Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Fast Scoring for PLDA with Uncertainty Propagation

Weiwei-LIN and Man-Wai MAK

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR of China
weiwei.lin@connect.polyu.hk, enmwmak@polyu.edu.hk

Abstract

By treating utterances as points in the i-vector space, i-vector/PLDA can achieve fast verification. However, this approach lacks the ability to cope with utterance-length variability. A method called uncertainty propagation (UP) that takes the uncertainty of i-vectors into account has been recently proposed to deal with this problem. However, the loading matrix for modeling utterance-length variability is session-dependent, making UP computationally expensive. In this paper, we demonstrate that utterance-length variability mainly affects the scale of the posterior covariance matrices. Based on this observation, we propose to substitute the session-dependent loading matrices by the ones trained from development data, where the selection of pre-computed loading matrices is based on a fast scalar comparison. This approach can reduce the computation cost of standard UP to the one comparable with the conventional PLDA. Experiments on the NIST 2012 Speaker Recognition Evaluation show that the proposed method can perform as good as the standard UP, but requires only 3.7% of the scoring time. The method also requires substantially less memory as compared with the standard UP, especially when the number of target speakers is large.

1. Introduction

The i-vector/PLDA framework [1–3] currently dominates the text-independent speaker verification domain. The method has also been extended to address noise variability in robust speaker verification [4, 5]. Its success relies on the assumption that utterances can be treated as points in the i-vector space irrespective of their duration. This assumption works well when utterances are sufficiently long. When presented with short utterances or utterances with varying durations, the performance of i-vector/PLDA systems will degrade rapidly.

In [6, 7], Kenny et al. proposed a sophisticated method called uncertainty propagation (UP) to deal with utterance-length variability. The method relies on the fact that an i-vector is the posterior mean of the latent variables conditioned on the Baum-Welch statistics and that the posterior covariance matrix represents the reliability of the i-vector. As a result, the longer the utterance, the smaller the posterior covariances. The idea of uncertainty propagation is to model the uncertainty in the estimated i-vector by a session-dependent loading matrix in the PLDA model.

This length-variability loading matrix, unlike speaker loading matrix or channel loading matrix, is session-dependent. As a result, the session-dependent terms in the PLDA scoring function cannot be pre-computed, which makes the scoring process computationally expensive. Beside computation cost, UP also requires to store the posterior covariance matrices of enrollment utterances. As a result, the memory consumption per utterance growths quadratically with the i-vector dimension. Both computation cost and memory consumption make this method unattractive in real applications.

Several attempts have been made to speed up the scoring process of UP. For example, Cumani et al. [8] proposed using MAP-estimated i-vectors for the target speakers and propagating the posterior covariances of test i-vectors to the PLDA model. The method works well when the test utterances are short and the target-speakers' utterances are long. In [9], the author proposed diagonalizing the matrices involved in scoring to reduce the computational complexity. This approach, although outperforms the conventional PLDA, still degrades performance of UP when the test utterances are very short.

In this paper, we propose a scoring method based on a factor analysis model that does not involve session-dependent loading matrices. To remove the session dependency, we trained multiple length-variability loading matrices from development data. Then via some simple metrics, length-variability loading matrices are selected to model the length-variability of target and test i-vectors during the scoring stage. By getting rid of session-dependent matrices, we can pre-compute the length-variability related terms and store them in a repository for retrieval during verification. Thus, the computational complexity of our method is the same as that of the conventional PLDA.

Experiments on the NIST 2012 SRE [10] show that the proposed method can perform as good as standard UP, provided that sufficient length-variability matrices are available for selection. The scoring time is only 3.7% of standard UP. The proposed method is also economical in memory consumption in that the memory for storing the pre-computed terms does not depend on the number of target speakers, which makes the method an ideal solution for large-scale speaker verification.

This paper is organized as follows. Section 2 describes the conventional i-vector/PLDA approach and explains the principle of uncertainty propagation. Emphases are made on the scoring functions and explanation of why UP is computationally expensive. In Section 3, we provide the details of the fast scoring methods. Experimental setup and results are presented in Section 4 and Section 5, respectively. Finally, we conclude our findings in Section 6.

2. I-vector/PLDA and Uncertainty Propagation

2.1. I-vector Extraction

The i-vector approach is based on the joint factor analysis [11], [12]. It can be viewed as a feature extraction process that maps

This work was in part supported by RGC of Hong Kong and The Hong Polytechnic University, Grant No. 152117/14E and 4-ZZEE.

doi: 10.21437/Odyssey.2016-5

a sequence of acoustic vectors in an utterance to a vector of low and fixed dimension. It assumes that the speaker- and channel-dependent GMM-supervectors [13] ($\boldsymbol{\beta}$'s) live in a low dimensional space:

$$\boldsymbol{\beta} = \mathbf{m} + \mathbf{T}\boldsymbol{\eta}, \tag{1}$$

where $\mathbf{m}$ is the speaker- and channel-independent GMM-supervector constructed by stacking up the means of a universal background model (UBM); $\mathbf{T}$ is a low-rank total variability matrix whose columns span the subspace where speaker- and channel-specific information varies; $\boldsymbol{\eta}$ is a latent variable which is assumed to follow a standard normal distribution. Given an utterance, its i-vector is the maximum-a-posteriori (MAP) estimate of the latent variable $\boldsymbol{\eta}$, which we denote as $\boldsymbol{\omega}$. Consider an utterance of T acoustic vectors $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ and an UBM of C mixture components: $\{\lambda_c, \mathbf{m}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$. To estimate the i-vector of this utterance, we compute the Baum-Welch statistics as follows [14]:

$$N_c = \sum_{t=1}^{T} \gamma_t(c) \tag{2}$$

$$\tilde{\mathbf{f}}_c = \sum_{t=1}^{T} \gamma_t(c)(\mathbf{x}_t - \mathbf{m}_c) \tag{3}$$

where

$$\gamma_t(c) = \frac{\lambda_c \mathcal{N}(\mathbf{x}_t; \mathbf{m}_c, \boldsymbol{\Sigma}_c)}{\sum_{c=1}^{C} \lambda_c \mathcal{N}(\mathbf{x}_t; \mathbf{m}_c, \boldsymbol{\Sigma}_c)}. \tag{4}$$

