On the use of i–vector posterior distributions in Probabilistic Linear Discriminant Analysis

Sandro Cumani, Oldřich Plchot, and Pietro Laface

***Abstract*—The i-vector extraction process is affected by several factors such as the noise level, the acoustic content of the observed features, the channel mismatch between the training conditions and the test data, and the duration of the analyzed speech segment. These factors influence both the i–vector estimate and its uncertainty, represented by the i–vector posterior covariance. This paper presents a new PLDA model that, unlike the standard one, exploits the intrinsic i–vector uncertainty. Since the recognition accuracy is known to decrease for short speech segments, and their length is one of the main factors affecting the i–vector covariance, we designed a set of experiments aiming at comparing the standard and the new PLDA models on short speech cuts of variable duration, randomly extracted from the conversations included in the NIST SRE 2010 extended dataset, both from interviews and telephone conversations. Our results on NIST SRE 2010 evaluation data show that in different conditions the new model outperforms the standard PLDA by more than 10% relative when tested on short segments with duration mismatches, and is able to keep the accuracy of the standard model for long enough speaker segments. This technique has also been successfully tested in the NIST SRE 2012 evaluation.**

***Index Terms*—I-vector extraction, I-vectors, probabilistic linear discriminant analysis, speaker recognition.**

I. Introduction

RECENT developments in speaker recognition technology have seen the success of systems based on a low–dimensional representation of a speech segment, the so–called "identity vector" or i–vector [2]. I–vector based techniques represent the state–of–the–art in speaker detection [3], [4], [5], [6], [7], [8], [9], [10], [11]. An i–vector is a compact representation of a Gaussian Mixture Model (GMM) supervector [12], which captures most of the GMM supervectors variability. It is obtained by a MAP point estimate of a posterior distribution [13].

Manuscript received October 08, 2013; revised December 28, 2013; accepted February 24, 2014. Date of publication February 26, 2014; date of current version March 06, 2014. The work of O. Plchot was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20015, by Czech Ministry of Education Project MSM0021630528, and by IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. This paper is an extended and revised version of a conference paper that appeared as [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Thomas Fang Zheng.

S. Cumani and P. Laface are with the Dipartimento di Automatica e Informatica, Politecnico di Torino, 10129 Torino, Italy (e-mail: sandro.cumani@polito.it; pietro.laface@polito.it).

O. Plchot is with the Brno University of Technology, 602 00 Brno, Czech Republic (e-mail: iplchot@fit.vutbr.cz).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2308473

Probabilistic Linear Discriminant Analysis (PLDA) [14] classifiers based on i–vectors are among the best models for speaker recognition. Some PLDA systems for the last NIST 2012 Speaker Recognition Evaluation and for the DARPA RATS project have been described in [15], [16], [17], [18], [19], [20]. The covariance of the distribution, which accounts for the "uncertainty" of the i–vector extraction process is, however, not exploited by the classifiers based on i–vectors, such as the ones based on cosine distance scoring [2], Probabilistic Linear Discriminant Analysis (PLDA) [14], or SVMs [7]. The i–vector covariance mainly depends on the zero–order statistics estimated on the Gaussian components of a Universal Background Model (UBM) for the set of observed features (see equation 3 in Section II), i.e., by the duration of the speech segments that are used for characterizing a speaker. Shorter segments tend to produce larger covariances, so that i–vector estimates become less reliable.

In [1] we presented a new PLDA model that incorporates the intrinsic uncertainty of the i–vector extraction process. This work revises and completes the theory, extends the set of experiments that have been performed to validate the new model, and analyzes the typical speaker detection scenario that allows the computational complexity of speaker recognition scoring to be reduced.

Our approach shows that the simple and effective PLDA framework can still be used even if a speech segment is no more mapped to a single i–vector but to its posterior distribution. In particular, we derive the formulation of the likelihood for a Gaussian PLDA model based on the i–vector posterior distribution, and propose a new PLDA model where the inter–speaker variability is assumed to have an utterance–dependent distribution. We show that it is possible to rely on the standard PLDA framework simply replacing the Gaussian PLDA likelihood definition.

Since segment duration is the main factor affecting the i–vector covariance, and short segments are known to produce less reliable i–vectors, our approach has been assessed using cuts of variable duration, collected from different channels, extracted from the NIST SRE 2010 extended core tests [21].

Our results show that the new model outperforms the standard PLDA when tested on short segments, particularly for training and test conditions with duration mismatch, without losing accuracy for long enough speaker segments.

An independent development of the same topic has been presented in [22]. Although the Gaussian PLDA models proposed are equivalent, we developed a more general framework, from which the Gaussian PLDA model has been derived, which also allows a more compact and effective scoring formulation. Since

the models are equivalent, the scoring functions compute the same scores. However, the derivations of the scoring function are different. Our formulation leads to a minimal change in the standard PLDA scoring function, allowing a simple and straightforward implementation. In particular, it clearly shows the interaction between the PLDA parameters and the i–vector covariance, so that the standard PLDA scoring can be used, provided that the i–vector covariance is added to the PLDA noise covariance. Moreover, the work of [22] is more focused on training the PLDA models with short segments, whereas efficiency in testing is our main concern.

In [23], [24], a formulation for the comparison of supervectors has been presented, which does not require the two–step approach, consisting in the extraction of i–vectors followed by their PLDA based classification. We comment on the similarities and differences of this approach with respect to [1], [22] in Section VI.

The paper is organized as follows: Section II recalls the i-vector extraction process. Section III presents the generative PLDA model using the i–vector distributions, and the gives the expression for the computation of the likelihood that a set of speech segments belong to the same speaker. Section IV focuses on the computation of the likelihood for a PLDA model based on the i–vector posterior distribution, with Gaussian priors. The new PLDA model, where the distribution of the inter–speaker variability is assumed to be utterance–dependent, is introduced in Section V. Section VI is devoted to the estimation of the parameters of this PLDA model. The important issue of i–vector length normalization is discussed in Section VII. A detailed analysis of the complexity of the PLDA and of the proposed approach is given in Section VIII, exploiting the optimizations allowed by some practical applications. The experimental results are given in Section IX, and our conclusions are drawn in Section X.

II. I–VECTOR MODEL

The i–vector model constrains the GMM supervector $\mathbf{s}$, representing both the speaker and inter–session characteristics of a given speech segment, to live in a single sub–space according to:

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

where $\mathbf{u}$ is the Universal Background Model (UBM), a GMM mean supervector, composed of C GMM components of dimension F. $\mathbf{T}$ is a low-rank rectangular matrix spanning the sub–space including important inter and intra–speaker variability in the supervector space, and $\mathbf{w}$ is a realization of a latent variable $\mathbf{W}$, of size M, having a standard normal prior distribution.

A Maximum-Likelihood estimate of matrix $\mathbf{T}$ is usually obtained by minor modifications of the Joint Factor Analysis approach [13]. Given $\mathbf{T}$, and the set of τ feature vectors $\mathcal{X} = \{\mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_\tau\}$ extracted from a speech segment, it is possible to compute the likelihood of $\mathcal{X}$ given the model (1), and a value for the latent variable $\mathbf{W}$. The i-vector $\boldsymbol{\phi}$, which represents the segment, is computed as the Maximum a Posteriori (MAP) point estimate of the variable $\mathbf{W}$, i.e., the mode $\boldsymbol{\mu}_\mathcal{X}$ of the posterior distribution $P_{\mathbf{W}|\mathcal{X}}(\mathbf{w})$. It has been shown in [13] that assuming a standard normal prior for $\mathbf{W}$, the posterior probability of $\mathbf{W}$ given the acoustic feature vectors $\mathcal{X}$ is Gaussian:

$$\mathbf{W}|\mathcal{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathcal{X}, \boldsymbol{\Gamma}_\mathcal{X}^{-1}), \tag{2}$$

with mean vector and precision matrix:

$$\begin{aligned}\boldsymbol{\mu}_\mathcal{X} &= \boldsymbol{\Gamma}_\mathcal{X}^{-1}\mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{f}_\mathcal{X} \\ \boldsymbol{\Gamma}_\mathcal{X} &= \mathbf{I} + \sum_{c=1}^{C} N_\mathcal{X}^{(c)}\mathbf{T}^{(c)^{\mathrm{T}}}\boldsymbol{\Sigma}^{(c)^{-1}}\mathbf{T}^{(c)},\end{aligned} \tag{3}$$

respectively. In these equations, $N_\mathcal{X}^{(c)}$ are the zero–order statistics estimated on the c-th Gaussian component of the UBM for the set of feature vectors in $\mathcal{X}$, $\mathbf{T}^{(c)}$ is the $F \times M$ sub-matrix of $\mathbf{T}$ corresponding to the c–th mixture component such that $\mathbf{T} = (\mathbf{T}^{(1)\mathrm{T}}, \ldots, \mathbf{T}^{(C)\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{f}_\mathcal{X}$ is the supervector stacking the first–order statistics $\mathbf{f}_\mathcal{X}^{(c)}$, centered around the corresponding UBM means:

$$\mathbf{f}_\mathcal{X}^{(c)} = \sum_t \left(\gamma_t^{(c)}\mathbf{x}_t\right) - N_\mathcal{X}^{(c)}\mathbf{m}^{(c)}, \tag{4}$$

$\boldsymbol{\Sigma}^{(c)}$ is the UBM c–th covariance matrix, $\boldsymbol{\Sigma}$ is a block diagonal matrix with matrices $\boldsymbol{\Sigma}^{(c)}$ as its entries, and $\gamma_t^{(c)}$ is the occupation probability of feature vector $\mathbf{x}_t$ for the c-th Gaussian component.

