PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS OF I–VECTOR POSTERIOR DISTRIBUTIONS

Sandro Cumani[1,2]*, Oldřich Plchot*[1] *and Pietro Laface*[2]

[1] *cumani, iplchot@fit.vutbr.cz* - Brno University of Technology, Czech Republic
[2] *sandro.cumani, pietro.laface@polito.it* - Politecnico di Torino, Italy

ABSTRACT

The i-vector extraction process is affected by several factors such as the noise level, the acoustic content of the observed features, 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 present a new PLDA model that, unlike the standard one, exploits the intrinsic i–vector uncertainty. Since short segments are known to decrease recognition accuracy, and segment duration is the main factor 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 female telephone extended core condition. Our results show that the new model outperforms the standard PLDA when tested on short segments, and keeps the accuracy of the latter for long enough utterances. In particular, the relative improvement is up to 13% for the EER, 5% for DCF08, and 2.5% for DCF10.

Index Terms— Speaker recognition, i–vector, PLDA

1. 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 [1]. An i–vector is a compact representation of a Gaussian Mixture Model (GMM) supervector [2], which captures most of the GMM supervectors variability. It is obtained by a MAP estimate of the mean of a posterior distribution [3]. 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 Probabilistic Linear Discriminant Analysis (PLDA) [4, 5].

The i–vector covariance essentially 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 2 in Section 2). These statistics are affected by several factors such as the noise level and the acoustic content of the observed features, but mainly depend on the number of the observed features, i.e., on the length of the speech segments that are used for characterizing a speaker. Shorter utterances tend to have larger covariances, so that i–vector estimates become less reliable.

Sandro Cumani is supported by the European Social Fund (ESF) in the project Support of Interdisciplinary Excellence Research Teams Establishment at BUT. This project is part of the IT4Innovations Centre of Excellence (CZ.1.05/1.1.00/02.0070).

Oldřich Plchot is partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015, by Czech Ministry of Education project No. MSM0021630528 and by IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

This paper presents a new PLDA model that incorporates the intrinsic uncertainty of the i–vector extraction process. In our approach, we show that it is possible to keep the simple and effective PLDA framework even if a speech segment is no more mapped to a single i–vector but to the i–vector extractor posterior distribution.

Since segment duration is the main factor affecting the i–vector covariance, and short segments are known to be less reliable, we tested our approach on the NIST SRE 2010 telephone extended core condition [6] standard tests, and on segments of small and variable duration in enrollment and test, evaluating also the effects of i–vector length normalization [7]. Our results show that, when tested on short segments, the new model outperforms the standard PLDA, and keeps the accuracy of the latter for long enough utterances.

The paper is organized as follows: Section 2 recalls the i-vector extraction process. Section 3 illustrates the generative PLDA model, and shows how to compute the likelihood that a set of utterances belong to the same speaker. In Section 4 we derive the formulation of the likelihood for the Gaussian PLDA model based on the i–vector extractor posterior distribution. Section 5 illustrates our new PLDA model, where the distribution of the inter–speaker variability is assumed to be utterance–dependent. Section 6 is devoted to the important issue of i–vector length normalization. Section 7 presents the experimental results, and in Section 8 we draw our conclusions.

2. I–VECTOR MODEL

I–vector based techniques represent the state–of–the–art in speaker verification [1, 8]. 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 subspace according to:

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

where $\mathbf{u}$ is the Universal Background Model (UBM) GMM mean supervector, with C GMM components of dimension F. $\mathbf{T}$ is a low-rank rectangular matrix spanning the subspace 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 [3]. Given $\mathbf{T}$ and a sequence of τ feature vectors $\mathcal{X} = \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_\tau$ extracted for a speech segment, it is possible to compute the likelihood of $\mathcal{X}$ given the model (1) and a value for variable $\mathbf{W}$. The i-vector $\boldsymbol{\phi}$ corresponding to the speech segment is computed as the Maximum a Posteriori (MAP) point estimate of the variable $\mathbf{W}$, i.e., the mean $\boldsymbol{\mu}_{\mathcal{X}}$ of the posterior distribution $P_{\mathbf{W}|\mathcal{X}}(\mathbf{w})$.

