is shown [8] that this technique improves Gaussianity of the i-vectors. Also, it involves a number of properties related to intersession compensation [9, 10] and contributes strongly to the model efficiency [11]. Using these normalized i-vectors, performance of the Gaussian and Heavy-Tailed PLDA models are comparable, the former being much faster both in training and in testing.

The goal of this paper is to study and reveal some limits of the G-PLDA modeling when applied in the i-vector field. After length-normalization (LN), data are closer to Gaussianity assumptions but they lie on the non-linear and finite surface of a hypersphere. This point caught our attention. We consider that this discrepancy between a non-linear data-manifold and a linear framework could reveal some limits of G-PLDA for modeling i-vectors distributions. In particular, homoscedasticity of the within-class variability is questionable, when data distribution is spherical: PLDA assumes that it exists a unique (thus speaker-independent) and linear parameter of within-class variability. We consider that spherical distributions cannot fulfill such an assumption.

First, we propose a new approach for estimating the PLDA metaparameters. Based on the pre-conditioning procedure and a deterministic estimate of parameters, this non-probabilistic approach helps to assess ability of a maximum likelihood (ML)-based approach to improve the goodness of fit, and reveals some lack of compliance of i-vector distributions with G-PLDA assumptions.

Second, we analyze homoscedasticity of G-PLDA modeling after LN. Any significant relation between the within-class variability of a speaker and another class parameter would violate the assumption of homoscedasticity. We compute posterior likelihoods of speaker and residual factors for each speaker of our development corpus and confirm the concern about linearity and homoscedasticity assumptions.

2. Gaussian-PLDA

2.1. Modeling

Introduced in [4], Gaussian Probabilistic Linear Discriminant Analysis (PLDA) is a generative i-vector model. The most common PLDA model in speaker verification assumes that each p-dimensional i-vector $\mathbf{w}$ of a speaker s can be decomposed as

$$\mathbf{w} = \mu + \mathbf{\Phi}\mathbf{y}_s + \varepsilon \tag{1}$$

The mean vector μ is a global offset, $\mathbf{\Phi}$ is a $p \times r$ matrix whose columns provide a basis for the eigenvoice subspace, the r-dimensional vector y_s is the speaker factor and ϵ is the residual term. Therefore, the speaker-specific part $\mu + \mathbf{\Phi}\mathbf{y}_s$ represents the between-speaker variability and is assumed to be tied across all utterances of the same speaker. G-PLDA assumes that all latent variables are statistically independent. Standard normal prior is assumed for the speaker factor y_s and normal prior for the residual term ϵ with mean 0 and full covariance matrix Λ. The maximum of likelihood (ML) point estimates of the model parameters are obtained from a large collection of development

data using an expectation-maximization (EM) algorithm as in [4].

Note that this approach is the simplified version of the original PLDA model: eigenchannels have been removed, as proposed in [5], since i-vectors are of sufficiently low dimension (400 to 600 usually) and since PLDA modeling does not show major improvement with eigenchannels.

2.2. Verification score

The speaker verification score, given the two i-vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ involved in a trial, is the likelihood-ratio

$$score = \log \frac{P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{tar}\right)}{P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{\text{non}}\right)} \tag{2}$$

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. Likelihoods of (2) can be decomposed as

$$\begin{aligned} P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{tar}\right) &= \int_y \prod_{i=1,2} P\left(\mathbf{w}_i | \mu + \mathbf{\Phi} y, \mathbf{\Lambda}\right) P\left(y | \mathbf{0}, \mathbf{I}\right) dy \\ P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{\text{non}}\right) &= \prod_{i=1,2} \int_y P\left(\mathbf{w}_i | \mu + \mathbf{\Phi} y, \mathbf{\Lambda}\right) P\left(y | \mathbf{0}, \mathbf{I}\right) dy \end{aligned} \tag{3}$$

For the G-PLDA case, all the marginal likelihoods are Gaussian and the score (2) can be evaluated analytically [12]

$$\begin{aligned} score\left(\mathbf{w}_1, \mathbf{w}_2\right) = &\log \mathcal{N}\left(\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{tot} & \mathbf{\Sigma}_{ac} \\ \mathbf{\Sigma}_{ac} & \mathbf{\Sigma}_{tot} \end{bmatrix}\right) \\ &- \log \mathcal{N}\left(\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{tot} & 0 \\ 0 & \mathbf{\Sigma}_{tot} \end{bmatrix}\right) \end{aligned} \tag{4}$$

where $\mathcal{N}()$ denotes normal density function and

$$\begin{cases} \mathbf{\Sigma}_{tot} = \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Lambda} \\ \mathbf{\Sigma}_{ac} = \mathbf{\Phi}\mathbf{\Phi}^t \end{cases} \tag{5}$$

The matrix $\mathbf{\Sigma}_{tot}$ consists of a probabilistic total covariance matrix, adding variabilities of the decomposition to take into account underlying assumptions of the probabilistic model. The matrix $\mathbf{\Sigma}_{ac}$ is the covariance matrix of speaker-dependent factor (across-class). It is worth noting that computing this score with test vectors depends solely on the knowledge of metaparameters $\mathbf{\Sigma}_{tot}$ and $\mathbf{\Sigma}_{ac}$ (speaker and residual factors of i-vectors of test have not to be computed).

