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

Bayesian Adaptation of PLDA Based Speaker Recognition to Domains with Scarce Development Data

Jesús Villalba, Eduardo Lleida

Communications Technology Group (GTC),
Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba, lleida}@unizar.es

Abstract

Recently, speaker verification based on i-vectors and PLDA has become state-of-the art. This approach relays on models whose parameters need to be estimated from a development database with a large number of speech segments and speakers. That is one of the reasons why it has been very successful on NIST evaluations where we have sufficient data available. However, when we need to do speaker verification in a domain where the development data is scarce, training accurate models is complicated. In this paper, we propose a method to do Bayesian adaptation of the PLDA parameters from a domain with sufficient development data to a domain with scarce development data. The method is based on the variational Bayes recipe. We perform experiments adapting models trained with the NIST databases to the EVALITA09 database. Results show interesting improvements.

1. Introduction

Recently, the i-vector approach has become state of the art in the speaker verification field. It provides a method to map a speech utterance to a low dimensional fixed length vector that retains the speaker identity information (i-vector) [1]. Great performance has been achieved modeling the i-vectors distributions by a generative model known as PLDA [2, 3, 4, 5]. Both, the i-vector extractor and PLDA, are models whose parameters need to be estimated from a development database with a large number of speakers and sessions. That does not pose a problem when working with NIST databases [6] where sufficient data is available. However, if we want to work with data with channels or languages different from NIST, training good models is a big challenge.

There are previous works that address the problem of database mismatch with PLDA models. In [7], dataset shift is prevented by normalizing each i-vector by its magnitude. In this manner, we make the development and test i-vector distributions more similar and more Gaussian shaped. It has been proven, that with this method we can achieve very good performance on several conditions of the NIST SRE10 dataset [8].

In [9, 8], authors presented a method to compute a fully Bayesian likelihood ratio integrating out the parameters of the PLDA model. This methods intends to take into account the uncertainly about the values of the model parameters and, in this way, to prevent over-fitting. However, this method has the side-effect that it also helps against dataset shift, because the predictive distributions that result, if you have a small amount of training data, are heavy-tailed.

In this work, we address the problem of database mismatch in a different manner. We are going to assume that we have a model trained with a large development database and a small amount of development data from the domain of interest. We present a Bayesian method to adapt the PLDA models from the original database to the target database. We have done experiments adapting models from the NIST dataset to the EVALITA09 dataset [10]. Besides, we compare our method with the i-vector length normalization.

The rest of the paper is organized as follows: Section 2 describes the i-vector extraction and PLDA approaches. Section 3 describes the Bayesian adaptation method and how to approximate the posterior distributions of the parameters of the PLDA model using variational Bayes. Section 4 describes our experimental setup and results. Finally, section 5 shows some conclusions.

2. i-Vector speaker recognition framework

2.1. i-Vector extractor

The i-vector extractor transforms a sequence of features into a fixed length low dimensional vector that retains the identity information of the signal. The i-vector approach has become state of the art for speaker verification [1]. This technique is based on the Factor Analysis (FA) approach. The idea is that each speech utterance can be modeled by a Gaussian Mixture Model (GMM). We can concatenate the means of the GMM components to form a supervector $\mathbf{M}$. Then we can write the speaker and channel dependent supervector as

$$\mathbf{M} = \mathbf{m} + \mathbf{W}\phi \tag{1}$$

where $\mathbf{m}$ is the Universal Background Model (UBM) GMM mean supervector, $\mathbf{W}$ is a low rank matrix whose columns span the subspace of maximum variability and ϕ is a Gaussian distributed vector. For each speech utterance, we compute an i-vector as the MAP estimate of the latent variable ϕ. The $\mathbf{W}$ matrix is estimated from a large development database by Maximum Likelihood and Minimum Divergence iterations [11].

2.2. Two-covariance model

We can model the i-vectors distribution with the *two-covariance model* introduced in [12]. It is a generative model which supposes that an (observed) i-vector ϕ of speaker s can be written as the sum of two hidden variables:

$$\phi = \mathbf{y}_s + \epsilon \tag{2}$$

where $\mathbf{y}_s$ is called the *speaker identity variable* and ϵ the *channel offset*. The identity variable remains constant between different observations of the speaker, but the channel offset changes. The model $\mathcal{M}$ is defined by the following two probability distributions:

$$P(\mathbf{y}|\mathcal{M}) = \mathcal{N}\left(\mathbf{y}|\mu, \mathbf{B}^{-1}\right) \quad (3)$$

$$P(\phi|\mathbf{y}, \mathcal{M}) = \mathcal{N}\left(\phi|\mathbf{y}, \mathbf{W}^{-1}\right) \quad (4)$$

where $\mathcal{N}$ denotes a Gaussian distribution; μ is the speakers mean; $\mathbf{B}^{-1}$ is the between speaker covariance matrix and $\mathbf{W}^{-1}$ is the within speaker covariance matrix. The parameters of the model μ, $\mathbf{B}$ and $\mathbf{W}$ need to be estimated by Maximum Likelihood from a development database. As $\mathbf{B}$ and $\mathbf{W}$ are full rank matrices we need a number of development speakers and segments larger than the i-vector dimension. See [12] for a closed-form expression of the likelihood-ratio between the target and non-target hypothesis.

2.3. SPLDA

The SPLDA model is a simplified version of the PLDA introduced in [2]. This a generative model that assumes that i-vector ϕ of speaker s can be written as:

$$\phi = \mu + \mathbf{V}\mathbf{y}_s + \epsilon \quad (5)$$

where μ is a speaker independent term, $\mathbf{V}$ is a low rank matrix of eigenvoices, $\mathbf{y}_s$ is the speaker factors vector, and ϵ is a channel offset.