In [3], it has been shown that, assuming a standard Normal prior for $\mathbf{W}$, the posterior probability of $\mathbf{W}$ given the acoustic feature

vectors $\mathcal{X}$ is Gaussian $\mathbf{W}|\mathcal{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{X}}, \boldsymbol{\Gamma}_{\mathcal{X}}^{-1})$, with mean vector and precision matrix:

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{X}} &= \boldsymbol{\Gamma}_{\mathcal{X}}^{-1}\mathbf{T}^{*}\boldsymbol{\Sigma}^{-1}\mathbf{f}_{\mathcal{X}} \\ \boldsymbol{\Gamma}_{\mathcal{X}} &= \mathbf{I} + \sum_{c=1}^{C} N_{\mathcal{X}}^{(c)}\mathbf{T}^{(c)^{*}}\boldsymbol{\Sigma}^{(c)^{-1}}\mathbf{T}^{(c)}\,,\end{aligned} \tag{2}$$

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 $\mathcal{X}$, $\mathbf{f}_{\mathcal{X}}$ is the supervector stacking the first–order statistics $\mathbf{f}_{\mathcal{X}}^{(c)}$, centered around the corresponding UBM means, $\boldsymbol{\Sigma}^{(c)}$ is the UBM c–th covariance matrix, $\boldsymbol{\Sigma}$ is a block diagonal matrix having the matrices $\boldsymbol{\Sigma}^{(c)}$ as its entries, $\mathbf{T}^{(c)}$ is the sub-matrix of $\mathbf{T}$ corresponding to the c–th mixture component, and $\gamma_t^{(c)}$ is the c-th occupation probability of feature vector $\mathbf{x}_t$.

3. PLDA WITH I–VECTOR POSTERIORS

State–of–the–art performance has been obtained by using i–vectors with generative models based on PLDA. In the PLDA framework, Factor Analysis is applied to describe the i–vector generation process. In particular, an i–vector is considered a random variable $\boldsymbol{\Phi}$ whose generation process can be described in terms of latent variables. Different PLDA models exist [5, 4, 9], which use different numbers of hidden variables as well as different priors. All PLDA models for speaker recognition, however, represent the speaker identity in terms of a latent variable $\mathbf{Y}$ which is assumed to be tied across all utterances of the same speaker. Usually, inter–speaker variability is represented by utterance–dependent hidden variables $\mathbf{X}_i$, which are assumed to be i.i.d. with respect to the utterances. The most common PLDA model considers an i–vector $\boldsymbol{\phi}$ as the sum of different terms [4]:

$$\boldsymbol{\phi} = \mathbf{m} + \mathbf{U}\mathbf{y} + \mathbf{V}\mathbf{x} + \mathbf{e} \tag{3}$$

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 subspaces for $\mathbf{y}$ and $\mathbf{x}$, respectively. Since i–vectors are assumed independent given the hidden variables, the likelihood that a set of n utterances belong to the same speaker (hypothesis H_s) can be computed as:

$$\begin{aligned}l\left(u_1 \dots u_n |\, H_s\right) &= P_{\boldsymbol{\Phi}_1 \dots \boldsymbol{\Phi}_n | H_s}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n) \\ &= \int_{\mathbf{y}} \int_{\mathbf{x}_1} \cdots \int_{\mathbf{x}_n} \prod_i \left[P_{\boldsymbol{\Phi}_i|\mathbf{Y},\mathbf{X}_i}\left(\boldsymbol{\phi}_i|\mathbf{y},\mathbf{x}_i\right) P_{\mathbf{X}_i}\left(\mathbf{x}_i\right) d\mathbf{x}_i \right] \\ &\cdot P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}\,,\end{aligned} \tag{4}$$