The posterior covariance matrix $\mathrm{cov}(\boldsymbol{\eta}, \boldsymbol{\eta})$ and the i-vector $\boldsymbol{\omega}$ of the utterance are given by:

$$\mathrm{cov}(\boldsymbol{\eta}, \boldsymbol{\eta}) = \mathbf{L}^{-1} = \left(\mathbf{I} + \sum_{c=1}^{C} N_c \mathbf{T}_c^{\mathsf{T}} \boldsymbol{\Sigma}_c^{-1} \mathbf{T}_c \right)^{-1} \tag{5}$$

$$\boldsymbol{\omega} = \mathrm{cov}(\boldsymbol{\eta}, \boldsymbol{\eta}) \sum_{c=1}^{C} \mathbf{T}_c^{\mathsf{T}} \boldsymbol{\Sigma}_c^{-1} \tilde{\mathbf{f}}_c, \tag{6}$$

where $\mathbf{L}$ is the posterior precision matrix.

2.2. Probabilistic Linear Discriminant Analysis

After i-vector extraction, channel compensation is applied to suppress the undesired variability in the i-vectors. The low dimensionality of i-vectors makes it possible to apply statistical methods that are not practical in the high-dimensional supervector space. The most popular method is probabilistic linear discriminant analysis (PLDA). Early PLDA is based on Student's t distributions because it was found empirically that i-vectors follow a heavy-tailed distribution [3]. It was found later that Gaussian PLDA with length-normalized i-vectors performs equally well [15]. Because of its low computational requirements, Gaussian PLDA is preferred in practice.

2.2.1. I-Vector Preprocessing for PLDA

To use Gaussian PLDA, we first need to preprocess the i-vectors, which involves two steps. First, we need to whiten i-vectors using the matrix $\mathbf{W}$ learned from training set. It can be written as:

$$\boldsymbol{\omega}^{\text{wht}} = \mathbf{W}^{\mathsf{T}}(\boldsymbol{\omega} - \bar{\boldsymbol{\omega}}) \tag{7}$$

where $\bar{\boldsymbol{\omega}}$ is the mean of training i-vectors, $\boldsymbol{\omega}^{\text{wht}}$ is the whitened i-vector and $\mathbf{W}$ is a transformation matrix. $\mathbf{W}$ can be obtained by the Cholesky decomposition of the within-class covariance matrix of training i-vectors [16]. Then we apply length-normalization to i-vectors individually:

$$\boldsymbol{\omega}^{\text{l-norm}} = \frac{\boldsymbol{\omega}^{\text{wht}}}{\|\boldsymbol{\omega}^{\text{wht}}\|}. \tag{8}$$

It is also customary to apply linear discriminant analysis (LDA) followed by within-class covariance normalization (WCCN) [17] to the length-normalized i-vectors. Let us denote a matrix $\mathbf{P}$ as the transformation matrix that combines whitening, LDA and WCCN, then the whole pre-processing step can be written as:

$$\mathbf{w} = \frac{\mathbf{P}(\boldsymbol{\omega} - \bar{\boldsymbol{\omega}})}{\|\boldsymbol{\omega}^{\text{wht}}\|}, \tag{9}$$

where $\mathbf{w}$ is a preprocessed i-vector for Gaussian PLDA modeling.

2.2.2. Gaussian PLDA Modelling

In Gaussian PLDA, a preprocessed i-vector $\mathbf{w}_{i,j}$ from the j-th session of speaker i is considered generated from a factor analysis model:

$$\mathbf{w}_{i,j} = \boldsymbol{\mu} + \mathbf{V}\mathbf{h}_i + \mathbf{G}\mathbf{z}_{i,j} + \boldsymbol{\epsilon}_{i,j}, \tag{10}$$

where $\boldsymbol{\mu}$ is the global mean of i-vectors, the column of $\mathbf{V}$ defines the speaker subspace where speaker factor $\mathbf{h}_i$ varies, the column of $\mathbf{G}$ defines the channel subspace where channel factor $\mathbf{z}_{i,j}$ varies, and $\boldsymbol{\epsilon}_{i,j}$ is the residual noise that is not captured by both subspaces. Both $\mathbf{h}_i$ and $\mathbf{z}_{i,j}$ are assumed to follow a standard normal prior. $\boldsymbol{\epsilon}_{i,j}$ is assumed to follow a Gaussian distribution with zero mean and diagonal covariance matrix $\boldsymbol{\Sigma}$.[1] The model can be divided into two parts: (1) the speaker-dependent part $\boldsymbol{\mu}+\mathbf{V}\mathbf{h}_i$ describing the inter-speaker variability, which remains unchanged for the same speaker; (2) the session-dependent part $\mathbf{G}\mathbf{z}_{i,j} + \boldsymbol{\epsilon}_{i,j}$ describing the intra-speaker variability, which varies from utterances to utterances even for utterances from the same speaker.