We assume the following priors for the variables:

$$P(y) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{I}) \quad (6)$$

$$P(\epsilon|\mathcal{M}) = \mathcal{N}\left(\epsilon|\mathbf{0}, \mathbf{W}^{-1}\right) \quad (7)$$

where $\mathcal{N}$ denotes a Gaussian distribution; and $\mathbf{W}$ is the within class precision matrix. The parameters μ, $\mathbf{V}$ and $\mathbf{W}$ are trained from a development database by ML and MD iterations. This can be seen as a variant of the two-covariance model were the speaker covariance is not full rank. In order to train this model, we need a development database with a number of speakers larger than the number of speaker factors.

2.4. i-vector length normalization

Length normalization intends to apply a transform to the non-Gaussian i-vectors in order to make them more Gaussian. In this way, we can go on using the simple and computationally efficient Gaussian models with good performance. The results presented in [7] show that, for high dimensional data, it can be achieved by just normalizing the i-vectors by their magnitude.

$$\hat{\phi} = \frac{\phi}{\|\phi\|} \quad (8)$$

The i-vectors need to be centered and whitened before the length normalization. Thus, the normalized i-vectors are evenly distributed around a unitary hypersphere and we can say that they have an almost Gaussian distribution. Otherwise, if the i-vectors were very far from the origin, the normalization would project all of them into a small region of the hypersphere making them less discriminative.

3. Bayesian adaptation of the PLDA model

Here, we explain how to adapt the parameters of the two-covariance model from a domain with a large amount of development data to a domain with scarce development data. We start by introducing some notation.

3.1. Notation

From now on we will call *prior* database to the database with a large amount of development data, and *target* database to the database of the domain of interest. The whole prior database i-vectors are denoted by $\mathbf{\Phi}_{\mathrm{d}}$, while the target database i-vectors are denoted by $\mathbf{\Phi}_{\mathrm{t}}$. We shall also use $\mathbf{\Phi}$ to refer in general to any of these datasets.

Let θ_{d} be the labelling of the prior dataset. It partitions the N_d i-vectors into M_d speakers. Let θ_{t} be the labelling of the target dataset. It partitions the N_t i-vectors into M_t speakers. Let θ be any of the previous labellings.

Let $\mathbf{Y}_{\mathrm{d}}$ and $\mathbf{Y}_{\mathrm{t}}$ respectively denote the hidden speaker identity variables of the prior and target sets. $\mathbf{Y}$ can be used to refer to any of them.

Finally, we define $\mathcal{M} = (\mu, \mathbf{B}, \mathbf{W})$ and $\mathcal{M}_y = (\mu, \mathbf{B})$.

3.2. Bayesian adaptation

Following a Bayesian treatment, instead of assuming fixed values for μ, $\mathbf{B}$ and $\mathbf{W}$, we work with probability distributions for the model parameters. For doing that, we need priors for the model parameters, $P\left(\mathcal{M}_y|\Pi_{\mathcal{M}_y}\right)$ and $P(\mathbf{W}|\Pi_{\mathbf{W}})$, and calculate the posterior distribution of the model given the data, the labelling and the priors:

$$P\left(\mathcal{M}|\mathbf{\Phi}, \theta, \Pi_{\mathcal{M}_y}, \Pi_{\mathbf{W}}\right) \quad (9)$$

Figure 1 shows the graphical representation of this model. $\mathbf{\Phi}$, are observed variables; μ, $\mathbf{B}$, $\mathbf{W}$ and $\mathbf{Y}$ are hidden variables; and θ, $\Pi_{\mathcal{M}_y}$ and $\Pi_{\mathbf{W}}$ are deterministic parameters. The plates indicate that we have M speakers with N_i segments each.

The Bayesian adaptation method consists of three stages. First, we estimate the posterior distribution of the parameters of the model given the prior database $P(\mathcal{M}|\mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \Pi)$. In this stage, we assume no prior knowledge about the parameters of the model. We do that using a non-informative prior Π. We will talk more in detail about non-informative priors in the following sections.

Second, we compute the posterior distribution of the model given the target data $P(\mathcal{M}|\mathbf{\Phi}_{\mathrm{t}}, \theta_{\mathrm{t}}, \Pi_{\mathrm{d}})$ where we assume that

$$P(\mathcal{M}|\Pi_{\mathrm{d}}) = P(\mathcal{M}|\mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \Pi)\ . \quad (10)$$

That is, we use the posterior distribution given the prior data as prior to compute the posterior given the target data.

Finally, we take point estimates of μ, $\mathbf{B}$, $\mathbf{W}$ by computing their expected values given the target posterior distribution.

Unfortunately, even for the simple two-covariance model, the posteriors involved cannot be expressed in closed form. We propose to use a variational Bayes (VB) approach to calculate approximate posteriors. In the next sections, we present the VB solutions for the two-covariance model assuming two different types of model priors: non-informative and conjugate.

