Multi-session PLDA Scoring of I-vector for Partially Open-Set Speaker Detection

Kong Aik Lee, Anthony Larcher, Chang Huai You, Bin Ma, Haizhou Li

Multi-session PLDA Scoring of I-vector for Partially Open-Set Speaker Detection

Kong Aik Lee, Anthony Larcher, Chang Huai You, Bin Ma, Haizhou Li

Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
{kalee, alarcher, echyou, mabin, hli}@i2r.a-star.edu.sg

Abstract

This paper advocates the use of *probabilistic linear discriminant analysis* (PLDA) for partially open-set detection task with multiple i-vectors enrollment condition. Also referred to as speaker verification, the speaker detection task has always been considered under an open-set scenario. In this paper, a more general partially open-set speaker detection problem in considered, where the imposters might be one of the known speakers previously enrolled to the system. We show how this could be coped with by modifying the definition of the alternative hypothesis in the PLDA scoring function. We also look into the impact of the conditional-independent assumption as it was used to derive the PLDA scoring function with multiple training i-vectors. Experiments were conducted using the NIST 2012 Speaker Recognition Evaluation (SRE'12) datasets to validate various points discussed in the paper.

Index Terms: speaker verification, multi-session training.

1. Introduction

Probabilistic linear discriminative analysis (PLDA) [1] has shown to be an effective model for disentangling speaker and channel variability in the i-vector space for text-independent speaker verification [2]. An i-vector is a low-dimensional vector containing both speaker and session information acquired from a speech segment [3]. The unwanted session variability could be due to the transmission channel, acoustic environment or phonetic content of the speech segment itself [4].

In this paper, we look into two aspects in generalizing the use of PLDA to a more general setting, namely multi-session training and partially open-set detection problem. It is customary to assume that only one i-vector is available per speaker during enrolment [2, 5]. Given multiple training utterances, one could easily accumulate the statistics over all these utterances to end up with one single i-vector. Another alternative that has shown to be viable is by taking the mean before feeding the i-vectors to the PLDA that follows [6, 7]. Nevertheless, neither of these solutions is optimal. They are actually undesirable short-cuts so as to stick with the single-session solution without having to deal with the problems related to the multi-session PLDA scoring, notably, the conditional independent assumption used in the derivation.

Another issue that has been brought to our attention, partly due to the recent NIST Speaker Recognition Evaluation 2012 (SRE'12), is the use of PLDA for the so-called *partially open-set* speaker detection task. Speaker detection (or verification) task has always been considered in an open-set scenario [8, 4]. The imposters (i.e., those falsely claiming to be valid users) are assumed to be unknown to the system. A more general condition one might consider is that the imposters could be one of the known speakers previously enrolled to the system. This leads to the partially open-set speaker detection problem. On the application side, the partially open-set problem is of particular interest when the test segment consists of speech from multiple speakers intervening in the audio recording, for example, in a meeting where some participants are known with some out-of-set or unseen speakers[1]. Effective use of the joint knowledge of known speakers could lead to a significant performance improvement.

The aim of this paper is twofold. First, we present a concise formulation for the multi-session PLDA, in which the scoring function could take arbitrary number of i-vectors as inputs. More importantly, we look into the impact of the conditional-independence assumption used in the derivation. Second, we extend the PLDA scoring function for partially open-set speaker detection task, taking into account the conventional open-set and closed set conditions as special cases.

In the following, we first present a brief overview of i-vector and PLDA in Section 2. Section 3 presents the scoring function for multi-session PLDA. In Section 4, we look into the use of PLDA for partially open-set detection task. Section 5 is dedicated to experiments and Section 6 concludes the paper.

