Digital Signal Processing 31 (2014) 93–101

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification

Padmanabhan Rajan [a,b,*,1], Anton Afanasyev [a], Ville Hautamäki [a], Tomi Kinnunen [a]

[a] Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Joensuu, Finland
[b] School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Himachal Pradesh, India

ARTICLE INFO

Article history:
Available online 9 May 2014

Keywords:
i-vector
Probabilistic linear discriminant analysis
Multiple enrollment
Speaker verification

ABSTRACT

The availability of multiple utterances (and hence, i-vectors) for speaker enrollment brings up several alternatives for their utilization with probabilistic linear discriminant analysis (PLDA). This paper provides an overview of their effective utilization, from a practical viewpoint. We derive expressions for the evaluation of the likelihood ratio for the multi-enrollment case, with details on the computation of the required matrix inversions and determinants. The performance of five different scoring methods, and the effect of i-vector length normalization is compared experimentally. We conclude that length normalization is a useful technique for all but one of the scoring methods considered, and averaging i-vectors is the most effective out of the methods compared. We also study the application of multicondition training on the PLDA model. Our experiments indicate that multicondition training is more effective in estimating PLDA hyperparameters than it is for likelihood computation. Finally, we look at the effect of the configuration of the enrollment data on PLDA scoring, studying the properties of conditional dependence and number-of-enrollment-utterances per target speaker. Our experiments indicate that these properties affect the performance of the PLDA model. These results further support the conclusion that i-vector averaging is a simple and effective way to process multiple enrollment utterances.

1. Introduction

The *i-vector* representation [1] followed by *probabilistic linear discriminant analysis* (PLDA) [2] has become state-of-the-art in speaker verification systems over the past few years. In a typical speaker verification trial, there are two i-vectors; one represents the enrollment utterance of a given speaker, and the other a test utterance. When speech utterances are represented as i-vectors, the speaker verification problem is simply to determine if the i-vectors share the same speaker information or not.

The most recent 2012 NIST speaker recognition evaluation (SRE) allows for *multiple* enrollment utterances (and hence i-vectors) for a given target speaker. In principle, the availability of more enrollment data can help in enhancing system performance, but it is not obvious how this can be achieved in practice. Although multiple i-vectors can be integrated directly into the PLDA model [3], approximate methods like i-vector averaging have been shown to be effective [4]. The PLDA model assumes statistical independence among enrollment i-vectors, which may be difficult to achieve in practice. Enrollment i-vectors from a given target speaker might share common attributes like acoustic content, transmission channel etc., thus invalidating the independence assumption. Scoring methods which do not have the independence assumption are often more effective in dealing with multiple enrollment i-vectors. Multiple enrollment utterances also occur in the context of *multicondition training*, which has been successful in improving noise robustness of both classical [5] and modern speaker recognition systems [6].

Previous studies on the i-vector PLDA system have investigated the effect of utterance duration [7] and mismatched duration [8,9]. The effect of using multiple speech sources (including telephone, interview and microphone speech) has been studied in [10]. Most of these studies have looked at the effect of variations beginning with the estimation of i-vector hyperparameters. On the other hand, in this paper, our focus is solely on the enrollment stage and generative model represented by PLDA; the i-vector hyperparameters are left unchanged.

* Corresponding author at: School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Himachal Pradesh, India.
E-mail addresses: padman@iitmandi.ac.in (P. Rajan), aafanasy@cs.uef.fi (A. Afanasyev), villeh@cs.uef.fi (V. Hautamäki), tkinnu@cs.uef.fi (T. Kinnunen).

[1] This work was done when Padmanabhan Rajan was a postdoctoral researcher at the University of Eastern Finland.

http://dx.doi.org/10.1016/j.dsp.2014.05.001
1051-2004/

In [11], the authors propose a multi-channel version of PLDA with channel-specific generative model for i-vectors. Varying utterance duration was compensated for by calibrating the PLDA score in [12], and by exploiting the uncertainty in the i-vector in [13]. The effect of noise on PLDA-based systems is studied in [14]. The experimental protocol in most of the above mentioned works involved a single i-vector for enrollment. Other studies have looked at explicitly incorporating multiple enrollment i-vectors into the PLDA model (the so-called 'by-the-book' scoring or multi-session scoring). The study [15] looked at multi-session scoring in with a partially open set speaker population, in the context of the NIST 2012 SRE. Further, [16] incorporated utterance duration as observation noise in a supervector generative model, leading to an investigation of different scoring methods using multiple enrollment utterances. Performance of i-vector averaging and multi-session scoring is also studied in [4], whereas [17] includes a comparison of score-averaging and multi-session scoring. In the context of face recognition, a scalable formulation of PLDA is described in detail in [18].

Despite the recent advances in PLDA-based speaker verification, an insightful survey of the basic scoring techniques is missing. The present study, targeted for practitioners, is intended to be a self-contained tutorial for PLDA scoring involving multiple enrollment utterances. It extends our preliminary study in [19]. The current study involves three major contributions. Firstly, we elaborate on the mathematical details involved in the practical computation of the determinants and inverses of the large matrices required by PLDA scoring. In particular, we concentrate our effort on a simplified version of PLDA, which is described in [20,21,11]. Our second contribution is a detailed experimental comparison of five straightforward scoring strategies. Three of them – multi-session scoring, i-vector averaging and score averaging, have been reported elsewhere but not compared within a single study. The two remaining methods, are maximum score and pooled-sessions scoring. We compare each scoring variant with and without i-vector length normalization [21] and provide a recommended choice for practitioners. Furthermore, we restrict our focus to PLDA scoring variants that require only the enrollment and test i-vectors. Alternatively, more advanced techniques may utilize either *a priori* known or estimated channel labels (for example, see [11]) to tackle the conditional independence assumption of standard PLDA scoring. As such, these techniques require either supplementary metadata or estimated channel/microphone labels produced by another classifier, leading to more complex design with increased computations or added human effort.

We also address the question regarding multicondition training: should multicondition training be applied to the enrollment i-vectors, the PLDA hyperparameter training, or both? The last and most interesting contribution, extending a previously noted problem of PLDA scoring dependency on the number of enrollment utterances [15,4], proposes to partially overcome the conditional independence assumption of PLDA. We address scoring in the situation when the number of enrollment utterances is not fixed but a random variable, providing new insights on the preferred ways to prepare enrollment i-vectors.

The topics studied in this paper are illustrated in Fig. 1. The experiments are carried out on two up-to-date sets of data: a subset of the I4U consortium dataset [22] and a subset of the NIST 2012 SRE data.

2. i-vector representation

In this section, we give an overview of the i-vector PLDA system utilized for the studies in this paper.