3. Conditioning

A pre-processing step applied before any i-vector modeling has been introduced in [8, 9], following WCC Normalization and cosine-scoring technique of [2]. I-vectors are whitened and length-normalized, in order to make them more Gaussian. The most commonly used whitening technique is a standardization and the transformation applied to an i-vector $\mathbf{w}$ can be summarized as follows:

$$\mathbf{w} \leftarrow \frac{\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w} - \mu\right)}{\left\|\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w} - \mu\right)\right\|} \tag{6}$$

First, data are standardized according to the mean μ and a variability matrix $\mathbf{A}$ of a training corpus. Then they are length-normalized, confining the i-vectors to the hypersphere of unit radius. Parameters are computed for the i-vectors present in the training corpus and applied to test i-vectors. The matrix $\mathbf{A}$ can be the total covariance matrix $\mathbf{\Sigma}$ or, as proposed in [2, 10], the within-class covariance matrix. We denote in this article $\mathbf{L\Sigma}$, $\mathbf{LW}$ these transformations (Length-normalization of standardized vectors according to $\mathbf{\Sigma}$ or $\mathbf{W}$).

In [8], it is shown that this technique improves Gaussianity of i-vectors. It reduces the gap between the underlying assumptions on the data distribution and the real distribution and also reduces the dataset shift between development and trial i-vectors [13]. Performance of a G-PLDA system with this preconditioning is competitive versus the HT-PLDA, when the latter shows a significant higher complexity. It is shown in [11] that this procedure mainly contributes to model efficiency. As proposed in [9], its two steps can be iterated. As a result, i-vectors tend to be simultaneously $\mathbf{A}$-standardized and length-normalized (magnitude 1), involving a number of properties related to intersession compensation. Some of them are detailed in [9, 10].

4. Proposed approach

4.1. Deterministic estimation of G-PLDA parameters

G-PLDA scoring of (4) is based solely on the determination of metaparameters $\mathbf{\Sigma}_{tot}$ and $\mathbf{\Sigma}_{ac}$. Factors $\mathbf{y}_s$ and ε have not to be computed. Given a training corpus $\mathcal{T}$ comprised of i-vectors of S speakers, we denote by $\mathbf{\Sigma}$ the total covariance matrix of $\mathcal{T}$ and $\mathbf{B}$ the between-class covariance matrix of $\mathcal{T}$ defined by

$$\mathbf{B} = \sum_{s=1}^{S} \frac{n_s}{n} \left(\mathbf{w}_s - \mu\right)\left(\mathbf{w}_s - \mu\right)^t \tag{7}$$

where n_s is the number of utterances for speaker s, n is the total number of utterances of $\mathcal{T}$, $\mathbf{w}_s$ is the mean i-vector of s and μ represents the overall mean of $\mathcal{T}$. Let $\mathbf{W}$ denote the within-class covariance matrix of $\mathcal{T}$, equal to $\mathbf{\Sigma} - \mathbf{B}$.
Singular value decomposition of $\mathbf{B}$ provides a matrix $\mathbf{P}$ whose columns are the eigenvectors of $\mathbf{B}$ sorted by decreasing order of eigenvalues and a corresponding eigenvalue diagonal matrix $\mathbf{\Delta}$, such that $\mathbf{B} = \mathbf{P}\mathbf{\Delta}\mathbf{P}^t$. Given a rank $r < p$, the r-range principal between-class variability can be summarized into the $p \times p$ covariance matrix $\mathbf{B}_{1:r}$ defined by

$$\mathbf{B}_{1:r} = \mathbf{P}_{1:r}\mathbf{\Delta}_{1:r}\mathbf{P}_{1:r} \tag{8}$$

where $\mathbf{P}_{1:r}$ denotes the $p \times r$ matrix comprised of the first r columns of $\mathbf{P}$ and $\mathbf{\Delta}_{1:r}$ the $r \times r$ top-left block of $\mathbf{\Delta}$.