where $P_{\boldsymbol{\Phi}_1 \dots \boldsymbol{\Phi}_n | H_s}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n)$ is the joint distribution of the i–vectors given the "same speaker" hypothesis H_s, $P_{\mathbf{X}}(\mathbf{x})$ and $P_{\mathbf{Y}}(\mathbf{y})$ are the prior distributions for $\mathbf{X}$ and $\mathbf{Y}$, and $P_{\boldsymbol{\Phi}|\mathbf{Y},\mathbf{X}}\left(\boldsymbol{\phi}|\mathbf{y},\mathbf{x}\right)$ is the conditional distribution of an i–vector given the hidden variables, which is related to the distribution $P_{\mathbf{E}}(\mathbf{e})$ of the noise term by $P_{\boldsymbol{\Phi}|\mathbf{Y},\mathbf{X}}\left(\boldsymbol{\phi}|\mathbf{y},\mathbf{x}\right) = P_{\mathbf{E}}(\boldsymbol{\phi} - \mathbf{m} - \mathbf{U}\mathbf{y} - \mathbf{V}\mathbf{x})$. Since speaker factors are assumed independent, given a set of n enrollment utterances $u_{e_1} \dots u_{e_n}$ for a target speaker and a set of m test utterances belonging to a (single) unknown speaker $u_{t_1} \dots u_{t_m}$, the speaker verification log–likelihood ratio s can be computed, using (4), as:

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

The standard i–vector, which is extracted by MAP point estimate of the posterior distribution of $\mathbf{W}$ given $\mathcal{X}$, and then used by PLDA, ignores the intrinsic uncertainty of its estimate. However, it is well known, for example, that i–vectors extracted from short utterances do not capture the speaker characteristic as well as i–vectors extracted from long utterances. 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" of $\mathbf{W}$ given $\mathcal{X}$, where we assume that every utterance is no more mapped to a single i–vector but to the i–vector extractor posterior distribution of $\mathbf{W}|\mathcal{X}$. Thus, $\mathcal{X}$ is mapped to i–vector $\boldsymbol{\phi}$ according to probability distribution $P_{\mathbf{W}|\mathcal{X}}(\boldsymbol{\phi})$.

The PLDA model allows computing the likelihood of an utterance given a realization of the random variable $\mathbf{W}|\mathcal{X}$, which is a mapping of the utterance features $\mathcal{X}$. The likelihood of a set of utterances, thus, can be evaluated by integrating the classical PLDA likelihood over all i–vectors that the utterances can generate as:

$$\begin{aligned}l\left(u_1 \dots u_n | H_s\right) &= \int_{\boldsymbol{\phi}_1} \cdots \int_{\boldsymbol{\phi}_n} l\left(u_1 \dots u_n | H_s, \mathbf{W}_1 = \boldsymbol{\phi}_1, \dots,\right. \\ &\left.\mathbf{W}_n = \boldsymbol{\phi}_n\right) \prod_i \left[P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right] \\ &= \int_{\boldsymbol{\phi}_1} \cdots \int_{\boldsymbol{\phi}_n} P_{\boldsymbol{\Phi}_1 \dots \boldsymbol{\Phi}_n | H_s}\left(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n\right) \prod_i \left[P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right]\end{aligned} \tag{5}$$

where the first term is the likelihood of the utterances according to the classical PLDA model given the realizations $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_n$ of the i–vector posterior random variables, computed as in (4), and the second term is the likelihood that the i–vectors $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_n$ were mapped to utterances $u_1, \dots, u_n$ according to the i–vector extractor model. Replacing (4) in (5) we can rewrite the likelihood as:

$$\begin{aligned}&l\left(u_1 \dots u_n | H_s\right) = \\ &= \int_{\boldsymbol{\phi}_1} \cdots \int_{\boldsymbol{\phi}_n} \int_{\mathbf{y}} \int_{\mathbf{x}_1} \cdots \int_{\mathbf{x}_n} \prod_i \left[P_{\boldsymbol{\Phi}_i|\mathbf{Y},\mathbf{X}_i}\left(\boldsymbol{\phi}_i|\mathbf{y},\mathbf{x}_i\right)\right. \\ &\left. \cdot P_{\mathbf{X}_i}(\mathbf{x}_i) P_{\mathbf{W}_i|\mathcal{X}_i}\left(\boldsymbol{\phi}_i\right) d\mathbf{x}_i d\boldsymbol{\phi}_i \right] P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}\,.\end{aligned} \tag{6}$$

It is worth noting that, by replacing the posterior for $\mathbf{W}|\mathcal{X}$ with a delta distribution centered in the posterior mean $\delta(\boldsymbol{\mu}_{\mathcal{X}})$, we return to the original PLDA model using MAP–estimated i–vectors.