3.3. VB with non-informative priors

3.3.1. Non-informative prior

We can assume a non-informative prior (Jeffreys prior) for the parameters μ, $\mathbf{B}$ and $\mathbf{W}$ of the Gaussian distributions [13]. A

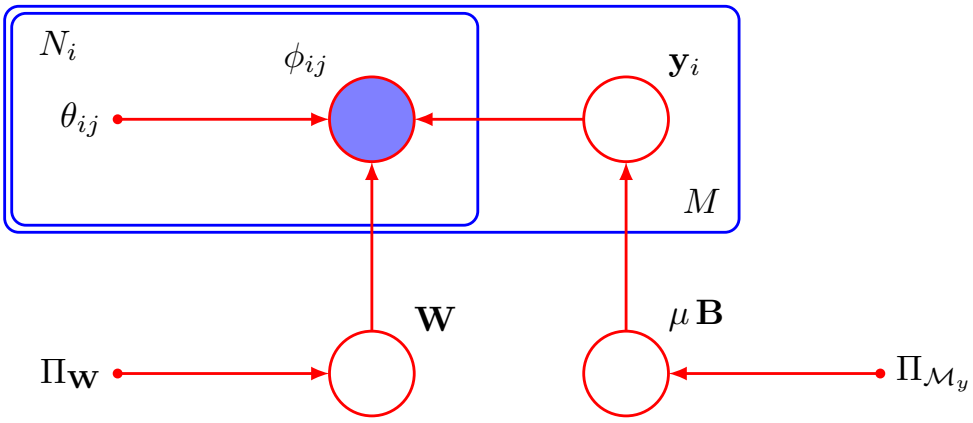

Figure 1: *Graphical model of the two-covariance model.*

non-informative prior encodes the absence of information about μ, $\mathbf{B}$ and $\mathbf{W}$ other than the training data. With this prior no Gaussian should be preferred over others and it should be invariant to any translation or scaling of the measurement space. These conditions are satisfied by this distribution:

$$P(\mathcal{M}|\Pi) = P\left(\mu, \mathbf{B}|\Pi_{\mathcal{M}_y}\right) P\left(\mathbf{W}|\Pi_{\mathbf{W}}\right) \tag{11}$$

$$P\left(\mu, \mathbf{B}|\Pi_{\mathcal{M}_y}\right) = P\left(\mu|\mathbf{B}, \Pi_{\mathcal{M}_y}\right) P\left(\mathbf{B}|\Pi_{\mathcal{M}_y}\right) \tag{12}$$

$$= \lim_{k\to 0} \mathcal{N}\left(\mu|\mu_0, (k\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}|\mathbf{B}_0/k, k\right) \tag{13}$$

$$= \alpha \left|\frac{\mathbf{B}}{2\pi}\right|^{1/2} |\mathbf{B}|^{-(d+1)/2} \tag{14}$$

$$P\left(\mathbf{W}|\Pi_{\mathbf{W}}\right) = \lim_{k\to 0} \mathcal{W}\left(\mathbf{W}|\mathbf{W}_0/k, k\right) \tag{15}$$

$$= \alpha\, |\mathbf{W}|^{-(d+1)/2} \tag{16}$$

where $\mathcal{W}$ denotes a Wishart distribution and d the dimensionality of μ. Since this density does not integrate to 1, it is improper and the symbol α is used to denote a normalizing constant which approaches zero. Note that using an improper prior does not mean that the posterior will be improper.

3.3.2. VB distributions

Our VB solution approximates the joint posterior distribution for the hidden variables and model parameters by a factorized distribution of the form:

$$P(\mathcal{M}, \mathbf{Y}|\mathbf{\Phi}, \theta, \Pi) \approx q(\mathcal{M}, \mathbf{Y}) = q(\mathcal{M})\, q(\mathbf{Y}) \tag{17}$$

which ignores any posterior dependencies between the speaker variables $\mathbf{Y}$ and the model $\mathcal{M}$. Note that we are not making further factorizing assumptions or restricting the functional form of the individual factors.

According to variational Bayes theory [14], given a set of visible variables $\mathbf{X}$ and hidden variables $\mathbf{Z}$, the optimum value of the factoring distribution $q_j^*(\mathbf{Z}_j)$ is given by

$$\ln q_j^*(\mathbf{Z}_j) = \mathrm{E}_{i\neq j}\left[\ln P(\mathbf{X}, \mathbf{Z})\right] + \text{const} \tag{18}$$

This equation means that the log of the optimum solution for factor q_j is estimated by taking the expectation of the log joint distribution over all hidden and visible variables with respect to all other factors $q_{i\neq j}$. The additive constant is needed to normalize the distribution to integrate to one.

VB is an iterative procedure. We first initialize the factors and then cycle $q_j(\mathbf{Z}_j)$ through the factors re-estimating each one using (18) until convergence.

We can use the rules described in [14] on the graphical model of Figure 1 to determine the dependencies between the model variables. Those rules allow to write the joint distribution of all variables as

$$P(\mathbf{\Phi}, \mathcal{M}, \mathbf{Y}|\theta, \Pi) = P(\mathbf{\Phi}|\mathbf{Y}, \mathbf{W}, \theta)\, P(\mathbf{Y}|\mathcal{M}_y) \\ P\left(\mathcal{M}_y|\Pi_{\mathcal{M}_y}\right) P\left(\mathbf{W}|\Pi_{\mathbf{W}}\right) . \tag{19}$$

Now, applying (18), it is straightforward to obtain our variational distributions.

The optimum for the factor $q(\mathbf{Y})$ is given by a product of Gaussian distributions:

$$q^*(\mathbf{Y}) = \prod_{i=1}^{M} q^*(\mathbf{y}_i) \tag{20}$$

$$q^*(\mathbf{y}_i) = \mathcal{N}\left(\mathbf{y}_i|\mathbf{L}_i^{-1}\gamma_i, \mathbf{L}_i^{-1}\right) \tag{21}$$

$$\mathbf{L}_i = \mathrm{E}_{\mathcal{M}}[\mathbf{B}] + n_i \mathrm{E}_{\mathcal{M}}[\mathbf{W}] \tag{22}$$

$$\gamma_i = \mathrm{E}_{\mathcal{M}}[\mathbf{B}\mu] + \mathrm{E}_{\mathcal{M}}[\mathbf{W}] \sum_{\phi\in\mathcal{S}_i} \phi \tag{23}$$

where the speaker identity variables $\mathbf{y}_i$ are independent. Note that we have not forced that in any way but it originates naturally from the original factorization that we have chosen.