To estimate G-PLDA metaparameters $\mathbf{\Phi}$ and $\mathbf{\Lambda}$, we propose the following procedure:

- Apply $\mathbf{LW}$-Conditioning of Equation (6), eventually iterated.
- Replace EM-ML based estimates of speaker and residual parameters by the following direct expressions:

$$\begin{cases} \mathbf{\Sigma}_{tot} &= \mathbf{\Sigma} \\ \mathbf{\Sigma}_{ac} &= \mathbf{B}_{1:r} \end{cases} \tag{9}$$

4.2. Justification of the approach and conformity to G-PLDA assumptions

Ignoring the probabilistic constraints of G-PLDA, we presume that the most relevant metaparameters $\mathbf{\Sigma}_{tot}$ and $\mathbf{\Sigma}_{ac}$ (by only considering information of the training set and assuming that

exists a r-range eigenvoices subspace) are the total covariance matrix $\mathbf{\Sigma}$ and the part $\mathbf{B}_{1:r}$ of within-speaker variability due to the principal axes of $\mathbf{B}$. In order to assess conformity of this non-probabilistic approach to G-PLDA assumptions (standardity of $\mathbf{y}_s$, statistical independence between $\mathbf{y}_s$ and ε, Gaussianity of factors), its factors have first to be expressed. As matrix $\mathbf{P}$ of (8) is orthogonal, we use the equality

$$\mathbf{I} = \mathbf{P}\mathbf{P}^t = \mathbf{P}_{1:r}\mathbf{P}_{1:r}^t + \mathbf{P}_{(r+1):p}\mathbf{P}_{(r+1):p}^t \tag{10}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and $\mathbf{P}_{(r+1):p}$ is the $p \times (p-r)$ matrix comprised of the last $(p-r)$ columns of $\mathbf{P}$, to decompose an i-vector $\mathbf{w}$ of a speaker s as follows:

$$\begin{aligned}\mathbf{w} = \mu &+ \left[\mathbf{P}_{1:r}\mathbf{P}_{1:r}^t\left(\mathbf{w}_s - \mu\right)\right] \\ &+ \left[\mathbf{P}_{(r+1):p}\mathbf{P}_{(r+1):p}^t\left(\mathbf{w}_s - \mu\right) + \mathbf{w} - \mathbf{w}_s\right]\end{aligned} \tag{11}$$

where $\mathbf{w}_s$ is the mean i-vector of speaker s. Factor and matricial parameters can be expressed as

$$\begin{cases}\mathbf{y}_s &= \mathbf{\Delta}_{1:r}^{-\frac{1}{2}}\mathbf{P}_{1:r}^t\left(\mathbf{w}_s - \mu\right) \\ \mathbf{\Phi} &= \mathbf{P}_{1:r}\mathbf{\Delta}_{1:r}^{\frac{1}{2}} \\ \varepsilon &= \mathbf{P}_{(r+1):p}\mathbf{P}_{(r+1):p}^t\left(\mathbf{w}_s - \mu\right) + \mathbf{w} - \mathbf{w}_s\end{cases} \tag{12}$$

A straightforward computation shows that the covariance matrix of $\mu + \mathbf{\Phi}\mathbf{y}_s$ is equal to $\mathbf{B}_{1:r}$, as desired. Note that computing factors $\mathbf{y}_s$ and ε requires the knowledge of the speaker mean-vector $\mathbf{w}_s$, which is unknown for test vectors. G-PLDA assumes that the speaker factor $\mathbf{y}_s$ is standardized and that all latent variables are statistically independent. Considering solely data from the development corpus, a straightforward computation shows that $\mathbf{y}_s$, as defined in (12), has a mean equal to 0 and a covariance matrix equal to the identity matrix $\mathbf{I}$. Moreover, to obtain covariance matrices of (5) from (3), only nullity of the covariance between $\mathbf{y}_s$ and ε (a necessary condition of independence) is required. As $\mathbf{y}_s$ and ε are centered, this value is equal to

$$\begin{aligned}\mathbf{E}\left[\mathbf{y}_s\varepsilon^t\right] = &\ \mathbf{\Delta}_{1:r}^{-\frac{1}{2}}\mathbf{P}_{1:r}^t\mathbf{B}\mathbf{P}_{(r+1):p}\mathbf{P}_{(r+1):p}^t \\ &+ \mathbf{E}\left[\mathbf{\Delta}_{1:r}^{-\frac{1}{2}}\mathbf{P}_{1:r}^t\left(\mathbf{w}_s - \mu\right)\left(\mathbf{w} - \mathbf{w}_s\right)^t\right]\end{aligned} \tag{13}$$

As $\mathbf{P}_{1:r}^t\mathbf{B}\mathbf{P}_{(r+1):p} = \mathbf{0}$, the first term is null. The second term is equal to $\mathbf{\Delta}_{1:r}^{-\frac{1}{2}}\mathbf{P}_{1:r}^t\mathbf{E}_s\left[\left(\mathbf{w}_s - \mu\right)\mathbf{E}_{\mathbf{w}\in s}\left[\mathbf{w} - \mathbf{w}_s\right]\right] = \mathbf{0}$, since $\mathbf{E}_{\mathbf{w}\in s}\left[\mathbf{w} - \mathbf{w}_s\right] = 0$.