In this work we consider only PLDA with Gaussian priors, because these models have shown to be accurate and effective with respect to more computational expensive models such as the Heavy–Tailed PLDA [4, 7]. Moreover, we will assume that the noise term $\mathbf{E}$ has full covariance matrix, so that the term $\mathbf{V}\mathbf{x}$ and $\mathbf{e}$ in (3) can be merged. Thus, an i–vector $\boldsymbol{\phi}$ is defined as:

$$\boldsymbol{\phi} = \mathbf{m} + \mathbf{U}\mathbf{y} + \mathbf{e}\,. \tag{7}$$

4. GAUSSIAN PLDA MODEL

The Gaussian PLDA approach assumes that the speaker factors and the residual noise priors are Gaussian, in particular:

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

where $\boldsymbol{\Lambda}$ is the precision matrix of noise $\mathbf{E}$. According to (7), the conditional distribution of an i–vector random variable $\boldsymbol{\Phi}$ given a

value $\mathbf{y}$ for the speaker identity $\mathbf{Y}$ is:

$$\mathbf{\Phi}|\left(\mathbf{Y}=\mathbf{y}\right) \sim \mathcal{N}(\mathbf{m}+\mathbf{U}\mathbf{y}, \mathbf{\Lambda}^{-1}) \,. \tag{8}$$

The likelihood that a set of n utterances belong to the same speaker can be computed by (4) ignoring the channel factors:

$$\begin{aligned} l(u_1 \dots u_n|H_s) &= P_{\mathbf{\Phi}_1 \dots \mathbf{\Phi}_n}(\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_n|H_s) \\ &= \int_{\mathbf{y}} \prod_i P_{\mathbf{\Phi}_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \,. \end{aligned} \tag{9}$$

Introducing the full i–vector posterior as in (5), we get:

$$\begin{aligned} l\left(u_1 \dots u_n|H_s\right) &= \int_{\boldsymbol{\phi}_i} \cdots \int_{\boldsymbol{\phi}_n} \int_{\mathbf{y}} P_{\mathbf{Y}}(\mathbf{y}) \\ &\cdot \prod_i \left[P_{\mathbf{\Phi}_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right] d\mathbf{y} \\ &= \int_{\mathbf{y}} P_{\mathbf{Y}}(\mathbf{y}) \prod_i \left[\int_{\boldsymbol{\phi}_i} P_{\mathbf{\Phi}_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i \right] d\mathbf{y} \,, \end{aligned}$$

The inner integral can be computed as:

$$\begin{aligned} &\int_{\boldsymbol{\phi}_i} P_{\mathbf{\Phi}_i|\mathbf{Y}}(\boldsymbol{\phi}_i|\mathbf{y}) P_{\mathbf{W}_i|\mathcal{X}_i}(\boldsymbol{\phi}_i) d\boldsymbol{\phi}_i = \\ &\int_{\boldsymbol{\phi}_i} \frac{1}{(2\pi)^{\frac{D}{2}} \left|\mathbf{\Lambda}^{-1}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\phi}_i - \mathbf{m} - \mathbf{U}\mathbf{y})^T \mathbf{\Lambda} (\boldsymbol{\phi}_i - \mathbf{m} - \mathbf{U}\mathbf{y})} \\ &\cdot \frac{1}{(2\pi)^{\frac{D}{2}} \left|\mathbf{\Gamma}_i^{-1}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\phi}_i - \boldsymbol{\mu}_i)^T \mathbf{\Gamma}_i (\boldsymbol{\phi}_i - \boldsymbol{\mu}_i)} d\boldsymbol{\phi}_i \,, \end{aligned} \tag{10}$$

where $\boldsymbol{\mu}_i$ and $\mathbf{\Gamma}_i$ are the mean and precision matrix of $\mathbf{W}_i|\mathcal{X}_i$ computed as in (2). Integral (10) can be interpreted as the convolution of two Gaussian distributions, leading to:

$$\begin{aligned} l(u_1 \dots u_n|\mathbf{Y}=\mathbf{y}) = &\frac{1}{(2\pi)^{\frac{D}{2}} \left|\mathbf{\Lambda}^{-1} + \mathbf{\Gamma}_i^{-1}\right|^{\frac{1}{2}}} \\ &\cdot e^{(\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y})^T \left(\mathbf{\Lambda}^{-1} + \mathbf{\Gamma}_i^{-1}\right)^{-1} (\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y})} \,. \end{aligned} \tag{11}$$