The optimum for the factor $q(\mathcal{M})$ is again a product of factors

$$q^*(\mathcal{M}) = q^*(\mathcal{M}_y)\, q^*(\mathbf{W}) \tag{24}$$

The factor $q^*(\mathcal{M}_y)$ is a Gaussian-Wishart distribution.

$$q^*(\mathcal{M}_y) = \mathcal{N}\left(\mu|\overline{\mathbf{y}}, (M\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right) \tag{25}$$

where we have defined

$$\overline{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i] \tag{26}$$

$$\mathbf{S}_y = \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] - M\overline{\mathbf{y}}\overline{\mathbf{y}}^T \tag{27}$$

We have to remark that for this distribution to be proper we need the number of speakers M to be larger than the i-vectors dimensionality.

The factor $q^*(\mathbf{W})$ is Wishart distributed

$$q^*(\mathbf{W}) = \mathcal{W}\left(\mathbf{W}|\mathbf{S}_\phi^{-1}, N\right) \quad \text{if } N > d \tag{28}$$

where

$$\overline{\phi}_i = \frac{1}{n_i}\sum_{\phi\in\mathcal{S}_i}\phi \tag{29}$$

$$\mathbf{R}_\phi = \sum_{j=1}^{N}\phi_j\phi_j^T \tag{30}$$

$$\mathbf{S}_\phi = \sum_{i=1}^{M}\sum_{\phi\in\mathcal{S}_i}\mathrm{E}_{\mathbf{Y}}\left[(\phi-\mathbf{y}_i)(\phi-\mathbf{y}_i)^T\right] \tag{31}$$

$$= \mathbf{R}_\phi + \sum_{i=1}^{M} N_i\left(\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] - \mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\right]\overline{\phi}_i^T - \overline{\phi}_i\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\right]^T\right) \tag{32}$$

In order to estimate the parameters of $q(\mathbf{Y})$ and $q(\mathcal{M})$ we still need to evaluate some additional expectations. Using the properties of the Gaussian and Wishart distributions [14] we have

$$\mathrm{E}_{\mathcal{M}}[\mathbf{B}] = M\mathbf{S}_y^{-1} \tag{33}$$

$$\mathrm{E}_{\mathcal{M}}[\mathbf{B}\mu] = M\mathbf{S}_y^{-1}\overline{\mathbf{y}} \tag{34}$$

$$\mathrm{E}_{\mathcal{M}}[\mathbf{W}] = N\mathbf{S}_\phi^{-1} \tag{35}$$

$$\mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i] = \mathbf{L}_i^{-1}\gamma_i \tag{36}$$

$$\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] = \mathbf{L}_i^{-1} + \mathbf{L}_i^{-1}\gamma_i\gamma_i^T\mathbf{L}_i^{-1} \tag{37}$$

3.4. VB with conjugate priors

3.4.1. Conjugate prior

Now, we want to take the posterior distribution that we have got with the non-informative prior and use it as an informative prior to compute a new posterior given the target database. Our model prior is now given by the following approximation:

$$P(\mathcal{M}|\Pi_{\mathrm{d}}) = P(\mathcal{M}|\mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}\Pi) \approx q_{\mathrm{d}}(\mathcal{M}) \tag{38}$$

where $q_{\mathrm{d}}(\mathcal{M})$ is the variational factor of $\mathcal{M}$ conditioned on the prior data. As we got in equations 24, 25 and 28:

$$q_{\mathrm{d}}(\mathcal{M}) = q_{\mathrm{d}}(\mathcal{M}_y)\, q_{\mathrm{d}}(\mathbf{W}) \tag{39}$$

$$q_{\mathrm{d}}(\mathcal{M}_y) = \mathcal{N}\left(\mu|\overline{\mathbf{y}}_{\mathrm{d}}, (\beta_{\mathrm{d}y}\mathbf{B})^{-1}\right)\mathcal{W}\left(\mathbf{B}|\mathbf{S}_{\mathrm{d}y}^{-1}, \nu_{\mathrm{d}y}\right) \tag{40}$$

$$q_{\mathrm{d}}(\mathbf{W}) = \mathcal{W}\left(\mathbf{W}|\mathbf{S}_{\mathrm{d}\phi}^{-1}, \nu_{\mathrm{d}\phi}\right) \tag{41}$$

where $\beta_{\mathrm{d}y} = \nu_{\mathrm{d}y} = M_{\mathrm{d}} > d$ and $\nu_{\mathrm{d}\phi} = N_{\mathrm{d}} > d$.

These distributions are conjugate priors for the Gaussian distribution. This is very convenient because they should produce posteriors that are again Gaussian Wishart.

3.4.2. Variational Distributions

Now, we use again the factorization given by equation (17) along with equations (18) and (19) to compute the variational distributions given the conjugate prior.

The optimum for the factor $q(\mathbf{Y})$ is the same as for the non-informative prior.