Normal priors are assumed for speaker and session factors. Gaussian shape of factors can be estimated by analyzing square-norms of their standardized versions [8]. Indeed, the square-length of vectors drawn from a standard Gaussian distribution follows a χ^2 distribution with number of degrees of freedom equal to the dimension of the vector. Figure 1 presents histograms of square-norm distributions of standardized factors $\mathbf{y}_s$ (left panel) and ε (right panel). As the model assumes $y_s \sim \mathcal{N}(0, \mathbf{I})$ and $\varepsilon \sim \mathcal{N}(0, \mathbf{\Lambda})$, the standardized versions of y_s and ε are assumed to be y_s and $\mathbf{\Lambda}^{-\frac{1}{2}}\varepsilon$. Thus, Figure 1 displays $\|y_s\|^2$ (left panel) and $\left\|\mathbf{\Lambda}^{-\frac{1}{2}}\varepsilon\right\|^2$ (right panel) for the development set (thick line) and the evaluation set NIST SRE10 male telephone (thin line). Also, Figure 1 depicts the pdfs of χ^2 distributions with r and p degrees of freedom (dashed line). Three systems are considered:

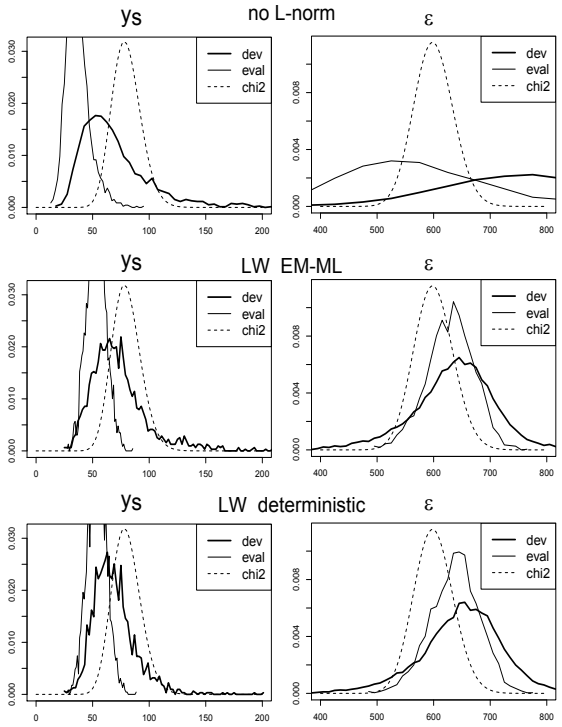

Figure 1: Histograms of the square norm distributions of G-PLDA factors. Graphs also depict the pdfs of χ^2 distributions with r degrees of freedom for $\mathbf{y}_s$ and p for ε.

1. No LN procedure has been carried out before G-PLDA modeling. Development and evaluation i-vectors are centered only, by subtracting the mean vector of the development set,
2. **LW**-normalization (2 iterations) followed by EM-ML based G-PLDA modeling.
3. **LW**-normalization (2 iterations) followed by the proposed deterministic approach.

Three observations could be gathered from this analysis. First, G-PLDA factors extracted from initial i-vectors have a non-symmetric and heavy-tailed distribution. Moreover, a severe dataset shift between development and evaluation i-vectors occurs. Second, after LN, distributions better match the χ^2 distributions. In particular, non-Gaussian behavior and dataset shift of the residual factor are significantly reduced. Lastly, histograms of the two post-LN approaches are similar, for both factors. This last result shows that the deterministic estimate improves the Gaussian shape of factors and reduces the mismatch between development and evaluation datasets in a similar manner to ML technique.

5. Experimental setup

Evaluation was performed on NIST SRE08 core conditions 6 and 7 male only, corresponding to telephone-telephone (respectively All and English-only) enrollment-verification trials, SRE10 extended condition 5 male only, corresponding to telephone-telephone enrollment-verification trials and SRE12 core conditions 4 and 5 male only, corresponding to telephone

Table 1: Comparison of performance between deterministic and EM-ML approaches, without or with pre-conditioning.

