Fast Scoring of Full Posterior PLDA Models

1

Fast scoring of Full Posterior PLDA models

Sandro Cumani

***Abstract*—A low–dimensional representation of a speech segment, the so–called i–vector, in combination with Probabilistic Linear Discriminant Analysis (PLDA) models, is the current state–of–the–art in speaker recognition. An i–vector is a compact representation of a Gaussian Mixture Model (GMM) supervector, which captures most of the GMM supervectors variability. It is usually obtained by a MAP estimate of the mean of a posterior distribution. A new PLDA model has been recently presented that, unlike the standard one, exploits the intrinsic i–vector uncertainty. This approach, referred to in this paper as Full Posterior Distribution PLDA (FP–PLDA), is particularly effective for speaker detection of short and variable duration speech segments. It is, however, computationally far more expensive than standard PLDA, making it unattractive for real applications. This paper presents three simplifications of FP–PLDA based on approximate diagonalizations of matrices involved in FP–PLDA scoring. Using in sequence these approximations allows obtaining computational costs comparable to PLDA models, with only a small performance degradation with respect to the more accurate, but less efficient, FP–PLDA models. In particular, up to 10% better performance than PLDA is obtained, with similar computational complexity, on short speech segments of variable duration, randomly extracted from the interviews and telephone conversations included in the NIST SRE 2010 extended dataset. The benefits of the proposed diagonalization approaches have also been confirmed on a short utterance text–independent verification task, where approximately 43% and 34% improvement of the EER and minimum DCF08, respectively, has been obtained with respect to PLDA.**

***Index Terms*—Speaker Recognition, I-vectors, I-vector extraction, Probabilistic Linear Discriminant Analysis.**

I. Introduction

Probabilistic Linear Discriminant Analysis (PLDA) [1] classifiers based on i–vectors [2] are among the best models for speaker recognition [3], [4], [5], [6], [7], [8], [9]. Some PLDA systems for the last NIST 2012 Speaker Recognition Evaluation and for the DARPA RATS project have been described in [10], [11], [12], [13], [14], [15]. Standard PLDA, however, does not exploit the covariance of the i–vector distribution, which accounts for the "uncertainty" of the i–vector extraction process. This uncertainty is affected by the length of the speech segments that are used for characterizing a speaker. Shorter utterances tend to produce larger covariances, so that i–vector estimates become less reliable.

A new PLDA model has been recently proposed [16], [17], [18], which incorporates the intrinsic uncertainty of the i–vector extraction process. In this model, referred to as Full Posterior Distribution PLDA (FP–PLDA), the inter–speaker variability has an utterance dependent distribution. Similar approaches have shown to outperform PLDA on short variable duration segments [18], [17], [19]. The main drawback of all these models is their computational complexity.

In [18], the complexity of the PLDA and FP–PLDA implementations has been analyzed, and an Asymmetric FP–PLDA (AFP–PLDA) approach has been proposed, which allows obtaining a substantial complexity reduction in a practical detection scenario where test utterances are short but the target utterances have long duration. FP–PLDA and AFP–PLDA are more accurate than standard PLDA, but they are far more expensive, and the AFP–PLDA is only useful in presence of long target utterances. Thus, in this work we present three different techniques for the simplification of FP–PLDA, based on the diagonalization of some matrices that are appear both in i–vector extraction and in scoring, suitable for scenarios involving short target and test utterances. The advantage of the proposed Diagonalized FP–PLDA approach is that better performance than PLDA can be obtained with comparable scoring complexity, while memory requirements for storing the target speaker representations are greatly reduced with respect to FP-PLDA. These techniques have been tested using two different datasets. The first set includes cuts of variable duration, extracted from conversations recorded from different channels included NIST SRE 2010 extended core tests [20]. This dataset is the same used for assessing the performance of the FP–PLDA approach in [18]. The second set of experiments has been performed on a short utterance text–independent verification task. The application of the diagonalization operations presented in this work dramatically speeds–up test segment scoring with respect to FP–PLDA, and allows obtaining a system that is almost as fast as a PLDA system, but sensibly more accurate for short utterances.

The paper is organized as follows: in order to make the paper self–contained, the i-vector extraction process, and the FP–PLDA model are recalled in Sections II and III, respectively. A detailed analysis of the FP–PLDA and standard PLDA complexity is given in Section IV, Section V illustrates the proposed methods to simplify FP–PLDA, and compares the computational complexity of these approaches, showing that scoring costs can be dramatically reduced allowing the approximate FP–PLDA to be almost computationally inexpensive as PLDA. The experimental results are given in Section VI, and conclusions are drawn in Section VII.

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:

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

Sandro Cumani is with the Dipartimento di Automatica e Informatica, Politecnico di Torino, 10143 Torino, Italy (e-mail: cumani@polito.it .

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 an M-dimensional realization of a latent variable $\mathbf{W}$, 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 [21]. Given $\mathbf{T}$, and the set of τ feature vectors $\mathcal{X} = \{\mathbf{x}_1\mathbf{x}_2 \dots \mathbf{x}_\tau\}$ 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 which represents the segment, is computed as the Maximum a Posteriori (MAP) point estimate of the variable $\mathbf{W}$, i.e., as the mean $\boldsymbol{\mu}_\mathcal{X}$ of the posterior distribution $P_{\mathbf{W}|\mathcal{X}}(\mathbf{w})$. It has been shown in [21] 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 precision matrix and mean vector:

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

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 $\mathbf{T} = \left(\mathbf{T}^{(1)T}, \dots, \mathbf{T}^{(C)T}\right)^T$, 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. Gaussian Full Posterior Distribution PLDA model

An utterance u is represented in the standard Gaussian PLDA model by the i–vector posterior mean $\boldsymbol{\mu}$, which is assumed to be the combination of three terms:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{U}\mathbf{y} + \mathbf{e} \ , \tag{5}$$

where $\mathbf{m}$ is the i–vector mean, $\mathbf{y}$ is a speaker factor sampled from a normal prior distribution, matrix $\mathbf{U}$ typically constrains the speaker factor to be of lower dimension than the i-vectors, and the residual noise term $\mathbf{e}$ is the realization of a Gaussian distributed random variable $\mathbf{E}$ with full precision matrix $\boldsymbol{\Lambda}$, i.e.:

$$\mathbf{Y} \sim \mathcal{N}(0, \mathbf{I}) \ , \quad \mathbf{E} \sim \mathcal{N}(0, \boldsymbol{\Lambda}^{-1}) \ . \tag{6}$$

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 [16], [17], [18] the PLDA model has been extended to exploit this additional information. This new model, referred to as PLDA based on the "Full Posterior Distribution" of $\mathbf{W}|\mathcal{X}$, assumes that an i–vector can be described as:

$$\boldsymbol{\mu}_i = \mathbf{m} + \mathbf{U}\mathbf{y} + \overline{\mathbf{e}}_i \ , \tag{7}$$

where the difference with equation (5) is that the residual noise $\mathbf{e}$ has been replaced by the utterance–dependent term $\overline{\mathbf{e}}_i$, sampled from the utterance–dependent distribution $\overline{\mathbf{E}}_i$. The prior distributions for the residual noise and speaker factor are given by:

$$\overline{\mathbf{E}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_{eq,i}^{-1}) \ , \tag{8}$$

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ , \tag{9}$$

respectively, where $\boldsymbol{\Gamma}_i$ is the precision matrix produced by the i–vector extractor, and the equivalent precision matrix $\boldsymbol{\Lambda}_{eq,i}$ is:

$$\boldsymbol{\Lambda}_{eq,i} = \left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Gamma}_i^{-1}\right)^{-1} \ . \tag{10}$$