2. PLDA modeling of i-vector

2.1. I-vector extraction

The central idea of i-vector extraction is to represent variable-length utterances with fixed-length low-dimensional vectors for the classifiers that followed. The fundamental assumption is that the feature vector sequence of an utterance was generated from a session-specific GMM. Furthermore, the mean supervector (i.e., obtained by stacking the means from all mixtures) of each session, $\mathbf{m}_r$, is constrained to lie in a low dimensional subspace $\mathbf{T}$ with origin $\mathbf{m}$, as follows

$$\mathbf{m}_r = \mathbf{m} + \mathbf{T}\mathbf{x}_r \,. \tag{1}$$

The matrix $\mathbf{T}$, referred to as the total variability matrix, models the speaker and session variations learned from a training set. An i-vector is then taken as the posterior mean of the latent variable $\mathbf{x}_r$, representing both the speaker and session information of an utterance [3]. Notice that the rank of the matrix $\mathbf{T}$, and therefore the dimensionality of the i-vectors, is usually taken to be a small fraction of the original supervector.

2.2. Probabilistic linear discriminant analysis

In PLDA, speaker and session variability is modeled with separate subspaces in order to tease apart the contribution of session variability from that of the speaker. The fixed-length nature of the i-vector allows this to be done relatively easier than in the acoustic space.

---

[1] More precisely, here we are referring to the speaker tracking task (usually speaker segmentation followed by detection), which aims at localizing a particular speaker in an audio recording [9].

Let $\phi_{l,r}$ be an i-vector representing the *r*-th session of the *l*-th speaker. PLDA assumes that the i-vector is generated from a linear-Gaussian model [10], as follows

$$\phi_{l,r} = \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_l + \mathbf{G}\mathbf{w}_{l,r} + \boldsymbol{\varepsilon}_{l,r}. \quad (2)$$

Here, the low-rank rectangular matrices, **F** and **G**, model the subspaces corresponding to the speaker and session variability, respectively. The vectors $\mathbf{h}_l$ and $\mathbf{w}_{l,r}$ quantify the observed deviations from the mean $\boldsymbol{\mu}$ due the changes of speaker or different sessions of the same speaker. The remaining variation is described by the residual noise term $\boldsymbol{\varepsilon}_{l,r} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$.

Given a fixed set of parameters, θ_{PLDA} = {**F**, **G**, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$} we can see from (2) that an i-vector $\phi_{l,r}$ is determined by the speaker-specific vector $\mathbf{h}_l$ and the session specific vector $\mathbf{w}_{l,r}$, both assumed to be normally distributed. Notice also the same vector $\mathbf{h}_l$ is shared across all sessions from the same speaker. In probabilistic term, we write (2) as

$$p(\phi_{l,r}) = \mathcal{N}(\phi_{l,r} \mid \boldsymbol{\mu}, \mathbf{F}\mathbf{F}^T + \mathbf{G}\mathbf{G}^T + \boldsymbol{\Sigma}), \quad (3)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma} = \mathbf{F}\mathbf{F}^T + \mathbf{G}\mathbf{G}^T + \boldsymbol{\Sigma}$ denotes the global mean and covariance of all the i-vectors. Notice that the ranks of the matrices, **F** and **G**, are bounded by the dimensionality of the i-vector. The parameters θ_{PLDA} = {**F**, **G**, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$} of the PLDA model are estimated using the expectation maximization (EM) algorithm. Details about training procedure used in this paper could be found in [5].

3. Multi-session PLDA

The ultimate motivation of training the PLDA model is to use it for explaining new observations, i.e., the i-vector ϕ_r. For brevity, we have dropped the speaker-dependent index l. Figure 1 illustrates the idea in the form of graphical model. The number *R* of observed i-vectors $\{\phi_r\}_{r=1}^{R}$ are made dependent (indicated by the horizontal and downward arrows) on the latent variables $\mathbf{w}_r$, for *r* = 1, 2, …, *R*, each characterizing an individual session while sharing the same speaker-dependent latent variable $\mathbf{h}$.