Conditioning	No length-norm.				$\mathbf{L\Sigma}$		$\mathbf{LW}$			
Parameter estimation	EM-ML		deterministic		EM-ML		EM-ML		deterministic	
NIST-SRE male evaluation	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)
2008 tel. all (*det 6*)	0.317	6.52	0.342	7.44	0.277	5.24	0.279	4.92	0.289	4.92
2008 tel. eng. (*det 7*)	0.180	3.19	0.176	4.10	0.121	1.82	0.116	1.59	0.128	1.59
2010 tel. extended (*det 5 ext*)	0.599	5.97	0.601	6.76	0.496	2.40	0.483	2.28	0.487	2.28
2012 tel. with added noise (*det 4*)	0.463	5.37	0.463	5.30	0.438	3.53	0.412	3.21	0.452	3.21
2012 tel. noisy environment (*det 5*)	0.673	7.27	0.639	19.88	0.419	2.48	0.394	2.24	0.433	2.40

with added noise and noisy environment. These last two conditions enable to evaluate the proposed approach on noisy utterances. EM-ML-based and deterministic estimations of G-PLDA metaparameters have been evaluated with LIA configuration : the feature extraction, 512-components GMM-UBM functionalities and i-vector extraction configurations for NIST SRE08, SRE10 evaluations are described in [11]. For SRE12 evaluation, a gender-dependent LIA i-vector extractor was trained on data from NIST SRE04,05,06,08,10, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, giving 1879 speakers in 25024 segments of speech. PLDA model was trained by merging two datasets: first, i-vectors of the same dataset than for extraction, second, i-vectors of target speakers SRE12 and their noisy versions. For each clean segment of a target speaker, two noisy versions (6dB and 15dB) are generated, following the method suggested in [14]. Channel factor ε of G-PLDA is kept full and speaker factor is fixed to its optimal value in terms of performance. For SRE12 multi-cut enrollment, scoring is performed by averaging all the enrollment i-vectors of each target speaker after LN. The size of i-vectors is 400.

6. Results

Table 1 summarizes results of the tests performed on the NIST SRE conditions described above, in terms of Equal Error Rate and normalized minimum Detection Cost Function as defined by NIST for SRE08 and SRE10 evaluations (DCF 2010 is applied for SRE12). Evaluations have been carried without or with LN, and with EM-ML or deterministic approach. Also, results of $\mathbf{L\Sigma}$-conditioning (standardization according to the total covariance matrix followed by LN) are reported in columns 5,6 of Table 1, since this transformation is the most commonly implemented. Comparison of systems without or with conditioning recalls the necessity of LN to handle i-vectors in a Gaussian framework: it yields 49% and 24% relative improvements in average EER and minDCF. Comparison of systems with $\mathbf{L\Sigma}$ or $\mathbf{LW}$ conditioning recalls the relevance of the latter, remarked in [10]. $\mathbf{LW}$-conditioning leads to 8.5% and 3.6% relative improvements in average EER and minDCF.

Without LN, the deterministic approach degrades performance. Accuracy degradation can even be considerable (see EER of last condition, SRE12 noisy environment). EM-ML approach brings the expected improvement in performance in a probabilistic context. However, after $\mathbf{LW}$ conditioning, deterministic approach yields similar performance across the reported conditions, relative to EM-ML, in terms of EER as minDCF (with a slight gap in terms on minDCF, except for last conditions in noisy environment). It even outperforms, in terms of EER, the commonly used $\mathbf{L\Sigma}$-based second system.

To better assess independence of these results to the upstream configuration, we carried out the same NIST SRE10 evaluation condition with i-vectors we already used in [10], provided by BUT laboratory [15]. EERs and minDCFs of EM-ML vs deterministic approach are respectively equal to $(1.04\%, 0.31)$ vs $(1.04\%, 0.30)$ for male set, and to $(1.78\%, 0.33)$ vs $(1.75\%, 0.33)$ for female set, confirming the previous observations.

Results of the deterministic approach recall the relevance of $\mathbf{LW}$ conditioning. After this procedure, i-vectors fit better the Gaussian model. But they also reveal some limits of G-PLDA statistical modeling, in terms of goodness of fit. The next section addresses this issue.

