HANDLING I-VECTORS FROM DIFFERENT RECORDING CONDITIONS USING MULTI-CHANNEL SIMPLIFIED PLDA IN SPEAKER RECOGNITION

Jesús Villalba, Eduardo Lleida

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

ABSTRACT

In this work, we address the problem of having i-vectors that have been produced in different channel conditions. Traditionally, this problem has been handled training the LDA covariance matrices pooling the data of all the conditions or averaging the covariance matrices of each condition in different ways. We present a PLDA variant that we call, multi-channel SPLDA, where the speaker space distribution is common to all i-vectors and the channel space distribution depends on the type of channel where the segment has been recorded. We test our approach on the telephone part of the NIST SRE10 extended condition where we added some additive noises to the test segments. We compare results of a SPLDA model trained only with clean data, SPLDA trained with pooled noisy and clean data and our MCSPLDA model.

Index Terms— speaker recognition, PLDA, i-vector, multi-channel, generative

1. INTRODUCTION

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 [2–4].

This paper addresses the problem of how to model the i-vector distributions when we have segments recorded over different types of channels or noisy environments. The standard PLDA model describes the inter-session variability between the i-vectors of a given speaker by a unique within class covariance matrix. Intuition tells us that session variability is very dependent on the channel conditions. Therefore, we propose a PLDA variant with different within class covariance matrices for each channel.

The problem of multi-channel speaker recognition has been addressed before. The works in [5–10], present similar approaches. Some kind of LDA projection is applied to telephone and microphone i-vectors to project them into a common space. Then, i-vectors are classified using cosine similarity or PLDA. The main difference is the method to estimate the LDA projection matrix. In [5], LDA is trained pooling all the telephone and microphone data or averaging the telephone and microphone between and within class covariance matrices. In [6], authors project the i-vectors using a PLDA where the covariance of the residual term is trained only on telephone data and the eigenchannel matrix is trained to handle the variability included in the microphone data that is not already included in the telephone data. In [7], i-vectors are projected using heavy tail PLDA trained on telephone and microphone. In [8–10], several ways of estimating and averaging the between and within class covariance matrices are studied.

A different approach is adopted in [11] where standard PLDA is trained using pooled clean and noisy data. In [12], authors train three conditioned PLDA models (telephone, microphone and telephone+microphone). Then, in the classification phase they are treated as components of a mixture of PLDA and Bayesian fusion of scores is implemented. In in [13], several PLDA variants are explored (condition dependent, pooled PLDA, tied PLDA) and the scores fused.

In this work, we present a variant of Prince's tied PLDA [2] where each i-vector is modeled by the same between class covariance but a different within class covariance matrix depending on the type of channel. This model can also be seen as a mixture of PLDA models where the speaker component is tied to be the same across the components. This framework allows pooling all the data available to estimate the PLDA parameters in such a way that the speaker space is estimated with all the data and the channel spaces are estimated only with the data of their corresponding channel.

The rest of the paper is organized as follows: Section 2 describes the standard PLDA approach. Section 3 describes our multi-channel PLDA framework. Section 5 describes our experimental setup and results on the SRE10 extended condition where we added some additive noises, we compare the multi-channel PLDA with a standard PLDA trained pooling all the data. Finally, in section 6, we discuss the results.

2. SPLDA

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

$$\phi = \mu + \mathbf{V}\mathbf{y}_i + \epsilon \tag{1}$$

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

We assume Gaussian priors for the variables:

$$P(\mathbf{y}_i) = \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathbf{I}) \tag{2}$$

$$P(\epsilon|\mathcal{M}) = \mathcal{N}\left(\epsilon|\mathbf{0}, \mathbf{W}^{-1}\right) \tag{3}$$

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. We call $\mathcal{M}$ to the set of all the model parameters.