III. PLDA WITH I–VECTOR POSTERIORS

Excellent performance has been reported on the last NIST Speaker Recognition Evaluation campaigns [21], [25] for systems using i–vectors with generative models based on PLDA. A PLDA system models the underlying distribution of the speaker and channel components of the i–vectors in a generative framework. From these distributions it is possible to evaluate the likelihood ratio between the same "speaker" hypothesis (H_s) and "different speakers" hypothesis (H_d) for sets of i–vectors. In particular, in the PLDA framework, Factor Analysis is applied to describe the i–vector generation process. An i–vector is considered a random variable Φ whose generation process can be described in terms of a set of latent variables. Different PLDA models exist [26], [14], which use different numbers of hidden variables as well as different priors. All PLDA models for speaker recognition [14], [4], however, represent the speaker identity in terms of a latent variable $\mathbf{Y}$ which is assumed to be tied across all segments of the same speaker. Usually, inter–speaker variability for a speech segment $\mathcal{X}_i$ is represented by hidden variable $\mathbf{X}_i$. The hidden variables $\mathbf{X}_i$ are assumed to be i.i.d. with respect to the speech segments.

In the most common PLDA model, an i–vector $\boldsymbol{\phi}$ is the sum of multiple terms [14]:

$$\boldsymbol{\phi} = \mathbf{m} + \mathbf{U}\mathbf{y} + \mathbf{V}\mathbf{x} + \mathbf{e} \tag{5}$$

where $\mathbf{m}$ is the i–vector mean, $\mathbf{y}$ is a realization of the speaker identity variable $\mathbf{Y}$, $\mathbf{x}$ is the realization of channel variable $\mathbf{X}$ and $\mathbf{e}$ is the realization of the residual noise $\mathbf{E}$. The role of matrices $\mathbf{U}$ and $\mathbf{V}$ is to constrain the dimension of the sub–spaces for $\mathbf{y}$ and $\mathbf{x}$, respectively.

Since i–vectors are assumed independent given the hidden variables, the likelihood that a set of n speech segments $\mathcal{X}_1 \dots \mathcal{X}_n$ belongs to the same speaker (hypothesis H_s) can be computed as:

$$\begin{aligned} l(\mathcal{X}_1 \dots \mathcal{X}_n|H_s) &= P_{\Phi_1 \dots \Phi_n|H_s}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n) \\ &= \int_{\mathbf{y}} \int_{\mathbf{x}_1} \cdots \int_{\mathbf{x}_n} \prod_{i=1}^{n} \left[P_{\Phi_i|\mathbf{Y},\mathbf{X}_i}(\boldsymbol{\phi}_i|\mathbf{y},\mathbf{x}_i) P_{\mathbf{X}_i}(\mathbf{x}_i) d\mathbf{x}_i \right] \\ &\cdot P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}, \end{aligned} \tag{6}$$

where Φ_i is the i–vector extracted from segment $\mathcal{X}_i$, $P_{\Phi_1 \dots \Phi_n|H_s}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n)$ is the pdf of the joint distribution of the i–vectors given the same speaker hypothesis H_s, $P_X(\mathbf{x})$ and $P_{\mathbf{Y}}(\mathbf{y})$ are the prior distributions for $\mathbf{X}$ and $\mathbf{Y}$, respectively. $P_{\Phi|\mathbf{Y},\mathbf{X}}(\boldsymbol{\phi}|\mathbf{y},\mathbf{x})$ is the conditional distribution of an i–vector given the hidden variables. It is related to the distribution $P_{\mathbf{E}}(\mathbf{e})$ of the noise term by $P_{\Phi|\mathbf{Y},\mathbf{X}}(\boldsymbol{\phi}|\mathbf{y},\mathbf{x}) = P_{\mathbf{E}}(\boldsymbol{\phi} - \mathbf{m} - \mathbf{U}\mathbf{y} - \mathbf{V}\mathbf{x})$.

Since speaker factors are assumed independent, given a set of n enrollment segments $\mathcal{X}_{e_1} \dots \mathcal{X}_{e_n}$ for a target speaker, and a set of m test segments of a single unknown speaker $\mathcal{X}_{t_1} \dots \mathcal{X}_{t_m}$, the speaker verification log–likelihood ratio s can be computed as:

$$s = \log \frac{l(\mathcal{X}_{e_1} \dots \mathcal{X}_{e_n}, \mathcal{X}_{t_1} \dots \mathcal{X}_{t_m}|H_s)}{l(\mathcal{X}_{e_1} \dots \mathcal{X}_{e_n}|H_s)\, l(\mathcal{X}_{t_1} \dots \mathcal{X}_{t_m}|H_s)}. \tag{7}$$

The standard i–vector, which is extracted by MAP point estimate of the posterior distribution of $\mathbf{W}$ given $\mathcal{X}$, and classified by PLDA, does not embed the intrinsic uncertainty of its estimate. However, it is well known that i–vectors extracted from short segments do not capture the speaker characteristic as well as i–vectors extracted from long segments. Since the uncertainty associated with the extraction process of the i-vector, which is represented by its posterior covariance, is not taken into account by the usual PLDA models, in this work we extend the model to exploit this additional information. We refer to this new model as the PLDA based on the "Full Posterior Distribution" (FPD–PLDA) of $\mathbf{W}$ given $\mathcal{X}$. In this model we assume that every segment $\mathcal{X}$ is no more mapped to a single i–vector but to the i–vector extractor distribution $\mathbf{W}|\mathcal{X}$. Thus, $\mathcal{X}$ is mapped to i–vector $\boldsymbol{\phi}$ according to the probability distribution $P_{\mathbf{W}|\mathcal{X}}(\boldsymbol{\phi})$.

The PLDA model allows computing the likelihood of a speech segment given a realization of the random variable $\mathbf{W}|\mathcal{X}$. The likelihood of a set of segments $\mathcal{X}_1 \dots \mathcal{X}_n$, thus, can be evaluated by integrating the classical PLDA likelihood over all the i–vectors that these segments can generate as:

$$\begin{aligned} l(\mathcal{X}_1 \dots \mathcal{X}_n|H_s) &= \int_{\boldsymbol{\phi}_1} \cdots \int_{\boldsymbol{\phi}_n} l(\mathcal{X}_1 \dots \mathcal{X}_n|H_s, \mathbf{W}_1 = \boldsymbol{\phi}_1, \\ &\qquad \dots, \mathbf{W}_n = \boldsymbol{\phi}_n) \prod_{i=1}^{n} \left[P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right] \\ &= \int_{\boldsymbol{\phi}_1} \cdots \int_{\boldsymbol{\phi}_n} P_{\Phi_1 \dots \Phi_n|H_s}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n) \prod_{i=1}^{n} \left[P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right], \end{aligned} \tag{8}$$

where the first factor is the likelihood of the segments according to the classical PLDA model given the realizations $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_n$ of the i–vector posterior random variables, computed by (6), and the second factor is the likelihood that the i–vectors $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_n$ are mapped to segments $\mathcal{X}_1, \dots, \mathcal{X}_n$, respectively, according to the i–vector extractor model.

Replacing (6) in (8), the likelihood can be rewritten as:

$$\begin{aligned} l(\mathcal{X}_1 \dots \mathcal{X}_n|H_s) = \int_{\boldsymbol{\phi}_1} \cdots \int_{\boldsymbol{\phi}_n} \int_{\mathbf{y}} \int_{\mathbf{x}_1} \cdots \int_{\mathbf{x}_n} \\ \prod_{i=1}^{n} \left[P_{\Phi_i|\mathbf{Y},\mathbf{X}_i}(\boldsymbol{\phi}_i|\mathbf{y},\mathbf{x}_i) \right. \\ \left. \cdot P_{\mathbf{X}_i}(\mathbf{x}_i) P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i)\, d\mathbf{x}_i d\boldsymbol{\phi}_i \right] P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}. \end{aligned} \tag{9}$$

It is worth noting that, if the posterior for $\mathbf{W}|\mathcal{X}$ is replaced by a delta distribution centered in the posterior mean $\delta(\boldsymbol{\mu}_{\mathcal{X}})$, the likelihood of the original PLDA model using MAP–estimated i–vectors, given by (6), is obtained.

IV. Gaussian PLDA model

In this work we consider only PLDA with Gaussian priors, because this model has shown to be as accurate and more effective than other more expensive models, such as the Heavy–Tailed PLDA [14], provided that the i–vectors are properly length–normalized [27]. Moreover, we will assume that the noise term $\mathbf{E}$ has full covariance matrix, so that the terms $\mathbf{V}\mathbf{x}$ and $\mathbf{e}$ in (5) can be merged. Thus, in our approach an i–vector $\boldsymbol{\phi}$ is defined as:

$$\boldsymbol{\phi} = \mathbf{m} + \mathbf{U}\mathbf{y} + \mathbf{e}. \tag{10}$$

The Gaussian PLDA approach assumes that the speaker factors and the residual noise priors are Gaussian, i.e.:

$$\mathbf{Y} \sim \mathcal{N}(0, \mathbf{I}), \quad \mathbf{E} \sim \mathcal{N}(0, \boldsymbol{\Lambda}^{-1}), \tag{11}$$

where $\boldsymbol{\Lambda}$ is the precision matrix of noise $\mathbf{E}$. According to (10) and (11), the conditional distribution of an i–vector random variable Φ given a value $\mathbf{y}$ for the speaker identity $\mathbf{Y}$ is:

$$\boldsymbol{\Phi}|(\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}(\mathbf{m} + \mathbf{U}\mathbf{y}, \boldsymbol{\Lambda}^{-1}). \tag{12}$$

Ignoring the channel factors, which in our model are embedded in the noise term, the likelihood that the n speech segments $\mathcal{X}_1 \dots \mathcal{X}_n$ belong to the same speaker can be computed by means of a simplified expression of (6) as:

$$\begin{aligned} l(\mathcal{X}_1 \dots \mathcal{X}_n|H_s) &= P_{\Phi_1 \dots \Phi_n}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n|H_s) \\ &= \int_{\mathbf{y}} \prod_{i=1}^{n} P_{\Phi_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}. \end{aligned} \tag{13}$$