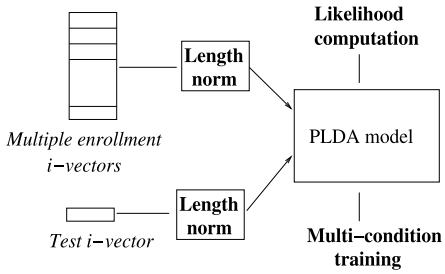

Fig. 1. Figure representing i-vector PLDA representation. Experimental studies in this paper are indicated in **boldface.**

2.1. The i-vector representation

The i-vector representation [1] is a fixed-length representation of speech utterances, which usually consist of variable number of acoustic feature vectors. Given an $FM \times 1$ supervector of means $\boldsymbol{\mu}$ from a universal background model (UBM), a speaker and recording specific supervector $\mathbf{s}$ is assumed to be of the form

$$\mathbf{s} = \boldsymbol{\mu} + \mathbf{Tw}. \tag{1}$$

Here, the acoustic feature vector is F-dimensional, the UBM has M components, $\mathbf{T}$ is an $FM \times D$ low-rank matrix whose columns span the major variability in the supervector space, and $\mathbf{w}$ is a $D \times 1$ dimensional latent vector with a standard normal distribution; i.e. $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The i-vector representation of an utterance is defined as the mean of the posterior distribution of $\mathbf{w}$, given the utterance. To estimate the i-vector, cepstral coefficients extracted from the speech utterance are represented in terms of zero- and first-order Baum–Welch statistics, with respect to the UBM. $\mathbf{T}$ is the i-vector extractor, and the resulting i-vectors are of much lower dimension (typically between 400 and 600) than the supervector. The UBM and the i-vector extractor are estimated from appropriate training corpora. Methods to train the i-vector extractor and estimate the i-vectors can be found in [1,23].

3. PLDA model

Originally applied to face recognition [2], PLDA has been applied successfully to specify a generative model of the i-vector representation [20]. For the ith speaker, the i-vector $\mathbf{w}_{i,j}$ representing the jth recording can be represented as,

$$\mathbf{w}_{i,j} = \mathbf{m} + \mathbf{Sx}_i + \mathbf{Gy}_{i,j} + \boldsymbol{\epsilon}_{i,j}. \tag{2}$$

Here, $\mathbf{m} + \mathbf{Sx}_i$ is the speaker-dependent part, and $\mathbf{Gy}_{i,j} + \boldsymbol{\epsilon}_{i,j}$ is the recording-dependent part. $\mathbf{m}$ is a global offset, $\mathbf{S}$ is a set of basis vectors for the speaker subspace, representing *between-speaker* variability, and $\mathbf{G}$ is a set of basis vectors representing the channel subspace, representing *within-speaker* variability. The remaining residual variability is represented by $\boldsymbol{\epsilon}_{i,j}$. The latent variables $\mathbf{x}$ and $\mathbf{y}$ are assumed to have standard normal distributions, and respectively represent a particular speaker and channel. The residual term $\boldsymbol{\epsilon}$ is assumed to have a normal distribution with a diagonal covariance matrix.

In this paper, we focus on a simplified variant of PLDA [20], termed as either *Gaussian PLDA* [21] or *simplified PLDA* [11]. Here, the within-speaker variability is modeled by a full-covariance residual term, which allows us to omit the channel subspace. The generative model for the i-vector is now represented by

$$\mathbf{w}_{i,j} = \mathbf{m} + \mathbf{Sx}_i + \boldsymbol{\epsilon}_{i,j}. \tag{3}$$

The residual term $\boldsymbol{\epsilon}$ representing the within-speaker variability is assumed to have a normal distribution with full covariance matrix

$\boldsymbol{\Sigma}$. A special case of the simplified PLDA model where the speaker factors $\mathbf{S}$ is full-rank is termed as the two-covariance model in [24,25].

3.1. Length normalization

Although the PLDA model assumes Gaussian behavior, there is empirical evidence that channel- and speaker- effects result in i-vectors that are non-Gaussian. By replacing the Gaussian assumptions of the PLDA model with a Student's t-distribution, improved performance was obtained in [20]. Since these are more complicated to apply in practice, a straightforward non-linear transformation of the i-vectors was proposed in [21]. This involves whitening the i-vectors followed by normalizing their length. This technique, called *radial Gaussianisation*, restores the Gaussian assumptions of the PLDA model, and is a popular pre-processing step.[2] It is believed that session variability affects only the i-vector length, and hence length normalization improves robustness to session effects.

4. Likelihood computation

We next examine various scoring strategies for utilizing the PLDA model to get a likelihood ratio for a given speaker verification trial.

4.1. Two i-vector scoring

Given two i-vectors $\mathbf{w}_1$ (for enrollment) and $\mathbf{w}_t$ (for test), the PLDA framework forms the verification score $s_{\text{lin}}(\mathbf{w}_1, \mathbf{w}_t)$ by determining the likelihood ratio given by,

$$s_{\text{lin}}(\mathbf{w}_1, \mathbf{w}_t) = \frac{p(\mathbf{w}_1, \mathbf{w}_t|H_1)}{p(\mathbf{w}_1|H_0)p(\mathbf{w}_t|H_0)}. \tag{4}$$

Here, the hypothesis H_1 indicates that both i-vectors come from the same speaker (and hence have the same speaker identity variable $\mathbf{x}$ in Eq. (3)), and H_0 indicates they come from different speakers (and hence have independently drawn $\mathbf{x}$). Given the Gaussian assumptions above, and following [3], the log likelihood ratio can be computed in closed form as,

$$\begin{aligned} s_{\log}(\mathbf{w}_1, \mathbf{w}_t) = {} & \log \mathcal{N}\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_t\end{bmatrix}; \begin{bmatrix}\mathbf{m}\\ \mathbf{m}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T\\ \mathbf{S}\mathbf{S}^T & \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\end{bmatrix}\right) \\ & - \log \mathcal{N}\left(\mathbf{w}_1; \mathbf{m}, \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\right) \\ & - \log \mathcal{N}\left(\mathbf{w}_t; \mathbf{m}, \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\right). \end{aligned} \tag{5}$$

After straightforward algebra, this turns out to be,

$$\begin{aligned} s_{\log}(\mathbf{w}_1, \mathbf{w}_t) = {} & \begin{bmatrix}\mathbf{w}_1^T & \mathbf{w}_t^T\end{bmatrix} \begin{bmatrix}\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T\\ \mathbf{S}\mathbf{S}^T & \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{w}_1 & \mathbf{w}_t\end{bmatrix} \\ & - \mathbf{w}_1^T\left[\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\right]^{-1}\mathbf{w}_1 - \mathbf{w}_t^T\left[\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\right]^{-1}\mathbf{w}_t \\ & + C, \end{aligned} \tag{6}$$

where all the constant terms have been incorporated into C, and can be omitted for a given PLDA model.