It is well known that, for this model, the posterior of the speaker variables $\mathbf{y}_i$ is a Gaussian distribution given by

$$P(\mathbf{y}_i|\mathbf{\Phi}_i, \mathcal{M}) = \mathcal{N}\left(\mathbf{y}_i|\mathbf{L}^{-1}\gamma, \mathbf{L}^{-1}\right) \tag{4}$$

where

$$\mathbf{L} = \mathbf{I} + N_i\mathbf{V}^T\mathbf{W}\mathbf{V} \tag{5}$$

$$\gamma = \mathbf{V}^T\mathbf{W}\overline{\mathbf{F}}_i \tag{6}$$

and N_i are the zeroth order statistics and $\overline{\mathbf{F}}_i$ are the first order statistics centered in μ, for a speaker i.

3. MULTI-CHANNEL SPLDA

3.1. Model description

The SPLDA model can be modified to take into account the fact that each i-vector can be generated by a different channel. Now, we assume that an i-vector ϕ of speaker i, generated in a channel k can be written as:

$$\phi|_{z_k=1} = \mu_k + \mathbf{V}\mathbf{y}_i + \epsilon_k \tag{7}$$

where μ_k is a channel dependent mean, $\mathbf{V}$ is the eigenvoices matrix, $\mathbf{y}_i$ is the speaker factor vector, and ϵ_k is a channel offset with channel dependent precision matrix $\mathbf{W}_k$. We define $\mu = \{\mu_k\}_{k=1}^K$ and $\mathbf{W} = \{\mathbf{W}_k\}_{k=1}^K$.

Furthermore, $\mathbf{z}$ is a variable that indicates the type of channel that generates ϵ. It is a 1-of-K binary vector with elements z_k for $k = 1, \ldots, K$ where z_k is equal to 1 if ϕ has been generated by channel k and 0 otherwise. For simplicity, we assume that we have some kind of channel detector that provides the type of channel or, at least, the probability of $\mathbf{z}$, $P(\mathbf{z})$. We could use the own MCPLDA model to compute $P(\mathbf{z})$ but we find more convenient to assume that it is given.

Figure 1 shows the Bayesian network that depicts this model. Note that this is equivalent to a mixture of PLDA models where $\mathbf{V}$ and $\mathbf{y}$ are tied across the components of the mixture. Using this model, we intend to keep a channel independent speaker space ($\mathbf{V}$), given that speaker are human beings that should not change depending on the recording environment. Besides, this model forces that the speaker variable $\mathbf{y}_i$ is unique regardless of the channel.

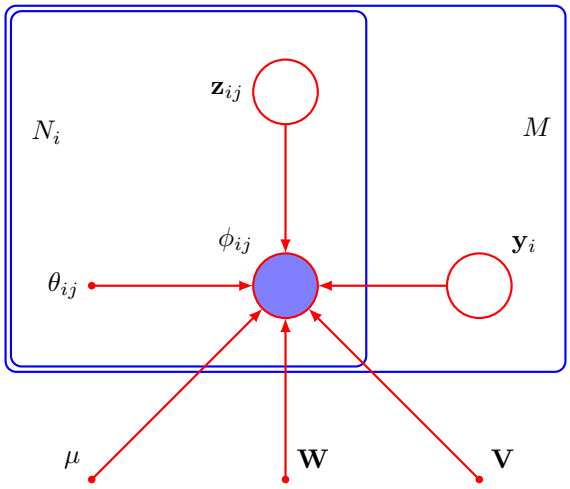

Fig. 1. BN for multichannel SPLDA model.