For the optimum for the factor $q(\mathcal{M})$ has a similar form to the non-informative case.

$$q^*(\mathcal{M}) = q^*(\mathcal{M}_y)\, q^*(\mathbf{W}) \tag{42}$$

$$q^*(\mathcal{M}_y) = \mathcal{N}\left(\mu|\overline{\mathbf{y}}', (\beta_y'\mathbf{B})^{-1}\right)\mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu_y'\right) \tag{43}$$

$$q^*(\mathbf{W}) = \mathcal{W}\left(\mathbf{W}|\mathbf{S}_\phi'^{-1}, \nu_\phi'\right)\ . \tag{44}$$

In the previous equations we have defined

$$\overline{\mathbf{y}} = \frac{1}{M_{\mathrm{t}}}\sum_{i=1}^{M_{\mathrm{t}}}\mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i] \tag{45}$$

$$\mathbf{S}_y = \sum_{i=1}^{M_{\mathrm{t}}}\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] - M_{\mathrm{t}}\overline{\mathbf{y}}\overline{\mathbf{y}}^T \tag{46}$$

$$\mathbf{S}_\phi = \sum_{i=1}^{M_{\mathrm{t}}}\sum_{\phi\in\mathcal{S}_i}\mathrm{E}_{\mathbf{Y}}\left[(\phi-\mathbf{y}_i)(\phi-\mathbf{y}_i)^T\right] \tag{47}$$

$$\beta_y' = \beta_{\mathrm{d}y} + M_{\mathrm{t}} \tag{48}$$

$$\nu_y' = \nu_{\mathrm{d}y} + M_{\mathrm{t}} \tag{49}$$

$$\overline{\mathbf{y}}' = \frac{1}{\beta_y'}\left(\beta_{\mathrm{d}y}\overline{\mathbf{y}}_{\mathrm{d}} + M_{\mathrm{t}}\overline{\mathbf{y}}\right) \tag{50}$$

$$\mathbf{S}_y' = \mathbf{S}_{\mathrm{d}y} + \mathbf{S}_y + \frac{\beta_{\mathrm{d}y}M_{\mathrm{t}}}{\beta_y'}(\overline{\mathbf{y}}-\overline{\mathbf{y}}_{\mathrm{d}})(\overline{\mathbf{y}}-\overline{\mathbf{y}}_{\mathrm{d}})^T \tag{51}$$

$$\nu_\phi' = \nu_{\mathrm{d}\phi} + N_{\mathrm{t}} \tag{52}$$

$$\mathbf{S}_\phi' = \mathbf{S}_{\mathrm{d}\phi} + \mathbf{S}_\phi\ . \tag{53}$$

Finally, we need the expectations

$$\mathrm{E}_{\mathcal{M}}[\mu] = \overline{\mathbf{y}}' \tag{54}$$

$$\mathrm{E}_{\mathcal{M}}[\mathbf{B}] = \nu_y'\mathbf{S}_y'^{-1} \tag{55}$$

$$\mathrm{E}_{\mathcal{M}}[\mathbf{B}\mu] = \nu_y'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}' \tag{56}$$

$$\mathrm{E}_{\mathcal{M}}[\mathbf{W}] = \nu_\phi'\mathbf{S}_\phi'^{-1} \tag{57}$$

Once, we have the posterior distributions we approximate μ, $\mathbf{B}$ and $\mathbf{W}$ by their expectations given by equations (54), (55) and (57). Then, we plug them into the likelihood ratio formula during the test phase.

The parameters $\beta_{\mathrm{d}y} = \nu_{\mathrm{d}y} = M_{\mathrm{d}}$ and $\nu_{\mathrm{d}\phi} = N_{\mathrm{d}}$ control the weight of the prior on the posterior. If we would want the target data to have more influence on the posterior we could lower the values of M_{d} and N_{d} manually. If we do that, we had to adjust $\mathbf{S}_{\mathrm{d}y}$ and $\mathbf{S}_{\mathrm{d}\phi}$ so that prior expectations of $\mathbf{B}$ and $\mathbf{W}$ given by equations (33) and (35) remain unchanged.

4. Experiments

4.1. Experimental setup

4.1.1. EVALITA09 Dataset

We performed speaker verification experiments on the EVALITA09 dataset. EVALITA is an evaluation of natural language processing and speech tools for Italian. We have chosen this database because it has guidelines [10] that are similar to the ones of the NIST SRE [6]. The data is recorded from landline (PSTN) or mobile (GSM) telephone channels. Recordings are in Italian language, including speakers uniformly selected from all regions. It includes several groups of data:

- UBM data: Speech data recorded by 30 male and 30 female speakers, during 20 sessions (10 PSTN calls + 10

GSM calls). The total durations of speech is 1200 minutes (∼1 minute per call). Calls were provided cut into small segments, thus we have 18000 short speech segments (9000 male + 9000 female).

- Training data: Data for speaker enrollment. It has 50 male and 50 female speakers. 6 training conditions are considered:
  - TC1: PSTN short (1 PSTN call, ∼1 minute per client).
  - TC2: GSM short (1 GSM call, ∼1 minute per client).
  - TC3: PSTN long (3 PSTN calls, ∼3 minutes per client).
  - TC4: GSM long (3 GSM calls, ∼3 minutes per client).
  - TC5: mixed short (1 PSTN + 1 GSM calls, ∼2 minutes per client).
  - TC6: mixed long (3 PSTN + 3 GSM calls, ∼6 minutes per client).
- Test data: Two test conditions are considered with 2071 trials each:
  - TS1: short (1 sequence of digits; ∼10 seconds).
  - TS2: long (1 sequence of digits, 4 short sentences, 2 isolated words; ∼30 seconds).

4.1.2. NIST Dataset

Speech from SRE04, SRE05 and SRE06 is used to estimate the prior distribution of the parameters of the two-covariance model. It includes 529 male and 729 female speakers with a total of 7410 male and 9920 female phone calls. Each phone call has around 2 minutes of speech. This database presents a wide variety of transmission channels and telephone handsets.

4.1.3. i-Vector extractor

We used 400 dimensional i-vectors. The i-vector extractor uses 20 short-time Gaussianized MFCC plus deltas and double deltas and a 2048 component diagonal covariance UBM. The UBM and the i-vector extractor are gender dependent and they were trained with data from the NIST dataset.