4.2. Multi-session scoring

Eq. (6) gives a scoring formula to compare two i-vectors. When *multiple* i-vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ are available for enrollment, and $\mathbf{w}_t$ is the test i-vector, the scoring function can be generalized as follows:

$$s_{\text{lin}}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N, \mathbf{w}_t) = \frac{p(\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_N, \mathbf{w}_t|H_1)}{p(\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_N|H_0)p(\mathbf{w}_t|H_0)}. \tag{7}$$

As earlier, the hypothesis H_1 represents the sharing of the same speaker variable between all the i-vectors, and H_0 represents the test i-vector having an independently drawn speaker variable. To evaluate this expression, we first form an expression for the likelihood when i-vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ share the same speaker variable $\mathbf{x}$, by extending the case in [3]. We can write

$$\begin{aligned} & p\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ \vdots\\ \mathbf{w}_N\end{bmatrix}\right) \\ & = \mathcal{N}\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ \vdots\\ \mathbf{w}_N\end{bmatrix}; \begin{bmatrix}\mathbf{m}\\ \mathbf{m}\\ \vdots\\ \mathbf{m},\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T & \ldots & \mathbf{S}\mathbf{S}^T\\ \mathbf{S}\mathbf{S}^T & \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T & \ldots & \mathbf{S}\mathbf{S}^T\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T & \ldots & \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\end{bmatrix}\right) \end{aligned} \tag{8}$$

Computing the likelihood in Eq. (8) requires inverting and computing the determinant of an $N \times N$ block matrix. Subtracting the common mean $\mathbf{m}$ from all enroll and test i-vectors, and utilizing the lemmas in Appendix A, we can write the log likelihood from Eq. (8) as,

$$\begin{aligned} \log p\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ \vdots\\ \mathbf{w}_N\end{bmatrix}\right) = {} & \frac{-ND}{2}\log(2\pi) - \frac{N-1}{2}\log|\boldsymbol{\Sigma}| \\ & - \frac{1}{2}\log\left|\boldsymbol{\Sigma}+N\mathbf{S}\mathbf{S}^T\right| - \frac{1}{2}\sum_{i=1}^{N}\mathbf{w}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{w}_i \\ & - \frac{1}{2}\left(\sum_{i=1}^{N}\mathbf{w}_i\right)^T\mathbf{K}_N\left(\sum_{i=1}^{N}\mathbf{w}_i\right), \end{aligned} \tag{9}$$

where

$$\mathbf{K}_N = -\left(\boldsymbol{\Sigma}+N\mathbf{S}\mathbf{S}^T\right)^{-1}\mathbf{S}\mathbf{S}^T\boldsymbol{\Sigma}^{-1}.$$

Noting that the numerator in Eq. (7) shares the same speaker variable $\mathbf{x}$ for both the enroll and test i-vectors, and applying Eq. (8), we can write the log likelihood ratio for the multi-session case as

$$\begin{aligned} & s_{\log}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N, \mathbf{w}_t) \\ & = \log \mathcal{N}\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ \vdots\\ \mathbf{w}_N\\ \mathbf{w}_t\end{bmatrix}; \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T & \ldots & \mathbf{S}\mathbf{S}^T\\ \mathbf{S}\mathbf{S}^T & \Sigma+\mathbf{S}\mathbf{S}^T & \ldots & \mathbf{S}\mathbf{S}^T\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T & \ldots & \mathbf{S}\mathbf{S}^T\\ \mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T & \ldots & \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\end{bmatrix}\right) \end{aligned}$$

[2] Although length normalization is one of the steps of the radial Gaussianisation process, the latter is popularly called just 'length normalization'.

$$
\begin{aligned}
&-\log\mathcal{N}\left(\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ \vdots\\ \mathbf{w}_N\end{bmatrix};\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0},\end{bmatrix},\right.\\
&\qquad\left.\begin{bmatrix}\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T & \dots & \mathbf{S}\mathbf{S}^T\\ \mathbf{S}\mathbf{S}^T & \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T & \dots & \mathbf{S}\mathbf{S}^T\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{S}\mathbf{S}^T & \mathbf{S}\mathbf{S}^T & \dots & \boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\end{bmatrix}\right)\\
&-\log\mathcal{N}\left(\mathbf{w}_t;\mathbf{0},\boldsymbol{\Sigma}+\mathbf{S}\mathbf{S}^T\right) \qquad (10)
\end{aligned}
$$

Eq. (10) can be evaluated by applying Eq. (9) with $N+1$, N and 1 i-vectors respectively. Additionally, terms that do not depend on the enroll and test i-vectors can be removed. This yields the following simplified expression for the log likelihood ratio:

$$
\begin{aligned}
&s_{\log}(\mathbf{w}_1,\mathbf{w}_2,\dots,\mathbf{w}_N,\mathbf{w}_t)\\
&=-\delta(N+1)-\left(\sum_{i=1}^{N}\mathbf{w}_i+\mathbf{w}_t\right)^T\mathbf{K}_{N+1}\left(\sum_{i=1}^{N}\mathbf{w}_i+\mathbf{w}_t\right)\\
&\quad+\delta(N)+\left(\sum_{i=1}^{N}\mathbf{w}_i\right)^T\mathbf{K}_N\left(\sum_{i=1}^{N}\mathbf{w}_i\right)\\
&\quad+\delta(1)+\frac{1}{2}\mathbf{w}_t^T\mathbf{K}_1\mathbf{w}_t+C, \qquad (11)
\end{aligned}
$$

where we denoted the log determinant as

$$\delta(N)=\log\left|\boldsymbol{\Sigma}+N\mathbf{S}\mathbf{S}^T\right|$$

and C represents the remaining constant terms that can be omitted. It can be shown that Eq. (6) is a special case of Eq. (11) with $N=1$.

4.3. Alternatives to multi-session scoring

The likelihood ratio score for multiple enrollment i-vectors, as defined mathematically (Eq. (7)), assumes the enrollment i-vectors to be statistically independent, given the speaker identity. The independence assumption is for mathematical convenience, rather than reflecting physical reality. For instance, different i-vectors obtained from the same target speaker might have more in common than just the speaker identity (for instance, acoustic environment or transmission channel). In general, i-vectors derived from human speech signals cannot be considered truly statistically independent. This is the reason why the multi-session likelihood ratio computation will be sub-optimal in a practical setting. As a result, other heuristic scoring methods are used for handling multiple enrollment i-vectors.