The model in Fig. 1 explains a given set of i-vectors as if they belong to the same speaker. That is, all observations are tied to the same latent variable **h**, and they are conditionally independent given **h** [10]. The likelihood of the model can be computed by using the result in (3), with a slight twist, as follows:

$$p(\phi_1, \phi_2, \ldots, \phi_R \mid \theta_{\text{PLDA}}) = \mathcal{N}\left(\begin{bmatrix} \phi_1 \\ \vdots \\ \phi_R \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{bmatrix}, \Omega_R \Omega_R^{\mathrm{T}} + \mathbf{S}_R \right) \quad (4)$$

where

$$\Omega_R = \begin{bmatrix} \mathbf{F} & \mathbf{G} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{F} & 0 & \cdots & \mathbf{G} \end{bmatrix} \text{ and } \mathbf{S}_R = \begin{bmatrix} \boldsymbol{\Sigma} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\Sigma} \end{bmatrix}.$$

Taking the logarithm of (4) and breaking down the composite terms into component matrices, we represent (4) in a more convenient form, as follow

$$\log p(\phi_1, \phi_2, \ldots, \phi_R \mid \theta_{\text{PLDA}}) = \frac{1}{2}\left[\sum_{r=1}^{R} \mathbf{F}^{\mathrm{T}}\mathbf{J}\mathbf{y}_r\right]^{\mathrm{T}} \mathbf{K}_R \times \left[\sum_{r=1}^{R} \mathbf{F}^{\mathrm{T}}\mathbf{J}\mathbf{y}_r\right] + \frac{1}{2}\alpha(R) - \frac{1}{2}\sum_{r=1}^{R} \mathbf{y}_r^{\mathrm{T}}\mathbf{J}\mathbf{y}_r - Rc \quad (5)$$

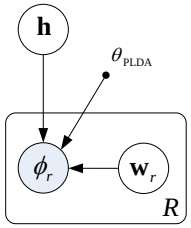

Figure 1: Graphical model illustrating the use of a PLDA model with parameter θ_{PLDA} = {**F**, **G**, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$} for explaining new observations. The shaded circles represent the i-vectors ϕ_r as the observed variables as opposed to the latent variables $\mathbf{h}$ and $\mathbf{w}_r$ used to represent the speaker identity and session variation. The box denotes *R* observations.

Here, $\mathbf{y}_r = (\phi_r - \boldsymbol{\mu})$ is the centralized i-vector of the *r*-th session, $\alpha(R) = \log|\mathbf{K}_R|$ is the matrix log-determinant depending on the number of sessions *R*, and $c = 0.5 \times (D\log(2\pi) - \log|\mathbf{J}|)$ is a scalar which holds constant for a given PLDA model. The two precision matrices involved are defined as follows

$$\mathbf{J} = \left[\mathbf{G}\mathbf{G}^{\mathrm{T}} + \Sigma\right]^{-1}$$
$$\mathbf{K}_R = \left[R\mathbf{F}^{\mathrm{T}}\mathbf{J}\mathbf{F} + I\right]^{-1} \quad (6)$$

We use the general form as given in (5) to derive the scoring function for speaker detection tasks in the next section. As we shall see, only the first and second terms are relevant for scoring, while the remaining terms will be canceled off. For the special case of $R = 1$, where only one i-vector is given, (5) reduces to the simple evaluation of (3) by taking its logarithm.

Both the PLDA training [5] and multi-session scoring are now available via the recent release of the open-source toolkit Alize 3.0 [11].

4. PLDA for partially open-set detection task

Speaker detection or verification is a binary classification problem, where a decision has to be made between two classes with respect to a decision threshold (i.e., a likelihood-ratio test). To this end, the detection score is taken as the log-likelihood ratio between two hypotheses $\{H_0, H_1\}$:

$$s(\phi_t) = \log\left[\frac{p(\phi_t \mid H_0)}{p(\phi_t \mid H_1)}\right]. \quad (7)$$

The null hypothesis H_0 says that a test segment, represented by the i-vector ϕ_t, is from the target speaker while the alternative H_1 hypothesizes the opposite. Mathematically, H_0 is represented by a model that characterizes the target speaker. In a completely open-set scenario, the alternative hypothesis represents any yet unseen *out-of-set* speakers (i.e., those other than the known speakers). This was conventionally achieved using a background model like the UBM [12].