7. Limits of Gaussian linear modeling for i-vectors

While Gaussian PLDA applied to whitened and length-normalized i-vectors has shown its efficiency, previous outcomes raise an issue which relates to the relevance of a linear Gaussian approach with i-vectors. Model training consists of fitting a parametric model to the training set. The $\mathbf{LW}$-conditioning and deterministic estimation are built solely on covariance parameters of this latter, ignoring the aims of Gaussianity and generalization to new observations, in particular from unknown speakers. Moreover, the eigenvoice basis is orthogonal, which can limit the accuracy of the estimate. By increasing metaparameter likelihoods, EM-ML algorithm should enhance the Gaussian modeling. Previous results show that EM-ML based approach does not bring the expected improvement of performance. After LN, data are closer to probabilistic assumptions but they lie on the non-linear and finite surface of a hypersphere. We consider that this inconsistency between a non-linear data-manifold and a linear Gaussian framework could reveal some limits of G-PLDA modeling for i-vectors. In particular, the assumption that there exists a linear and speaker-independent metaparameter $\mathbf{\Lambda}$ of within-class variability is questionable.

7.1. Heteroscedasticity of residue

G-PLDA model assumes homoscedasticity of speaker-classes (i.e., that their distributions share the common covariance matrix $\mathbf{\Lambda}$). It is noticed in [16] that two Gaussian distributions laying on a p-sphere can only be homoscedastic if the mean feature vector of one class is the same as that of the others up to a sign [1]. Equality of covariance, when model is homoscedastic, assumes that the modeling errors (the residue between the actual variability of a class s and the metaparameter $\mathbf{\Lambda}$) are uncorrelated, normally distributed and that their variances do not vary with the effects being modeled. Any significant relation between this residue and another class parameter would violate the assumption of random prior. Such a relation could penalize the goodness of fit of G-PLDA modeling.

[1] except when the common covariance matrix is the identity matrix or a multiple of the identity matrix.

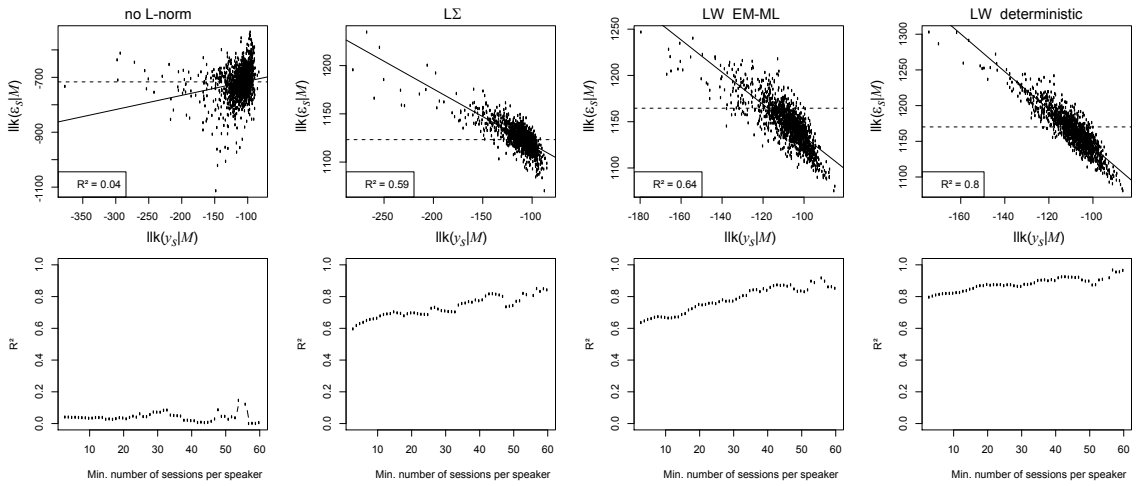

Figure 2: Likelihoods of speaker and residual factors computed on PLDA development speaker-classes, without or with length-normalization and using the different approaches. Figure also depicts the coefficient of determination $\mathbf{R}^2$ between the likelihoods as a function of the minimal number of utterances per speaker.

7.1.1. Property of length-normalized vectors

First, we present a simple relation between the variance of a training speaker-class and the position of its centroïd, once all the i-vectors are length-normalized.

Let $\{\mathbf{w}_i^s\}_{i=1}^{n_s}$ denote the set of n_s i-vectors of a speaker s and $\overline{\mathbf{w}_s}$ its mean vector. The variance of this sample is equal to:

$$\frac{1}{n_s}\sum_{i=1}^{n_s}\left\|\mathbf{w}_i^s-\overline{\mathbf{w}_s}\right\|^2=\frac{1}{n_s}\sum_{i=1}^{n_s}\left\|\mathbf{w}_i^s\right\|^2-\left\|\overline{\mathbf{w}_s}\right\|^2 \tag{14}$$

$$=1-\left\|\overline{\mathbf{w}_s}\right\|^2 \tag{15}$$

as all the vectors have a norm of 1. This equality shows that a link exists between the position of a speaker-class and its variance. G-PLDA model assumes that for any speaker and, thus, for any speaker-dependent term $\boldsymbol{\mu}+\boldsymbol{\Phi}\mathbf{y}_s$, the within-class variability can be expressed by a unique speaker-independent covariance matrix $\boldsymbol{\Lambda}$. Equation (15) shows a dependency between the dispersion of a class and its position in regard to the origin, expressed by $\left\|\overline{\mathbf{w}_s}\right\|^2$. To better assess this phenomenon and observe whether or not it occurs with G-PLDA class-factors y_s and ε, the next paragraph analyzes heteroscedasticity of the residue ε.