Because i-vectors are of low dimension, it is feasible to absorb the intra-speaker variability into the full covariance matrix $\boldsymbol{\Sigma}$, which results in the simplified Gaussian PLDA model:

$$\mathbf{w}_{i,j} = \boldsymbol{\mu} + \mathbf{V}\mathbf{h}_i + \boldsymbol{\epsilon}_{i,j}, \tag{11}$$

where $\boldsymbol{\epsilon}_{i,j} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

2.2.3. Gaussian PLDA Scoring

Given a test i-vector $\mathbf{w}_t$ and a target speaker's i-vector $\mathbf{w}_s$, the log-likelihood ratio of the same-speaker hypothesis to different-speaker hypothesis can be computed by [15]:

$$\begin{aligned} \text{score} &= \log \left[\frac{p(\mathbf{w}_s, \mathbf{w}_t | \text{same-speaker})}{p(\mathbf{w}_s, \mathbf{w}_t | \text{different-speakers})} \right] \\ &= \frac{1}{2}\mathbf{w}_s^{\mathsf{T}} \boldsymbol{\Phi} \mathbf{w}_s + \mathbf{w}_s^{\mathsf{T}} \boldsymbol{\Psi} \mathbf{w}_t + \frac{1}{2}\mathbf{w}_t^{\mathsf{T}} \boldsymbol{\Phi} \mathbf{w}_t + \text{const} \end{aligned} \tag{12}$$

where

$$\begin{aligned} \boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{tot}^{-1} - (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\ \boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \end{aligned} \tag{13a}$$

$$\boldsymbol{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^{\mathsf{T}} \qquad \boldsymbol{\Sigma}_{tot} = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}. \tag{13b}$$

Note that Eq. 13 can be pre-computed and only Eq. 12 needs to be evaluated during verification. As a result, PLDA scoring is

[1] Do not confuse this covariance matrix with $\boldsymbol{\Sigma}_c$'s in Eq. 4.

very efficient.

2.3. Uncertainty Propagation

Despite the huge success of the i-vector/PLDA framework, its underlying assumptions limit its applications. First, in i-vector extraction the duration of utterances is totally ignored, i.e., utterances are represented by vectors of fixed dimension regardless of their duration. Recall that an i-vector is the MAP estimate of latent variable $\boldsymbol{\eta}$; the accuracy of such estimate depends on the number of acoustic vectors. By ignoring durations, all i-vectors are treated as equally reliable. Secondly, in PLDA modelling, it is assumed that all of the intra-speaker variabilities are represented by the covariance matrix $\boldsymbol{\Sigma}$, which is the same across all i-vectors. This is apparently not a satisfactory assumption because short utterances have more severe intra-speaker variabilities than long utterances. It is undesirable to model both short and long utterances by the same covariance matrix ($\mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}$).

To better accommodate utterance-length variability, a modified PLDA is proposed in [6]. The basic idea is to tightly couple i-vector extraction and PLDA modelling by propagating the uncertainty during i-vector extraction into the PLDA model. Recall that the posterior covariance matrix in Eq. 5 represents the uncertainty of the MAP point-estimate in i-vector extraction. The shorter the utterance, the larger the posterior covariances. By propagating this information into PLDA and using a loading matrix to model the variability due to duration variation, this PLDA model can better handle the length-variability than the conventional PLDA model.

2.3.1. I-Vector Preprocessing for PLDA with UP

Because of the i-vector preprocessing steps in Section 2.2.1, we also need to apply the equivalent steps to the posterior covariance matrix. If only linear transformation $\mathbf{P}$ is applied to an i-vector, we can obtain its transformed posterior covariance matrix by:

$$\text{cov}(\mathbf{P}\boldsymbol{\eta}, \mathbf{P}\boldsymbol{\eta}) = \mathbf{P}\mathbf{L}^{-1}\mathbf{P}^{\mathsf{T}}. \tag{14}$$

We denote this covariance matrix as $\mathbf{U}\mathbf{U}^{\mathsf{T}}$, i.e., $\text{cov}(\mathbf{P}\boldsymbol{\eta}, \mathbf{P}\boldsymbol{\eta}) = \mathbf{U}\mathbf{U}^{\mathsf{T}}$ where $\mathbf{U}$ will be *propagated* to the PLDA model (see Section 2.3.2) to reflect the uncertainty of this i-vector. If length-normalization is applied to i-vectors, we can no longer find an exact preprocessed covariance matrix. Nevertheless, it can be approximated by [6]:

$$\mathbf{U}\mathbf{U}^{\mathsf{T}} \leftarrow \frac{\mathbf{P}\mathbf{L}^{-1}\mathbf{P}^{\mathsf{T}}}{\|\boldsymbol{\omega}^{\text{wht}}\|^2}. \tag{15}$$

While it is a crude approximation, we will show later that the approximation is good enough for UP to achieve very good performance.

2.3.2. Gaussian PLDA Modelling with UP

To propagating i-vectors' uncertainty into the PLDA model, an extra *session-dependent* space $\mathbf{U}$ is added to the model to reflect duration variability:

$$\mathbf{w}_{i,j} = \boldsymbol{\mu} + \mathbf{V}\mathbf{h}_i + \mathbf{U}_{i,j}\mathbf{z}_{i,j} + \boldsymbol{\epsilon}_{i,j}, \tag{16}$$

where $\mathbf{U}_{i,j}$ is the Cholesky decomposition of the posterior covariance matrix.[2] We refer to the space spanned by the column vectors of $\mathbf{U}$ as the length-variability space and $\mathbf{z}_{i,j}$ as the length-variability factor, which is assumed to follow a standard normal distribution. It should be noted that, unlike the speaker subspace or the channel subspace, the length-variability space is session-dependent.

[2]To simply naming, we refer $\mathbf{U}\mathbf{U}^{\mathsf{T}}$ to as posterior covariance matrix or covariance matrix in the rest of this paper unless stated otherwise.