In [16], [18] it has been shown that the likelihood that a set of n utterances $u_1 \dots u_n$, i.e., of i–vectors $\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n$, belongs to the same speaker, can be computed according to the FP–PLDA model as:

$$\begin{aligned} &\log P(\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n | H_s) = \\ &\sum_i \left[\frac{1}{2}\log|\boldsymbol{\Lambda}_{eq,i}| - \frac{M}{2}\log 2\pi - \frac{1}{2}(\boldsymbol{\mu}_i - \mathbf{m})^T \boldsymbol{\Lambda}_{eq,i}(\boldsymbol{\mu}_i - \mathbf{m})\right] \\ &- \frac{1}{2}\log|\boldsymbol{\Lambda}_y| + \frac{1}{2}\boldsymbol{\mu}_y^T \boldsymbol{\Lambda}_y \boldsymbol{\mu}_y - \frac{S}{2}\log 2\pi \ , \end{aligned} \tag{11}$$

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

$$\begin{aligned} \boldsymbol{\Lambda}_y &= \mathbf{I} + \sum_i \mathbf{U}^T \boldsymbol{\Lambda}_{eq,i} \mathbf{U} \\ \boldsymbol{\mu}_y &= \boldsymbol{\Lambda}_y^{-1} \mathbf{U}^T \sum_i \boldsymbol{\Lambda}_{eq,i} \left(\boldsymbol{\mu}_i - \mathbf{m}\right) \ . \end{aligned} \tag{12}$$

This equation is exactly the same required by the PLDA model, just replacing in FP–PLDA precision matrix $\boldsymbol{\Lambda}$ appearing in PLDA by $\boldsymbol{\Lambda}_{eq,i}$, which accounts for an utterance–dependent i–vector precision matrix.

IV. Complexity analysis

Given a set of n enrollment utterances $u_{e_1} \dots u_{e_n}$ for a target speaker, and a set of m test utterances $u_{t_1} \dots u_{t_m}$ of a single unknown speaker, the speaker verification log–likelihood ratio s is:

$$s = \log \frac{l\left(u_{e_1} \dots u_{e_n}, u_{t_1} \dots u_{t_m} | H_s\right)}{l\left(u_{e_1} \dots u_{e_n} | H_s\right) l\left(u_{t_1} \dots u_{t_m} | H_s\right)} \ , \tag{13}$$

where H_s is the hypothesis that the two set of utterances belong to the same speaker.

The complexity of the log–likelihood computation accounts for three separate contributions. The first contribution is given by the operations that can be independently performed on

each utterance, which will be referred as per–utterance costs (excluding i–vector extraction costs). The second contribution involves all operations that can be independently performed on the set of utterances for a speaker (either the target or the test speaker), but do not depend on the number of utterances in the set. These operations will be referred to as per–speaker operations, or per–target and per–test operations wherever the distinction is relevant. The final contribution, the per–trial complexity, is given by the operations which jointly involve the target and test sets. This distinction is not relevant for naïve scoring implementations, but is relevant, instead, in scenarios with a fixed set of target speakers, because the per–target terms can be precomputed, and per–test terms need to be computed only once regardless of the number of target speakers. It is worth noting that the per–utterance complexity should also account for the complexity of the i–vector extraction. The computation of the i–vector covariance matrix, for each utterance, has complexity $O(M^3)$ [22], where M is the i–vector dimension. This complexity dominates the per–set costs, because, as shown in Section IV-A and IV-B, and summarized in Table I, both PLDA and FP–PLDA have lower per–set costs.

Replacing (11) in (13), the speaker verification log–likelihood ratio for a target set E and a test set T can be computed as:

$$\begin{aligned} llr(E,T) &= \log \frac{l(E,T|H_s)}{l(E|H_s)l(T|H_s)} \\ &= \sigma(E,T) - \sigma(E) - \sigma(T) + \frac{S}{2}\log 2\pi \ , \end{aligned} \quad (14)$$

where the scoring function σ is defined as:

$$\sigma(G) = -\frac{1}{2}\log\left|\mathbf{\Lambda}_{y|G}\right| + \frac{1}{2}\boldsymbol{\mu}_{y|G}^T\mathbf{\Lambda}_{y|G}\boldsymbol{\mu}_{y|G} \ . \quad (15)$$

and

$$\mathbf{\Lambda}_{y|G} = \mathbf{I} + \sum_{i\in G}\mathbf{U}^T\mathbf{\Lambda}_{eq,i}\mathbf{U} \quad (16a)$$

$$\boldsymbol{\mu}_{y|G} = \mathbf{\Lambda}_y^{-1}\mathbf{U}^T\sum_{i\in G}\mathbf{\Lambda}_{eq,i}\left(\boldsymbol{\mu}_i - \mathbf{m}\right) \quad (16b)$$

are the posterior parameters of $\mathbf{Y}$ conditioned on the i–vectors in the set G. 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.

A. Complexity of the standard Gaussian PLDA

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

$$\mathbf{\Lambda}_{y|(E,T)} = \mathbf{I} + (n_E + n_T)\mathbf{U}^T\mathbf{\Lambda}\mathbf{U} \quad (17a)$$

$$\begin{aligned} \boldsymbol{\mu}_{y|(E,T)} &= \mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{U}^T\mathbf{\Lambda}\left(\sum_{i\in E}\left(\boldsymbol{\mu}_i - \mathbf{m}\right) + \sum_{i\in T}\left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right) \\ &= \mathbf{\Lambda}_{y|(E,T)}^{-1}\left(\mathbf{F}_E + \mathbf{F}_T\right) \ , \end{aligned} \quad (17b)$$

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:

$$\mathbf{F}_E = \mathbf{M}\sum_{i\in E}\left(\boldsymbol{\mu}_i - \mathbf{m}\right) \ , \ \mathbf{F}_T = \mathbf{M}\sum_{i\in T}\left(\boldsymbol{\mu}_i - \mathbf{m}\right) \ , \quad (18)$$

and $\mathbf{M} = \mathbf{U}^T\mathbf{\Lambda}$ is an $S \times M$ matrix, where S is the PLDA speaker sub–space dimension. Using these definitions, the scoring function $\sigma(E,T)$ can be rewritten as:

$$\begin{aligned} \sigma(E,T) = &-\frac{1}{2}\log\left|\mathbf{\Lambda}_{y|(E,T)}\right| + \mathbf{F}_E^T\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_T \\ &+ \frac{1}{2}\mathbf{F}_T^T\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_T + \frac{1}{2}\mathbf{F}_E^T\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_E \ . \end{aligned} \quad (19)$$

Computing the projected statistics (18) has per–utterance complexity $O(NM)$, where N is the number of utterances in the set, and a per–set complexity $O(MS)$.

TABLE I
COMPARISON OF THE COMPLEXITY OF TWO IMPLEMENTATIONS OF PLDA AND OF FPD–PLDA. PER–UTTERANCE COSTS SHOULD BE MULTIPLIED BY THE NUMBER OF UTTERANCES N OF A GIVEN SPEAKER. PER–TEST AND PER–TRIAL COSTS DO NOT DEPEND ON THE NUMBER OF SPEAKER UTTERANCES. THE COSTS IN THIS TABLE ARE RELATED ONLY TO PLDA, I.E., EXCLUDING THE CONTRIBUTION OF I–VECTOR EXTRACTION.