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

$$l(u_1 \dots u_n|\mathbf{Y}=\mathbf{y}) = \prod_i P_{\mathbf{M}_i|\mathbf{Y}}(\boldsymbol{\mu}_i|\mathbf{y}) \,, \tag{12}$$

where $\mathbf{M}_i|\mathbf{y}$ is distributed as in (11). The likelihood that the utterances belong to the same speaker is then given by:

$$l(u_1 \dots u_n|H_s) = \int_{\mathbf{y}} \prod_i P_{\mathbf{M}_i|\mathbf{Y}}(\boldsymbol{\mu}_i|\mathbf{y}) P_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \,. \tag{13}$$

Comparing (13) and (9) we see that the models are equivalent except for the form of the conditional likelihood. Thus, we can derive simple expressions for parameter training and for computing speaker verification log–likelihood scores. In particular, training the PLDA parameters can be performed by adapting the EM algorithm that is used for estimating the standard PLDA model parameters [4]. However, since we assume that the training utterances are long enough so that MAP approximation is accurate, in the following we will focus on the computation of speaker verification log–likelihood ratios.

5. GAUSSIAN PLDA POSTERIORS

The same steps used for the standard Gaussian PLDA model can be followed for deriving the log–likelihood of a set of utterances belonging to the same speaker, just using the modified likelihood in (11). Our new PLDA model can then be described as:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{U}\mathbf{y} + \overline{\mathbf{e}} \,, \tag{14}$$

as in (7), but the distribution of the residual noise $\overline{\mathbf{E}}$ is utterance–dependent. The i–vector associated to the utterance u_i is again the mean $\boldsymbol{\mu}_i$ of the i–vector posterior $\mathbf{W}_i|\mathcal{X}_i$, while the priors of the PLDA parameters are given by:

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

where

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

In the following, to simplify the notation we will refer to distributions without explicitly naming the corresponding hidden variable, i.e., 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 $\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n$ (i.e., of a set of n utterances $u_1 \dots u_n$), we observe that the joint log–likelihood of the i–vectors and the hidden variables is:

$$\begin{aligned} log P(\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n, \mathbf{y}|H_s) &= \sum_i \log P(\boldsymbol{\mu}_i|\mathbf{y}) + \log P(\mathbf{y}) \\ &= \sum_i \left[-\frac{1}{2} (\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y})^T \mathbf{\Lambda}_{eq,i} (\boldsymbol{\mu}_i - \mathbf{m} - \mathbf{U}\mathbf{y}) \right] \\ &+ \frac{1}{2}\mathbf{y}^T\mathbf{y} + k \,, \end{aligned} \tag{15}$$

where k is a constant collecting terms that do not depend on $\mathbf{y}$. Equation (15) shows that the posterior distribution for $\mathbf{y}$ given a set of i–vectors is once again Gaussian $\mathbf{y}|\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n \sim \mathcal{N}(\boldsymbol{\mu}_y, \mathbf{\Lambda}_y^{-1})$, with parameters:

$$\begin{aligned} \mathbf{\Lambda}_y &= \mathbf{I} + \sum_i \mathbf{U}^T \mathbf{\Lambda}_{eq,i} \mathbf{U} \\ \boldsymbol{\mu}_y &= \mathbf{\Lambda}_y^{-1} \mathbf{U}^T \sum_i \mathbf{\Lambda}_{eq,i} (\boldsymbol{\mu}_i - \mathbf{m}) \,. \end{aligned} \tag{16}$$