2.3.3. Uncertainty Propagation Scoring

Given a test i-vector $\mathbf{w}_t$ and a target speaker's i-vector $\mathbf{w}_s$ and their corresponding posterior covariance matrix $\mathbf{U}_t\mathbf{U}_t{}^{\mathsf{T}}$ and $\mathbf{U}_s\mathbf{U}_s{}^{\mathsf{T}}$, the log-likelihood ratio of the same-speaker hypothesis to different-speaker hypothesis can be computed by [18]:

$$\begin{aligned}\text{score} &= \log\left[\frac{p(\mathbf{w}_s, \mathbf{w}_t|\mathbf{U}_s\mathbf{U}_s{}^{\mathsf{T}}, \mathbf{U}_t\mathbf{U}_t{}^{\mathsf{T}}, \text{same-speaker})}{p(\mathbf{w}_s, \mathbf{w}_t|\mathbf{U}_s\mathbf{U}_s{}^{\mathsf{T}}, \mathbf{U}_t\mathbf{U}_t{}^{\mathsf{T}}, \text{different-speakers})}\right] \\ &= \frac{1}{2}\left[\mathbf{w}_s^{\mathsf{T}}\mathbf{A}_{s,t}\mathbf{w}_s + 2\mathbf{w}_s^{\mathsf{T}}\mathbf{B}_{s,t}\mathbf{w}_t + \mathbf{w}_t^{\mathsf{T}}\mathbf{C}_{s,t}\mathbf{w}_t\right] + D_{s,t}\end{aligned} \tag{17}$$

where

$$\mathbf{A}_{s,t} = \boldsymbol{\Sigma}_s^{-1} - (\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \tag{18a}$$

$$\mathbf{B}_{s,t} = \boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \tag{18b}$$

$$\mathbf{C}_{s,t} = \boldsymbol{\Sigma}_t^{-1} - (\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \tag{18c}$$

$$D_{s,t} = -\frac{1}{2}\log\begin{vmatrix}\boldsymbol{\Sigma}_s & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Sigma}_t\end{vmatrix} + \frac{1}{2}\log\begin{vmatrix}\boldsymbol{\Sigma}_s & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_t\end{vmatrix} \tag{18d}$$

$$\boldsymbol{\Sigma}_t = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{U}_t\mathbf{U}_t^{\mathsf{T}} + \boldsymbol{\Sigma} \tag{18e}$$

$$\boldsymbol{\Sigma}_s = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{U}_s\mathbf{U}_s^{\mathsf{T}} + \boldsymbol{\Sigma} \tag{18f}$$

$$\boldsymbol{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^{\mathsf{T}} \tag{18g}$$

It is apparent from Eq. 18 that scoring in uncertainty propagation is much more computationally expensive than that in conventional PLDA, because Eq. 18(a–e) involve terms dependent on the test utterance. Unlike the speaker loading matrix $\mathbf{V}$ and the channel loading matrix $\mathbf{G}$, $\mathbf{U}_t$ are session-dependent, which forbids us to perform pre-computation as in Eq. 12 and Eq. 13. Besides, it is also necessary to store the loading matrix $\mathbf{U}_s$ (or the covariance matrix $\mathbf{U}_s\mathbf{U}_s^{\mathsf{T}}$), which results in considerable memory consumption. To reduce computation burden during verification, we need a method that allows us to perform pre-computation as much as possible while still be able to propagate the i-vector uncertainty to the PLDA model.

3. Fast Scoring for Uncertainty Propagation

3.1. Similarity in Posterior Covariance Matrices

Eq. 5 suggests that the posterior covariance matrices quantify the uncertainty of i-vectors through the zero-order sufficient statistics N_c's, which in turn are proportional to the utterance length. Therefore, if two utterances are of approximately the same duration, the posterior covariance matrices should be very similar.

To verify this conjecture, we selected 2000 utterances from development data and equally divided them into two sets: $\mathbb{U}_A$ and $\mathbb{U}_B$. For each utterance in $\mathbb{U}_A$, we truncated it to 1000 speech frames (after VAD). For each utterance in $\mathbb{U}_B$ we progressively truncated the speech regions from the end to produce short utterances comprising 9000 frames down to 1000 frames at an interval of 1000 frames. As a result, each utterance in $\mathbb{U}_B$ produces nine truncated utterances. Then, we compute

System	Group Identity Indicator
Sys. 1	Utterance length (after VAD)
Sys. 2	The mean of the diagonal elements of $\mathbf{U}\mathbf{U}^{\mathsf{T}}$
Sys. 3	The largest eigenvalue of $\mathbf{U}\mathbf{U}^{\mathsf{T}}$

Table 1: The group identity indicators used by the 3 systems to quantify the characteristics of the posterior covariance matrices.

the posterior covariance matrices of the truncated utterances and denoted them as $\mathbf{\Gamma}_{A,i}^{(k)}$ and $\mathbf{\Gamma}_{B,i}^{(l)}$ where $i = 1, \ldots, 1000$ $k = 1000$, and $l \in \{1000, 2000, \ldots, 9000\}$. This procedure effectively creates utterance pairs with variable utterance-length differences. For example, the utterance-length difference between $\mathbf{\Gamma}_{A,i}^{(1000)}$ and is $\mathbf{\Gamma}_{B,i}^{(5000)}$ is 4000 frames. Finally, we measured the distance between $\mathbf{\Gamma}_{A,i}^{(k)}$ and $\mathbf{\Gamma}_{B,i}^{(l)}$ by [19]:

$$d(\mathbf{\Gamma}_{A,i}^{(k)}, \mathbf{\Gamma}_{B,i}^{(l)}) = \sqrt{\frac{\text{trace}\left\{\left(\mathbf{\Gamma}_{A,i}^{(k)} - \mathbf{\Gamma}_{B,i}^{(l)}\right)^{\mathsf{T}}\left(\mathbf{\Gamma}_{A,i}^{(k)} - \mathbf{\Gamma}_{B,i}^{(l)}\right)\right\}}{\text{trace}\left\{\mathbf{\Gamma}_{A,i}^{(k)\,\mathsf{T}}\mathbf{\Gamma}_{A,i}^{(k)} + \mathbf{\Gamma}_{B,i}^{(l)\,\mathsf{T}}\mathbf{\Gamma}_{B,i}^{(l)}\right\}}}, \tag{19}$$

where $k = 1000$, $l \in \{1000, 2000, \ldots, 9000\}$, and $i = 1, \ldots, 1000$.

Fig. 1 shows a box plot of the distance $d(\cdot, \cdot)$ against the length-difference between the two utterances. The central mark inside each box indicates the median distance of 1000 pairs, and the bottom and top edges of the box indicate the 25th and 75th percentiles of distances, respectively. The whiskers extend to the most extreme non-outliers, and the outliers are represented by the '+' symbol [20].