The assumption that a known speaker might also appear as an imposter in other detection trials leads to a partially open-set speaker detection problem. From the modeling perspective, this essentially boils down to reformulating the alternative hypothesis. In addition to the unseen out-of-set speakers as already considered in the open-set case, H_1 now includes all known speakers except the one considered as the target in a specific trial. In the following, we show how this could be formulated using PLDA scoring model presented in Section 2.

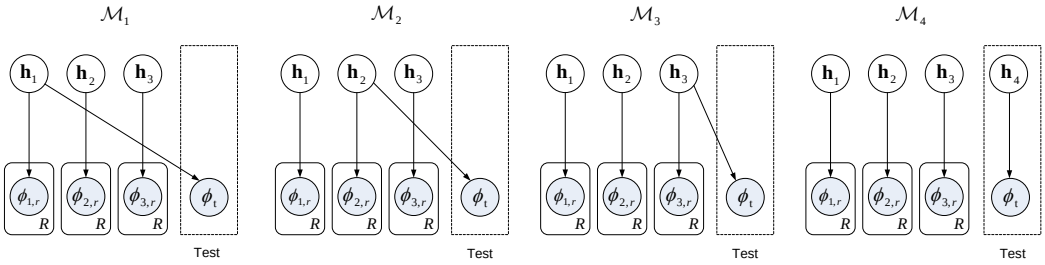

Figure 2: Model comparison in an open-set task illustrated for the case of $N=3$ speakers in the target set. For brevity, the session-specific variable $\mathbf{w}_{l,r}$ is not shown and each target speaker is assumed to have *R* enrollment sessions given by the i-vector $\{\phi_{l,r}\}_{r=1}^{R}$. Each model $\mathcal{M}_l$ represents a hypothesis about the identity of the test segment ϕ_t.

4.1. Score conditioning for partially open-set detection task

Let N be the number of target speakers. Considering an open set scenario, we dedicate one model for each target speaker and an additional model to represent the out-of-set option. This amounts to $(N+1)$ models as illustrated in Fig. 2 for the case of $N=3$. The model $\mathcal{M}_l$, where $l=1,2,\ldots,N$, says that the test i-vector ϕ_t is from the *l*-th speaker compared to other speakers in the target set. This is indicated by the arrows extending from the same speaker-specific latent variable $\mathbf{h}_l$ to the observations $\phi_{l,r}$ and ϕ_t, where $\{\phi_{l,r}\}_{r=1}^{R}$ are the i-vectors pertaining to training segments of the *l*-th speaker. The out-of-set model $\mathcal{M}_{N+1}$ represents the proposition that the test ϕ_t is generated by some yet unseen speakers other than the N target speakers known by the system. This is explained by the arrow between the test ϕ_t and a latent variable $\mathbf{h}_{N+1}$ representing the out-of-set option.

Let L_l be the likelihood of the model $\mathcal{M}_l$ (we shall deal with the likelihood computation in the next section). For the case of speaker identification, we simply pick the model with the highest likelihood. For the case of detection [13], we form the log-likelihood ratio between the null and alternative hypotheses, defined in (7), as follows:

$$s_l(\phi_t)=\log\left[\frac{L_l(\phi_t)}{(P_{\text{Known}})\frac{1}{N-1}\sum_{k\neq l}L_k(\phi_t)+(1-P_{\text{Known}})L_{N+1}(\phi_t)}\right]. \quad (8)$$

The likelihood of the null hypothesis in the numerator is given by L_l. In the denominator, the alternative hypothesis consists of two terms. The first term accounts for the joint knowledge of all known speakers other than the *l*-th target, while the second term is the likelihood of the out-of-set model. The probability P_{Known} controls the balance between these two terms. Clearly, (8) falls back to the open-set case by setting $P_{\text{Known}}=0$, while $P_{\text{Known}}=1$ leads to the closed-set scenario at the other end. Any value between these two extremes leads to the partially open-set detection task.