The likelihood that a set of utterances belong to the same speaker can be written as:

$$P(\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n|H_s) = \frac{P(\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n|\mathbf{y}_0) P(\mathbf{y}_0)}{P(\mathbf{y}_0|\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n)} \,, \tag{17}$$

where $\mathbf{y}_0$ can be freely chosen as long as the denominator is defined. Setting for convenience $\mathbf{y}_0 = \mathbf{0}$, from (17) and (16) we finally get:

$$\begin{aligned} &\log P(\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_n|H_s) = \\ &\sum_i \left[\frac{1}{2} \log |\mathbf{\Lambda}_{eq,i}| - \frac{D}{2} \log 2\pi - \frac{1}{2} (\boldsymbol{\mu}_i - \mathbf{m})^T \mathbf{\Lambda}_{eq,i} (\boldsymbol{\mu}_i - \mathbf{m}) \right] \\ &- \frac{1}{2} \log |\mathbf{\Lambda}_y| + \frac{1}{2} \boldsymbol{\mu}_y^T \mathbf{\Lambda}_y \boldsymbol{\mu}_y \,. \end{aligned} \tag{18}$$

Table 1: Comparison of GPLDA and Full Posterior GPLDA for complete conversations and randomly chosen cuts of different duration. FP refers to full posterior GPLDA, LN, and PLN to Length Normalization and Projected Length Normalization, respectively.

Cuts	Enroll: Full - Test: Full			Enroll: 10-30s - Test: 10-30s			Enroll: 3-60s - Test: 3-60s			Enroll: 10-30s - Test: 3-60s		
System	EER %	DCF08	DCF10	EER %	DCF08	DCF10	EER %	DCF08	DCF10	EER %	DCF08	DCF10
PLDA	3.59	0.154	0.401	9.56	0.482	0.932	10.94	0.464	0.836	10.12	0.466	0.900
PLDA-LN	**1.99**	**0.100**	**0.339**	7.37	0.382	0.860	7.13	0.339	0.771	7.25	0.360	0.830
FP	3.51	0.150	0.392	8.21	0.409	0.885	7.80	0.377	0.802	7.89	0.392	0.845
FP-LN	2.03	**0.100**	0.346	**6.80**	**0.354**	**0.828**	**6.21**	**0.324**	**0.753**	**6.29**	**0.341**	**0.810**
FP-PLN	2.03	**0.100**	0.346	6.82	**0.354**	**0.828**	**6.21**	**0.324**	**0.753**	**6.29**	**0.341**	**0.810**

6. I–VECTOR PRE–PROCESSING

A pre–processing step, which involves i–vector whitening followed by length normalization [10, 7], 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. 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 $\widehat{\mathbf{W}}$ is not Gaussian distributed, it is not possible to rely on the simple derivations of Section 4, and 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 $\boldsymbol{\phi}_0$ that can be approximated by its first order Taylor expansion around the i–vector itself:

$$F(\boldsymbol{\phi}) = F(\boldsymbol{\phi}_0) + J_F(\boldsymbol{\phi}_0)(\boldsymbol{\phi} - \boldsymbol{\phi}_0) + o(\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\|) \,, \quad (19)$$

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:

$$\widehat{F}(\boldsymbol{\phi}) = F(\boldsymbol{\phi}_0) + J_F(\boldsymbol{\phi}_0)(\boldsymbol{\phi} - \boldsymbol{\phi}_0) = \mathbf{u} + \frac{(\mathbf{I} - \mathbf{u}\mathbf{u}^T)}{\|\boldsymbol{\phi}_0\|}\boldsymbol{\phi} \quad (20)$$

where $\mathbf{u} = \frac{\boldsymbol{\phi}_0}{\|\boldsymbol{\phi}_0\|}$ and $\mathbf{I}$ is the identity matrix. Assuming that i–vector posterior covariances are small enough, we can replace length normalization by the linear transformation (20) computed around the i–vector posterior mean $\boldsymbol{\mu}_{\mathcal{X}}$. The extension to the full i–vector posterior consists in modifying the mean and covariance of the posterior distribution of $\mathbf{W}|\mathcal{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{X}}, \boldsymbol{\Gamma}_{\mathcal{X}}^{-1})$ as:

$$\widehat{\mathbf{W}} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_{\mathcal{X}}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}, \frac{1}{\|\boldsymbol{\mu}_{\mathcal{X}}\|^2}(\mathbf{I} - \mathbf{u}_{\mathcal{X}}\mathbf{u}_{\mathcal{X}}^T)\boldsymbol{\Gamma}_{\mathcal{X}}^{-1}(\mathbf{I} - \mathbf{u}_{\mathcal{X}}\mathbf{u}_{\mathcal{X}}^T)\right) \,, \quad (21)$$

where $\mathbf{u}_{\mathcal{X}} = \frac{\boldsymbol{\mu}_{\mathcal{X}}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}$.

Since the projection matrix $(\mathbf{I} - \mathbf{u}_{\mathcal{X}}\mathbf{u}_{\mathcal{X}}^T)$ and matrix $\mathbf{I}$ differ for a single eigen–value, $(\mathbf{I} - \mathbf{u}_{\mathcal{X}}\mathbf{u}_{\mathcal{X}}^T)$ can be well approximated by the identity matrix, and $\widehat{\mathbf{W}}$ as:

$$\overline{\mathbf{W}} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_{\mathcal{X}}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|}, \frac{\boldsymbol{\Gamma}_{\mathcal{X}}^{-1}}{\|\boldsymbol{\mu}_{\mathcal{X}}\|^2}\right) \,. \quad (22)$$

In the next section we will refer to (21) as "Projected Length Normalization" (PLN), and to (22) as "Length Normalization" (LN).

7. EXPERIMENTAL RESULTS

The proposed PLDA model aims at compensating inaccuracy and mismatch in i–vector estimates of short and variable duration speech segments. Thus, a dataset has been defined that consists of speech segments, from NIST SRE10 female tel–tel extended core condition, which were cut to obtain segments of variable duration in the range 10–30 and 3–60 seconds, respectively.

A gender dependent i–vector extractor based on 60–dimensional cepstral features and a 2048–component full covariance gender independent UBM was used for the experiments. The UBM and i–vector extractor were trained using the same data described in [8] and set to produce 400–dimensional i–vector posteriors. PLDA was trained with a 120–dimensional speaker subspace.

Table 1 summarizes the results of the tests performed on the standard NIST SRE 2010 female tel–tel extended core condition, in terms of percent Equal Error Rate and normalized minimum Detection Cost Function (DCF) as defined by NIST for SRE08 and SRE10 evaluations [6]. The standard Gaussian PLDA and the new Full Posterior PLDA systems are compared using complete conversations as well as cuts of different duration.

The first set of tests refers to standard NIST SRE10 female tel–tel extended core condition, without cuts. The aim of these experiments, performed without and with i–vector length normalization, was to verify that the new model does not introduce any degradation for long utterances. Length normalization for the Full Posterior (FP) PLDA was performed as described in Section 6. As expected, the standard and the FP systems give the same performance. In both cases, length normalization is crucial to obtain the best results.

The systems were then tested using cuts of variable durations for enrollment and test segments. For cuts in the range 10 to 30 seconds of speech used both in enrollment and testing, there is of course a performance degradation, but the FP model performs better than standard GPLDA, showing a significant improvement in terms of EER and slight improvement in terms of DCF. It is worth noting that length normalization plays again an important role also for the FP model, and that length normalization and projected length normalization give similar results. The other results in Table 1, referring to different training and test cuts, show the same trend.

8. CONCLUSIONS

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

9. REFERENCES

[1] 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.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 31–44, 2000.

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

[4] 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`.

[5] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for inferences about identity," in *Proceedings of 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[6] "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`.

[7] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i–vector length normalization in speaker recognition systems," in *Proc. of Interspeech 2011*, 2011, pp. 249–252.

[8] N. Brümmer, L. Burget, P. Kenny, P. Matějka, E. de Villiers, M. Karafiát, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 2010," in *Proc. NIST 2010 Speaker Recognition Evaluation*, 2010.

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

[10] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proceedings of ICSLP 2006*, 2006, pp. 1471–1474.