Evidently, when the two utterances have equal length (Utterance-Length Difference $= 0$), the distance between the two posterior covariance matrices is very small, suggesting that the two covariance matrices are very similar. On the other hand, when the length-difference increases, the two matrices become dissimilar. Fig. 1 suggests that we may use utterance length to quantify the characteristics of posterior covariance matrices or session-dependent loading matrices $\mathbf{U}_{i,j}$. Therefore, for i-vectors that are estimated from utterances of approximately the same duration, the length-variabilities of these i-vectors could be modelled by the same loading matrix despite of the differences in channel- or speaker-specific information.

3.2. Fast Scoring for PLDA with UP

Motivated by the above observation, we divided the time between 1 and 42 seconds into a number of equal-length intervals and grouped the i-vectors into these intervals according to their utterance duration. As a result, the utterances in each group span a limited range of durations. For each group, we chose the $\mathbf{U}$ that corresponds to the middle of the interval as the length-variability loading matrix for that group. Through this procedure, we obtained a number of length-variability loading matrices, which can be used for enrolment as well as scoring the test i-vectors.

We denote k as the group index of i-vector $\mathbf{w}_{i,j}$ and $\mathbf{U}_k$ as the corresponding length-variability loading matrix. The factor analysis model in Eq. 16 becomes:

$$\mathbf{w}_{i,j}^{(k)} = \boldsymbol{\mu} + \mathbf{V}\mathbf{h}_i + \mathbf{U}_k\mathbf{z}_{i,j} + \boldsymbol{\epsilon}_{i,j}. \tag{20}$$

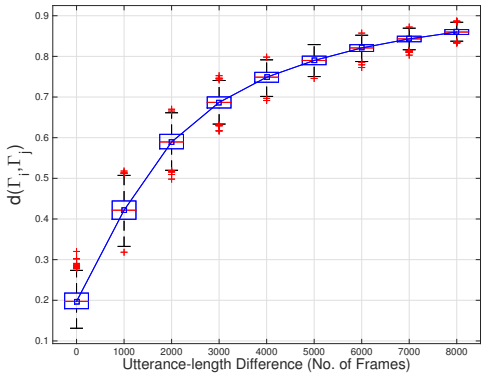

Figure 1: The distance (Eq. 19) between two covariance matrices with utterance-length difference ranging from 0 to 8000 frames (after VAD).

The major difference between Eq. 20 and Eq. 16 is the sources of the length-variability matrices $\mathbf{U}$. In Eq.16, $\mathbf{U}_{i,j}$ is the Cholesky decomposition of the posterior covariance matrix of i-vector $\mathbf{w}_{i,j}$. On the other hand, $\mathbf{U}_k$ in Eq. 20 is the Cholesky decomposition of the posterior covariance matrix of an i-vector derived from development data, where the posterior covariance matrix is very *similar* to that of $\mathbf{w}_{i,j}$. Because $\mathbf{U}_k$ is not session dependent, we can pre-compute all of the $\mathbf{U}_k$'s and their associated terms before verification.

Let us denote $\mathbf{w}_s$ and $\mathbf{w}_t$ as the i-vector of a target speaker and a test utterance, respectively. Suppose that the target speaker's i-vector belongs to group m and its corresponding length-variability matrix is $\mathbf{U}_m$, which is selected from a group of length-variability matrices $\mathcal{U} = \{\mathbf{U}_k; k = 1, \ldots, K\}$. Similarly, assume that the test i-vector belongs to group n and its corresponding length-variability matrix is $\mathbf{U}_n$, which is also selected from $\mathcal{U}$. Then, the log-likelihood ratio score in Eq. 17 becomes:

$$\text{score} = \frac{1}{2}\mathbf{w}_s^{\mathsf{T}}\mathbf{A}_{m,n}\mathbf{w}_s + \mathbf{w}_s^{\mathsf{T}}\mathbf{B}_{m,n}\mathbf{w}_t + \frac{1}{2}\mathbf{w}_t^{\mathsf{T}}\mathbf{C}_{m,n}\mathbf{w}_t + D_{m,n}. \tag{21}$$

The terms $\mathbf{A}_{m,n}, \mathbf{B}_{m,n}, \mathbf{C}_{m,n}$ and $D_{m,n}$ are retrieved from a repository that stores all of the pre-computed $\mathbf{A}_{p,q}, \mathbf{B}_{p,q}, \mathbf{C}_{p,q}$ and $D_{p,q}$, which are:

$$\mathbf{A}_{p,q} = \mathbf{\Sigma}_p^{-1} - (\mathbf{\Sigma}_p - \mathbf{\Sigma}_{ac}\mathbf{\Sigma}_q^{-1}\mathbf{\Sigma}_{ac})^{-1} \tag{22a}$$

$$\mathbf{B}_{p,q} = \mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{ac}(\mathbf{\Sigma}_q - \mathbf{\Sigma}_{ac}\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{ac})^{-1} \tag{22b}$$

$$\mathbf{C}_{p,q} = \mathbf{\Sigma}_q^{-1} - (\mathbf{\Sigma}_q - \mathbf{\Sigma}_{ac}\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{ac})^{-1} \tag{22c}$$

$$D_{p,q} = -\frac{1}{2}\log\begin{vmatrix}\mathbf{\Sigma}_p & \mathbf{\Sigma}_{ac} \\ \mathbf{\Sigma}_{ac} & \mathbf{\Sigma}_q\end{vmatrix} + \frac{1}{2}\log\begin{vmatrix}\mathbf{\Sigma}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_q\end{vmatrix}, \tag{22d}$$

where

$$\mathbf{\Sigma}_p = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{U}_p\mathbf{U}_p^{\mathsf{T}} + \mathbf{\Sigma} \qquad p = 1, \ldots, K \tag{22e}$$

$$\mathbf{\Sigma}_q = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{U}_q\mathbf{U}_q^{\mathsf{T}} + \mathbf{\Sigma} \qquad q = 1, \ldots, K \tag{22f}$$

$$\mathbf{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^{\mathsf{T}}. \tag{22g}$$