Introducing the full i–vector posterior we get:

$$\begin{aligned} l\left(\mathcal{X}_1 \ldots \mathcal{X}_n | H_s\right) &= \int_{\boldsymbol{\phi}_i} \cdots \int_{\boldsymbol{\phi}_n} \int_{\mathbf{y}} P_{\mathbf{Y}}(\mathbf{y}) \\ &\quad \cdot \prod_{i=1}^{n} \left[P_{\boldsymbol{\Phi}_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right] d\mathbf{y} \\ &= \int_{\mathbf{y}} P_{\mathbf{Y}}(\mathbf{y}) \prod_{i=1}^{n} \left[\int_{\boldsymbol{\phi}_i} P_{\boldsymbol{\Phi}_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) \right. \\ &\quad \left. \cdot \, P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right] d\mathbf{y}, \end{aligned}$$

According to the Gaussian assumptions given in (2) and (11), the inner integral can be computed as:

$$\begin{aligned} &\int_{\boldsymbol{\phi}_i} P_{\boldsymbol{\Phi}_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \\ &= \int_{\boldsymbol{\phi}_i} \frac{1}{(2\pi)^{\frac{M}{2}} \left|\boldsymbol{\Lambda}^{-1}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\phi}_i - \mathbf{m} - \mathbf{U}\mathbf{y})^{\mathrm{T}} \boldsymbol{\Lambda} (\boldsymbol{\phi}_i - \mathbf{m} - \mathbf{U}\mathbf{y})} \\ &\quad \cdot \frac{1}{(2\pi)^{\frac{M}{2}} \left|\boldsymbol{\Gamma}_i^{-1}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\phi}_i - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Gamma}_i (\boldsymbol{\phi}_i - \boldsymbol{\mu}_i)} d\boldsymbol{\phi}_i, \end{aligned} \tag{14}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Gamma}_i$ are the mean and precision matrix of $\mathbf{W}_i|\mathcal{X}_i$ computed as in (3). Integral (14) can be interpreted as the convolution of two Gaussian distributions, leading to:

$$\begin{aligned} l(\mathcal{X}_1 \ldots \mathcal{X}_n | \mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{n} &\frac{1}{(2\pi)^{\frac{M}{2}} \left|\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}\right|^{\frac{1}{2}}} \\ &\cdot e^{(\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y})^{\mathrm{T}} \left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}\right)^{-1} (\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y})}. \end{aligned} \tag{15}$$

The result in (15) can be interpreted as the likelihood of a standard PLDA model where a segment is mapped to the mean $\boldsymbol{\mu}_i$ of the i–vector posterior $\mathbf{W}_i|\mathcal{X}_i$, but the PLDA conditional likelihood is segment–dependent, i.e., the residual noise $\mathbf{E}_i$ in the PLDA model (11), is replaced by the segment–dependent noise $\overline{\mathbf{E}}_i$ distributed as $\overline{\mathbf{E}}_i \sim \mathcal{N}(\mathbf{0}, [\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}])$. Indeed, the right side of equation (15) is a Gaussian pdf for $\boldsymbol{\mu}_i$. Considering every $\boldsymbol{\mu}_i$ as a realization of a random variable $\mathbf{M}_i$, the conditional likelihood of a set of n speech segments can be written as:

$$l(\mathcal{X}_1 \ldots \mathcal{X}_n | \mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{n} P_{\mathbf{M}_i|\mathbf{Y}}(\boldsymbol{\mu}_i|\mathbf{y}), \tag{16}$$

where $\mathbf{M}_i|\mathbf{Y}$ is distributed as $\mathcal{N}(\mathbf{m} + \mathbf{U}\mathbf{y}, [\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}])$. The likelihood that the segments belong to the same speaker is then given by:

$$l(\mathcal{X}_1 \ldots \mathcal{X}_n | H_s) = \int_{\mathbf{y}} \prod_{i=1}^{n} P_{\mathbf{M}_i|\mathbf{Y}}(\boldsymbol{\mu}_i|\mathbf{y}) P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}. \tag{17}$$

Comparing (17) and (13) it can be observed that the two models differ only for the parameters of their conditional likelihoods. Due to the similarity of these two models, simple expressions can be derived for estimating the parameters of the Full Posterior Distribution model, and for computing the speaker verification log–likelihood scores according to this model. In particular, PLDA can be trained by adapting the EM algorithm which estimates the standard PLDA model parameters [14].

V. Scoring with Gaussian PLDA Posteriors

The log–likelihood that a set of segments belongs to the same speaker can be obtained by means of the same steps followed for the standard Gaussian PLDA model, just using the modified likelihood in (15). The new PLDA model can be described as:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{U}\mathbf{y} + \bar{\mathbf{e}}, \tag{18}$$

as in (10), but with an segment–dependent distribution of the residual noise $\overline{\mathbf{E}}$. The i–vector associated to speech segment $\mathcal{X}_i$ is again the mean $\boldsymbol{\mu}_i$ of the i–vector posterior $\mathbf{W}_i|\mathcal{X}_i$, but the priors of the PLDA parameters are given by:

$$\overline{\mathbf{E}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_{eq,i}^{-1}), \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{19}$$

where

$$\boldsymbol{\Lambda}_{eq,i} = \left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}\right)^{-1}. \tag{20}$$

In the following, to simplify the notation, we will refer to distributions without explicitly naming the corresponding hidden variable, e.g., we will write $P(\mathbf{y})$ rather than $P_{\mathbf{Y}}(\mathbf{y})$.

In order to compute the likelihood of a set of n i–vectors $\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n$ (i.e., of the set of speech segments $\mathcal{X}_1 \ldots \mathcal{X}_n$), we observe that the joint log–likelihood of the i–vectors and the hidden variables is:

$$\begin{aligned} \log P(\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n, \mathbf{y} | H_s) &= \sum_{i=1}^{n} \log P(\boldsymbol{\mu}_i|\mathbf{y}) + \log P(\mathbf{y}) \\ &= \sum_{i=1}^{n} \left[-\frac{1}{2} (\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y})^{\mathrm{T}} \right. \\ &\quad \left. \cdot \, \boldsymbol{\Lambda}_{eq,i} (\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y}) \right] \\ &\quad + \frac{1}{2} \mathbf{y}^{\mathrm{T}} \mathbf{y} + k, \end{aligned} \tag{21}$$

where k is a constant collecting terms that do not depend on $\mathbf{y}$. Equation (21) shows that the posterior distribution of $\mathbf{y}$ given a set of i–vectors is Gaussian:

$$\mathbf{y} | \boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Lambda}_y^{-1}), \tag{22}$$

with parameters:

$$\begin{aligned} \boldsymbol{\Lambda}_y &= \mathbf{I} + \sum_{i=1}^{n} \mathbf{U}^{\mathrm{T}} \boldsymbol{\Lambda}_{eq,i} \mathbf{U} \\ \boldsymbol{\mu}_y &= \boldsymbol{\Lambda}_y^{-1} \mathbf{U}^{\mathrm{T}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{eq,i} (\boldsymbol{\mu}_i - \mathbf{m}). \end{aligned} \tag{23}$$

The likelihood that a set of segments belongs to the same speaker can be written as:

$$P(\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n | H_s) = \frac{P(\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n | \mathbf{y}_0) P(\mathbf{y}_0)}{P(\mathbf{y}_0 | \boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n)}, \tag{24}$$

where $\mathbf{y}_0$ can be freely chosen provided that the denominator is non–zero. Setting for convenience $\mathbf{y}_0 = \mathbf{0}$, so that $\mathbf{U}\mathbf{y}_0 = \mathbf{0}$, from (22), and (23) we finally get:

$$\begin{aligned}\log P(\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n | H_s) = \sum_{i=1}^{n} \Bigg[& \frac{1}{2} \log |\boldsymbol{\Lambda}_{eq,i}| - \frac{M}{2} \log 2\pi \\ & - \frac{1}{2} (\boldsymbol{\mu}_i - \mathbf{m})^{\mathrm{T}} \boldsymbol{\Lambda}_{eq,i} (\boldsymbol{\mu}_i - \mathbf{m}) \Bigg] \\ & - \frac{1}{2} \log |\boldsymbol{\Lambda}_y| + \frac{1}{2} \boldsymbol{\mu}_y^{\mathrm{T}} \boldsymbol{\Lambda}_y \boldsymbol{\mu}_y - \frac{S}{2} \log 2\pi, \end{aligned} \tag{25}$$

where M is the i–vector dimension, and S is the speaker factor dimension.

VI. PLDA Parameter Estimation

The model presented in (18) allows obtaining a simple expression for computing the log–likelihood ratio of a speaker recognition trial. However, it does not allow the update formulas to be easily derived. An equivalent expression of (18), where the contributions of the i–vector posterior covariance and of the residual noise are decoupled, is more effective for the estimation of model parameters [22]. To this extent, the segment–dependent residual term $\overline{\mathbf{E}}_i$ can be written as:

$$\overline{\mathbf{E}}_i = \mathbf{C}_i \mathbf{X}_i + \mathbf{E}, \tag{26}$$

where $\mathbf{C}_i$ is the Cholesky decomposition $\mathbf{C}_i \mathbf{C}_i^{\mathrm{T}} = \boldsymbol{\Gamma}_i^{-1}$, $\mathbf{X}_i$ is a standard Gaussian distributed random variable, $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{E}$ is the PLDA residual term introduced in (11). The corresponding PLDA model is then given by:

$$\boldsymbol{\phi}_i = \mathbf{m} + \mathbf{U}\mathbf{y} + \mathbf{C}_i \mathbf{x}_i + \mathbf{e}_i, \tag{27}$$

where $\mathbf{x}_i$ is a realization of $\mathbf{X}_i$. It is worth noting that (27) formally corresponds to the PLDA model in (5) with the channel sub–space matrix $\mathbf{V}$ replaced by a segment–dependent matrix $\mathbf{C}_i$. The same steps followed to derive the EM algorithm for the PLDA model (5) can be easily modified to estimate the parameters of the FPD–PLDA model. The details of the derivation of the EM algorithm can be found in [22].