3.2. Posterior of the hidden variables

The main thing needed to implement our PLDA model is to compute the posterior distribution of the speaker variables $\mathbf{y}_i$. To do that, we find convenient to define the sufficient statistics for speaker i and channel k as

$$N_{ik} = \sum_{j=1}^{N_i} P(z_{ijk} = 1) \tag{8}$$

$$\mathbf{F}_{ik} = \sum_{j=1}^{N_i} P(z_{ijk} = 1)\,\phi_{ij} \tag{9}$$

where N_i is the number of i-vectors of speaker i and $P(z_{ijk} = 1)$ is the probability for ϕ_{ij} to be generated by channel k. Besides, the channel centered statistics are defined as:

$$\overline{\mathbf{F}}_{ik} = \mathbf{F}_{ik} - N_{ik}\mu_k\ . \tag{10}$$

It can be shown that, the posterior of the hidden variables is a Gaussian distribution given by

$$P(\mathbf{y}_i|\mathbf{\Phi}_i, \mathbf{z}_i, \mathcal{M}) = \mathcal{N}\left(\mathbf{y}_i|\mathbf{L}^{-1}\gamma, \mathbf{L}^{-1}\right) \tag{11}$$

where

$$\mathbf{L} = \mathbf{I} + \sum_{k=1}^{K} N_{ik} \mathbf{V}^T \mathbf{W}_k \mathbf{V} \tag{12}$$

$$\gamma = \sum_{k=1}^{K} \mathbf{V}^T \mathbf{W}_k \overline{\mathbf{F}}_{ik} \tag{13}$$

Note that equation 11 does not average the channel covariances to estimate the expectation of $\mathbf{y}_i$. Instead, the channel dependent first order statistics $\overline{\mathbf{F}}_{ik}$ are multiplied by the precision matrix of their corresponding channel $\mathbf{W}_k$ and, then, summed. That is different to other approaches like [10] where they use an averaged within class covariance. In theory, this model should do a robust estimation of the speaker identity variable when we have several i-vectors produced by different channels.

To estimate the parameters of the model we use this posterior to compute the expectations needed when maximizing the EM auxiliary function. We do ML and MD iterations.

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. For high dimensional data, it can be achieved by just normalizing the i-vectors by their magnitude.

$$\hat{\phi} = \frac{\phi}{\|\phi\|} \tag{14}$$

The results presented in [14] show that this technique boosts the performance of the PLDA.

The i-vectors need to be centered and whitened before the length normalization. Thus, the length 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.

There are several ways of doing centering and whitening. As the speaker variable $\mathbf{y}_i$ of the PLDA has a standard Gaussian prior, we do centering and whitening computing the expectation of $\mathbf{y}_i$ given the i-vector ϕ and the PLDA model. If we use, our multi-channel PLDA, we get channel dependent centering and whitening:

$$\hat{\mathbf{y}} = \left(\mathbf{I} + \sum_{k=1}^{K} P(z_k = 1) \mathbf{V}^T \mathbf{W}_k \mathbf{V} \right)^{-1} \sum_{k=1}^{K} P(z_k = 1) \mathbf{V}^T \mathbf{W}_k (\phi - \mu_k) \tag{15}$$

Note, that we use soft values for the channel posteriors $P(z_k = 1)$.

Thus, we train a MCSPLDA model on the no-normalized i-vectors and use it to do vector-wise dimensionality reduction. Then, we train another MCSPLDA on the length-normalized i-vectors and use it for classification.

5. EXPERIMENTS

5.1. Development and evaluation dataset

The dataset used in these experiments is part of the development dataset that we created for NIST SRE12. We evaluate our approach on the SRE10 extended dataset telephone-telephone (det5) with some additive noises added to the test segments, the enrollment segments are kept clean. To training the PLDA models we used data from NIST SRE04 to SRE08 and augmented with noise-corrupted versions. The UBM and i-vector extractor are trained with clean data only.

We considered three SNR levels: clean, 15dB and 6dB. We have used two types of noises: HVAC[1] and babble. Babble noises were created summing 1000 conversations from previous evaluations following NIST SRE12 guidelines. Different noise samples were added to the development and evaluation datasets. When adding the noise to the files, the power of the noise and speech signals was estimated using a psophometric filter and a VAD. The noise added to telephone segments is filtered by a simulated telephone channel.