4.1.4. PLDA

We show results training PLDA models with the NIST dataset, the EVALITA09 dataset or adapting models from NIST to EVALITA09. The models are gender dependent. For the adaptation case, first, we compute the posterior distribution of the parameters of the two-covariance model given the NIST data and a non-informative prior. Then, we use that posterior distribution as prior to compute the posterior distribution of the model parameters given the EVALITA09 UBM dataset.

4.1.5. Length normalization

For the cases where i-vector length normalization is used, the parameters needed to do the centering and whitening steps (mean and rotation matrix) are estimated in the same manner as the corresponding PLDA parameters. That is, trained from NIST, from EVALITA09 or adapted from NIST to EVALITA09.

Table 1: *EER(%)/minDCF TC6 TS2 vs. effective number of speakers (M) and segments (N) in the prior distribution.*

	male		female	
	EER	DCF	EER	DCF
NIST	2.99	0.098	1.37	0.089
EVITA09	6.08	0.279	7.02	0.232
Adapt actual M N	1.83	0.104	1.32	0.059
Adapt M401 N401	2.12	0.160	1.56	0.107
Adapt M401 N500	2.12	0.158	1.55	0.107
Adapt M401 N750	2.03	0.151	1.45	0.102
Adapt M401 N1500	1.77	0.141	1.32	0.090
Adapt M401 N3000	1.79	0.119	1.39	0.071
Adapt M401 N6000	1.83	0.106	1.35	0.061
Adapt M401 N9000	1.74	0.101	1.26	0.059
Adapt M401 N12000	1.68	0.089	1.26	0.058
Adapt M401 N15000	1.80	0.086	1.17	0.048
Adapt M401 N18000	1.79	0.081	1.17	0.048
Adapt M500 N401	2.17	0.160	1.56	0.112
Adapt M500 N500	2.16	0.160	1.56	0.112
Adapt M500 N750	2.12	0.154	1.50	0.107
Adapt M500 N1500	1.83	0.141	1.38	0.092
Adapt M500 N3000	1.79	0.121	1.39	0.071
Adapt M500 N6000	1.83	0.109	1.35	0.061
Adapt M500 N9000	1.80	0.104	1.31	0.059
Adapt M500 N12000	1.83	0.096	1.26	0.059
Adapt M500 N15000	1.80	0.089	1.18	0.051
Adapt M500 N18000	1.83	0.084	1.26	0.048

4.1.6. Score Normalization

Unless stated otherwise, we show results with s-norm [15]. We used the utterances from the EVALITA09 UBM dataset (9000 male + 9000 female) as cohorts.

4.2. Results

4.2.1. Number of effective number of development speakers and segments

In the first experiment, we compare the results of training the PLDA model with only NIST data, only EVALITA09 data and Bayesian adaptation. For the Bayesian adaptation we compare different configurations where we manually tune the values of M and N in the prior Gaussian-Whishart distribution. We call effective number of development speakers and segments to M and N. We must point out that the prior distribution is always trained using the actual number of speakers and segments of the development dataset, and we tune M and N afterwards. Tuning M and N has a double effect: it changes the width of the prior, and, at the same time, it changes the weight of the prior on the posterior. Thus, if we assign low M and N values, the EVALITA09 data has more influence on the posterior

In all cases, except when training with EVALITA09 only, we have used the two-covariance model. In the case of training with EVALITA09 we have used a SPLDA model with 25 speaker factors. We have chosen 25 factors because the number of factors must be smaller than the number of development speakers (30). We cannot use the two-covariance model either because it need a number of development speakers larger than the i-vector dimension (400). We have not used length normalization.

Results for the TC6 TS2 condition are shown in Table 1.

Table 2: *EER(%)/minDCF TC6 TS2 vs. adapted parameters*

	male		female	
	EER	DCF	EER	DCF
NIST	2.99	0.098	1.37	0.089
EVITA09	6.08	0.279	7.02	0.232
Adapt μ	2.96	0.096	1.37	0.086
Adapt $\mu\mathbf{B}$	2.86	0.073	1.39	0.087
Adapt $\mu\mathbf{BW}$	1.80	0.086	1.17	0.048

The results are given in terms of EER and normalized minimum Decision Cost Functions as defined by the EVALITA09 guidelines [10] ($C_{Miss} = 10, C_{FA} = 1, P_T = 0.5$). We observe that training with only EVALITA09 largely degrades the performance. We think that it is mainly due to the low number of speakers that forces us to use a low number of speaker factors. On the other hand the Bayesian adaptation produces a nice improvement over training only with NIST.

We achieve the optimum performance tuning the values of effective speakers (M) and segments (N) of the prior distribution. The values of M and N start from 401 so that the prior distributions can be proper. Tuning M does not change the results much because it is still much larger than the number of EVALITA09 speakers ($M >> 30$). On the contrary, we get an improvement tuning N. We could think that having 9000 development segments in EVALITA09 we could train $\mathbf{W}$ only with EVALITA09. If that were the case we should get the best performance with a low value of N. However, taking N $1.5-2$ times larger than the number of EVALITA09 segments achieves better performance. For example, if we take $M = 401$ and $N = 15000$ we have an EER improvement of 40% for males and 14% for females; and an DCF improvement of 12% for males and 46% for females.

From now on we will use $M = 401$ and $N = 15000$ for all the experiments.

4.2.2. Adapting different parameters

Here, we compare the result of adapting different parameters of the two-covariance model: μ; μ and $\mathbf{B}$; and μ, $\mathbf{B}$, and $\mathbf{W}$. Results for condition TC6 TS2 are shown in Table 2. For males, adapting μ and $\mu\mathbf{B}$ produce small improvements; and for females, not improvement at all. To get a clear improvement we need to adapt $\mathbf{W}$.