The group identities m and n are determined using their duration, as illustrated in Fig. 2. The group identity of $\mathbf{w}_s$ can be determined before verification. We only need to determine the group identity of test i-vector $\mathbf{w}_t$ and then evaluate Eq. 21

Method	K	Male (CC2)					
		EER(%)			minDCF		
		Sys. 1	Sys. 2	Sys. 3	Sys. 1	Sys. 2	Sys. 3
PLDA with UP by Fast Scoring	20	6.21	7.02	6.17	0.640	0.685	0.654
	25	6.07	6.35	6.00	0.635	0.658	0.646
	30	5.96	6.07	5.93	0.632	0.632	0.648
	35	6.45	5.97	5.91	0.633	0.631	0.643
	40	5.91	5.93	5.85	0.641	0.641	0.649
	45	5.95	5.89	5.96	0.633	0.642	0.636
PLDA	-	7.77			0.654		
PLDA with UP	-	5.75			0.644		

Table 2: The performance of PLDA, PLDA with standard UP, and PLDA with UP using the proposed fast scoring methods on the truncated utterances in CC2 of NIST 2012 SRE. K is the number of matrices $\mathbf{U}$'s in the repository. See Table 1 for the group-identity indicators used by the three systems.

during verification.

3.3. Other Possible Identity Indicators

The method in Section 3.2 has a problem when length normalization is included in i-vector preprocessing. This is because in Eq. 15 the posterior covariance matrix is scaled by the i-vector length. As length normalization is a non-linear function, it is unclear how the utterance length is related to the covariance matrix $\mathbf{U}\mathbf{U}^{\mathsf{T}}$ in Eq. 15. Therefore, using durations to determine length-variability groups may not be appropriate. Therefore, we propose two other measures that directly compare the transformed covariance matrices in Eq. 15. To this end, we define a scalar α to quantify the characteristics of the matrix $\mathbf{U}\mathbf{U}^{\mathsf{T}}$:

$$\alpha = f(\mathbf{U}\mathbf{U}^{\mathsf{T}}), \tag{23}$$

where this α could be:

1) *The mean of the diagonal elements of* $\mathbf{U}\mathbf{U}^{T}$. Because the posterior covariance matrix $\mathbf{U}\mathbf{U}^{\mathsf{T}}$ is almost diagonal, the mean of the diagonal elements is a compact representation of it;

2) *The largest eigenvalue of* $\mathbf{U}\mathbf{U}^{T}$. If the largest eigenvalues of two posterior covariance matrices are close, then the covariance matrices represent the same degree of variability;

We divided development i-vectors into a number of groups such that each group spans a limited range of α. For each group, we chose the $\mathbf{U}$ corresponding to the middle the α-interval as the length-variability loading matrix of that group. In this way, we obtained a number of length-variability matrices $\mathbf{U}_k$'s together with their corresponding α_k's ($k = 1, \ldots, K$). The group identities of target speaker's i-vector $\mathbf{w}_s$ and test i-vector $\mathbf{w}_t$ is then determined by:

$$m = \arg\min_{k\in\{1,\ldots,K\}} |\alpha_k - \alpha(s)| \tag{24a}$$

$$n = \arg\min_{k\in\{1,\ldots,K\}} |\alpha_k - \alpha(t)| \tag{24b}$$

where $\alpha(s)$ and $\alpha(t)$ are derived from their posterior matrices $\mathbf{U}_s\mathbf{U}_s^{\mathsf{T}}$ and $\mathbf{U}_t\mathbf{U}_t^{\mathsf{T}}$ using Eq. 23.

4. Experimental Setup

4.1. Speech Data and Performance Metrics

Speech files from NIST 2005–2012 SRE were used for system development and performance evaluation. The speech regions of each file were determined by a two-channel voice activity detector [21]. For each 10ms in the speech regions, we used a 25-ms Hamming window to extract 19 Mel frequency cepstral coefficients (MFCC) and its log-energy plus their first and second derivatives, which is followed by cepstral mean normalization and feature warping [22]. This procedure results in a 60-dimensional acoustic vector per 10 ms. To simulate utterances with arbitrary duration, the speech regions of each utterance were concatenated and then truncated randomly at a length ranging from 1 to 42 seconds.

System performance was based on the truncated speech segments of Common Condition 2 of NIST 2012 SRE (core set, male speakers). Equal error rate (EER), minimum detection cost function (minDCF) in NIST 2012 SRE and detection error rate trade-off (DET) curves [23] were used as performance metrics.

4.2. I-vector Extraction and PLDA Model Training

A gender-dependent UBM with 1024 Gaussian components and a total variability matrix with 500 total factors were trained using the full-length microphone and telephone utterances (after VAD) from NIST 2005-2008 SREs. Then, 500-dimensional i-vectors were extracted. WCCN whitening [16] followed by length-normalization (LN) [15] were applied to reduce the heavy-tailed behavior of i-vectors. Then LDA was applied to suppress intra-speaker variability and reduce i-vectors' dimension to 200. Another WCCN was then applied to reduce the effect of high within-class variability in the LDA-projected space. The parameters of WCCN, LDA, PLDA and PLDA with uncertainty propagation were trained using the truncated telephone and microphone utterances in NIST 2006–2010 SRE.

In fast scoring, the length-variability loading matrices were obtained from the truncated telephone utterances in NIST 2006–2010 SRE. Based on the three approaches to quantifying the characteristics of posterior covariance matrices in Sections 3.2 and 3.3, we have three PLDA-UP systems: Sys. 1, Sys. 2, and Sys. 3. Table 1 shows the group identity indicators used by these systems. We varied the values of K in Eq. 24 from 20 to 45.