5.2. Speaker recognition system configuration

As features, we used 20 short-time Gaussianized MFCC with deltas and double deltas. We trained diagonal covariance, gender dependent UBM with 2048 components with data from NIST SRE04 to SRE06 without noise added. We used an i-vector extractor of 600 dimensions trained on NIST SRE04 to SRE06 without noisy versions also.

We reduce the i-vector dimensionality to 400 using SPLDA or MCSPLDA. That has the side effect of centering and whitening the i-vectors. Then, we do length normalization of the i-vectors. Finally, we evaluate the likelihood ratio of the trials using SPLDA or MCSPLDA.

We compare three PLDA models: SPLDA trained on clean data only, SPLDA trained on pooled noisy and clean data and MCSPLDA considering three channel conditions (clean, 15dB and 6dB). The PLDA models used for dimensionality reduction and classification are matched.

We also show results for the fusion of SPLDA trained with the pool of all noisy data and the MCSPLDA. The fusion is performed in a k-fold fashion. The score matrices are divided into 50 blocks and each block is fused using a fusion function trained with the blocks that does not share any enrollment or test segment with it.

[1] We downloaded the noises from Freesound.org

5.3. Condition detection

To estimate the values of $P(z_k)$ needed by the MCSPLDA model, we use some quality measures $\mathbf{Q}$: Signal-to-noise-ratio, modulation index, spectral entropy and log-likelihood. Detailed explanation of how these measures are computed can be found in [15]

For each noise level we train a mixture of 8 Gaussians. Then $P(z_k)$ is computed as:

$$P(z_k = 1|\mathbf{Q}) = \frac{w_k P(\mathbf{Q}|z_k = 1)}{\sum_{k=1}^{K} w_k P(\mathbf{Q}|z_k = 1)} \quad (16)$$

where we choose $w_{clean} = w_{15dB} = w_{6dB} = 1/3$. We train these models using NIST SRE04 to SRE08 with noise added. The accuracy of this classifier for this dataset is around 99%.

5.4. Results

Table 1 shows results for the female part of NIST SRE10 extended det5 condition. We show EER, minDCF 2010 ($P_{\mathcal{T}} = 0.001$) and actDCF. Except for the fusion, that has the side effect of calibrating the scores, we do not apply any calibration to the scores. The actual costs are obtained with the scores straight from the PLDA.

For the clean test, the best configuration is the SPLDA trained with clean data. The systems trained with noise suffer some degradation. The SPLDA with pooled training is better in EER and the MCSPLDA is better in minDCF, but with small difference. It is interesting noting that the actDCF of the MCSPLDA is much better. It seems that the MCSPLDA produces likelihood ratios that are naturally better calibrated.

For the noisy tests, we obtained a great improvement of the systems trained with noise data over the system trained only with clean data. However, the difference between the SPLDA and the more complicated MCSPLDA is small. The SPLDA performs better in the EER part of the DET curve and the MCSPLDA performs better in DCF2010 part. Again, the MCSPLDA produces better calibrated likelihood ratios than the SPLDA. We can get a small gain from the fusion but not very significant.

We expected a better performance of the MCSPLDA model compared to the SPLDA trained with pooled noises. It is reasonable that using a dedicated within class covariance for each channel should be better than having only one trying to compensate all kinds of variability. We hypothesize that the difference between channel spaces is not so big so a unique channel matrix trained with more data can be more robust and perform well in multi-condition scenarios.