4.2.3. Length normalization

Table 3 shows results on condition TC6 TS2 using length normalized i-vectors. Experiments on NIST databases have shown that length normalization boosts performance and makes score

Table 3: *EER(%)/minDCF TC6 TS2 with Lnorm*

		male		female	
		EER	DCF	EER	DCF
No s-norm	NIST	3.28	0.146	1.61	0.113
	EVITA09	5.60	0.245	6.43	0.247
	Adapt $\mu\mathbf{BW}$	1.15	0.091	1.35	0.106
s-norm	NIST	2.23	0.100	1.25	0.055
	EVITA09	4.92	0.193	6.20	0.236
	Adapt $\mu\mathbf{BW}$	0.93	0.068	1.18	0.077

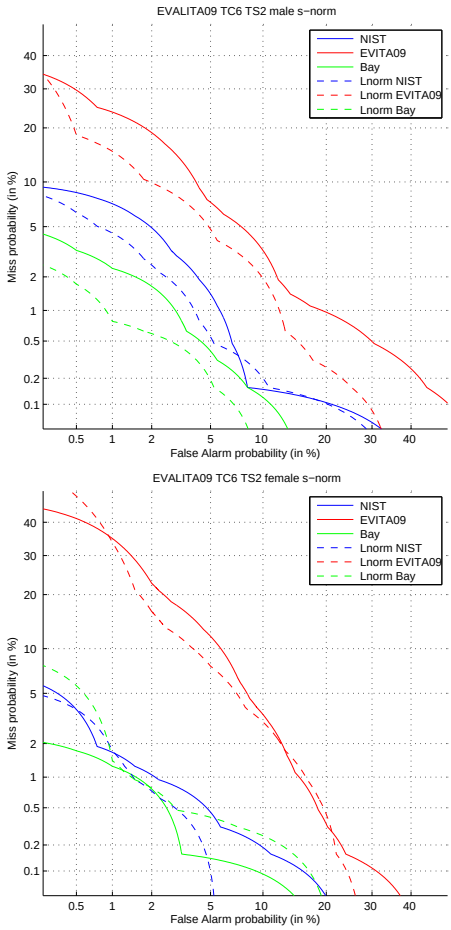

Figure 2: *DET curves TC6 TS2 for male (top) and female (bottom)*

normalization unnecessary [7, 5]. It has been claimed that it is mainly due to the fact that length normalization reduces mismatch between the development and test databases. However, for this database, we observe that length normalization improves performance but not as much as for NIST databases. Besides score normalization is needed to achieve optimum performance.

4.3. Results all conditions

Tables 4 and 5 show results on all EVALITA09 conditions without and with length normalization respectively. We observe something unexpected; length normalization does not achieve the best results in all conditions. For training conditions TC1-3 and TC5 females, with length normalization, we have a degradation with the adapted model respect to NIST model. However, for those conditions, the best results are for the adapted model without length normalization. In the rest of conditions the best results are achieved combining both length normalization and model adaptation.

The overall conclusion that we draw from these tables is that conditions with longer training or test data like TC6 TS2 get more benefit from length normalization and we can combine both techniques. Otherwise, using only model adaptation is better.

Table 4: *EER(%)/minDCF Multiple conditions without Lnorm*

		male		female	
		EER	DCF	EER	DCF
TC1 TS1	NIST	10.50	0.566	9.09	0.465
	EVITA09	17.05	0.667	12.75	0.594
	Adapt μ**BW**	9.15	0.441	7.89	0.458
TC1 TS2	NIST	6.02	0.352	4.28	0.146
	EVITA09	10.42	0.628	10.62	0.426
	Adapt μ**BW**	4.88	0.326	3.87	0.193
TC2 TS1	NIST	17.51	0.666	13.10	0.563
	EVITA09	18.59	0.740	18.37	0.785
	Adapt μ**BW**	13.67	0.612	9.97	0.527
TC2 TS2	NIST	11.79	0.417	5.64	0.318
	EVITA09	13.66	0.621	14.06	0.586
	Adapt μ**BW**	9.92	0.383	5.02	0.254
TC3 TS1	NIST	9.54	0.482	8.08	0.458
	EVITA09	13.94	0.637	12.51	0.562
	Adapt μ**BW**	7.44	0.400	6.60	0.407
TC3 TS2	NIST	4.38	0.243	3.60	0.137
	EVITA09	9.54	0.521	8.54	0.321
	Adapt μ**BW**	3.69	0.231	2.90	0.159
TC4 TS1	NIST	17.35	0.659	13.30	0.565
	EVITA09	16.30	0.749	16.21	0.719
	Adapt μ**BW**	12.45	0.618	8.63	0.403
TC4 TS2	NIST	10.84	0.443	4.44	0.272
	EVITA09	12.46	0.561	10.73	0.522
	Adapt μ**BW**	8.41	0.392	3.89	0.246
TC5 TS1	NIST	9.54	0.399	6.57	0.383
	EVITA09	14.03	0.545	12.84	0.588
	Adapt μ**BW**	6.41	0.308	4.36	0.234
TC5 TS2	NIST	3.81	0.189	1.58	0.089
	EVITA09	8.32	0.294	9.32	0.354
	Adapt μ**BW**	2.60	0.154	1.53	0.076
TC6 TS1	NIST	8.11	0.376	7.11	0.314
	EVITA09	12.59	0.499	12.03	0.512
	Adapt μ**BW**	5.12	0.297	3.81	0.192
TC6 TS2	NIST	2.99	0.098	1.37	0.089
	EVITA09	6.08	0.279	7.02	0.232
	Adapt μ**BW**	1.80	0.086	1.17	0.048