The SV–PLDA approach in [24] and the FPD–PLDA approach [1], [22], essentially differ because the former obtains the precision matrix $\boldsymbol{\Gamma}_i$ by means of a Maximum Likelihood Estimation process, whereas the latter relies on Maximum A Posteriori estimation. MAP estimation simply leads to the presence of the identity matrix $\mathbf{I}$ in the formulation of the precision matrix $\boldsymbol{\Gamma}_i$ in (23), which is missing in the so called π–vector definition (see equations (8) and (14) in Section 2.6 of [24], where the π–vector and i–vector definitions are compared). SV–PLDA is in principle an elegant and attractive single step MLE approach, compared with the standard two–step approach consisting in i–vector extraction followed by PLDA classification. Unfortunately SV–PLDA too has to rely on length normalization [27] in order to obtain good results. Casting length normalization in the middle of the generative framework makes SV–PLDA less coherent. In our opinion, a two–step approach is preferable because it gives the freedom of using different assumptions about the distribution of the i–vectors, which could be obtained by different extractors, and then used as features for the PLDA classifier. Moreover, in our experience no better accuracy was obtained by the elimination of the $\mathbf{I}$ term from the precision matrix $\boldsymbol{\Gamma}_i$.

VII. I–vector Pre–Processing

A pre–processing step, which involves i–vector whitening followed by length normalization [27], is required to achieve state-of-the-art results using i–vectors with Gaussian PLDA models. While it is easy to understand length normalization applied to i–vectors, different interpretations of length normalization lead to different normalizations of the posterior covariance matrices. This section presents three different interpretations of length normalization, and shows their effect on the normalization of the full i–vector posterior.

A straightforward approach consists in replacing the i–vector distribution $\mathbf{W}|\mathcal{X}$ by $\widehat{\mathbf{W}} = \frac{\mathbf{W}|\mathcal{X}}{\|\mathbf{W}|\mathcal{X}\|}$, which forces all realizations of $\widehat{\mathbf{W}}$ to lie on the unit sphere. However, since the resulting random variable $\widehat{\mathbf{W}}$ would not be Gaussian distributed, it would not be possible to rely on the simple derivations of Section IV, and to avoid the higher complexity introduced by the use of a non Gaussian distribution.

We implemented a second approach, where length normalization is considered a non–linear transformation $F(\boldsymbol{\phi}_0)$ of the observed i–vector $\boldsymbol{\phi}_0$, which can be approximated by its first order Taylor expansion around the i–vector itself:

$$F(\boldsymbol{\phi}) = F(\boldsymbol{\phi}_0) + J_F(\boldsymbol{\phi}_0)(\boldsymbol{\phi} - \boldsymbol{\phi}_0) + o(\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\|), \tag{28}$$

where $J_F(\boldsymbol{\phi}_0)$ is the Jacobian of F computed in $\boldsymbol{\phi}_0$ and F is the function $F(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$. Developing the Jacobian, the linear transformation which best approximates the length normalization function around the i–vector is given by:

$$\widehat{F}(\boldsymbol{\phi}) = F(\boldsymbol{\phi}_0) + J_F(\boldsymbol{\phi}_0)(\boldsymbol{\phi} - \boldsymbol{\phi}_0) = \mathbf{v} + \frac{(\mathbf{I} - \mathbf{v}\mathbf{v}^{\mathrm{T}})}{\|\boldsymbol{\phi}_0\|} \boldsymbol{\phi} \tag{29}$$

where $\mathbf{v} = \frac{\boldsymbol{\phi}_0}{\|\boldsymbol{\phi}_0\|}$ and $\mathbf{I}$ is the identity matrix.

The extension to the full i–vector posterior consists in computing the first order Taylor expansion of F centered at the posterior distribution mean $\boldsymbol{\mu}_{\mathcal{X}}$, and applying the resulting linear transformation to the i–vector posterior $\mathbf{W}|\mathcal{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{X}}, \boldsymbol{\Gamma}_{\mathcal{X}}^{-1})$. The expansion of F is:

$$\widehat{F}(\boldsymbol{\mu}_{\mathcal{X}}) = \mathbf{v}_{\mathcal{X}} + \frac{(\mathbf{I} - \mathbf{v}_{\mathcal{X}} \mathbf{v}_{\mathcal{X}}^{\mathrm{T}})}{\|\boldsymbol{\mu}_{\mathcal{X}}\|} \boldsymbol{\mu}_{\mathcal{X}} = \mathbf{v}_{\mathcal{X}} + \mathbf{A}\boldsymbol{\mu}_{\mathcal{X}}, \tag{30}$$

where $\mathbf{v}_{\mathcal{X}} = \frac{\boldsymbol{\mu}_{\mathcal{X}}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}$ and $\mathbf{A} = \frac{(\mathbf{I} - \mathbf{v}_{\mathcal{X}} \mathbf{v}_{\mathcal{X}}^{\mathrm{T}})}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}$. Thus, the transformed distribution is given by:

$$\begin{aligned}\widehat{\mathbf{W}} & \sim \mathcal{N}\left(\widehat{F}(\boldsymbol{\mu}_{\mathcal{X}}), \mathbf{A}\boldsymbol{\Gamma}_{\mathcal{X}}^{-1}\mathbf{A}^{\mathrm{T}}\right) \\ & \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_{\mathcal{X}}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}, \frac{1}{\|\boldsymbol{\mu}_{\mathcal{X}}\|^2} (\mathbf{I} - \mathbf{v}_{\mathcal{X}} \mathbf{v}_{\mathcal{X}}^{\mathrm{T}}) \boldsymbol{\Gamma}_{\mathcal{X}}^{-1} (\mathbf{I} - \mathbf{v}_{\mathcal{X}} \mathbf{v}_{\mathcal{X}}^{\mathrm{T}})\right), \end{aligned} \tag{31}$$

Expression (31) can be further approximated as:

$$\overline{\mathbf{W}} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_{\mathcal{X}}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}, \frac{\boldsymbol{\Gamma}_{\mathcal{X}}^{-1}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|^2}\right). \tag{32}$$

In the experimental section we show that these linearizations of the length normalization are effective. In particular, the approximation (32) allows a simplification of (31) without incurring in any performance degradation. We will refer to (31) as "Projected Length Normalization" (FPD1), and to (32) as "Length Normalization" (FPD2).

VIII. Complexity Analysis

The straightforward implementations of classical PLDA and FPD–PLDA have similar computational complexity. However, in practical scenarios some of the terms required for the evaluation of the PLDA log–likelihood ratio (7) can be pre–computed. These pre-computations allow fast test scoring, at the cost of a slight increase of the memory requirements for the PLDA model and for the target models. Unfortunately, some of these optimizations cannot be done for FPD–PLDA, which is thus a more accurate but slower approach. In the following we analyze the computational complexity of PLDA and FPD–PLDA implementations optimized for the most common scenario. This scenario consists of a speaker detection task where the system has to score several test sets, whose number of segments is known in advance, against a fixed set of target speakers. In particular, each set of segments of a single test speaker has to be verified against the segments of a known, fixed, set of target speakers. Since all targets are known in advance, target–dependent optimizations can be performed offline. The NIST 2012 SRE evaluation [25] follows this protocol. However, even for the previous evaluations, where each trial has to be scored independently it is possible to speed–up the scoring for the complete evaluation, without violating its rules, because all target segments are indeed known in advance.

In this scenario, as will be shown in sub–section VIII-B and VIII-C, a smart implementation of PLDA allows some of the terms required for the evaluation of the speaker verification log–likelihood ratio to be pre–computed, thus the per–trial scoring complexity is greatly reduced. Different optimizations are possible for FPD–PLDA depending on the duration of the trial segments. For short segments, FPD–PLDA does not allow the pre–computation of most of the terms of the scoring function, thus its complexity cannot be reduced. However, if the target segments are long enough, their i–vector posteriors can be safely approximated by their MAP point estimates, and the per–trial complexity of the proposed technique can be reduced.

A. Log–likelihood Computation

The complexity of the log–likelihood computation accounts for two separate contributions. The first contribution is the complexity of operations which can be independently performed on target or test sets, which will be referred to as per–target and per–test terms, respectively. The second contribution is the per–trial complexity, i.e. the complexity of the terms which jointly involve the target and the test sets. This distinction is not relevant for the naïve scoring implementations, but is relevant, instead, in the "fixed set of target speakers scenarios" because the per–target terms can be pre–computed, and per–test terms need to be computed only once regardless of the number of target speakers.

We will analyze both per–test and per–trial complexity of the PLDA and FPD–PLDA models. It is worth noting that the complexity of a complete system should account also for the complexity of the extraction of the acoustic features and of the i–vectors. The computation of the i–vector covariance matrix, for each segment, has complexity $O(M^3)$ [28], which, as we will see, dominates the other costs.