System	Complexity		
	Per–utterance	Per–test	Per–trial
PLDA Naïve	M	MS	S^3
PLDA Optmized	M	MS	S
FPD–PLDA	M^3	M^2S	S^3

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. For standard PLDA, these computations have a per–trial complexity of $O(S^3)$ because the term $\mathbf{U}^T\mathbf{\Lambda}\mathbf{U}$ can be precomputed. Given $\mathbf{\Lambda}_{y|(E,T)}^{-1}$, computing $\sigma(E,T)$ has per–trial 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 $\Lambda_{y|(E,T)}$ dominates the scoring costs. However, (17a) shows that in standard PLDA this factor depends only on the number of the target and test utterances. Since each set of target utterances E_k, and the number of test utterances n_T are known, the corresponding $\mathbf{\Lambda}_{y|(E_k,T)}^{-1}$ and its log–determinant can be precomputed. Moreover, since the statistics $\mathbf{F}_{E_k}$ are also known in advance, also the terms of the scoring function $\frac{1}{2}\mathbf{F}_{E_k}^T\mathbf{\Lambda}_{y|(E_k,T)}^{-1}$ can be precomputed. It is worth noting that these terms are small S–sized vectors. Since the term depending only on the test statistics $\mathbf{F}_T$ can 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 $O(S)$, and each term $\sigma(E_k)$ can be easily precomputed. Given the statistics, the term $\sigma(T)$ has a per–set complexity of $O(S^2)$. The overall per–utterance and per–set cost, including statistics computations, are then $O(NM)$ and $O(MS)$, respectively, whereas the per–trial cost is $O(S)$.

B. Full–Posterior PLDA

The main difference between standard PLDA and FP–PLDA is that in PLDA $\mathbf{\Lambda}_{y|(E,T)}$ depends just on the number of i–vectors in the two sets, whereas in FP–PLDA it also depends on the covariance of each i–vector in the target and test sets E and T (see (16a) and 10). This does not allow applying to FP–PLDA the optimizations for speaker detection with known targets, illustrated in the previous sub–section.

The speaker variable posterior parameters can still be written as:

$$\mathbf{\Lambda}_{y|(E,T)} = \mathbf{I} + (\mathbf{\Lambda}_{eq,E} + \mathbf{\Lambda}_{eq,T}) \tag{20a}$$

$$\boldsymbol{\mu}_{y|(E,T)} = \mathbf{\Lambda}_{y}^{-1} (\mathbf{F}_{eq,E} + \mathbf{F}_{eq,T}) \ , \tag{20b}$$

where

$$\mathbf{F}_{eq,G} = \mathbf{U}^T \sum_{i \in G} \mathbf{\Lambda}_{eq,i} (\boldsymbol{\mu}_i - \mathbf{m}) \tag{21}$$

$$\mathbf{\Lambda}_{eq,G} = \mathbf{U}^T \left(\sum_{i \in G} \mathbf{\Lambda}_{eq,i} \right) \mathbf{U} \ , \tag{22}$$

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

$$\begin{aligned} \sigma(E,T) = &- \frac{1}{2} \log \left| \mathbf{\Lambda}_{y|(E,T)}^{-1} \right| + \frac{1}{2} \mathbf{F}_{eq,E}^T \mathbf{\Lambda}_{y|(E,T)}^{-1} \mathbf{F}_{eq,E} \\ &+ \frac{1}{2} \mathbf{F}_{eq,T}^T \mathbf{\Lambda}_{y|(E,T)}^{-1} \mathbf{F}_{eq,T} + \mathbf{F}_{eq,E}^T \mathbf{\Lambda}_{y|(E,T)}^{-1} \mathbf{F}_{eq,T} \ . \end{aligned} \tag{23}$$

Computing the posterior parameters (20a) has a complexity $O(NM^3)+O(M^2S)$, mainly due to the computation of $\mathbf{\Lambda}_{eq,i}$, which is much higher than the $O(NM)+O(MS)$ complexity of the standard PLDA approach. However, these computations are required only for a new target or a new test speaker. These per–utterance and per–set costs are comparable to the costs $O(NM^3)$ of the i–vector extraction [22]. Given the statistics, $\mathbf{\Lambda}_{y|(E,T)}$ can be computed with complexity $O(S^2)$ and its inversion has a $O(S^3)$ complexity. 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 precomputed as in standard PLDA, the per–trial complexity does not reduce for the fixed set of target speakers scenario.

V. Approximated Full–Posterior PLDA

The FP–PLDA model allows improving the recognition performance [16], [17], [18], however, we have shown that the per–trial score computation complexity of FP–PLDA greatly increases compared to the standard PLDA approach. In this section we introduce three simplifications of FP–PLDA for fast scoring trying to keep small the impact on the its accuracy.

A. Diagonalized i–vector posterior

The first simplification consists in approximating the i–vector posterior covariance by the diagonal matrix:

$$\mathbf{\Gamma}_i^{-1} \leftarrow (\mathbf{\Gamma}_i \circ \mathbf{I})^{-1} \ , \tag{24}$$

where $\circ$ is the element-wise product operator, and $\mathbf{I}$ is an identity matrix of the same dimension of $\mathbf{\Gamma}_i$. However, a much better approximation can be obtained by an approximated simultaneous diagonalization of the terms composing the i–vector posterior covariance matrix as proposed in [22]. In particular, let's define an approximate $\hat{\mathbf{\Gamma}}_{\mathcal{X}}$ as:

$$\hat{\mathbf{\Gamma}}_{\mathcal{X}} = \sum_c \omega_c \mathbf{T}^{(c)^T} \mathbf{\Sigma}^{(c)^{-1}} \mathbf{T}^{(c)} \tag{25}$$

where each zero–order statistics $N_{\mathcal{X}}^{(c)}$ in (3) is replaced by ω_c, the weight of the c–th component of the UBM. Let also $\mathbf{U_T} \Sigma_{\mathbf{T}} \mathbf{U}_{\mathbf{T}}^T$ be the eigen–decomposition of $\hat{\mathbf{\Gamma}}_{\mathcal{X}}$. By applying the following linear transformations to the i–vector model:

$$\begin{aligned} \hat{\mathbf{w}} &= \mathbf{U}_T^T \mathbf{w} \\ \hat{\mathbf{T}} &= \mathbf{T} \mathbf{U}_T \ , \end{aligned} \tag{26}$$

it can be easily verified that the corresponding i–vector posterior parameters are given by:

$$\begin{aligned} \hat{\mathbf{\Gamma}}_{\mathcal{X}} &= \mathbf{U}_T^T \mathbf{\Gamma}_{\mathcal{X}} \mathbf{U}_T \\ \hat{\boldsymbol{\mu}}_{\mathcal{X}} &= \mathbf{U}_T^T \boldsymbol{\mu}_{\mathcal{X}} \ , \end{aligned} \tag{27}$$

i.e., that the i–vector posterior distribution corresponds to a linear transformation of the original i–vector distribution. It is worth noting that, since both PLDA and FP–PLDA results are invariant to linear transformations of the i–vector space (provided that the parameters are estimated in the transformed space as well), the use of i–vectors computed as in (27) has no impact on standard PLDA and on FP–PLDA. Moreover, as long as the utterance zero–order statistics have the same distribution of the UBM weights, $\hat{\mathbf{\Gamma}}_{\mathcal{X}}$ is a diagonal matrix. In general, $\hat{\mathbf{\Gamma}}_{\mathcal{X}}$ is not diagonal, but, as shown in [22], zeroing its off–diagonal elements provides an acceptable approximation of the original i–vector covariace. An even better diagonalization could be obtained through Heteroscedastic Linear Discriminant Analysis (HLDA) [22], but in our experience the simpler eigen–decomposition approach already provides accurate enough results.