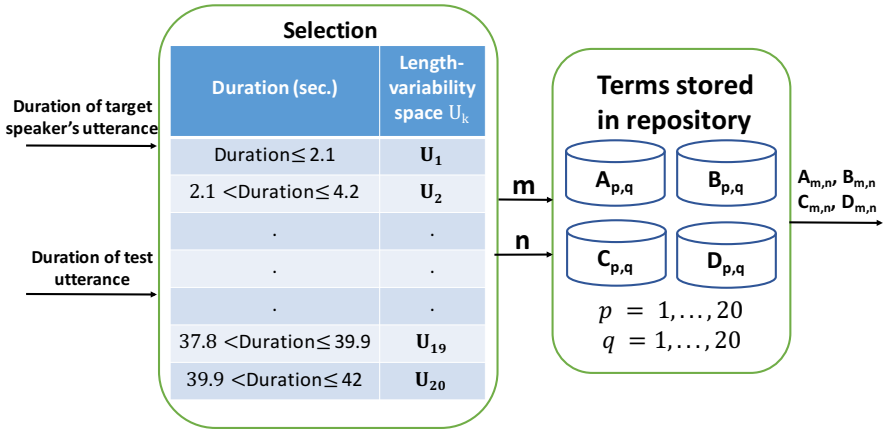

Figure 2: Determination of group identities of target speaker's i-vector and test i-vector based on their utterance duration. In the figure, $K = 20$ in Eq. 24.

Method	K	Male (CC2)			
		EER	minDCF	Time	Mem.
PLDA	-	7.77	0.654	412.7	0.01
PLDA-UP	-	5.75	0.644	14763.9	1.09
Sys. 2	20	7.02	0.685	490.0	0.17
	25	6.35	0.658	517.8	0.28
	30	6.07	0.632	541.1	0.40
	35	5.97	0.631	534.8	0.55
	40	5.93	0.641	545.0	0.72
	45	5.89	0.642	536.1	0.90

Table 3: EER, minDCF, scoring time (in seconds) and memory consumption (in gigabytes) of PLDA, PLDA with standard UP, and PLDA-UP with fast scoring (Sys. 2).

5. Results and Analysis

5.1. Performance of Fast Scoring Systems

Table 2 shows the EER and the minDCF achieved by conventional PLDA, PLDA with UP, and the three fast scoring systems. Fig. 3 shows the DET curves of the fast scoring systems with the number of length-variability loading matrices varies from 20 to 45. Three observations can be made from Table 2 and Fig. 3:

1) The choice of group-identity indicators does affect the performance of the proposed method. When the number of length-variability loading matrices is small (Graph 1), using the largest eigenvalue as the group-identity indicator outperforms using the other two indicators. The performance gaps between the three systems become closer when the number of length-variability loading matrices increases. For Sys. 1, however, when the number of loading matrices K increases from 30 to 35, a non-trivial increase in EER is observed. The performance dips of Sys. 1 are also observed in the DET curves of Graph 4 and Graph 5 in Fig. 3. This kind of performance dip does not happen in the other two systems, which suggests that using duration as group-identity indicators is inappropriate.

2) The performance gap between PLDA-UP with and without fast scoring depends on the number of length-variability loading matrices. When K is small (e.g., $K = 20$), although all of the three systems outperform conventional PLDA, they are still not as good as PLDA-UP. When K is larger than 25, there is no significant difference between PLDA-UP and the proposed fast scoring method (except for the problematic Sys. 1). For Sys. 2 and Sys. 3, although their EER are slightly higher than that of PLDA-UP, their minDCF are lower. The DET curves also suggest that Sys. 2 and Sys. 3 are as good as PLDA-UP.

3) Performance of PLDA-UP with fast scoring becomes saturated when the number of length-variability loading matrices is over 25. This suggests that, as far as performance is concerned, we only need a small number of length-variability loading matrices.

From the observations above, we conclude that the proposed fast scoring method can perform as good as standard UP, provided that the number of length-variability loading matrices are large enough. Because Sys. 1 is problematic under some situations and Sys. 3 requires extra computation to perform eigen-decomposition, we will focus on Sys. 2 in the sequel.

5.2. Running Time and Memory Consumption

Table 3 shows the running time and memory consumption of conventional PLDA, PLDA with standard UP, and PLDA-UP with fast scoring (Sys. 2).[3] We can see that although UP has an overwhelming advantage in performance, its computational cost is also overwhelming. Specifically, standard UP takes almost 35 times longer to finish the scoring. Beside, the memory required to store the covariance matrices of enrollment utterances also poses a problem.

The fast scoring that we proposed avoids to evaluate Eq. 18 during verification. Therefore, its computation complexity is the same as conventional PLDA. Besides, the proposed method only needs to store a modest number of the pre-computed terms: $\mathbf{A}_{p,q}, \mathbf{B}_{p,q}, \mathbf{C}_{p,q}, D_{p,q}$, where $p, q = 1, \ldots, K$. Table 3 shows that the fast scoring system only takes slightly more time than conventional PLDA. As shown in Table 3, when $K = 30$, fast

[3]The running time is the total time for the whole evaluation in CC2 of NIST 2012 SRE.