Since we compute the speaker variable $\mathbf{y}$ posteriors on different sets, we explicitly condition the parameters of the posterior distributions of $\mathbf{y}$ (23) to a generic set G as:

$$\begin{aligned}\boldsymbol{\Lambda}_{y|G} &= \mathbf{I} + \sum_{i\in G} \mathbf{U}^{\mathrm{T}} \boldsymbol{\Lambda}_{eq,i} \mathbf{U} \\ \boldsymbol{\mu}_{y|G} &= \boldsymbol{\Lambda}_{y}^{-1} \mathbf{U}^{\mathrm{T}} \sum_{i\in G} \boldsymbol{\Lambda}_{eq,i} \left(\boldsymbol{\mu}_i - \mathbf{m}\right). \end{aligned} \tag{33}$$

The indexes of the sum in this equation, and in the following equations, are to be interpreted as running over all the segments of the set. Replacing (25) in (7), the speaker verification log–likelihood ratio for a target set E and a test set T can be written as:

$$\begin{aligned} llr(E,T) &= \log \frac{l(E,T|H_s)}{l(E|H_s) l(T|H_s)} \\ &= -\frac{1}{2}\log\left|\boldsymbol{\Lambda}_{y|(E,T)}\right| + \frac{1}{2}\boldsymbol{\mu}_{y|(E,T)}^{\mathrm{T}} \boldsymbol{\Lambda}_{y|(E,T)} \boldsymbol{\mu}_{y|(E,T)} \\ &\quad + \frac{1}{2}\log\left|\boldsymbol{\Lambda}_{y|(E)}\right| - \frac{1}{2}\boldsymbol{\mu}_{y|(E)}^{\mathrm{T}} \boldsymbol{\Lambda}_{y|E} \boldsymbol{\mu}_{y|(E)} \\ &\quad + \frac{1}{2}\log\left|\boldsymbol{\Lambda}_{y|(T)}\right| - \frac{1}{2}\boldsymbol{\mu}_{y|(T)}^{\mathrm{T}} \boldsymbol{\Lambda}_{y|(T)} \boldsymbol{\mu}_{y|(T)} \\ &\quad + \frac{S}{2}\log 2\pi \\ &= \sigma(E,T) - \sigma(E) - \sigma(T) + \frac{S}{2}\log 2\pi, \end{aligned} \tag{34}$$

where the scoring function σ is defined as:

$$\sigma(G) = -\frac{1}{2}\log\left|\boldsymbol{\Lambda}_{y|(G)}\right| + \frac{1}{2}\boldsymbol{\mu}_{y|(G)}^{\mathrm{T}} \boldsymbol{\Lambda}_{y|(G)} \boldsymbol{\mu}_{y|(G)}. \tag{35}$$

Since the computation of $\sigma(E)$ and $\sigma(T)$ cannot be more expensive than the computation of $\sigma(E,T)$, we restrict our analysis to this term of the log–likelihood ratio.

B. Complexity of the Standard Gaussian PLDA

As shown in Section V, standard PLDA corresponds to a FPD–PLDA with $\boldsymbol{\Gamma}_i^{-1} = \mathbf{0}$ for all i–vectors. Thus, $\boldsymbol{\Lambda}_{eq,i} = \boldsymbol{\Lambda}$ for all i–vectors, and the speaker variable posterior parameters become:

$$\begin{aligned} \boldsymbol{\Lambda}_{y|(E,T)} &= \mathbf{I} + (n_E + n_T)\mathbf{U}^{\mathrm{T}} \boldsymbol{\Lambda} \mathbf{U} \\ \boldsymbol{\mu}_{y|(E,T)} &= \boldsymbol{\Lambda}_{y|(E,T)}^{-1} \mathbf{U}^{\mathrm{T}} \boldsymbol{\Lambda} \left(\sum_{i\in E} \left(\boldsymbol{\mu}_i - \mathbf{m}\right) + \sum_{i\in T} \left(\boldsymbol{\mu}_i - \mathbf{m}\right) \right) \\ &= \boldsymbol{\Lambda}_{y|(E,T)}^{-1} \left(\mathbf{F}_E + \mathbf{F}_T\right), \end{aligned} \tag{36}$$

where n_E and n_T are the number of target and test segments respectively, $\mathbf{F}_E$ and $\mathbf{F}_T$ are the projected first order statistics defined as:

$$\mathbf{F}_E = \mathbf{M} \sum_{i\in E} \left(\boldsymbol{\mu}_i - \mathbf{m}\right), \quad \mathbf{F}_T = \mathbf{M} \sum_{i\in T} \left(\boldsymbol{\mu}_i - \mathbf{m}\right), \tag{37}$$

and $\mathbf{M} = \mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}$ is a $S \times M$ matrix. Using these definitions, the scoring function $\sigma(E,T)$ can be rewritten as:

$$\sigma(E,T) = -\frac{1}{2}\log\left|\mathbf{\Lambda}_{y|(E,T)}\right| + \mathbf{F}_E^{\mathrm{T}}\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_T + \frac{1}{2}\mathbf{F}_T^{\mathrm{T}}\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_T + \frac{1}{2}\mathbf{F}_E^{\mathrm{T}}\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_E. \quad (38)$$

Computing the projected statistics (37) has complexity $O(NM + MS)$, where N is the number of speech segments in the set. It is worth noting that the $\mathbf{F}_E$ and $\mathbf{F}_T$ statistics are per–speaker computations because they can be computed for the target and test sets independently.

1) Naïve Scoring Implementation: The computation of the score function $\sigma(E,T)$, given the $\mathbf{F}_G$ statistics, requires computing $\mathbf{\Lambda}_{y|(E,T)}^{-1}$ and its log–determinant. These computations have complexity $O(S^3)$ because, for standard PLDA, the term $\mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{U}$ can be precomputed. Given $\mathbf{\Lambda}_{y|(E,T)}^{-1}$, scoring $\sigma(E,T)$ has complexity $O(S^2)$. The same considerations apply to the less expensive computation of $\sigma(E)$ and $\sigma(T)$. Thus, the overall per–trial complexity is $O(S^3)$.

2) Speaker Detection with Known, Fixed, Target Sets: In the naïve implementation, the computation and inversion of $\mathbf{\Lambda}_{y|(E,T)}$ dominates the scoring costs. However, in standard PLDA this factor depends on the number $(n_T + n_E)$ of the target and test segments only (36). Since each set of target segments E_k, and the number of test segments n_T, are known, it is possible to pre–compute the corresponding $\mathbf{\Lambda}_{y|(E_k,T)}^{-1}$, and its log–determinant. Moreover, since the statistics $\mathbf{F}_{E_k}$ are also known in advance, the terms of the scoring function $\frac{1}{2}\mathbf{F}_{E_k}^{\mathrm{T}}\mathbf{\Lambda}_{y|(E_k,T)}^{-1}$ can be pre–computed. It is worth noting that these terms are small S–sized vectors. Since the term depending only on the test statistics $\mathbf{F}_T$ must be evaluated just once for the whole set of K targets, its computation has a per–test, rather than a per–trial, cost. Every function $\sigma(E_k,T)$ can be computed in $O(S)$, each term $\sigma(E_k)$ can be easily pre–computed. Given the statistics, the term $\sigma(T)$ has a per–speaker complexity of $O(S^2)$. The overall per–speaker cost, including statistics computations, is then $O(NM + MS)$, whereas the per–trial cost is $O(S)$.

C. Full Posterior Distribution PLDA

The main difference between the standard PLDA and the FPD–PLDA approach is that in PLDA $\mathbf{\Lambda}_{y|(E,T)}$ depends just on the number of i–vectors in the set (36), whereas in FPD–PLDA it also depends on the covariance of each i–vector in the test set T (see (33)). This does not allow applying to FPD–PLDA the optimizations illustrated in the previous section.

The speaker variable posterior parameters can still be written as:

$$\begin{aligned}\mathbf{\Lambda}_{y|(E,T)} &= \mathbf{I} + (\mathbf{\Lambda}_{eq,E} + \mathbf{\Lambda}_{eq,T}) \\ \boldsymbol{\mu}_{y|(E,T)} &= \mathbf{\Lambda}_y^{-1}(\mathbf{F}_{eq,E} + \mathbf{F}_{eq,T}),\end{aligned} \quad (39)$$

TABLE I
COMPARISON OF THE LOG–LIKELIHOOD COMPUTATION COMPLEXITY FOR THREE IMPLEMENTATIONS OF PLDA. PER–SEGMENT COSTS SHOULD BE MULTIPLIED BY THE NUMBER OF SEGMENTS N OF A GIVEN SPEAKER. PER–SPEAKER COSTS DO NOT DEPEND ON THE NUMBER OF SPEAKER SEGMENTS. THESE COSTS REFER TO PLDA ONLY, WITHOUT CONSIDERING THE CONTRIBUTION OF I–VECTOR EXTRACTION

System	Per–segment costs	Per–speaker fixed costs	Per–trial costs
Naïve PLDA	M	MS	S^3
Optmized PLDA	M	MS	S
Standard FPD–PLDA	M^3	M^2S	S^3
Asymm. FPD–PLDA	M^3	M^2S	S^2

where

$$\begin{aligned}\mathbf{F}_{eq,G} &= \mathbf{U}^{\mathrm{T}}\sum_{i\in G}\mathbf{\Lambda}_{eq,i}(\boldsymbol{\mu}_i - \mathbf{m}) \\ \mathbf{\Lambda}_{eq,G} &= \mathbf{U}^{\mathrm{T}}\left(\sum_{i\in G}\mathbf{\Lambda}_{eq,i}\right)\mathbf{U},\end{aligned}$$

and the scoring function $\sigma(E,T)$ can be rewritten as:

$$\sigma(E,T) = -\frac{1}{2}\log\left|\mathbf{\Lambda}_{y|(E,T)}^{-1}\right| + \frac{1}{2}\mathbf{F}_{eq,E}^{\mathrm{T}}\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_{eq,E} + \frac{1}{2}\mathbf{F}_{eq,T}^{\mathrm{T}}\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_{eq,T} + \mathbf{F}_{eq,E}^{\mathrm{T}}\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_{eq,T}. \quad (40)$$

Computing the posterior parameters (39) has a complexity $O(NM^3 + M^2S)$, mainly due to the computation of $\mathbf{\Lambda}_{eq,i}$, and is much higher than the $O(NM + MS)$ complexity of standard PLDA approach. However, these computations are required only for a new target or test speaker. These costs are comparable to the costs $O(NM^3)$ of the i–vector extraction [28]. Given the statistics, $\mathbf{\Lambda}_{y|(E,T)}$ can be computed with complexity $O(S^2)$ and its inversion complexity is $O(S^3)$. The computation of the remaining terms requires $O(S^2)$, thus the overall per–trial complexity is $O(S^3)$. Since the posterior parameter $\mathbf{\Lambda}_{y|(E,T)}$ cannot be pre–computed as in standard PLDA, the per–trial complexity is the same also for the fixed set of target speakers scenarios.