Using a diagonal i–vector posterior covariance allows significant memory savings for storing the target models ($O(M)$ rather than $O(M^2)$). However, although the i–vector posterior covariance is diagonal, matrix $\mathbf{\Lambda}_{eq,i}$ of (10) remains a full matrix. This approach alone, thus, does not give any computational advantage with respect to the standard FP–PLDA, it only saves the memory necessary for storing the FP–PLDA parameters of a set of target speakers.

B. Diagonalized Residual Covariance

The second term that can be diagonalized in order to speedup scoring is the covariance $\mathbf{\Lambda}_{eq,i}$ of the residual term $\overline{\mathbf{E}}_i$ (8). Diagonalization allows avoiding an expensive matrix inversion. In particular, the precision matrix $\mathbf{\Lambda}$ of the PLDA residual term $\mathbf{E}$ can be eigen–decomposed as $\mathbf{\Lambda} = \mathbf{V_\Lambda} \mathbf{D_\Lambda} \mathbf{V_\Lambda}^T$, where $\mathbf{V_\Lambda}$ is an orthogonal matrix, and $\mathbf{D_\Lambda}$ is a diagonal matrix. The precision matrix of $\overline{\mathbf{E}}_i$ can be written as:

$$\begin{aligned} \mathbf{\Lambda}_{eq,i} &= \left(\mathbf{\Lambda}^{-1} + \mathbf{\Gamma}_i^{-1} \right)^{-1} \\ &= \left(\mathbf{V_\Lambda} \mathbf{D_\Lambda}^{-1} \mathbf{V_\Lambda}^T + \mathbf{\Gamma}_i^{-1} \right)^{-1} \end{aligned}$$

$$= \mathbf{V}_{\mathbf{\Lambda}}\Big(\mathbf{D}_{\mathbf{\Lambda}}^{-1} + \mathbf{V}_{\mathbf{\Lambda}}^T\mathbf{\Gamma}_i^{-1}\mathbf{V}_{\mathbf{\Lambda}}\Big)^{-1}\mathbf{V}_{\mathbf{\Lambda}}^T \ . \quad (28)$$

The proposed approximation consists in replacing the term $\mathbf{V}_{\mathbf{\Lambda}}^T\mathbf{\Gamma}_i^{-1}\mathbf{V}_{\mathbf{\Lambda}}$ by a diagonal matrix $\mathbf{V}_{\mathbf{\Lambda}}^T\mathbf{\Gamma}_i^{-1}\mathbf{V}_{\mathbf{\Lambda}} \circ \mathbf{I}$.

In order to analyze the scoring complexity using this approximation, let's define:

$$\mathbf{\Lambda}_{eq,i}^D = \Big(\mathbf{D}_{\mathbf{\Lambda}}^{-1} + \mathbf{V}_{\mathbf{\Lambda}}^T\mathbf{\Gamma}_i^{-1}\mathbf{V}_{\mathbf{\Lambda}} \circ \mathbf{I}\Big)^{-1} \ , \quad (29)$$

so that the approximated $\mathbf{\Lambda}_{eq,i}$ can be rewritten as:

$$\mathbf{\Lambda}_{eq,i} = \mathbf{V}_{\mathbf{\Lambda}}\mathbf{\Lambda}_{eq,i}^D\mathbf{V}_{\mathbf{\Lambda}}^T \ . \quad (30)$$

The statistics $\mathbf{F}_{eq,E}$ and $\mathbf{F}_{eq,T}$ can be computed by simply replacing (30) in (21). The approximated speaker identity posterior covariance (20a) can thus be rewritten, from (20a), (22), and (30), as:

$$\mathbf{\Lambda}_{y|(E,T)} = \mathbf{I} + \mathbf{U}^T\mathbf{V}_{\mathbf{\Lambda}}\left(\mathbf{\Lambda}_{eq,E}^D + \mathbf{\Lambda}_{eq,T}^D\right)\mathbf{V}_{\mathbf{\Lambda}}^T\mathbf{U} \ , \quad (31)$$

where

$$\mathbf{\Lambda}_{eq,E}^D = \sum_{i\in E}\mathbf{\Lambda}_{eq,i}^D \ , \qquad \mathbf{\Lambda}_{eq,T}^D = \sum_{i\in T}\mathbf{\Lambda}_{eq,i}^D \ . \quad (32)$$

Thus, $\mathbf{\Lambda}_{y|(E,T)}$ depends on the covariance of the i–vectors only through the diagonal statistics $\mathbf{\Lambda}_{eq,E}^D$ and $\mathbf{\Lambda}_{eq,T}^D$.

It is worth noting that, since the i–vector posterior covariance becomes smaller for longer utterances, the effects of this approximation become negligible, and the exact PLDA solution is recovered, whenever the test utterances are long enough.

C. Diagonalized Speaker Identity Posterior

A third approximation, which further decreases the scoring complexity, consists in the diagonalization of the speaker identity posterior covariances.

Let's eigen–decompose the term $\mathbf{U}^T\mathbf{\Lambda}\mathbf{U}$ in (17a) as:

$$\mathbf{U}^T\mathbf{\Lambda}\mathbf{U} = \mathbf{V}_{\mathbf{Y}}\mathbf{D}_{\mathbf{Y}}\mathbf{V}_{\mathbf{Y}}^T \ , \quad (33)$$

where $\mathbf{V}_{\mathbf{Y}}$ and $\mathbf{D}_{\mathbf{Y}}$ are an orthogonal and a diagonal matrix, respectively. The speaker identity posterior covariance is then given by:

$$\begin{aligned}\mathbf{\Lambda}_{y|(E,T)}^{-1} &= \Big(\mathbf{I} + (n_E + n_T)\,\mathbf{V}_{\mathbf{Y}}\mathbf{D}_{\mathbf{Y}}\mathbf{V}_{\mathbf{Y}}^T\Big)^{-1} \\ &= \mathbf{V}_{\mathbf{Y}}(\mathbf{I} + (n_E + n_T)\,\mathbf{D}_{\mathbf{Y}})^{-1}\mathbf{V}_{\mathbf{Y}}^T \ , \end{aligned} \quad (34)$$

where factor $\mathbf{I} + (n_E + n_T)\,\mathbf{D}_{\mathbf{Y}}$ is diagonal.

A similar decomposition of $\mathbf{U}^T\mathbf{\Lambda}\mathbf{U}$ can be applied to the approximated speaker identity posterior covariance of (31) obtaining:

$$\begin{aligned}\mathbf{\Lambda}_{y|(E,T)}^{-1} &= \Big(\mathbf{I} + \mathbf{V}_{\mathbf{Y}}\left(\widehat{\mathbf{D}}_{eq,E} + \widehat{\mathbf{D}}_{eq,T}\right)\mathbf{V}_{\mathbf{Y}}^T\Big)^{-1} \\ &= \mathbf{V}_{\mathbf{Y}}\Big(\mathbf{I} + \Big(\widehat{\mathbf{D}}_{eq,E} + \widehat{\mathbf{D}}_{eq,T}\Big)\Big)^{-1}\mathbf{V}_{\mathbf{Y}}^T \ , \end{aligned} \quad (35)$$

where

$$\begin{aligned}\widehat{\mathbf{D}}_{eq,E} &= \mathbf{V}_{\mathbf{Y}}^T\mathbf{U}^T\mathbf{\Lambda}_{eq,E}\mathbf{U}\mathbf{V}_{\mathbf{Y}} \\ \widehat{\mathbf{D}}_{eq,T} &= \mathbf{V}_{\mathbf{Y}}^T\mathbf{U}^T\mathbf{\Lambda}_{eq,T}\mathbf{U}\mathbf{V}_{\mathbf{Y}} \ . \end{aligned} \quad (36)$$

Since, in contrast with standard PLDA, the matrices $\widehat{\mathbf{D}}_{eq,E}$ and $\widehat{\mathbf{D}}_{eq,T}$ are not diagonal, the proposed simplification consists in replacing these terms by the corresponding diagonal matrices:

$$\mathbf{D}_{eq,E} = \widehat{\mathbf{D}}_{eq,E} \circ \mathbf{I} \ , \ \ \mathbf{D}_{eq,T} = \widehat{\mathbf{D}}_{eq,T} \circ \mathbf{I} \ . \quad (37)$$

Similarly to the diagonalized residual covariance approach, the impact of this approximation becomes irrelevant with the increase of the utterance duration, because the contribution of the i–vector posterior covariance becomes negligible compared to the PLDA residual noise covariance.