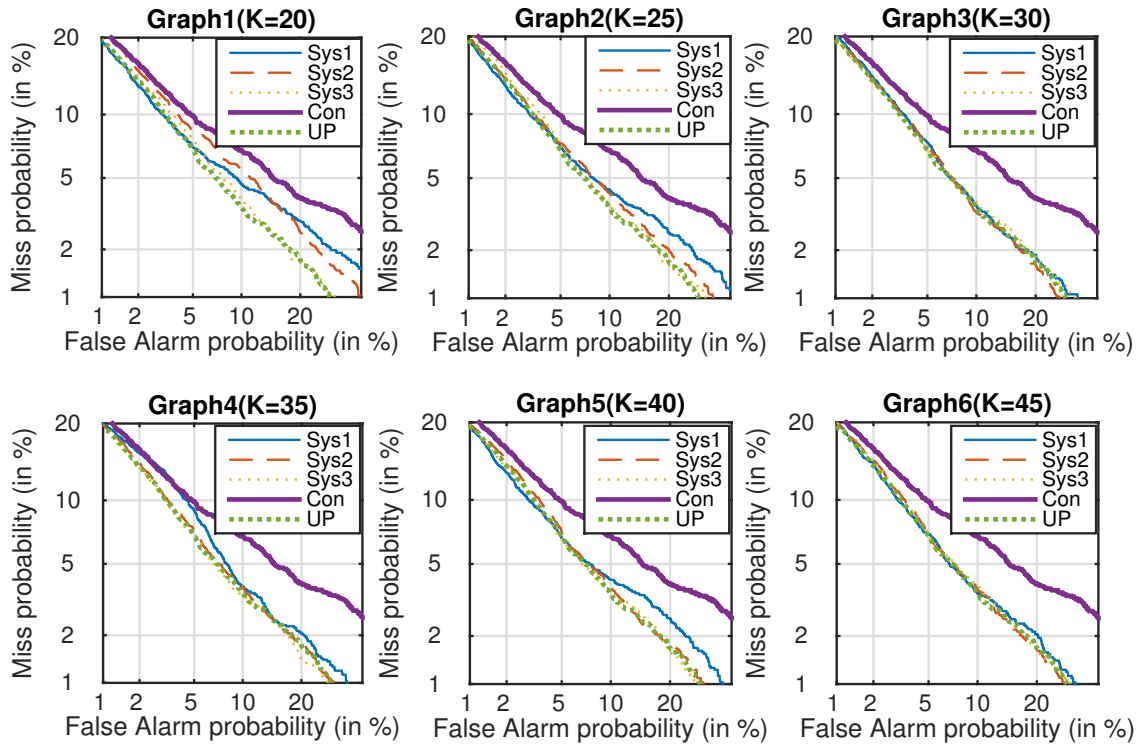

Figure 3: DET curves of three fast scoring systems with the number of length-variability loading matrices (K) varies from 20 to 45. DET curves of conventional PLDA and and PLDA with UP are also plotted as a reference. The characteristics of $\mathbf{U}\mathbf{U}^{\mathsf{T}}$ are quantified by α in Eq. 23. *Sys. 1*: α = utterance-length; *Sys. 2*: α = mean of the diagonal elements of $\mathbf{U}\mathbf{U}^{\mathsf{T}}$; *Sys. 3*: α = largest eigenvalue of $\mathbf{U}\mathbf{U}^{\mathsf{T}}$. *Con*: Conventional PLDA; *UP*: PLDA with UP but no fast scoring (Eq. 17).

scoring only consumes 37% memory that standard UP needs for performing the evaluation under CC2 and the scoring time is 3.7% of the standard UP, while the performance is as good as standard UP. In other words, the proposed method saves both computation resource and memory space of speaker verification systems while maintain the state-of-art performance on utterances with arbitrary length.

6. Conclusions

This paper proposed a fast scoring method for PLDA with uncertainty propagation. By substituting the session-dependent loading matrix with the one trained from development data, the proposed method enables us to pre-compute terms that reflect the length-variability in i-vectors. Thus, the computational burden during the verification can be greatly reduced. Experiments on NIST 2012 SRE show that the proposed method can perform as good as standard UP with a tiny fraction of scoring time that UP takes. Besides, the memory consumption of the proposed method does not increase quadratically with the i-vector dimension as would be the case in standard UP. This is an advantage for speaker verification systems that have a large number of target speakers.

7. References

[1] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Peng Li, Yun Fu, U. Mohammed, J.H. Elder, and S.J.D. Prince, "Probabilistic models for inference about identity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 144–157, 2012.

[3] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010, pp. 1–14.

[4] Man-Wai Mak, Xiaomin Pang, and Jen-Tzung Chien, "Mixture of PLDA for noise robust i-vector speaker verification," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 24, no. 1, pp. 130–142, 2016.

[5] Na Li and Man-Wai Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 10, pp. 1648–1659, 2015.

[6] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Mohammad Jahangir Alam, and Pierre Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7649–7653.

[7] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "I-vector/PLDA variants for text-dependent speaker recognition," Tech. Rep., CRIM, June 2013.

[8] Sandro Cumani, Oldrich Plchot, and Pietro Laface, "On the use of i–vector posterior distributions in probabilistic linear discriminant analysis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 4, pp. 846–857, 2014.

[9] Sandro Cumani, "Fast scoring of full posterior PLDA models," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 2036–2045, 2015.

[10] NIST, "The NIST year 2012 speaker recognition evaluation plan," `http://www.nist.gov/itl/iad/mig/sre12.cfm`.

[11] Patrick Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[12] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.

[13] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[14] Patrick Kenny, "A small footprint i-vector extractor," in *Proc. Odyssey*, 2012, vol. 2012.

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

[16] M. McLaren, M.I. Mandasari, and D.A. Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012, pp. 55–61.

[17] Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition.," in *Interspeech*, 2006.

[18] Man-Wai Mak, "Lecture notes on uncertainty propagation for i-vector/PLDA speaker verification," `http://www.eie.polyu.edu.hk/~mwmak/papers/UncertaintyProp.pdf`, 2015.

[19] Ian L Dryden, Alexey Koloydenko, and Diwei Zhou, "Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging," *The Annals of Applied Statistics*, pp. 1102–1123, 2009.

[20] MathWorks, "Box plot documentation," `http://www.mathworks.com/help/stats/boxplot.html`.

[21] Man-Wai Mak and Hon-Bill Yu, "A study of voice activity detection techniques for nist speaker recognition evaluations," *Computer Speech & Language*, vol. 28, no. 1, pp. 295–313, 2014.

[22] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, Jun. 2001, pp. 213–218.

[23] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki, "The DET curve in assessment of detection task performance," Tech. Rep., DTIC Document, 1997.