Table 5: *EER(%)/minDCF Multiple conditions with Lnorm*

		male		female	
		EER	DCF	EER	DCF
TC1 TS1	NIST	10.48	0.537	8.05	0.417
	EVITA09	14.47	0.607	12.38	0.635
	Adapt μ**BW**	12.16	0.680	11.72	0.622
TC1 TS2	NIST	5.27	0.287	3.86	0.159
	EVITA09	9.52	0.560	11.38	0.543
	Adapt μ**BW**	8.01	0.467	6.38	0.313
TC2 TS1	NIST	15.65	0.565	11.96	0.637
	EVITA09	17.28	0.674	15.47	0.785
	Adapt μ**BW**	14.94	0.700	14.22	0.663
TC2 TS2	NIST	10.68	0.492	5.92	0.295
	EVITA09	11.07	0.494	12.18	0.494
	Adapt μ**BW**	10.32	0.496	7.27	0.378
TC3 TS1	NIST	9.02	0.441	7.21	0.422
	EVITA09	11.98	0.554	11.87	0.548
	Adapt μ**BW**	9.67	0.536	8.90	0.494
TC3 TS2	NIST	3.96	0.247	3.07	0.136
	EVITA09	8.31	0.399	8.39	0.384
	Adapt μ**BW**	5.08	0.318	3.38	0.272
TC4 TS1	NIST	14.66	0.545	11.90	0.518
	EVITA09	15.10	0.678	13.75	0.637
	Adapt μ**BW**	11.95	0.622	10.37	0.478
TC4 TS2	NIST	9.98	0.401	4.15	0.253
	EVITA09	10.42	0.439	9.56	0.407
	Adapt μ**BW**	7.94	0.402	3.77	0.272
TC5 TS1	NIST	8.50	0.325	6.61	0.318
	EVITA09	10.94	0.468	10.73	0.579
	Adapt μ**BW**	5.88	0.328	8.04	0.362
TC5 TS2	NIST	3.30	0.156	1.22	0.058
	EVITA09	5.79	0.248	8.19	0.355
	Adapt μ**BW**	2.79	0.146	2.14	0.129
TC6 TS1	NIST	7.28	0.285	6.29	0.280
	EVITA09	9.63	0.409	10.58	0.477
	Adapt μ**BW**	3.81	0.224	5.85	0.273
TC6 TS2	NIST	2.23	0.100	1.25	0.055
	EVITA09	4.92	0.193	6.20	0.236
	Adapt μ**BW**	0.93	0.068	1.18	0.077

Figure 2 shows DET curves for the condition TC6 TS2. We observe that the improvement of the Bayesian approach is consistent along different operating points.

5. Conclusions

We have presented a method to adapt a PLDA i-vector classifier from a domain with a large amount of development data to a domain with scarce development data. This method consists in computing the posterior distribution of the parameters of the model given the data of the larger database. Then we use that posterior as prior to compute the posterior distribution of the parameters given the data of the domain of interest. These posterior distributions can be estimated approximately for the particular case of the two-covariance model using a variational Bayes procedure.

We have done experiments adapting PLDA from the NIST database to the EVALITA09 database. We have shown results on all EVALITA09 training and test conditions that indicate that this technique improves the performance of the system. We have seen that the improvement is mainly due to the adaptation of the $\mathbf{W}$ matrix that describes the channel space.

We have compared this method with the length normalization. We have seen that, for this dataset, conditions with longer training or test data get better results combining both techniques. Otherwise, using only model adaptation is better than length normalization. We have seen that, here, length normalization needs score normalization to achieve optimum performance, contrary to what happens when we do the test on the telephone conditions of NIST datasets.

As future work, we want extend our work doing Bayesian adaptation of the UBM and the i-vector extractor. Besides, we want to try this technique on other datasets.

6. Acknowledgment

This work has been supported by the Spanish Government through national projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

7. References

[1] Najim Dehak, Patrick Kenny, Redah Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis For Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2010.

[2] Simon J D Prince and James H Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," *IEEE International Conference on Computer Vision*, , no. iii, pp. 1–8, 2007.

[3] Patrick Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[4] Pavel Matejka, Ondrej Glembek, Fabio Castaldo, Md Jahangir Alam, Patrick Kenny, Lukas Burget, and Jan Cernocky, "Full-Covariance UBM and Heavy-Tailed PLDA in I-Vector Speaker Verification," in *IEEE International Conference on Acoustics Speech and Signal Processing 2011*, Prague, 2011.

[5] Mohammed Senoussaoui, Patrick Kenny, Niko Brummer, Edward De Villiers, and Pierre Dumouchel, "Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition," *Interspeech 2011*, pp. 1–19, 2011.

[6] NIST Speech Group, "NIST Speaker Recognition Evaluation," 2010.

[7] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Interspeech 2011*, Florence, 2011, pp. 249–252.

[8] Jesús Villalba, Niko Brummer, and Eduardo Lleida, "Fully Bayesian Likelihood Ratios vs i-vector Length Normalization in Speaker Recognition Systems," in *SRE11 Speaker Recognition Workshop*, Atlanta, 2011.

[9] Jesús Villalba and Niko Brummer, "Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance," in *Interspeech 2011*, Florence, 2011, pp. 28–31.

[10] Guido Aversano, "Evalita 2009 Speaker Identity Verification - Application Track Guidelines," 2009.

[11] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[12] Niko Brummer and Edward De Villiers, "The Speaker Partitioning Problem," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[13] Thomas Minka, "Inferring a Gaussian distribution," Tech. Rep., MIT media Lab, 1998.

[14] Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, 2006.

[15] Mohammed Senoussaoui, Patrick Kenny, Najim Dehak, and Pierre Dumouchel, "An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.