D. Diagonalized FP–PLDA

The three diagonalization approaches illustrated in the previous sub–sections can be efficiently combined in order to sensibly speed–up the computation of the FP–PLDA log–likelihood ratios. This section illustrates the sequence of steps for an efficient computation of the scoring function $\sigma(E,T)$ of a fully Diagonalized FP–PLDA. A comparison of the complexity of different diagonalization approaches is also provided in Table II. The details of the derivation of these complexities are given in the Appendix.

The standard FP–PLDA solution is obtained by replacing in all the presented approximations the diagonalizing operator $\circ\ \mathbf{I}$ by the operator $\circ\ \mathbf{1}$, where $\mathbf{1}$ is a matrix of ones. We define, for each possible diagonalization, a matrix operator $\mathbf{Q}$ such that $\mathbf{Q} = \mathbf{I}$ if the diagonalization is applied, and $\mathbf{Q} = \mathbf{1}$ otherwise. We will denote by $\mathbf{Q}_{\mathbf{\Gamma}}, \mathbf{Q}_{\mathbf{\Lambda}}, \mathbf{Q}_{\mathbf{Y}}$ the operators associated to i–vector covariance diagonalization, residual covariance diagonalization, and speaker identity posterior covariance diagonalization, respectively.

In order to derive a computational efficient formulation of the scoring function $\sigma(E,T)$ (23), it is worth expanding one of its terms involving the first–order statistics (21), for example the second term $\mathbf{F}_{eq,E}^T\mathbf{\Lambda}_{y|(E,T)}^{-1}\mathbf{F}_{eq,T}$, as shown at the top of next page.

By defining the i–vector covariance as:

$$\Big(\mathbf{\Gamma}_i^D\Big)^{-1} = (\mathbf{\Gamma}_i \circ \mathbf{Q}_{\mathbf{\Gamma}})^{-1} \ , \quad (44)$$

and setting

$$\mathbf{W} = \mathbf{V}_{\mathbf{Y}}^T\mathbf{U}^T\mathbf{V}_{\mathbf{\Lambda}} \ , \quad (45)$$

as in (43), the steps for a fast computation of the scoring function $\sigma(E,T)$ can be summarized as follows:

1) For each utterance compute the term:

$$\mathbf{\Gamma}_{\mathbf{\Lambda},i}^{-1} = \mathbf{V}_{\mathbf{\Lambda}}^T\Big(\mathbf{\Gamma}_i^D\Big)^{-1}\mathbf{V}_{\mathbf{\Lambda}} \circ \mathbf{Q}_{\mathbf{\Lambda}} \quad (46)$$

in (28), and the diagonalized approximation of the equivalent precision matrix:

$$\mathbf{\Lambda}_{eq,i}^D = \Big(\mathbf{D}_{\mathbf{\Lambda}}^{-1} + \mathbf{\Gamma}_{\mathbf{\Lambda},i}^{-1}\Big)^{-1} \quad (47)$$

2) For each set G compute the projected statistics (42):

$$\hat{\mathbf{F}}_{eq,G} = \mathbf{W}\sum_{i\in G}\mathbf{\Lambda}_{eq,i}^D\mathbf{V}_{\mathbf{\Lambda}}^T\left(\boldsymbol{\mu}_i - \mathbf{m}\right) \ , \quad (48)$$

and the diagonalized approximation of the cumulative

$$\mathbf{F}_{eq,E}^T \boldsymbol{\Lambda}_{y|(E,T)}^{-1} \mathbf{F}_{eq,T} = \tag{38}$$

$$\text{from (21)} \quad \left(\mathbf{U}^T \sum_{i\in E} \boldsymbol{\Lambda}_{eq,i} \left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right)^T \boldsymbol{\Lambda}_{y|(E,T)}^{-1} \left(\mathbf{U}^T \sum_{i\in T} \boldsymbol{\Lambda}_{eq,i} \left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right) = \tag{39}$$

$$\text{from (35)} \quad \left(\mathbf{U}^T \sum_{i\in E} \boldsymbol{\Lambda}_{eq,i} \left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right)^T \mathbf{V}_{\mathbf{Y}} \left(\mathbf{I} + \left(\widehat{\mathbf{D}}_{eq,E} + \widehat{\mathbf{D}}_{eq,T}\right)\right)^{-1} \mathbf{V}_{\mathbf{Y}}^T \left(\mathbf{U}^T \sum_{i\in T} \boldsymbol{\Lambda}_{eq,i} \left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right) = \tag{40}$$

$$\text{from (30)} \quad \left(\mathbf{V}_{\mathbf{Y}}^T \mathbf{U}^T \sum_{i\in E} \mathbf{V}_{\boldsymbol{\Lambda}} \boldsymbol{\Lambda}_{eq,i}^D \mathbf{V}_{\boldsymbol{\Lambda}}^T \left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right)^T \left(\mathbf{I} + \left(\widehat{\mathbf{D}}_{eq,E} + \widehat{\mathbf{D}}_{eq,T}\right)\right)^{-1} \left(\mathbf{V}_{\mathbf{Y}}^T \mathbf{U}^T \sum_{i\in T} \mathbf{V}_{\boldsymbol{\Lambda}} \boldsymbol{\Lambda}_{eq,i}^D \mathbf{V}_{\boldsymbol{\Lambda}}^T \left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right) = \tag{41}$$

$$\left(\sum_{i\in E} \boldsymbol{\Lambda}_{eq,i}^D \mathbf{V}_{\boldsymbol{\Lambda}}^T \left(\boldsymbol{\mu}_i - \mathbf{m}\right)\right)^T \mathbf{W}^T \left(\mathbf{I} + \left(\widehat{\mathbf{D}}_{eq,E} + \widehat{\mathbf{D}}_{eq,T}\right)\right)^{-1} \mathbf{W} \sum_{i\in E} \boldsymbol{\Lambda}_{eq,i}^D \mathbf{V}_{\boldsymbol{\Lambda}}^T \left(\boldsymbol{\mu}_i - \mathbf{m}\right) \ , \tag{42}$$

$$\text{where} \quad \mathbf{W} = \mathbf{V}_{\mathbf{Y}}^T \mathbf{U}^T \mathbf{V}_{\boldsymbol{\Lambda}} \tag{43}$$

TABLE II
COMPARISON OF THE COMPLEXITY OF APPROXIMATED FULL–POSTERIOR–PLDA DIAGONALIZATION APPROACHES.