D. Asymmetric Full Posterior Distribution PLDA

In some applications the target speaker segments have long enough duration, so that replacing the corresponding i–vector posterior distribution by a MAP point estimate has a negligible impact on the term $\mathbf{\Lambda}_{eq,E}$. In this case, it is possible to narrow the complexity gap between standard PLDA and FPD–PLDA, because the i–vector covariance is taken into account only for the test segments. Thus, we refer to this approach as Asymmetric Full Posterior Distribution PLDA. Since MAP–approximated i–vectors are used for the target speakers, the computational complexity of $\sigma(E)$ becomes equivalent to the one of the standard PLDA. The per–trial complexity with respect to the standard FPD–PLDA approach can be reduced because the same test set is scored against a fixed set of target speakers. In

TABLE II
TRAINING AND TEST CONDITIONS OF THE NIST 2010 EVALUATION

Condition	Female target / non-target trials	Male target / non-target trials	Training	Test	Channel
1	2326 / 449138	1978 / 346857	interview	interview	same microphone
2	8152 / 157394	6932 / 121558	interview	interview	different microphones
3	1958 / 334438	2031 / 303412	interview	telephone	
4	1751 / 392467	1886 / 364308	interview	microphone	
5	3704 / 233077	3465 / 175873	telephone	telephone	different numbers

TABLE III
RESULTS FOR THE CORE EXTENDED NIST SRE2010 FEMALE TESTS IN TERMS OF % EER, $\mathrm{minDCF08} \times 1000$ AND $\mathrm{minDCF10} \times 1000$ USING TRAINING LISTS AND PLDA MODELS. LABEL "TEL" AND "TEL+MIC" REFER TO THE DATASETS USED FOR TRAINING THE PLDA, INCLUDING OR NOT MICROPHONE DATA. "STD" AND "FPD" LABELS REFER TO STANDARD PLDA AND FPD–PLDA, RESPECTIVELY. I–VECTOR POSTERIOR LENGTH–NORMALIZATION IS PERFORMED BY MEANS OF (32)

List	Train	Test	cond2			cond3			cond4			cond1			cond5		
			EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10
tel	Std	Std	4.2	224	641	2.5	113	445	1.7	102	411	2.0	84	346	2.0	100	339
tel	Std	FPD	3.9	214	638	2.3	111	462	1.6	101	419	1.7	81	346	2.0	100	346
tel	FPD	FPD	3.9	214	635	2.4	110	450	1.6	99	415	1.8	79	345	2.0	98	336
tel+mic	Std	Std	2.6	124	460	2.2	103	405	1.1	65	303	1.8	68	258	1.9	105	335
tel+mic	Std	FPD	2.3	114	455	2.1	103	402	1.0	60	296	1.7	63	254	2.0	103	347
tel+mic	FPD	FPD	2.3	112	455	2.0	100	396	1.0	59	288	1.6	60	253	2.0	101	344

TABLE IV
RESULTS FOR CUTS OF 3–60 SECOND TEST DATA, USING DIFFERENT LENGTH–NORMALIZATION APPROACHES. THE PLDA PARAMETERS ARE TRAINED USING BOTH MICROPHONE AND TELEPHONE DATA. LABELS "STD" AND "FPD" REFER TO STANDARD PLDA AND FPD–PLDA, RESPECTIVELY, AND THE NUMERIC SUFFIX OF FPD CORRESPONDS TO THE I–VECTOR POSTERIOR LENGTH–NORMALIZATION METHOD

Train	Test	cond2			cond3			cond4			cond1			cond5		
		EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10
Std	Std	9.1	384	812	7.8	368	832	7.3	312	695	7.0	273	630	6.7	337	729
Std	FPD1 (eq. 31)	6.7	327	791	6.1	343	838	5.2	259	676	4.8	232	603	6.2	322	722
Std	FPD2 (eq. 32)	6.7	328	791	6.2	343	838	5.2	259	676	4.7	232	603	6.2	323	722
FPD2	FPD2	6.5	327	796	6.3	355	837	5.0	255	676	4.6	229	601	6.3	328	731

particular, the covariance of the posterior of the speaker identity variable:

$$\mathbf{\Lambda}_{y|(E,T)} = \mathbf{I} + n_E \mathbf{U}^{\mathrm{T}} \mathbf{\Lambda} \mathbf{U} + \sum_{i \in T} \mathbf{U}^{\mathrm{T}} \mathbf{\Lambda}_{eq,i} \mathbf{U}, \qquad (41)$$

depends only on the test i–vector covariance, and on the number of target segments. If the number of target segments per speaker is fixed, computing the term $\mathbf{\Lambda}^{-1}_{y|(E_k,T)}$ for each target speaker becomes a per–test cost because it can be computed only once. Computing the score function, given $\mathbf{\Lambda}^{-1}_{y|(E_k,T)}$, has thus complexity $O(S^2)$.

Table I summarizes the results presented in this Section. The costs have been divided into per–segment costs, depending on the number N of segments in the set, per–speaker fixed costs, and the per–trial costs.

The FP-PLDA approach has a notably higher complexity that standard PLDA. The Asymmetric FPLDA reduces the per–trial cost by a factor S, speeding–up the scoring computation when the number of target speakers is high. However, the duration of the target segments affects the accuracy of the approximation, and possibly the performance gain with respect to standard PLDA.

IX. EXPERIMENTAL RESULTS

The proposed PLDA model aims at compensating duration mismatches in i–vector estimates. Thus, a dataset was defined that consists of speech segments, from NIST SRE10 extended core condition, which were cut, after Voice Activity Detection, to obtain segments of variable duration in the range 3–30, 10–30, 3-60, and 10–60 seconds, respectively. These sets of segments have been scored according to the official NIST SRE 2010 conditions 1–5 [21], which are summarized in Table II.

In these experiments, we used cepstral features, extracted using a 25 ms Hamming window. 19 Mel frequency cepstral coefficients together with log-energy were calculated every 10 ms. These 20-dimensional feature vectors were subjected to short time mean and variance normalization using a 3s sliding window. Delta and double delta coefficients were then computed using a 5-frame window giving 60-dimensional feature vectors. Segmentation was based on the BUT Hungarian phoneme recognizer and relative average energy thresholding. Also, short segments were pruned out, after which the speech segments were merged together.

The i–vector extractor was based on a 2048–component full covariance gender–independent UBM, trained using NIST SRE 2004–2006 data. Gender–dependent i–vector extractors for the reference system were trained using the data of NIST SRE 2004–2006, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2.

All these experiments were performed using i–vector posteriors with dimension $M = 400$. The PLDA was trained with a speaker variability sub–space of dimension $S = 120$, and full channel variability sub–space. Although both female and male speaker tests were performed, we report more detailed results on the female datasets only, because the NIST SRE 2010

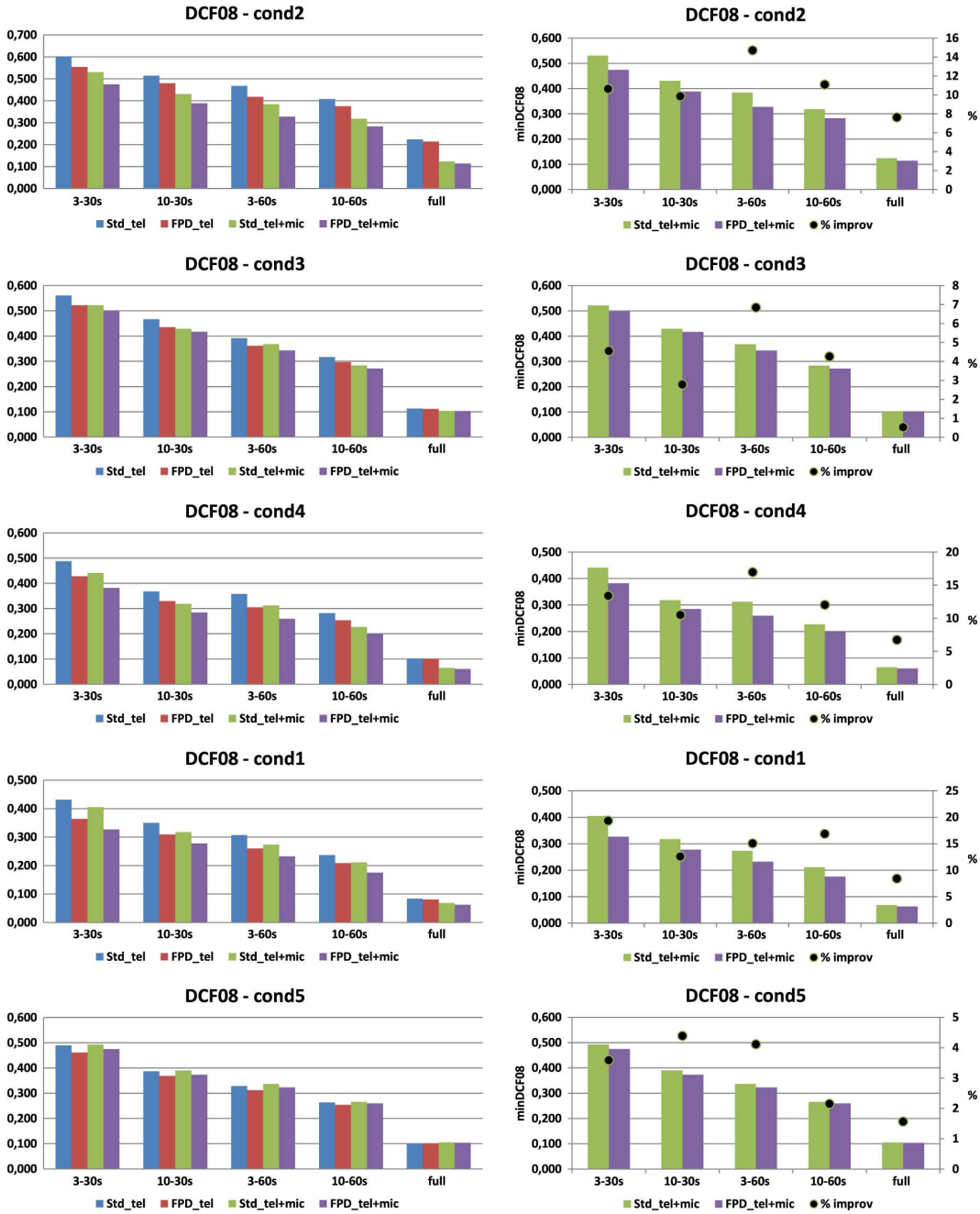

Fig. 1. Results for test cuts of variable duration, randomly chosen from the extended NIST SRE2010 female tests. On the left side, comparison of the minDCF08 obtained using PLDA and FPD–PLDA trained with different duration segments. On the right side, comparison of the minDCF08 obtained in the "tel+mic" condition, and % improvement shown as a black dot.

core test on female speakers is known to be more difficult, thus more often compared in the literature. The results on the male speakers confirm the ones reported for female speakers, as will be shown in Table VI.