A popular scheme is *i-vector averaging*, in which a single i-vector is obtained as the average of the enrollment i-vectors,

$$\mathbf{w}_{\text{avg}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i, \qquad (12)$$

and then scored using the two i-vector scoring $s_{\log}(\mathbf{w}_{\text{avg}},\mathbf{w}_t)$ described in Eq. (5). Another alternative is to use score fusion, in which the individual enrollment i-vectors are scored using the two i-vector scoring and then combined. One method for this is to use *score averaging*, defined as

$$s_{\text{avg}}=\frac{1}{N}\sum_{i=1}^{N}s_{\log}(\mathbf{w}_i,\mathbf{w}_t), \qquad (13)$$

and another is to use *max-scoring*, which is defined as

$$s_{\max}=\max_{1\le i\le N}s_{\log}(\mathbf{w}_i,\mathbf{w}_t). \qquad (14)$$

In contrast to utilizing multiple enrollment i-vectors, a single enrollment i-vector can be obtained from multiple enrollment utterances by *pooling sessions*. This is done by pooling acoustic feature vectors from all the utterances and estimating zeroth and first order Baum–Welch statistics. A single enrollment i-vector then is obtained as if only a single enrollment utterance was available, and scored using the two i-vector scoring.

The above methods provide alternatives to the multi-session scoring described in Eq. (7). We evaluate each of these methods later in the paper.

4.4. Computational complexity

The computational complexities of each of the scoring methods vary considerably. For the methods utilizing a single i-vector for enrollment (including i-vector averaging and pooling sessions), three matrix-vector products need to be computed. This assumes that the matrix inversion in Eq. (6) is constant across trials and can thus be pre-computed, giving a complexity of $O(D^2)$, where D is the i-vector dimension. For the score fusion, the two i-vector scoring needs to be repeated N times (N is the number of enrollment i-vectors in the given trial), giving a complexity of $O(ND^2)$. For the multi-session likelihood computation in Eq. (11), the log determinants and matrix inversions depend on N. Hence, in general, multi-session likelihood computation has a complexity of $O(D^3)$, representing the matrix inversion. If the number of enrollment utterances are known in advance, the inverses and determinants can be precomputed for each N, giving a complexity of $O(ND^2)$.

5. Experiments

5.1. Corpora for experiments

As part of the pre-evaluation activity for the NIST SRE 2012, the I4U consortium[3] developed a dataset based on previous years' NIST corpora. The EvalSet portion of the I4U dataset consists of data drawn from the SRE 2006, 2008 and 2010 corpora. The data has multiple channels and speaking styles, including telephone, microphone and interview data, as determined from the keys released by NIST. In addition to the utterances used as-such from these corpora (henceforth termed *original utterances*), noisy versions of each utterance were generated using FaNT.[4] For each utterance, two noisy versions at 6 dB and 15 dB signal-to-noise ratio (SNR) were generated using HVAC (heating, ventilation and air-conditioning) and crowd noises. Thus, the data has three distinct SNR levels. The number of enrollment utterances for target speakers varies from 3 to 108, with an average of 19 per speaker. More details about the I4U dataset is provided in [22].

We perform the task of 'speaker detection', as described by NIST in the evaluations prior to the year 2012 (see [26]). In the experiments in the later part of the paper, performance is reported in terms of the equal error rate (EER) and the normalized detection cost function (DCF) given as

$$\text{DCF}=C_{\text{Det}}/C_{\text{Default}},$$

where

$$C_{\text{Default}}=\min\left\{C_{\text{Miss}}\times P_{\text{Target}},C_{\text{FalseAlarm}}\times(1-P_{\text{Target}})\right\},$$

[3] The I4U consortium consists of nine universities and research institutes.

[4] FaNT – Filtering and Noise Adding Tool. Available: http://dnt.kr.hsnr.de/download.html.

Table 1
Summary of the data used for experiments, which is derived from the I4U EvalSet and SRE 2012 data.

	Male	Female
Enroll/eval I4U subset		
Num. target speakers	381	577
Num. enroll segments	15,057	21,903
Num. test speakers	302	459
Num. test segments	7,926	10,524
Num. target trials	7,926	10,524
Num. non-target trials	3,011,880	6,061,824
SRE12 condition 4 subset		
Num. target speakers	381	577
Num. target trials	1,497	2,580
Num. non-target trials	60,930	145,086
PLDA training (no common speakers with the above)		
Num. speakers	382	578
Num. training utt.	29,961	43,119

and

$$C_{\text{Det}} = C_{\text{Miss}} \times P_{\text{Miss|Target}} P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{\text{FalseAlarm|NonTarget}} \times (1 - P_{\text{Target}}),$$

and the parameters $C_{\text{Miss}} = C_{\text{FalseAlarm}} = 1$ and $P_{\text{Target}} = 0.001$. This gives $C_{\text{Default}} = 0.001$.

To adhere to the NIST protocol, it is ensured that the data utilized for the PLDA training has no common speakers with the enrollment and evaluation data. A portion of the I4U EvalSet, consisting of 382 male speakers and 578 female speakers is used to train the PLDA model. Data from the remaining speakers are used for enrollment and evaluation. A 'full-matrix' of scores (all evaluation segments against all enrollment speakers) is used to compute the metrics, and the statistics are summarized in Table 1.

Results are also reported on the noisy telephone condition (common condition 4) of the NIST 2012 data, using the SRE 2010 evaluation metrics as above. Evaluation is restricted to SRE 2012 trials involving claims from speakers appearing in the I4U enrollment data. This results in evaluation of a subset of the SRE 2012 corpus. Filelists of the data used for the experiments have been shared online.[5]

5.2. System description

The i-vector PLDA system used for our studies uses a standard Mel frequency cepstral coefficient (MFCC) front-end with 30 ms frame size and 15 ms shift. The MFCCs were obtained using a 27-channel mel-frequency filterbank followed by RASTA filtering, adding delta and double deltas, frame dropping using SAD [27] and utterance level cepstral mean and variance normalization (CMVN), in this order. The 1024-mixture UBM is trained with data from NIST 2004, 2005, 2006 SRE, whereas the i-vector extractor from NIST 2004, 2005, 2006, Fisher and Switchboard data. The i-vector dimension D is 600, with a gender-dependent UBM and i-vector extractor.

5.3. Effect of multicondition training

Multicondition training [5,6] is a popular method to enhance noise robustness in speaker verification systems. In a multicondition setup, multiple noisy versions of the training data are available.