7.1.2. Heteroscedasticity

Given a G-PLDA model $\mathcal{M}=(\mu,\boldsymbol{\Phi},\boldsymbol{\Lambda})$ computed on a development corpus $\mathcal{T}$, we denote by $\{\mathbf{w}_{s,i}\}_{i=1}^{n_s}$ the collection of n_s i-vectors from speaker s of $\mathcal{T}$ and $\mathbf{y}_s$, $\{\varepsilon_{s,i}\}_{i=1}^{n_s}$ the corresponding factors provided by $\mathcal{M}$. The model assumes that, for $i=1$ to n_s

$$\mathbf{w}_{s,i}=\mu+\boldsymbol{\Phi}\mathbf{y}_s+\varepsilon_{s,i} \tag{16}$$

We denote by $llk\left(\mathbf{y}_s|\mathcal{M}\right)$ the posterior log-likelihood of $\mathbf{y}_s$ according to $\mathcal{M}$, equal to

$$llk\left(\mathbf{y}_s|\mathcal{M}\right)=\log\mathcal{N}\left(\mathbf{y}_s|\,0,\mathbf{I}\right)$$

$$=-\frac{r}{2}\log\left(2\pi\right)-\frac{1}{2}\left\|\mathbf{y}_s\right\|^2 \tag{17}$$

Let $llk\left(\varepsilon|\mathcal{M}\right)$ denote the average posterior log-likelihood of the residue over all observations of s according to $\mathcal{M}$, given by

$$llk\left(\varepsilon|\mathcal{M}\right)=\frac{1}{n_s}\sum\nolimits_i\log\mathcal{N}\left(\mathbf{w}_{s,i}|\mu+\boldsymbol{\Phi}\mathbf{y}_s,\boldsymbol{\Lambda}\right)$$

$$=-\frac{p}{2}\log 2\pi-\frac{1}{2}\log\left|\boldsymbol{\Lambda}\right|-\frac{1}{2n_s}$$

$$\sum\nolimits_i\left(\mathbf{w}_{s,i}\ -\mu-\boldsymbol{\Phi}\mathbf{y}_s\right)^t\boldsymbol{\Lambda}^{-1}\left(\mathbf{w}_{s,i}\ -\mu-\boldsymbol{\Phi}\mathbf{y}_s\right) \tag{18}$$

Figure 2 top panel displays $llk\left(\mathbf{y}_s|\mathcal{M}\right)$ and $llk\left(\varepsilon|\mathcal{M}\right)$ of speaker-classes from our development corpus $\mathcal{T}$. Analysis has been carried out for systems 1, 3, 4, 5 of section 6: without normalization then ($\mathbf{L\Sigma}$, EM-ML), ($\mathbf{LW}$, EM-ML) and ($\mathbf{LW}$, deterministic). Without normalization (first graph), no relation occurs between the two likelihoods. The coefficient of determination $\mathbf{R}^2$, which indicates how well data points fit a line, is indicated in the figure and close to 0. The least-square line is displayed and close to be horizontal. If variances of $\boldsymbol{\Lambda}$ and ε exactly match, $llk\left(\varepsilon|\mathcal{M}\right)$ is equal to $-\frac{p}{2}\log 2\pi-\frac{1}{2}\log\left|\boldsymbol{\Lambda}\right|-\frac{p}{2}$. This value is plotted by a horizontal dashed line. Figure 2 shows that the series $llk\left(\varepsilon|\mathcal{M}\right)$ corresponds to its theoretical value before length-normalization.

After any LN technique (graphs 2 to 4), a strong linear relation occurs between the two likelihoods. All the $\mathbf{R}^2$ exceed 0.59. It can be objected that this result is due to the less informative training speaker-classes (training speakers with low amount of utterances). Figure 2 bottom panel displays the $\mathbf{R}^2$ series between the likelihoods (y-axis), only for speakers with a minimal number of utterances (x-axis). The intensity of the relation remains high and even slightly increases with the minimal amount of utterances, thus for the main training classes in terms of information.