Table III summarizes the results of the tests performed on the NIST SRE 2010 female extended conditions, including the core condition (cond5), in terms of percent Equal Error Rate and normalized minimum Detection Cost Function (DCF) as defined by NIST for SRE08 and SRE10 evaluations [21]. In this table, the PLDA and FPD–PLDA systems are compared using the original interview data, or telephone conversations, without any cut. Labels "tel" and "tel+mic" refer to the datasets used for training the PLDA parameters, including telephone data only, or additional microphone data. Labels "Std" and "FPD" refer to the standard and the Full Posterior Distribution PLDA, respectively. The first two rows give the baseline results, obtained using standard i–vectors trained on telephone data only, for the five NIST 2010 conditions. It can be observed that the matched conditions cond5 and cond1, tel–tel and int–int, respectively, achieve the best results, whereas the difficulty of the task decreases from cond2 to cond4. The same behavior is confirmed for the other experimental conditions, shown in the remaining lines, and for the other tests using variable duration segments. The new model not only keeps the accuracy of the standard model, as expected for long segments, but also shows an approximately 7% relative improvement in three conditions. The third row describes the effect of using the i–vector covariance also in training. As expected, since the training segments have long durations, the

TABLE V
RESULTS FOR CUTS OF VARIABLE DURATION TEST DATA, RANDOMLY CHOSEN FROM THE EXTENDED NIST SRE2010 FEMALE TESTS, IN TERMS OF % EER, $\mathrm{minDCF08} \times 1000$ AND $\mathrm{minDCF10} \times 1000$ USING DIFFERENT PLDA MODELS. THE PLDA PARAMETERS ARE TRAINED USING BOTH MICROPHONE AND TELEPHONE DATA, LABELS "STD", "FPD", AND "SV" REFER TO STANDARD PLDA, FPD–PLDA, AND SV–PLDA, RESPECTIVELY. I–VECTOR POSTERIOR LENGTH–NORMALIZATION IS PERFORMED BY MEANS OF (32)

Test	Duration	cond2			cond3			cond4			cond1			cond5		
		EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10
Std	3–30	12.4	531	921	11.3	521	915	11.1	441	864	9.8	405	794	10.6	493	915
FPD	3–30	9.8	474	901	9.3	498	929	8.3	382	849	7.6	327	756	9.7	475	912
SV	3–30	10.9	524	924	11.5	590	967	9.5	433	866	8.8	374	799	11.0	537	927
Std	10–30	9.0	431	890	8.6	429	900	6.6	318	820	7.0	317	707	7.6	390	856
FPD	10–30	7.7	388	873	7.5	417	893	5.7	285	785	5.5	278	650	7.2	373	836
SV	10–30	8.5	425	888	8.6	451	919	6.3	314	827	6.6	307	716	7.8	402	855
Std	3–60	9.1	384	812	7.8	368	832	7.3	312	695	7.0	273	630	6.7	337	729
FPD	3–60	6.7	328	791	6.2	343	838	5.2	259	676	4.7	232	603	6.2	323	722
SV	3–60	7.1	351	809	7.5	429	877	5.7	275	705	5.2	251	620	7.2	359	742
Std	10–60	7.0	318	787	5.0	283	777	4.7	227	636	4.9	211	558	4.9	265	701
FPD	10–60	5.7	283	761	4.8	271	806	3.9	200	603	4.1	176	555	4.7	260	693
SV	10–60	5.9	297	780	5.5	307	825	4.1	213	639	4.1	190	561	4.9	276	710
Std	Full	2.6	124	460	2.2	103	405	1.1	65	303	1.8	68	258	1.9	105	335
FPD	Full	2.3	114	455	2.1	103	402	1.0	60	296	1.7	63	254	2.0	103	347
SV	Full	2.3	115	454	2.0	104	410	1.0	62	298	1.6	63	260	2.1	104	349

TABLE VI
RESULTS FOR CUTS OF VARIABLE DURATION TEST DATA, RANDOMLY CHOSEN FROM THE EXTENDED NIST SRE2010 MALE TESTS. SEE TABLE V CAPTIONS

Test	Duration	cond2			cond3			cond4			cond1			cond5		
		EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10	EER	DCF 08	DCF 10
Std	3–30	8.3	379	825	9.8	448	923	8.9	364	766	6.0	280	697	9.4	436	857
FPD	3–30	6.2	325	795	8.0	432	929	6.6	308	747	4.3	224	641	8.6	419	842
Std	10–30	5.7	286	777	6.8	368	892	5.8	273	701	4.2	192	607	6.7	326	811
FPD	10–30	4.7	243	741	6.1	326	877	5.1	240	665	3.1	157	529	6.3	308	771
Std	3–60	5.8	259	645	6.3	284	753	5.9	247	596	4.5	182	464	6.3	286	692
FPD	3–60	4.1	204	605	5.6	276	819	4.1	194	540	3.0	136	402	5.3	269	697
Std	10–60	3.8	196	609	5.1	251	738	3.5	172	547	2.5	116	402	4.6	224	627
FPD	10–60	2.9	159	565	4.5	231	744	3.0	149	523	2.0	88	370	4.2	218	631
Std	Full	1.1	57	270	1.9	86	353	1.2	47	200	0.6	28	138	1.5	82	310
FPD	Full	0.9	47	249	1.7	83	356	1.1	45	192	0.5	24	121	1.4	84	319

results are similar to the ones reported in the second row. The last three rows show the effect of adding microphone data in training the PLDA parameters: sensible performance improvement is obtained, excluding, as expected, the matched tel–tel condition 5.

Since the system trained with the "tel" list performs worse than the one trained with the "tel+mic" list, all the remaining experiment on the NIST 2010 data, whenever not mentioned, have been performed with the latter. Table IV compares, in its first three rows, the performance of the PLDA and FPD–PLDA classifiers using the two length–normalization methods illustrated in Section VII on the 3–60 seconds cuts. The results of the last row show that there is no advantage in using the full i–vector posterior in training the PLDA models. The effect of the two length–normalization approaches is comparable, thus in the following we will present only the results obtained with the Projected Length Normalization (FPD2) (32).

The tests on variable duration cuts, randomly chosen from the extended NIST SRE2010 female set, are shown in Fig. 1 and Table V. Fig. 1 compares on its left side column the minDCF08 obtained using PLDA and FPD–PLDA trained with different data, either telephone data only or telephone and additional microphone data. The set of minDCF08 results are shown as a function of the training conditions and of the duration of the cuts. On the right side column, the minDCF08 results obtained using the parameters trained in the "tel+mic" condition for PLDA and FP–PLDA are compared. The figure also shows as black dots the percent improvement obtained by FPD–PLDA with respect to standard PLDA. Excluding the matched tel–tel condition 5, the PLDA models trained adding microphone data, indicated as "Std_tel+mic" on the legend, are always better than the corresponding models trained with telephone data only, and FPD–PLDA shows always a relative improvement, quite small for long enough segments, but up to 20% depending on the average duration of the small cuts.

Table V also reports the results for the SV–PLDA approach. Since the training segments are long, as we did with FPD–PLDA, the SV–PLDA model was trained using π–vectors without considering the π–vector covariances. For the same reason, the $\mathbf{T}$ matrix was not retrained. It seems that the ML estimation of the π–vectors is not as effective as the MAP estimation of i–vectors, making the FPD–PLDA approach more attractive.

The results given in Table VI confirm the quality of the our approach for male speakers.

Pooled results for female and male speakers are reported in Table VII for the NIST 2012 SRE evaluation experiments described below. In these experiments, the acoustic features

TABLE VII
NIST SRE 2012 EXTENDED SET: MINIMUM $C_{primary}$ COMPARISON OF FPD–PLDA AND ASYMMETRIC FPD–PLDA. THE NUMBERS ASSOCIATED TO THE CONDITIONS REFER TO THE MEAN DURATION OF THE SEGMENTS, AFTER VOICE ACTIVITY DETECTION, AND TO THE CORRESPONDING STANDARD DEVIATION

System	Condition 1 interview without added noise 45s – 41	Condition 2 phone call without added noise 56s – 48	Condition 3 interview with added noise 75s – 37	Condition 4 phone call phone with added noise 110s – 56	Condition 5 phone call from a noisy environment 57s – 48
PLDA	0.254	0.224	0.255	0.245	0.218
FPD–PLDA	0.259	0.209	0.245	0.243	0.195

were again 60–dimensional MFCCs, modeled with a 2048 components full–covariance UBM. The i–vector dimension was increased to $M = 600$. Moreover, Linear Discriminant Analysis was performed to reduce the i–vector dimensionality to 200, before applying i–vector whitening and length normalization. Since the resulting i–vectors are already small, no dimensionality reduction was applied for the speaker sub–space, i.e. the speaker sub–space was set to 200. The UBM was trained on speech segments taken from previous the NIST 2004, 2005, 2006, 2008 and 2010 evaluation corpora, and from the enrollment set of NIST 2012 evaluation. Additionally, the Fisher, Switchboard Phase 2 and Switchboard Cellular datasets were used to train the i–vector extractor and the PLDA parameters. Due to the enormous amount of trials involved in the evaluation (some tens of millions), we did not test the complete FPD–PLDA approach. Since NIST 2012 enrollment segments are on average quite long, we were able to test FPD–PLDA according to the Asymmetric FPD–PLDA approach illustrated in Section VIII-D. Moreover, we had empirical evidence that representing a target speaker by means of a single i–vector, computed as the average of all its i–vectors, provides higher accuracy with respect to the standard multi–session PLDA scoring. The same approach was, thus, followed for obtaining the FPD–PLDA scores.