Each utterance in the I4U dataset has two noisy versions, at 6 dB and 15 dB SNR. The multiple versions of the enrollment utterances can be used for multicondition training during PLDA hyperparameter estimation, during likelihood evaluation, or both. These cases are evaluated in Table 2. Following [6], *pooled multicondition training* is done to estimate the PLDA hyperparameters. Thus, this model assumes that all of the N enrollment i-vectors $\mathbf{w}_1, \ldots, \mathbf{w}_N$ for a given speaker are generated by the same hyperparameters in Eq. (3). In this experiment, for simplicity, we use i-vector averaging for estimating the enrollment i-vector.

Table 2
Effect of using multicondition training for likelihood computation, PLDA hyperparameter estimation, or both. 'MC' stands for multicondition training. The performance is in terms of EER (DCF).

MC enroll	MC PLDA hyperparam	Male	Female
No	No	2.22 (0.26)	2.02 (0.30)
Yes	No	1.86 (0.33)	2.23 (0.34)
No	Yes	1.39 (0.17)	1.32 (0.21)
Yes	Yes	**1.32** (**0.17**)	**1.32** (**0.20**)

Table 3
Comparison of SNR-wise analysis of matched and multicondition train/test. Analysis done on female data from I4U dataset. The performance is in terms of EER (DCF).

Enroll and PLDA	Test data		
	Orig.	15 dB	6 dB
Orig. only	**0.72** (**0.12**)	1.40 (0.21)	3.52 (0.53)
15 dB only	1.10 (0.17)	**1.28** (**0.17**)	2.16 (0.33)
6 dB only	1.58 (0.29)	1.58 (0.24)	**1.85** (**0.31**)
Orig. + 15 dB + 6 dB	1.01 (0.16)	1.35 (0.20)	2.06 (0.32)

As expected, multicondition training improves verification performance. An interesting observation is that multicondition training brings considerable improvement when applied to PLDA hyperparameter estimation. This indicates that, once the PLDA hyperparameters are estimated in this manner, multicondition enrollment does not provide major additional robustness to the system. We provide more insight into this observation later in the paper.

5.4. Effect of matched-SNR for PLDA training

Matched SNR conditions between enroll and test data are generally expected to perform better than mismatched conditions. To verify this, experiments were carried out on the female trials using the three SNRs of the evaluation data, and are tabulated in Table 3. I-vector averaging is used to compute the enrollment i-vector. For the first three rows, the enrollment data and PLDA training data comprises of a single SNR. In the last row, both PLDA training and enrollment is performed with multicondition data. The amount of enrollment/PLDA training data is the same for each row. It is to be noted that the number of enrollment i-vectors per speaker varies.

From each column of Table 3, we infer that matched conditions for enrollment/PLDA training and the test data result in better performance as opposed to multicondition (the last row). But since the operating noise condition/SNR is rarely known beforehand in practice, multicondition enrollment/training is an effective workaround. Multicondition training results in only a minor degradation in performance when compared to the matched SNR case.

5.5. Likelihood computation with multiple enrollment i-vectors

We next study the effect of i-vector length normalization and the various methods for computing the likelihood ratio, given multiple enrollment utterances. Performance obtained for the various methods outlined in Section 4 are given in Table 4. Based on the performance on the I4U dataset, the likelihood computation is also repeated on the noisy-telephone condition (common condition 4) of the NIST 2012 SRE dataset (see [28]).

From Table 4 we find that i-vector length normalization improves performance for all the scoring methods, except for the

[5] The filelists for the data used in this paper is available from: http://cs.uef.fi/~paddy/public/pldaDSP2014filelist.tgz.

Table 4
Effect of length normalization on different scoring methods. LN = i-vector length normalization. For the SRE 2012 data, LN is applied to all methods except the pooled sessions method. The performance is in terms of EER (DCF).

Method	I4U EvalSet		SRE 2012 cond. 4 subset
	No LN	With LN	
Male			
Multi-session	3.22 (0.40)	1.60 (0.18)	11.86 (0.68)
I-vec avg.	2.84 (0.32)	**1.32** (**0.17**)	**4.85** (**0.51**)
Score avg.	3.28 (0.40)	1.65 (0.28)	9.39 (0.82)
Max. score	2.75 (0.40)	1.34 (0.26)	10.18 (0.81)
Pooled session	**2.72** (**0.30**)	3.18 (0.30)	5.18 (0.65)
Female			
Multi-session	3.18 (0.43)	1.57 (0.21)	9.50 (0.71)
I-vec avg.	2.71 (**0.32**)	1.32 (**0.20**)	**3.62** (**0.51**)
Score avg.	3.15 (0.39)	1.76 (0.32)	5.01 (0.59)
Max. score	**2.48** (0.34)	**1.26** (0.28)	4.09 (0.58)
Pooled session	2.65 (0.33)	3.56 (0.36)	4.18 (0.60)

Table 5
Relative performance of whitening transformation and making i-vectors unit length. Analysis in terms of EER on i-vector averaging on I4U data.

Whitening	Length norm	Male	Female
No	No	2.84	2.71
No	Yes	1.63	1.67
Yes	No	2.84	2.70
Yes	Yes	1.32	1.32

pooled-sessions scoring. For i-vector averaging, the enrollment i-vectors are first length normalized, then averaged into a single i-vector. Multi-session scoring of i-vectors does not work as well as the rest of the scoring methods considered, confirming the independent observations of [15] and [4]. In particular, with length normalization enabled, i-vector averaging outperforms the other compared methods (except for the max-scoring method for the female case). The max-scoring method gives very similar performance, but is computationally more expensive. The score-averaging method is poorer in performance, and again, is computationally more expensive.

The pooled session scoring method degrades in performance when length normalization is applied. The pooled session enrollment i-vector is obtained from statistics derived from multiple sessions. Thus the pooled i-vector represents an average of multiple channels and acoustic content. Applying length normalization on this i-vector possibly results in a mismatch with the test i-vector, which is from a single session.

The different scoring methods give more variation in performance on the SRE 2012 data. Here again, i-vector averaging gives the best performance. The relative difference in performance between i-vector averaging and multi-session scoring is almost 60%. Max-scoring does not fare well in the male case, but does well for the female case. As inferred from the performance on I4U data, i-vector length normalization is applied to all methods except the pooled sessions method.

Applying length normalization involves whitening the i-vectors and then making them unit length. The relative merit of each step on i-vector averaging for the I4U data, in terms of EER, is tabulated in Table 5. From these results, we conclude that, when applied in isolation, making the i-vectors unit length is more effective than whitening them. This is due to the possible mismatch between the data used to estimate the whitening matrix and the enrollment i-vectors. Applying both steps provide the largest improvement.