Diagonalization			Complexity		
$\mathbf{Q}_{\boldsymbol{\Gamma}} = \mathbf{I}$	$\mathbf{Q}_{\boldsymbol{\Lambda}} = \mathbf{I}$	$\mathbf{Q}_{\mathbf{Y}} = \mathbf{I}$	Per–utterance	Per–set	Per–trial
no	no	no	NM^3	M^2S	S^3
yes	no	no	NM^3	M^2S	S^3
no	yes	no	NM^3	MS^2	S^3
yes	yes	no	NM^2	MS^2	S^3
no	no	yes	NM^3	M^2S	S
yes	yes	yes	NM^2	MS	S

equivalent precision matrix:

$$\boldsymbol{\Lambda}_{eq,G}^D = \sum_{i\in G} \boldsymbol{\Lambda}_{eq,i}^D \tag{49}$$

$$\mathbf{D}_{eq,G} = \mathbf{W} \boldsymbol{\Lambda}_{eq,G}^D \mathbf{W}^T \circ \mathbf{Q}_{\mathbf{Y}} \tag{50}$$

3) For each trial compute the diagonalized speaker identity posterior covariance:

$$\left(\boldsymbol{\Lambda}_{y|(E,T)}^D\right)^{-1} = \left(\mathbf{I} + \left(\mathbf{D}_{eq,E} + \mathbf{D}_{eq,T}\right)\right)^{-1} , \tag{51}$$

and finally the scoring function $\sigma(E,T)$ as:

$$\begin{aligned}\sigma(E,T) = &-\frac{1}{2}\log\left|\left(\boldsymbol{\Lambda}_{y|(E,T)}^D\right)^{-1}\right| \\ &+ \frac{1}{2}\hat{\mathbf{F}}_{eq,E}^T \left(\boldsymbol{\Lambda}_{y|(E,T)}^D\right)^{-1} \hat{\mathbf{F}}_{eq,E} \\ &+ \frac{1}{2}\hat{\mathbf{F}}_{eq,T}^T \left(\boldsymbol{\Lambda}_{y|(E,T)}^D\right)^{-1} \hat{\mathbf{F}}_{eq,T} \\ &+ \hat{\mathbf{F}}_{eq,E}^T \left(\boldsymbol{\Lambda}_{y|(E,T)}^D\right)^{-1} \hat{\mathbf{F}}_{eq,T}\end{aligned} \tag{52}$$

It is worth noting that equation (44) can be considered part of the i–vector extractor, and has direct implications on the complexity of the extractor. If $\mathbf{Q}_{\boldsymbol{\Gamma}} = \mathbf{1}$, the full covariance of the i–vector has to be computed with complexity $O(NM^3)$, whereas only the diagonal of the i–vector posterior is needed if $\mathbf{Q}_{\boldsymbol{\Gamma}} = \mathbf{I}$. In the latter case, approximated i–vector extractors can be used [23], [24], which allow i–vector extraction to be performed in $O(NM)$, and approximated diagonal i–vector posterior precisions to be computed in $O(NM)$.

Table II summarizes the complexity of different diagonal FP–PLDA approximations, according to the different settings of the diagonalizing operators $\mathbf{Q}_{\boldsymbol{\Gamma}}$, $\mathbf{Q}_{\boldsymbol{\Lambda}}$ and $\mathbf{Q}_{\mathbf{Y}}$. Combining different approximations notably reduces the computational complexity with respect to the individual contribution of each diagonalization. Applying the sequence of the proposed approaches reduces both the per–set and per–trial scoring computations, thus shrinking the computational gap between standard PLDA and FP–PLDA.

VI. EXPERIMENTAL RESULTS

Two set of experiments were performed for assessing the performance and speedup tradeoff of the proposed diagonalization techniques. The first one uses the same cuts of variable duration that were used for assessing the performance of the FP–PLDA approach in [18]. The cuts were extracted from conversations recorded from different channels included in the NIST SRE 2010 extended core tests [20]. These experiments were devoted to the assessment of the diagonalization techniques on a task including test utterances of variable duration (from 3 to 60 seconds). The Diagonalized FP–PLDA has also been tested on a short utterance text–independent

TABLE III
NIST SRE 2010 ENROLLMENT AND TEST CONDITIONS.

Condition	Female targets / non-target trials	Male targets / non-target trials	Enrollment	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 IV
RESULTS FOR THE CORE EXTENDED NIST SRE2010 FEMALE TESTS IN TERMS OF % EER, MINDCF08×1000 AND MINDCF10×1000 USING DIFFERENT MODELS. "STD" AND "FP" LABELS REFER TO STANDARD PLDA AND FP–PLDA, RESPECTIVELY.

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	2.6	124	460	2.2	103	405	1.1	65	303	1.8	68	258	1.9	105	335
Std	FPD	2.3	114	455	2.1	103	402	1.0	60	296	1.7	63	254	2.0	103	347
FPD	FPD	2.3	112	455	2.0	100	396	1.0	59	288	1.6	60	253	2.0	101	344

verification task including very short test utterances from a dataset completely different with respect to the NIST data that have been used for training the models. In particular, the dataset for the first set of experiments 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 [20], which are summarized in Table III. Cepstral features, extracted using a 25 ms Hamming window, have been used. 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.

The i–vector extractor is 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.

The dimension of the i–vector subspace was set to $M = 400$, and the i–vector posteriors were normalized according to the Projected Length Normalization:

$$\overline{\mathbf{W}} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_{\mathcal{X}}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}, \frac{\boldsymbol{\Gamma}_{\mathcal{X}}^{-1}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|^2}\right) . \tag{53}$$

introduced in [16]. 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 detailed results on the female datasets only, because the NIST SRE 2010 core test on female speakers is known to be more difficult, thus more often compared in the literature. Table IV 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 [20]. In this table, the PLDA and FPD–PLDA systems are compared using the original interview or telephone data without any cut. Labels "Std" and "FPD" refer to the standard and the Full Posterior Distribution PLDA, respectively.

The first row gives the baseline results using standard i–vectors 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 second row gives the baseline results using the Full Posterior Distribution PLDA model. The FPD–PLDA model not only keeps the accuracy of the standard model for long segments, as expected, 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 the FP–PLDA models. Training was done using the EM algorithm, as presented in [17] for a model equivalent to FP–PLDA (a proof of equivalence is given in Section VI of [18]). As expected, since the training utterances have long durations, the results are similar to the ones reported in the second row, thus, there is little advantage in using the full i–vector posterior in training the FP–PLDA models when long training utterances are available.

The results of the tests on variable duration cuts, randomly chosen from the extended NIST SRE2010 female set, are shown in Table V. The minimum DCF10 results, given for core extended tests, have not been reported for these and the remaining short duration experiments because their value is often too large to be meaningful [18]. Excluding the matched

TABLE V
RESULTS IN TERMS OF % EER AND MINDCF08×1000 OF STANDARD PLDA, FULL POSTERIOR PLDA, AND TWO DIAGONALIZATION APPROACHES FOR TEST DATA OF VARIABLE DURATION, RANDOMLY CHOSEN FROM CUTS OF THE EXTENDED NIST SRE2010 FEMALE TESTS. THE PLDA PARAMETERS ARE TRAINED USING BOTH MICROPHONE AND TELEPHONE DATA.