The results comparing standard PDLA and Asymmetric FPD–PLDA are given in Table VII in terms of minimum $C_{primary}$, the primary cost measure defined by NIST [25] for this evaluation. These results clearly show that although the Asymmetric FP–PLDA introduces some approximations, it is still able to outperform standard PLDA in most of the conditions. In particular, it gains for conditions 2 and 5, which include short and variable duration segments, whereas it obtains almost the same performance for the long duration segments of conditions 3 and 4. Condition 1 is an exception, we speculate that errors in voice activity and interviewee detection may lead to the estimation of an incorrect i–vector covariance posterior. This effect might not manifest itself on condition 3 because the average segment duration is higher.

The real–time contribution of the PLDA techniques evaluated in this work, with i–vector dimension $M = 400$, is compared in Table VIII. It reports the scoring time per trial required by standard PLDA, by Full Posterior PLDA, by Asymmetric FPD-PLDA, and by an optimized implementation of PLDA, respectively, as a function of the number of enrollment and test pairs which must be scored. All times are given in milliseconds. It is worth noting that the optimized PLDA fully exploits both the "fixed set of target speakers scenario" and the performance of optimized matrix–to-matrix operations for scoring multiple test segments, whereas the Asymmetric FPD-PLDA is optimized for scoring a single test against a fixed set of target speakers. The number of enrollment and test segments shown in Table VIII have been selected so that the contribution of each

TABLE VIII
REAL–TIME SCORING (IN MILLISECONDS PER TRIALS) FOR THE STANDARD PLDA, FULL POSTERIOR DISTRIBUTION PLDA, ASYMMETRIC FPD-PLDA, AND FOR AN OPTIMIZED IMPLEMENTATION OF PLDA

Enrollment segments	Test segments	Standard PLDA	FPD PLDA	Asymmetric FPD–PLDA	Opt. PLDA
1	1	0.44	102.42	55.99	0.4680
100	100	0.36	7.12	0.54	0.0014
1000	100	0.33	7.05	0.07	0.0007
1000	1000	0.33	6.42	0.07	0.0002

PLDA technique in a different application scenario could be appreciated. The first row presents the single trial scenario, where, excluding the pre–computation of $\mathbf{\Lambda}_y$ for standard PLDA, one cannot obviously perform any pre–computation or optimization that can be used for speeding–up the scoring of other pairs. This is the worst case for FPD–PLDA and Asymmetric FPD–PLDA, which are approximately 200 and 100 times slower than PLDA, respectively. It is also the worst case for our highly optimized PLDA implementation, which does not have any advantage in this scenario. A dramatic speedup with respect to single pair scoring is obtained, instead, by the optimized PLDA when the number of enrollment and test segments (i.e., of trial pairs) increases, as shown in the second row for 10000 trials. This is also true for FPD–PLDA and Asymmetric FPD–PLDA. Comparing the scoring times for the $100 - 100$ and $1000 - 100$ scenarios, shown in the second and third row, respectively, one can appreciate the importance of Asymmetric FPD-PLDA, which does not suffer the FPD–PLDA overhead for a 10 times larger enrollment set. Finally, these results, and the ones reported in the last row, show that whenever a large set of tests has to be performed against a large, but fixed, set of target speakers, FPD–PLDA is approximately 20 times slower than standard PLDA, whereas Asymmetric FPD-PLDA is four times faster. However, in this scenario, the optimized PLDA is more than two order of magnitude faster than any other technique.

It is worth noting that these are pure classification times that do not include i-vector extraction time, which depends on the length of the speech segment because the UBM statistics are collected frame by frame. Taking also into account the i-vector extraction time, the ratio of the single pair scoring time between FPD–PLDA and standard PLDA reduces from 200 to 4 approximately.

X. CONCLUSIONS

A PLDA model which exploits the uncertainty of the i-vector extraction process has been presented. We derived the formulation of the likelihood for a Gaussian PLDA model based on the i–vector posterior distribution, and illustrated a new PLDA model, where the inter–speaker variability is assumed to have an segment–dependent distribution, showing that we can rely

on the standard PLDA framework simply replacing the likelihood definition.

We have proposed two i–vector pre–processing techniques, and compared their effects on the system accuracy, showing that an approximate version of a linearized length normalization is effective.

The complexity of the PLDA and FPD–PLDA implementations have been analyzed, and an Asymmetric FPD–PLDA approach has been proposed, which allows obtaining a substantial complexity reduction in a practical detection scenario. The results obtained both on the extended core tests and on short cuts of different duration of the NIST 2010, and on the extended tests of NIST 2012 evaluations, confirm that the FPD–PLDA outperforms PLDA mostly for short variable duration test segments.

ACKNOWLEDGMENT

Computational resources for this work were provided by the high performance computing clusters of BUT Faculty of Information Technology, and by HPC@POLITO (http://www.hpc.polito.it).

REFERENCES

[1] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i–vector posterior distributions," in *Proc. ICASSP '13*, 2013, pp. 7644–7648.
[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front–end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
[3] P. Matějka, O. Glembek, F. Castaldo, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i–vector speaker verification," in *Proc. ICASSP '11*, 2011, pp. 4828–4831.
[4] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey '10*, 2010, pp. 194–201.
[5] M. Senoussaoui, P. Kenny, N. Brummer, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender–independent speaker recognition," in *Proc. INTERSPEECH '11*, 2011, pp. 25–28.
[6] J. Villalba and N. Brümmer, "Towards fully bayesian speaker recognition: Integrating out the between-speaker covariance," in *Proc. INTERSPEECH '11*, 2011, pp. 505–508.
[7] S. Cumani, N. Brümmer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1217–1227, Jun. 2013.
[8] S. Cumani, N. Brümmer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i–vector space," in *Proc. ICASSP '11*, 2011, pp. 4852–4855.
[9] T. Hasan and J. H. L. Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 842–853, Apr. 2013.
[10] V. Hautamaki, T. Kinnunen, F. Sedlak, K. Lee, B. Ma, and H. Li, "Sparse classifier fusion for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1622–1631, Aug. 2013.
[11] B. Srinivasan, L. Yuancheng, D. Garcia-Romero, D. Zotkin, and R. Duraiswami, "A symmetric kernel partial least squares framework for speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1415–1423, Jul. 2013.
[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Process.*, vol. 10, no. 1-3, pp. 31–44, 2000.
[13] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep. CRIM-06/08-13, 2005.
[14] P. Kenny, "Bayesian speaker verification with heavy–tailed priors," Keynote presentation, Odyssey 2010 The Speaker and Language Recognition Workshop 2010 [Online]. Available: http://www.crim.ca/perso/patrick.kenny/kenny_Odyssey2010.pdf
[15] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 Speaker Recognition Evaluation," in *Proc. INTERSPEECH '13*, 2013, pp. 1981–1985.
[16] R. Saeidi *et al.*, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. INTERSPEECH '13*, 2013, pp. 1986–1990.
[17] D. Colibro *et al.*, "Nuance–Politecnico di Torino 2012 NIST Speaker Recognition Evaluation system," in *Proc. INTERSPEECH '13*, 2013, pp. 1996–2000.
[18] J. Villalba, E. Lleida, A. Ortega, and A. Miguel, "The I3A Speaker Recognition system for NIST SRE12: Post-evaluation analysis," in *Proc. INTERSPEECH '13*, 2013, pp. 3689–3693.
[19] O. Plchot *et al.*, "Developing a speaker identification system for the DARPA RATS Project," in *Proc. ICASSP '13*, 2013, pp. 6768–6772.
[20] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, and Y. Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion," in *Proc. ICASSP '13*, 2013, pp. 6773–6777.
[21] "The NIST year 2010 speaker recognition evaluation plan," [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf
[22] P. Kenny, T. Stafylakis, J. A. P. Ouellet, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. ICASSP '13*, 2013, pp. 7649–7653.
[23] B. Borgstrom and A. McCree, "Supervector bayesian speaker comparison," in *Proc. ICASSP '13*, 2013, pp. 7693–7697.
[24] D. Garcia-Romero and A. McCree, "Subspace-constrained supervector plda for speaker verification," in *Proc. INTERSPEECH '13*, 2013, pp. 2479–2483.
[25] "The NIST year 2012 speaker recognition evaluation plan," [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf
[26] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[27] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i–vector length normalization in speaker recognition systems," in *Proc. Interspeech '11*, 2011, pp. 249–252.
[28] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of i–vector extraction," in *Proc. ICASSP '11*, 2011, pp. 4516–4519.

Sandro Cumani received the M.S. degree in Computer Engineering from the Politecnico di Torino, Torino, Italy, in 2008, and the Ph.D. degree in Computer and System Engineering of Politecnico di Torino in 2011. He worked at the Brno University of Technology, Czech Republic, and is now with Politecnico di Torino. His current research interests include machine learning, speech processing and biometrics, in particular speaker and language recognition.

Oldřich Plchot received the Master degree in Computer Science and Engineering from Brno University of Technology, Czech Republic, in 2007, and he is pursuing the Ph.D. degree in Computer Science and Engineering at the same University. His current research interests include machine learning, data engineering, speech processing and biometrics, in particular speaker and language recognition.

Pietro Laface received the M.S. degree in Electronic Engineering from the Politecnico di Torino, Torino, Italy, in 1973.

Since 1988 it has been full Professor of Computer Science at the Dipartimento di Automatica e Informatica of Politecnico di Torino, where he leads the speech technology research group. He has published over 120 papers in the area of pattern recognition, artificial intelligence, and spoken language processing. His current research interests include all aspects of automatic speech recognition and its applications, in particular speaker and spoken language recognition.