5.6. Factors affecting multi-session scoring

The above results suggest that multi-session scoring is inferior to other methods which do not process all the enrollment i-vectors

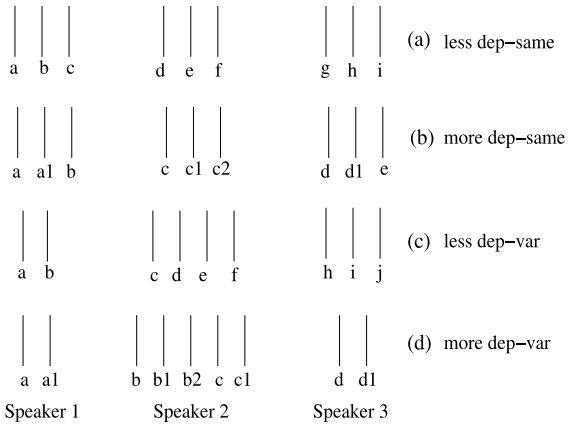

Fig. 2. Illustration of four different enrollment scenarios involving conditional dependence and number of enrollment utterances per speaker. Each row represents an enrollment scenario: (a) conditionally less-dependent, same number of enrollment utterances per target speaker (b) conditionally more-dependent, same number of utterances, (c) conditionally less-dependent, variable number utterances, (d) conditionally more-dependent, variable number of utterances.

simultaneously. To study this in more detail, we take two factors into account. The first is a by-product of the design of the I4U dataset: different target speakers have different numbers of enrollment utterances. This results in a different value of N in Eq. (7) for trials involving different target speakers. The second factor is that the PLDA model assumes that the i-vectors are conditionally independent given the latent speaker variable $\mathbf{x}$ [3].

The multicondition enrollment data in the I4U dataset do *not* satisfy the conditional independence assumption. This is because the individual noisy versions of an enrollment i-vector are derived from the same original utterance (by adding noise, as explained in Section 5.1). Hence, these i-vectors share more than just the same speaker identity, invalidating the independence assumption. On the other hand, enrollment i-vectors derived from different utterances (with different speech content, but from the same speaker) can be considered 'less-dependent' than the former.

Another factor is that the likelihood scores computed with different numbers of enrollment utterances exhibit different numerical ranges, making the scores inconsistent across trials [15,4]. This can be viewed as a score calibration problem. To study these two effects in greater detail, we derive a smaller dataset from the I4U EvalSet. This dataset consists of 106 female speakers, and consists of 5031 target trials and 528,255 non-target trials.

We simulate four different enrollment scenarios using this dataset, varying the properties of 'conditional dependence' and 'same number of enrollment utterances'. Enabling or disabling one of these properties results in a different enrollment scenario. To make the scenarios comparable, care is taken so that the *average* number of enrollment utterances per speaker is the same for all scenarios. These are illustrated for three speakers in Fig. 2. Each vertical line represents an enrollment utterance, and utterances labeled with the same character represent conditionally 'more-dependent' versions (for example, the utterance 'a' and 'a1' represent original and noisy versions of the same utterance, as used in multicondition training). Thus, the first row in Fig. 2 represents the scenario where all speakers have the same number (three, in this case) of conditionally less-dependent enrollment utterances. Similarly, the last row represents conditionally more-dependent, varying number of enrollment utterances (an average of three per speaker).

To simulate a statistically robust analysis, the enrollment utterances of a given target speaker is a random variable: it is a random subset of all the enrollment utterances of the speaker. Fifty random draws are made for each enrollment scenario, resulting in different

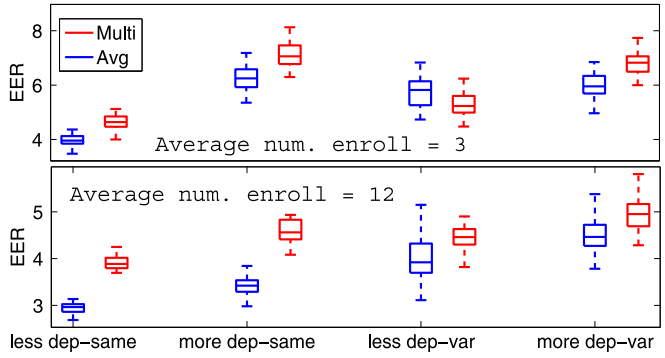

Fig. 3. Comparing performance of i-vector averaging and multi-session scoring. Box plots are shown for four enrollment scenarios, as given in Fig. 2. The top panel has 3 enrollment utterances on average per speaker, whereas bottom panel has 12.

enrollment subsets for target speakers. In other words, each random draw results in different utterances making up an enrollment scenario. Results of each scenario (processed from the fifty random draws), for i-vector averaging and multi-session scoring are plotted in Fig. 3, as box plots. The line inside the box represents the median, with the edges representing the 25 and 75 percentiles of the EERs observed, and the 'whiskers' represent the extreme values not considered outliers. Thus, the difference can be inferred as significant if boxes have no overlap along the vertical axis.

Comparing the top and bottom panels in Fig. 3, we find that the error rate reduces with increase in the average number of enrollment utterances, as expected. Having the same number of enrollment utterances, which are also conditionally less-dependent, for each target speaker is the optimal enrollment configuration. In the other cases, performance degrades gradually, with varying number of conditionally more-dependent enrollment utterances giving maximum error. Moreover, these results suggest that for smaller number of enrollment utterances per speaker, it is better to make them conditionally less-dependent as far as possible. In almost all the compared scenarios, i-vector averaging systematically outperforms multi-session scoring (the exception is the third scenario in the top panel). Moreover, both these scoring methods show similar trend in the different enrollment scenarios, meaning that the enrollment utterances need to be chosen with care for either scoring method.

6. Conclusions

We provided a review and an experimental evaluation of the i-vector PLDA framework in the context of multiple enrollment utterances. Our main findings, useful from a practical viewpoint, are:

1. **Applying multicondition training (Table 2):** Confirming the findings of [6], multicondition training is a useful technique to improve noise robustness. Applying it to the enrollment utterances (i.e. for likelihood computation) provided relative decrease of 16% and 9% in EER for males and females, respectively. Applying it to PLDA hyperparameter training stage instead provided corresponding relative decrements of 37% and 40%. Combining the two only increased computations without major added benefits. We therefore recommend applying multicondition data to PLDA training stage only.
2. **Multicondition versus matched-SNR training (Table 3):** When the operating SNR is not known in advance, multicondition training of the PLDA model is an effective way to add noise robustness. A relatively minor degradation of 8% EER is obtained in noisy conditions when multicondition data is used, when compared to the matched-SNR case.
3. **Length normalization (Table 4):** I-vector length normalization is a simple and effective technique, confirming the earlier findings reported by many others. Making i-vectors unit length provided the bulk of the improvement in length normalization. In our experiments, it provided relative decreases ranging from 50% to 40% in EER for all the scoring methods considered; the only exception was pooled-session scoring that was degraded by length-normalization.
4. **Choice of the scoring method (Table 4):** The performance of the scoring methods differ. The mathematically correct multi-session scoring, in general, did not perform consistently. On the I4U data, i-vector averaging reduced the EER by almost 16%, relative to multi-session scoring when length normalization was applied. Maximum-scoring provided performance comparable to i-vector averaging (in fact, the least error for the female case), but at the cost of increased computation. For SRE 2012 data, the various scoring methods exhibit considerable variation, with i-vector averaging providing improvement ranging from nearly 60% to 6% for male data, and from 60% to 11% for female data. Pooled session scoring did not perform consistently. Based on all these observations, for practitioners we recommend i-vector averaging.
5. **Scoring dependence on the enrollment utterances:** A closer look at factors affecting enrollment data reveals that conditional dependence and varying number of utterances per target speaker have a major impact on the performance. Having the same number of utterances per target speaker, which are also conditionally less-dependent, is a desirable configuration. Moreover, when having less enrollment utterances, it is useful to reduce their conditional dependence as much as possible. In practice, this means avoiding the use of both clean and noisy versions of the same utterance for likelihood computation. Since i-vector averaging sidesteps these issues, this is again a good reason for using it in practice.

This paper has provided insights into factors relevant for handling multiple enrollment i-vectors with probabilistic linear discriminant analysis. Future work will look at more effective measures on utilizing all available training and enrollment data, and the utilization of supplementary metadata.

Acknowledgments

This work was supported by the Academy of Finland (projects 253120, 253000).

Appendix A. Computational issues

To evaluate the logarithm of likelihood given by Eq. (8), we utilize two lemmas described below. We note that we need to compute the determinant and the inverse of a block matrix of size $N \times N$, where each block is a matrix of size $D \times D$. We exploit the special structure of the matrix in the computations. Since the number of enrollment i-vectors N is dependent on the target speaker, this expression has to be evaluated separately for each target speaker.

Lemma 1. *Let matrix* $\mathbf{M}$ *has the form*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}+\mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{B} & \mathbf{A}+\mathbf{B} & \dots & \mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B} & \mathbf{B} & \dots & \mathbf{A}+\mathbf{B} \end{bmatrix},$$

where $\mathbf{M}$ *is a block matrix of size* $N \times N$*, and each element of* $\mathbf{M}$ *is a matrix of size* $D \times D$*. Moreover, we assume that matrices* $\mathbf{A}$ *and* $N\mathbf{B}+\mathbf{A}$ *are invertible. Then the inverse of matrix* $\mathbf{M}$ *is*

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{P}+\mathbf{Q} & \mathbf{Q} & \dots & \mathbf{Q} \\ \mathbf{Q} & \mathbf{P}+\mathbf{Q} & \dots & \mathbf{Q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q} & \mathbf{Q} & \dots & \mathbf{P}+\mathbf{Q} \end{bmatrix},$$

where $\mathbf{P} = \mathbf{A}^{-1}$ *and* $\mathbf{Q} = -(N\mathbf{B}+\mathbf{A})^{-1}\mathbf{B}\mathbf{A}^{-1}$.

Proof. It is easy to see that the inverse matrix $\mathbf{M}^{-1}$ has to have the same form: it is invariant under all transformations replacing blocks inside diagonal or outside (to ensure one can apply such transformations to the identity $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$). Let us denote the non-diagonal blocks of $\mathbf{M}^{-1}$ by $\mathbf{Q}$ and diagonal blocks by $\mathbf{P}+\mathbf{Q}$. Then the following identity holds:

$$\begin{bmatrix} \mathbf{A}+\mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{B} & \mathbf{A}+\mathbf{B} & \dots & \mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B} & \mathbf{B} & \dots & \mathbf{A}+\mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{P}+\mathbf{Q} & \mathbf{Q} & \dots & \mathbf{Q} \\ \mathbf{Q} & \mathbf{P}+\mathbf{Q} & \dots & \mathbf{Q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q} & \mathbf{Q} & \dots & \mathbf{P}+\mathbf{Q} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} \end{bmatrix}$$

Multiplying, we obtain:

$$\mathbf{A}\mathbf{P} + \mathbf{A}\mathbf{Q} + \mathbf{B}\mathbf{P} + N\mathbf{B}\mathbf{Q} = \mathbf{I}$$

Substituting $\mathbf{P} = \mathbf{A}^{-1}$, we have,

$$\mathbf{A}\mathbf{Q} + \mathbf{B}\mathbf{P} + N\mathbf{B}\mathbf{Q} = \mathbf{0},$$

which gives

$$\mathbf{Q} = -(N\mathbf{B}+\mathbf{A})^{-1}\mathbf{B}\mathbf{A}^{-1}. \qquad \square$$

Lemma 2. *Let matrix* $\mathbf{M}$ *have the form*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}+\mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{B} & \mathbf{A}+\mathbf{B} & \dots & \mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B} & \mathbf{B} & \dots & \mathbf{A}+\mathbf{B} \end{bmatrix},$$

where $\mathbf{M}$ *is a block matrix of size* $N \times N$*, and each element of* $\mathbf{M}$ *is a matrix of size* $D \times D$*. Then the determinant* $|\mathbf{M}|$ *is equal to* $|\mathbf{A}|^{N-1}|\mathbf{A}+N\mathbf{B}|$.

Proof. We use the fact the determinant is invariant under elementary transformations. We subtract the second row from the others

$$\begin{bmatrix} \mathbf{A}+\mathbf{B} & \mathbf{B} & \mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{B} & \mathbf{A}+\mathbf{B} & \mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{B} & \mathbf{B} & \mathbf{A}+\mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{B} & \mathbf{B} & \mathbf{B} & \mathbf{A}+\mathbf{B} & \dots & \mathbf{B} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{B} & \mathbf{B} & \mathbf{B} & \mathbf{B} & \dots & \mathbf{A}+\mathbf{B} \end{bmatrix} \sim \begin{bmatrix} \mathbf{A} & -\mathbf{A} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{B} & \mathbf{A}+\mathbf{B} & \mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{0} & -\mathbf{A} & \mathbf{A} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{A} & \mathbf{0} & \mathbf{A} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & -\mathbf{A} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{A} \end{bmatrix}$$

Then we add the first row multiplied by $-\mathbf{A}^{-1}\mathbf{B}$ to the second row:

$$\begin{bmatrix} \mathbf{A} & -\mathbf{A} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}+2\mathbf{B} & \mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{0} & -\mathbf{A} & \mathbf{A} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{A} & \mathbf{0} & \mathbf{A} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & -\mathbf{A} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{A} \end{bmatrix}$$

Next we add all columns from third to Nth to the second:

$$\begin{bmatrix} \mathbf{A} & -\mathbf{A} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}+N\mathbf{B} & \mathbf{B} & \mathbf{B} & \dots & \mathbf{B} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{A} \end{bmatrix}$$

Being a block-triangular matrix, the determinant of is now equal to the product of the block determinants. Thus, $|\mathbf{M}| = |\mathbf{A}|^{N-1}|\mathbf{A}+N\mathbf{B}|$. $\square$

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (4) (2011) 788–798.