One thing to note in Fig. 2 and (8) is that PLDA allows the out-of-set class to be established in a systematic way by model comparison. We created $(N+1)$ models using training data from N target speakers. Except for the completely open-case case, where $P_{\text{Known}}=0$, the joint-knowledge from all models is used in forming the score for each trial.

4.2. Likelihood-ratio computation

Let $\Lambda_l=L_l/L_{N+1}$ be the likelihood ratio of the model $\mathcal{M}_l$ with respect to the out-of-set model $\mathcal{M}_{N+1}$. Looking at Fig. 3,

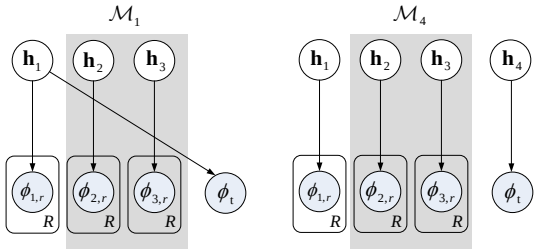

Figure 3: Likelihood ratio computation leads to a simpler implementation by which common components (with shaded background) in $\mathcal{M}_l$ cancel off those in $\mathcal{M}_{N+1}$. Shown above for $l=1$. Similar concept applies for $l=1,2,\ldots,N$ with respect to the common out-of-set model $\mathcal{M}_{N+1}$.

it could be seen that the computation of the likelihood ratio Λ_l is greatly simplified by cancelling off common terms. Notice that, the normalization also renders Λ_{N+1} equals to unity. Using these results in (8), we arrive at

$$s_l(\phi_t)=\log\left[\frac{\Lambda_l(\phi_t)}{(P_{\text{Known}})\frac{1}{N-1}\sum_{k\neq l}\Lambda_k(\phi_t)+(1-P_{\text{Known}})}\right]. \quad (9)$$

Given a test i-vector ϕ_t, we compute the likelihood ratio $\Lambda_l=L_l/L_{N+1}$ in log domain, as follows

$$\begin{aligned}\log\Lambda_l(\phi_t)&=\log p(\phi_t,\phi_{l,1},\ldots,\phi_{l,R}\,|\,\theta_{\text{PLDA}})\\&\quad-\log p(\phi_t\,|\,\theta_{\text{PLDA}})-\log p(\phi_{l,1},\ldots,\phi_{l,R}\,|\,\theta_{\text{PLDA}}).\end{aligned} \quad (10)$$

As before, each speaker is assumed to have R sessions of i-vectors, $\{\phi_{l,r}\}_{r=1}^{R}$, for enrollment. Here $R\geq 1$ could be different for individual speaker (more on this in Section 5). Notice also the first term in (10) corresponds to $\log L_l(\phi_t)$ while the remaining corresponds to $\log L_{N+1}(\phi_t)$.

The third term in (10) can be evaluated directly using (5). The first and second terms could be evaluated using the same formula by noticing that they are essentially cases with $R+1$ and 1 session(s), respectively. Using these, we arrive at

$$\begin{aligned}\log\Lambda_l(\phi_t)&=\frac{1}{2}\left[\sum_{r=1}^{R}\mathbf{y}_r+\mathbf{y}_t\right]^{\mathrm{T}}\mathbf{K}_{R+1}\left[\sum_{r=1}^{R}\mathbf{y}_r+\mathbf{y}_t\right]+\frac{1}{2}\alpha(R+1)\\&\quad-\frac{1}{2}\left[\sum_{r=1}^{R}\mathbf{y}_r\right]^{\mathrm{T}}\mathbf{K}_R\left[\sum_{r=1}^{R}\mathbf{y}_r\right]-\frac{1}{2}\alpha(R)\\&\quad-\frac{1}{2}\mathbf{y}_t^{\mathrm{T}}\mathbf{K}_1\mathbf{y}_t-\frac{1}{2}\alpha(1)\end{aligned} \quad (11)$$