Model	Duration (seconds)	cond2		cond3		cond4		cond1		cond5		average % improvement
		EER	DCF 08	EER	DCF 08	EER	DCF 08	EER	DCF 08	EER	DCF 08	
Standard	3–30	12.4	531	11.3	521	11.1	441	9.8	405	10.6	493	
Diagonal $\mathbf{Q_Y}$	3–30	10.0	482	9.5	504	8.5	388	7.7	331	9.6	481	13.5
Diagonal $\mathbf{Q_\Gamma}, \mathbf{Q_\Lambda}, \mathbf{Q_Y}$	3–30	10.5	511	10.3	544	8.9	406	8.1	366	10.2	503	8.0
FP–PLDA	3–30	9.8	474	9.3	498	8.3	382	7.6	327	9.7	475	14.6
Standard	10–30	9.0	431	8.6	429	6.6	318	7.0	317	7.6	390	
Diagonal $\mathbf{Q_Y}$	10–30	7.8	394	7.5	416	5.7	288	5.8	277	7.3	377	9.8
Diagonal $\mathbf{Q_\Gamma}, \mathbf{Q_\Lambda}, \mathbf{Q_Y}$	10–30	8.4	423	8.1	438	6.0	311	6.5	305	7.4	386	3.8
FP–PLDA	10–30	7.7	388	7.5	417	5.7	285	5.5	278	7.2	373	10.7
Standard	3–60	9.1	384	7.8	368	7.3	312	7.0	273	6.7	337	
Diagonal $\mathbf{Q_Y}$	3–60	6.8	330	6.1	346	5.3	256	4.7	232	6.2	324	17.2
Diagonal $\mathbf{Q_\Gamma}, \mathbf{Q_\Lambda}, \mathbf{Q_Y}$	3–60	6.9	352	6.9	375	5.4	271	5.0	244	6.5	333	12.5
FP–PLDA	3–60	6.7	328	6.2	343	5.2	259	4.7	232	6.2	323	17.3
Standard	10–60	7.0	318	5.0	283	4.7	227	4.9	211	4.9	265	
Diagonal $\mathbf{Q_Y}$	10–60	5.8	286	4.9	274	3.8	203	4.2	176	4.8	263	9.6
Diagonal $\mathbf{Q_\Gamma}, \mathbf{Q_\Lambda}, \mathbf{Q_Y}$	10–60	6.0	299	4.9	285	4.0	211	4.2	188	4.8	266	7.0
FP–PLDA	10–60	5.7	283	4.8	271	3.9	200	4.1	176	4.7	260	10.6
Standard	Full	2.6	124	2.2	103	1.1	65	1.8	68	1.9	105	
Diagonal $\mathbf{Q_Y}$	Full	2.4	115	2.2	103	1.0	61	1.7	64	2.1	104	3.2
Diagonal $\mathbf{Q_\Gamma}, \mathbf{Q_\Lambda}, \mathbf{Q_Y}$	Full	2.3	116	2.0	101	1.0	61	1.6	66	2.0	102	5.6
FPD	Full	2.3	114	2.1	103	1.0	60	1.7	63	2.0	103	5.0

TABLE VI
COMPARISON OF THE PERFORMANCE AND RELATIVE COMPUTATIONAL COST OF PLDA, FP–PLDA, AND DIAGONALIZED FP–PLDA, WITH $\mathbf{Q_\Gamma} = \mathbf{Q_\Lambda} = \mathbf{Q_Y} = I$, ON A SHORT UTTERANCE TEXT–INDEPENDENT VERIFICATION TASK.

Model	% EER	minDCF8	Model size (Kb)	Scoring time wrt PLDA
PLDA	13.6	612	3.5	1
FP–PLDA	8.2	421	81	22
Diagonalized PLDA	7.7	401	5.1	1.05

tel–tel condition 5, the FPD–PLDA always shows a relative improvement, quite small for long enough segments, but up to 20% depending on the average duration of the small cuts, and on the condition. Table V shows also the results for the two settings of the Diagonalized FPD–PLDA approach that are more relevant from an application viewpoint. The first setting approximates only the speaker identity posteriors, which allows reducing the per-trial cost in a speaker identification scenario where the target speakers are known in advance, so that their per–set and per–utterance costs are independent from the number of trials. The second setting, instead, is convenient in a speaker verification scenario, where the three proposed approximations are applied in sequence in order to minimize the memory and computation costs. The accuracy of the Diagonalized FP–PLDA decreases as a function of the number of the applied diagonalizing operators, but in the conditions in which the FP–PLDA technique shows most improvement, also the Diagonalized FP–PLDA performs very well, as indicated by the average percent improvement obtained by the EER and minimum DCF08 in all condition with respect to PLDA, reported in the last column.

A second set of experiments was conducted on a text–independent verification task. The dataset for these experiments was provided by NUANCE. It includes 308 female and 218 male speakers, contributing a total of 1177 and 849 utterances, respectively. The test consists in a single utterance selected between two short sentences only. The average duration of the test utterances for the two sentences, excluding silences, is 1.3 and 2.3 seconds, respectively. The number of true speaker and impostor trials is 4052 and 20494, respectively. Since this dataset did not include a specific development set for these tests, a gender–independent i–vector extractor was trained based on a 1024–component diagonal covariance gender–independent UBM. Both the i–vector extractor and the UBM have been trained using data from NIST SRE 2004–2010 and, additionally, the Switchboard II, Phases 2 and 3, and Switchboard Cellular, Parts 1 and 2 datasets, for a total of 66140 utterances. Every utterance was processed after Voice Activity Detection, extracting every 10 ms, 19 Perceptual Linear Predictive (PLP) coefficients, and the frame log-energy, on a 25 ms sliding Hamming window. This 20–dimensional feature vector was subjected to short time mean and variance normalization using a 3 s sliding window, and a 45-dimensional feature vector was obtained by stacking 18 PLP coefficients (c_1-c_{18}), 19 delta (Δc_0-Δc_{18}) and 8 double–delta ($\Delta\Delta c_0$-$\Delta\Delta c_7$) parameters. The i-vector dimension was fixed to $d = 400$. The PLDA model was trained with full–rank channel factors, and 200 dimensions for the speaker factors, using the NIST SRE 2004–2010 datasets, for a total of 48568 utterances of 3271 speakers. Also in this case length normalization was applied to the i–vectors.

Table VI shows the results obtained on this task, in terms of

percent Equal Error Rate, minimum Decision Cost Function DCF08 $\times 1000$, model size in KB, and the scoring times relative to PLDA scoring time. Although the test includes a small number of speakers and utterances, one can clearly appreciate comparing the plain PLDA and the FP–PLDA results in Table VI how valuable is the "uncertainty" information exploited by the FP–PLDA approach. FP–PLDA reduces the EER and the minimum DCF08 by approximately 40% and 31%, respectively. The Diagonalized FPD–PLDA approach not only improves the PLDA performance but, surprisingly, it gives better EER and DCF08 values with respect to FP–PLDA (43% and 34% better than PLDA, respectively). We should note, however, that the small changes between the FP–PLDA and the Diagonalized FP–PLDA have limited statistical significance. Particularly interesting is the comparison of the relative processing times of the three approaches. The scoring time of FP–PLDA is 22 times greater than PLDA, whereas the overhead of the Diagonalized FP–PLDA is just 5%. Although the scoring time is a small fraction of the processing time devoted to the i–vector extraction, fast scoring is important both for score normalization and for identification applications that require the same test segment to be compared with a large number of target speakers.

VII. Conclusions

The complexity of the PLDA and FPD–PLDA implementations have been analyzed, and a set of diagonalizing approximations has been proposed, which allows obtaining a substantial complexity reduction for trial scoring. In particular, by applying a sequence of diagonalization operators that approximate the matrices needed for i-vector scoring, it is possible to greatly enhance the scoring time for the FPD–PLDA approach while keeping most of the improvement of the FP–PLDA model in terms of recognition accuracy with respect to the standard PLDA approach. Other advantages of this approach are its reduced memory costs with respect to FP–PLDA. The proposed techniques also benefit from optimized i–vector extraction approaches, which avoid the computation of the i–vector covariance matrices [23], [24], further reducing the overall complexity of the system, and making the FPD approach suitable for embedded devices.

VIII. Acknowledgments

I would like to thank Prof. Pietro Laface for continuous support, Oldrich Plchot from Brno University for providing the i–vectors of the SRE 2010 experiments, and Claudio Vair and Daniele Colibro from NUANCE, for useful discussions, and for providing access to their data for the experiments on the short duration verification task.
Computational resources for this work were provided by the high performance computing cluster of Politecnico di Torino HPC@POLITO (http://www.hpc.polito.it).