[2] S.J.D. Prince, J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: Proceedings of International Conference on Computer Vision, 2007, pp. 1–8.

[3] S. Prince, Computer Vision: Models Learning and Inference, Cambridge University Press, 2012.

[4] J. Villalba, M. Diez, A. Varona, E. Lleida, Handling recordings acquired simultaneously over multiple channels with PLDA, in: Proceedings of Interspeech, 2013.

[5] J. Ming, T. Hazen, J. Glass, D. Reynolds, Robust speaker recognition in noisy conditions, IEEE Trans. Audio Speech Lang. Process. 15 (5) (2007) 1711–1723.

[6] D. Garcia-Romero, X. Zhou, C.Y. Espy-Wilson, Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2012, pp. 4257–4260.

[7] A. Kanagasundaram, R.J. Vogt, D.B. Dean, S. Sridharan, PLDA based speaker recognition on short utterances, in: Proceedings of Speaker Odyssey, 2012.

[8] A.K. Sarkar, D. Matrouf, P.M. Bousquet, J.F. Bonastre, Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification, in: Proceedings of Interspeech, 2012.

[9] P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, P. Dumouchel, PLDA for speaker verification with utterances of arbitrary duration, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[10] M. McLaren, D. van Leeuwen, Improved speaker recognition when using i-vectors from multiple speech sources, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2011, pp. 5460–5463.

[11] J. Villalba, L. Eduardo, Handling i-vectors from different recording conditions using multi-channel simplified PLDA in speaker recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6763–6767.

[12] T. Hasan, R. Saeidi, J.H.L. Hansen, D.A. van Leeuwen, Duration mismatch compensation for i-vector based speaker recognition systems, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[13] S. Cumani, O. Plchot, P. Laface, Probabilistic linear discriminant analysis of i-vector posterior distributions, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7644–7648.

[14] M. Mandasari, M. McLaren, D.A. van Leeuwen, The effect of noise on modern automatic speaker recognition systems, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2012.

[15] K.A. Lee, A. Larcher, C. You, B. Ma, H. Li, Multi-session PLDA scoring of i-vector for partially open-set speaker detection, in: Proceedings of Interspeech, 2013.

[16] D. Garcia-Romero, A. McCree, Subspace-constrained supervector PLDA for speaker verification, in: Proceedings of Interspeech, 2013.

[17] S. Yaman, J. Pelecanos, Using polynomial kernel support vector machines for speaker verification, IEEE Signal Process. Lett. 20 (9) (2013) 901–904.

[18] L. El Shafey, C. McCool, R. Wallace, S. Marcel, A scalable formulation of probabilistic linear discriminant analysis: applied to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1788–1794.

[19] P. Rajan, T. Kinnunen, V. Hautamäki, Effect of multicondition training on i-vector PLDA configurations for speaker recognition, in: Proceedings of Interspeech, 2013.

[20] P. Kenny, Bayesian speaker verification with heavy tailed priors, in: Proceedings of Speaker Odyssey, 2010.

[21] D. Garcia-Romero, C. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: Proceedings of Interspeech, 2011, pp. 249–252.
[22] R. Saeidi, et al., I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification, in: Proceedings of Interspeech, 2013.
[23] O. Glembek, L. Burget, P. Matejka, M. Karafiat, P. Kenny, Simplification and optimization of i-vector extraction, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2011, pp. 4516–4519.
[24] N. Brümmer, E. De Villiers, The speaker partitioning problem, in: Proceedings of the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, 2010.
[25] S. Cumani, N. Brummer, L. Burget, P. Laface, O. Plchot, V. Vasilakakis, Pairwise discriminative speaker verification in the i-vector space, IEEE Trans. Audio Speech Lang. Process. 21 (6) (2013) 1217–1227.
[26] NIST, The NIST year 2010 speaker recognition evaluation plan, http://www.itl.nist.gov/iad/mig/tests/sre/2010/, 2010.
[27] T. Kinnunen, P. Rajan, A practical, self adaptive voice activity detector for speaker verification with noisy telephone and microphone data, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
[28] NIST, The NIST year 2012 speaker recognition evaluation plan, http://www.nist.gov/itl/iad/mig/sre12.cfm, 2012.

Padmanabhan Rajan received the Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology, Madras in 2012. He has worked as a postdoctoral researcher at the University of Eastern Finland, Joensuu, Finland. He is currently a faculty member in the School of Computing and Electrical Engineering at Indian Institute of Technology, Mandi, India. His research interests are in speech processing and pattern recognition.

Anton Afanasyev received the M.Sc. degree in Mathematics from the Saint-Petersburg University, Russia in 2012. He has worked as a researcher at the Speech Technology Center and as a developer at Analog Micro Devices, Inc. Currently, he holds research position at the Intel Labs (Saint-Petersburg) and research position at the iBinom, Inc. His current research interests are bioinformatics, speaker recognition and programming languages.

Ville Hautamäki received the M.Sc. degree in Computer Science from the University of Joensuu, Finland in 2005. He received the Ph.D. degree in Computer Science from the same university in 2008. He has worked as a research fellow at the Institute for Infocomm Research, A*STAR, Singapore. Currently, he is post-doctoral researcher in University of Eastern Finland, funded by Academy of Finland. His current research interests are cluster analysis, speaker recognition and language recognition.

Tomi Kinnunen received the M.Sc., Ph.Lic. and Ph.D. degrees in Computer Science from the University of Joensuu (now University of Eastern Finland, UEF), Finland, in 1999, 2004 and 2005, respectively. From 2005 to 2007, he worked as an associate scientist at the Institute for Infocomm Research (I2R), Singapore. Since 2007, he has been with UEF. From 2010 to 2012, he was funded by a post-doc grant from Academy of Finland and he currently holds position of university researcher. He serves as an associate editor in *Digital Signal Processing* and is the chair of *Odyssey 2014: the Speaker and Language Recognition Workshop*. His primary research interests include speaker recognition, speech signal processing, pattern recognition and biometric person authentication.