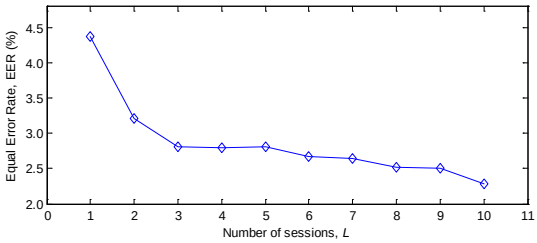

Figure 4: The EER reduces with increasing number of training i-vectors, *R*, from 1 to 10.

where the covariance matrix $\mathbf{K}_R$ and log-determinant $\alpha(R)$, as defined earlier, depend on the number of sessions *R*. For ease of notation, we have projected the centralized i-vector $\mathbf{y} = (\phi - \boldsymbol{\mu})$ such that $\mathbf{y} \leftarrow \mathbf{F}^{\mathrm{T}}\mathbf{J}\mathbf{y}$. Notice also, letting $R = 1$, (11) reduces to the special case of single-session training.

5. Experiments

Experiments were carried out on the core task of NIST SRE'12 using the equal error rate (EER) as the performance metric. We focus on *Common Condition 2* of the core task where the target speakers have number of training samples ranging from one to over hundred, while the test segments are telephone speech collected under relatively clean condition. We used gender-dependent setup, where the male and female UBMs consisting of 512 Gaussians (with full covariance matrices) were trained using data drawn from the SRE'04 dataset. Speech parameters used are 57-dimensional vector of *mel frequency cepstral coefficients* (MFCC) with first and second derivatives appended. The first-order sufficient statistics were whiten with respect to the covariance matrices of the UBM. Details could be found in [14]. The total variability matrix **T** in (1) consists of two subspaces, $\mathbf{T}_{\text{tel}}$ and $\mathbf{T}_{\text{mic}}$, trained in a decoupled manner, as described in [15]. The ranks of the matrices, $\mathbf{T}_{\text{tel}}$ and $\mathbf{T}_{\text{mic}}$, are 400 and 200, respectively. LDA was then used to reduce the dimension of the i-vector to 400. The ranks of the subspaces **F** and **G** are, 250 and 50, respectively.

First, we examine the effectiveness of the multi-session scoring function in (11). To this end, we change the number of training sessions (or i-vectors), *R*, for all speakers from 1 to 10 and test on the same data. Results are shown in Fig. 4. Progressive reduction in the EER can be observed with more training sessions used (i.e., *R* increases). This result confirms that the multi-session scoring function in (11) is valid. Figure 5 shows the distributions of the target and non-target scores for $R = 1$ and 4 sessions. Notice that, the score exhibits a larger range with larger *R*. This would be fine if all speakers are enrolled with the same number of sessions. However, for the case whereby target speakers are trained with different number of training sessions, such mismatch would cause inconsistency between the score produced by the speaker models. This by far is believed to attribute to the incorrect assumption of conditional independence as used in Section 3.

One solution is to restrict the same number of training sessions for all speakers. One could also take the mean of all i-vector and set *R* to 1 or the average number of sessions of all speakers. Another solution is to use score normalization (for instance, s-norm [2]). Here, we could consider tying speakers with the same number of sessions to have the same normalization. Nevertheless, the ultimate goal is to find the

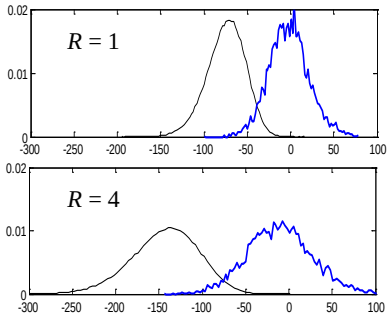

Figure 5: Distribution of target (right) and non-target (left) scores for $R = 1$ and 4 sessions.

Table I: The effect of proper conditioning of detection scores for open-set, $P_{\text{known}} = 0$, partially open-set, $P_{\text{known}} \in \{1/4,\ 1/2,\ 3/4\}$, and closed-set, $P_{\text{known}} = 1$, detection task.