Appendix

The contributions to the scoring complexity of each step of the algorithm presented in Section V-D, and the effects obtained by combining different approximations are detailed in the following sub–sections.

A. Standard FP–PLDA

Most of the steps detailed in V-D are redundant for standard FP–PLDA. However, the since the resulting asymptotic complexity does not change, we can use those steps as a reference for describing the contribution of the different approximations on the overall scoring complexity.

- Equation (46) has a complexity $O(NM^3)$.
- Since $\mathbf{\Gamma}_{\mathbf{\Lambda},i}^{-1}$ is a full matrix, equation (47) has a complexity $O(NM^3)$, and produces full $\mathbf{\Lambda}_{eq,i}^{D}$ matrices.
- The computation of the statistics in (48) and (49) have an overall complexity $O(NM^2)+O(MS)$, and $\mathbf{\Lambda}_{eq,G}^{D}$ is again a full matrix.
- The computation of $\mathbf{D}_{eq,G}$ in (50) has a complexity of $O(M^2S)$ and, again, results in a non–diagonal matrix.
- The per–trial term in equation (51) has a complexity $O(S^3)$.
- Finally, equation (52) can be computed in $O(S^2)$.

Combining all these steps gives an overall $O(NM^3)$ per–utterance complexity, $O(M^2S)$ per–set complexity, and $O(S^3)$ per–trial complexity.

B. Diagonalized i–vector covariance

Diagonalization of the i–vector posterior covariance corresponds to setting $\mathbf{Q}_{\mathbf{\Gamma}} = \mathbf{I}$.

- Although $\left(\mathbf{\Gamma}_i^D\right)^{-1}$ is diagonal equation (46) still requires $O(NM^3)$ operations.
- Since $\mathbf{\Gamma}_{\mathbf{\Lambda},i}^{-1}$ is full, all the remaining steps have the same complexity of the standard FP–PLDA.

The overall complexity is, therefore, the one given in previous sub–section.

C. Diagonalized residual covariance

The complexity of the diagonalized residual covariance approximation is related to the use the diagonalized i–vector covariance approximation. In particular:

- Equation (46) has complexity $O(NM^3)$. However, if $\left(\mathbf{\Gamma}_i^D\right)^{-1}$ is diagonal, $\mathbf{\Gamma}_{\mathbf{\Lambda},i}^{-1}$ can be evaluated in $O(NM^2)$ operations because only the diagonal of the right hand side of the equation is needed.
- Since $\mathbf{\Gamma}_{\mathbf{\Lambda},i}^{-1}$ is diagonal, (47) has a complexity $O(NM)$.
- The computations of the statistics in (48) requires $O(NM^2)+O(MS)$ operations.
- The terms in (49) can be computed in $O(NM)$ operations.
- Equation (50) has a per–set complexity $O(MS^2)$.
- Equation (51) has a per–trial complexity $O(S^3)$.
- Finally, equation (52) can be computed in $O(S^2)$.

Overall, the per–utterance and per–set complexity are $O(NM^3)$ and $O(MS^2)$, respectively, and the per–trial complexity is $O(S^3)$. However, if this approximation is preceded by the diagonalization of the i–vector posterior covariance, the per–utterance complexity decreases to $O(NM^2)$.

D. Diagonalized speaker identity posterior

Again, the complexity of this approximation depends on the sequential application of the first two diagonalizations. In particular:

- The complexity of equations (46) to (49) depends only on the previous approximations, and is not affected by the diagonalization of the speaker posterior covariance.
- Equation (50) has complexity $O(M^2S)$. However, it can be computed in $O(MS)$ if $\mathbf{\Lambda}^D_{eq,G}$ is diagonal, because only the diagonal of the right hand side of the equation is needed.
- Since $\mathbf{D}_{eq,G}$ is diagonal, equation (51) has a per–trial complexity $O(S)$.
- Finally, equation (52) can be computed in $O(S)$.

This approximation allows the per–trial complexity to be reduced from $O(S^3)$ of standard FP–PLDA to $O(S)$. The per–set complexity is also heavily dependent on the use of the previous approximations: it can be reduced to $O(MS)$ by using in sequence the three approximations.

REFERENCES

[1] P. Kenny, "Bayesian speaker verification with heavy–tailed priors," in *Keynote presentation, Odyssey 2010, The Speaker and Language Recognition Workshop*, 2010. Available at http://www.crim.ca/perso/patrick.kenny/kenny_Odyssey2010.pdf.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. 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] 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 *Proceedings of ICASSP 2011*, pp. 4828–4831, 2011.

[4] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey 2010*, pp. 194–201, 2010.

[5] M. Senoussaoui, P. Kenny, N. Brummer, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender–independent speaker recognition," in *Proceedings of INTERSPEECH 2011*, pp. 25–28, 2011.

[6] J. Villalba and N. Brümmer, "Towards fully bayesian speaker recognition: Integrating out the between-speaker covariance," in *Proceedings of INTERSPEECH 2011*, pp. 505–508, 2011.

[7] T. Hasan and J.H.L.Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 842–853, 2013.

[8] V. Hautamaki, T. Kinnunen, F. Sedlak, K. Lee, B. Ma, and H. Li, "Sparse classifier fusion for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1622–1631, 2013.

[9] B. Srinivasan, L. Yuancheng, D. Garcia-Romero, D. Zotkin, and R. Duraiswami, "A symmetric kernel partial least squares framework for speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1415–1423, 2013.

[10] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 Speaker Recognition Evaluation," in *Proceedings of INTERSPEECH 2013*, pp. 1981–1985, 2013.

[11] R. Saeidi and al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proceedings of INTERSPEECH 2013*, pp. 1986–1990, 2013.

[12] D. Colibro and al., "Nuance–Politecnico di Torino 2012 NIST Speaker Recognition Evaluation system," in *Proceedings of INTERSPEECH 2013*, pp. 1996–2000, 2013.

[13] J. Villalba, E. Lleida, A. Ortega, and A. Miguel, "The I3A Speaker Recognition system for NIST SRE12: Post-evaluation analysis," in *Proceedings of INTERSPEECH 2013*, pp. 3689–3693, 2013.

[14] O. Plchot and al., "Developing a speaker identification system for the DARPA RATS Project," in *Proceedings of ICASSP 2013*, pp. 6768–6772, 2013.

[15] 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 *Proceedings of ICASSP 2013*, pp. 6773–6777, 2013.

[16] S. Cumani, O. Plchot, and P. Laface, "Probabilistic Linear Discriminant Analysis of i–vector posterior distributions," in *Proceedings of ICASSP 2013*, pp. 7644–7648, 2013.

[17] P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proceedings of ICASSP 2013*, pp. 7649–7653, 2013.

[18] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in probabilistic linear discriminant analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 846–857, 2014.

[19] B. Borgstrom and A. McCree, "Supervector bayesian speaker comparison," in *Proceedings of ICASSP 2013*, pp. 7693–7697, 2013.

[20] "The NIST year 2010 speaker recognition evaluation plan." Available at http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.

[21] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," in *Technical report CRIM-06/08-13*, 2005.

[22] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of i–vector extraction," in *Proceedings of ICASSP 2011*, pp. 4516–4519, 2011.

[23] S. Cumani and P. Laface, "Memory and computation trade-offs for efficient i-vector extraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 934–944, 2013.

[24] S. Cumani and P. Laface, "Factorized sub-space estimation for fast and memory effective i-vector extraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 248–259, 2013.