Table 1. *Female coreext det5 condition with noise added*

SNR	System	EER(%)	minDCF	actDCF
Clean	SPLDA clean	**2.65**	**0.46**	**0.50**
	SPLDA pool	2.90	0.55	0.90
	MCSPLDA	3.03	0.53	0.62
	Fusion	2.80	0.52	0.54
15dB	SPLDA clean	4.62	0.69	0.77
	SPLDA pool	**2.43**	0.58	0.75
	MCSPLDA	3.45	0.57	0.65
	Fusion	2.83	**0.56**	**0.60**
6dB	SPLDA clean	9.85	0.90	0.97
	SPLDA pool	4.77	0.74	0.76
	MCSPLDA	4.91	0.69	**0.71**
	Fusion	**4.48**	**0.67**	0.82

6. DISCUSSION

In this work, we addressed the problem of handling i-vectors generated in multiple channel conditions. For that, we have presented a PLDA variant, that we call multi-channel SPLDA (MCSPLDA), where the speaker space distribution is common to all types of channels and the channel space distribution is different. This model can be seen as a mixture of PLDA where the eigenvoices matrix $\mathbf{V}$ and the speaker factors $\mathbf{y}$ are shared across the components of the mixture.

We showed results on the telephone part of the NIST SRE10 extended condition where we added some noises to the test segments. We compared results using a standard SPLDA trained with pooled noisy and clean data, and our MCSPLDA. We have seen that both clearly improve the performance in noisy conditions compared to a SPLDA trained only with clean data. However, there are very small differences of EER and minDCF between the SPLDA trained with pooled data and the MCSPLDA. Something interesting is that the likelihood ratios from the MCSPLDA are better calibrated producing clearly lower actDCF.

In the future, we plan to go on researching on this kind of models. An interesting approach would be combining the strength of having a robust channel space estimated with more data, as happens in the pooled SPLDA, with the strength of having dedicated channels spaces. We can do that using a Bayesian approach where we put a prior on the channel space and the conditioned channel spaces are adapted from that prior.

7. ACKNOWLEDGMENT

This work has been supported by the Spanish Government and the European Union (FEDER) through projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

8. 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] 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.

[5] 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.

[6] Najim Dehak, Zahi N Karam, Douglas A Reynolds, William M Campbell, and James R Glass, "A Channel-Blind System for Speaker Verification," *Proc ICASSP 2011*, pp. 4536–4539, 2011.

[7] Mohammed Senoussaoui, Patrick Kenny, Pierre Dumouchel, and Fabio Castaldo, "Well-calibrated heavy tailed Bayesian speaker verification for microphone speech," 2011.

[8] Mitchell McLaren and David Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *2011 IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, Prague (Czech Republic), 2011, Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands, pp. 5456–5459, IEEE.

[9] Mitchell McLaren and David Van Leeuwen, "To Weight or not to Weight: Source-Normalised LDA for Speaker Recognition using i-vectors," in *Interspeech 2011*, Florence (Italy), 2011.

[10] Mitchell McLaren and David Van Leeuwen, "Source-normalised LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 755–766, 2011.

[11] Yun Lei, Lukas Burget, Luciana Ferrer, Martin Graciarena, and Nicolas Scheffer, "Towards Noise-Robust Speaker Recognition Using Probabilistic Linear Discriminant Analysis," in *International Conference on Acoustics, Speech and Signal Processing ICASSP 2012*, Kyoto (Japan), Mar. 2012, pp. 4253–4256.

[12] Konstantin Simonchik, Timur Pekhovsky, Andrey Shulipa, and Anton Afanasyev, "Supervized Mixture of PLDA Models for Cross-Channel Speaker Verification," in *Interspeech 2012*, Portland (USA), 2012.

[13] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson, "Multicondition Training of Gaussian PLDA Models in i-Vector Space for Noise and Reverberation Robust Speaker Recognition," in *International Conference on Acoustics, Speech and Signal Processing ICASSP 2012*, Kyoto (Japan), Mar. 2012, pp. 4257–4260.

[14] 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.

[15] Jesús Villalba, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel, "Reliability Estimation of the Speaker Verification Decisions Using Bayesian Networks to Combine Information from Multiple Speech Quality Measures," in *IberSpeech 2012*, Madrid (Spain), 2012.