This significant relation between likelihoods of G-PLDA class-factors $\mathbf{y}_s$ and ε_s entails heteroscedasticity of the residue. The likelihood $llk\left(\varepsilon|\mathcal{M}\right)$ is equal, up to a constant, to a linear function of $\boldsymbol{\Lambda}^{-1}$. This relation shows that the residual variance is depending on the class position, which prevents an overall linear parameter $\boldsymbol{\Lambda}$ to be optimal. It should be replaced by

a class-dependent parameter $\mathbf{\Lambda}_s$, at least taking into account $\|\mu + \mathbf{\Phi}\mathbf{y}_s\|$ to fit to the actual class-dispersion.

8. Conclusion

This paper investigates the limits of linear and homoscedastic Gaussian PLDA modeling applied to spherical i-vector distributions. First, we propose a new deterministic approach for estimating PLDA metaparameters. This non-probabilistic approach enables to assess ability of a maximum likelihood (ML)-based approach to improve the goodness of fit. It turns out that results of this non-probabilistic approach are similar to those of the EM-ML approach in terms of EER, and close in terms of minimal DCF. We consider the inefficiency of the latter approach as an empirical evidence towards lack of compliance with the model assumptions. Therefore, we carried out an analysis of homoscedasticity, revealing a dependency between the dispersion of a class and its position in regard to the origin. This significant relation violates clearly the assumption of homoscedasticity.

Instead of dealing with severe within-class distortions and outlying observations by proposing non-Gaussian prior, as done in heavy-tailed PLDA, Gaussian PLDA draws on length-normalization to bring data onto a surface of high Gaussian likelihood. Length-normalization makes the development and trial i-vector distributions more similar and more Gaussian shaped, but i-vectors lie on the non-linear and finite surface of a hypersphere. As noticed in [16], assumptions of linearity and homoscedasticity cannot be fulfilled when data share a common norm and, thus, G-PLDA model cannot be optimal.

Also, these findings confirm the benefit of the **LW** conditioning. By standardizing data according to a within-class variability, techniques like WCCN [2] or **LW** [10] move this variability towards an isotropic model $\sigma\mathbf{I}$ [17]. As remarked in [10], this model does not favor any principal direction of session variability nor dependence between directions and, thus, alleviates the concern about heteroscedasticity.

9. Perspectives

Advances in i-vector modeling could be achieved by pursuing the HT-PLDA approach, which attempts to find out adequate priors, or by preserving length-normalization (as this technique has underlined importance of the directional information in the i-vector space) then modeling such representations using spherical distributions. The difficulty associated with spherical representations prompts researchers to model spherical data using Gaussian distributions. Approximations of covariance matrix have been proposed, based on "unscented transforms" [18] or first order Taylor expansion of the non-linear length-normalization [19]. However, the analysis of homoscedasticity presented in this paper shows that the overall within-class variability parameter should be replaced by a class-dependent parameter taking into account the local position of the class to fit to its actual dispersion. Such a non-linear model induces a complex density by passing the within-class variability parameter through a non-linear function. This model is harder to work with in practice. It requires further research, to marginalize over the hidden variables and provide a low time consuming scoring phase.

10. References

[1] 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.

[2] 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.

[3] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[4] 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.

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

[6] Mohammed Senoussaoui, Patrick Kenny, Niko Brummer, Edward de Villiers, and Pierre Dumouchel, "Mixture of PLDA models in i-vector space for gender independent speaker recognition," in *International Conference on Speech Communication and Technology*, 2011.

[7] Zia Khan and Frank Dellaert, "Robust generative subspace modeling: The subspace t distribution," Tech. Rep., GVU Center, College of Computing, Georgia, 2004.

[8] 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.

[9] 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.

[10] 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.

[11] Pierre-Michel Bousquet, Jean-François Bonastre, and Driss Matrouf, "Identify the benefits of the different steps in an i-vector based speaker verification system," in *CIARP*, Springer, Ed., 2013, vol. Part II, pp. 278–285.

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

[13] Carlos Vaquero, "Dataset Shift in PLDA based Speaker Verification," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2012.

[14] R. Saeidi and al., "I4u submission to NIST sre 2012: A large-scale collaborative effort for noise-robust speaker verification," in *International Conference on Speech Communication and Technology*, 2013.

[15] 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.

[16] Onur C. Hamsici and Aleix M. Martinez, "Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification," *The Journal of Machine Learning Research*, vol. 8, pp. 1583–1623, 2007.

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

[18] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md. Jahangir Alam, and Pierre Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *International Conference on Speech Communication and Technology*, 2013.

[19] Sandro Cumani, Oldřich Plchot, and Pietro Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2013.