P_{known}	Equal Error Rate (%)	
	w/o cond.	with cond.
0	2.5704	2.5704
1/4	2.5972	2.4340
1/2	2.6244	2.2981
3/4	2.6910	2.1757
1	2.7468	1.5774

right compensation factor (which obviously depends on *R*) for the multi-session scoring function in (11). These are some points for future research.

Next, we examine the score conditioning for partially open-set detection task. We follow the core condition as specified in SRE'12, where the "known" and "unknown" non-target score distributions are weighted according to P_{known}. Table I shows the EER at various values of P_{known}. One point to note here is that the known non-target set is more difficult than the unknown non-target set. This can be seen when P_{known} increases from 0 to 1, false alarm rate and therefore the EER increase as higher weight is given to the known non-target set. From the last column of Table I, it is clear that significant improvement could be obtained by using the prior information as given by P_{known} in score conditioning for partially open-set detection task. This amounts to 43% of relative improvement given $P_{\text{known}} = 1$.

6. Conclusions

We have shown that PLDA model can be systematically used to form the log-likelihood ratio for partially open-set and closed-set detection task. In this regard, open-set detection scores are conditioned based on the prior probability and the joint information of the known non-targets leading to significant performance improvement. We also look into various aspects when multiple sessions are available for speaker enrollment. The remaining challenge is to find a suitable compensation factor for the multi-session scoring function which balances out the complexity from speakers with different number of enrollment sessions.

7. Acknowledgements

The proposal presented in this paper was partly inspired by the discussion during the NIST SRE'12 workshop in Orlando.

8. References

[1] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. International Conference on Computer Vision*, 2007.

[2] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in *Proc. Odyssey: Speaker and Language Recognition Workshop*, Jun. 2010.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, May 2011.

[4] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: from features to supervectors,” Speech Communication, vol. 52, no. 1, pp. 12-40, Jan. 2010.

[5] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, “PLDA modeling in i-vector and supervector space for speaker verification,” in *Proc. INTERSPEECH*, 2012, paper 198.

[6] H. Li, B. Ma, K. A. Lee, C. H. You, H. Sun, and A. Larcher, “IIR system description for the NIST 2012 speaker recognition evaluation,” in *NIST SRE'12 Workshop*, Orlando, Dec. 2012.

[7] N. Brümmer, A. Swart, L Burget, S. Cumani, O. Glembek, M. Karafiát, P.Matejka, O.Plchot, M. Soufifar, J. Silovský, P. Kenny, J. Alam, P. Dumouchel, P. Ouellet, M. Senoussaoui and T. Stafylakis, “ABC System description for NIST SRE 2012,” in *NIST SRE'12 Workshop*, Orlando, Dec. 2012.

[8] D. A. Reynolds, “An overview of automatic speaker recognition technology,” in *Proc. ICASSP*, 2002, pp. IV-4072 – IV-4075.

[9] F. Bimbot, “Automatic speaker recognition,” in Language and Speech Processing, J Mariani, Ed., Wiley, 2009.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] A. Larcher, J. -F. Bonastre, B. Fauve, K. A. Lee, C. Levy, H. Li, J. S. D. Mason, J. -Y. Parfait, “ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition,” submitted, *INTERSPEECH*, 2013.

[12] D.A. Reynolds, T.F. Quatieri, and R.B. Dumn, “Speaker verification using adapted Gaussian mixture model,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

[13] N. Brummer and D. Leeuwen, “On calibration of language recognition scores,” in *Proc. IEEE Odyssey: Speaker Lang. Recognition Workshop*, 2006, pp. 1-8.

[14] P. Kenny, “A small foot-print i-vector extractor,” in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 1- 6.

[15] M. Senoussaoui, P. Kenny, N. Dehak, P. Dumouchel, “An i-vector extractor suitable for speaker recognition with both microphone and telephone speech,” in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010, pp. 28- 3.