

Fast Scoring for PLDA with Uncertainty Propagation via I-vector Grouping

Wei-wei Lin^a, Man-Wai Mak^a, Jen-Tzung Chien^b

^a*Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University*

^b*Dept. of Electrical and Computer Engineering,
The National Chiao Tung University*

Abstract

The i-vector/PLDA framework has gained huge popularity in text-independent speaker verification. This approach, however, lacks the ability to represent the reliability of i-vectors. As a result, the framework performs poorly when presented with utterances of arbitrary duration. To address this problem, a method called uncertainty propagation (UP) was proposed to explicitly model the reliability of an i-vector by an utterance-dependent loading matrix. However, the utterance-dependent matrix greatly complicates the evaluation of likelihood scores. As a result, PLDA with UP, or PLDA-UP in short, is far more computational intensive than the conventional PLDA. In this paper, we propose to group i-vectors with similar reliability, and for each group the utterance-dependent loading matrices are replaced by a representative one. This arrangement allows us to pre-compute a set of representative matrices that cover all possible i-vectors, thereby greatly reducing the computational cost of PLDA-UP while preserving its ability in discriminating the reliability of i-vectors. Experiments on NIST 2012 SRE show that the proposed method can perform as good as the PLDA with UP while the scoring time is only 3.18% of it.

Keywords: Speaker verification, i-vector/PLDA, Uncertainty Propagation, duration mismatch.

1. Introduction

Recent years have witnessed the significant advances in text-independent speaker recognition. With the state-of-the-art techniques like i-vector, PLDA and DNN acoustic models, an EER of 0.59% on NIST speaker recognition evaluation has been reported [1]. Despite of these great advances, short-utterance speaker recognition remains a great challenge, as evident by a number of studies showing that system performance degrades rapidly when only short utterances are available [2, 3, 4]. However, in real applications users may not be willing to provide long utterances, especially during verification.

It has now become clear that naive applications of advanced text-independent methods, such as i-vector/PLDA, to short-utterance speaker verification could result in performance even poorer than that of the GMM and HMM modeling [5, 6]. One of the problems associated with short-utterance speaker verification is duration mismatch, where the length of enrolment utterances and test utterances are very different. Hasan *et al.* [7] assumed that duration mismatches can cause a shift in PLDA scores and proposed a duration-dependent quality measure function to compensate for the shift. Kanagasundaram *et al.* [4] compared joint factor analysis (JFA), i-vector PLDA, and i-vectors equipped with various subspace projections and variance normalization techniques under short-utterance scenarios. They found that no significant performance difference between JFA and i-vector PLDA when the enrollment and test utterances are very short and that JFA and PLDA offer marginally better performance than i-vectors with LDA followed by WCCN. Li *et al.* [8] noticed that for GMM-UBM systems, when both enrollment and test utterances are very short, the Gaussian components covered by the test utterances will not be properly trained during enrollment. To address this problem, they proposed to distribute speech signals into a number of phonetic sub-regions and model speakers within the sub-regions by region-specific GMMs.

A special concern for i-vector/PLDA is that it has no ability to represent the reliability of i-vectors. This problem is especially severe in short-utterance

31 speaker verification. Recall that an i-vector is a maximum-a-posteriori (MAP)
32 estimate of the latent variable in a factor analysis model. For short utterances,
33 the number of acoustic frames is not enough to estimate i-vectors reliably. By
34 ignoring the time dimension, i-vectors estimated from long and short utterances
35 are essentially treated as equally reliable. Kenny *et al.* [9] proposed to tightly
36 couple i-vector extraction with PLDA modelling instead of treating them as two
37 separated procedures. Specifically, the posterior covariance matrix of the latent
38 factor is propagated into the PLDA model by introducing an extra loading ma-
39 trix to represent the reliability of the i-vector. The method is called uncertainty
40 propagation (UP) and the modified PLDA model is called PLDA-UP in this
41 paper.

42 The extra loading matrix in PLDA-UP is utterance-dependent. As a result,
43 the scoring of PLDA-UP is much more computationally intensive than conven-
44 tional PLDA. Besides, PLDA-UP also requires to store the posterior covariance
45 matrices of target-speakers' i-vectors, which is much more memory consuming
46 than storing the i-vectors alone. Thus, both computational cost and memory
47 consumption restrict the applications of PLDA-UP. To reduce the computational
48 cost of PLDA-UP, Cumani *et al.* [10] proposed using MAP-estimated i-vectors
49 to represent target speakers and propagating the posterior covariance matrix of
50 test utterances into the PLDA model. This method relies on the assumption
51 that enrolment utterances tend to be long. In [11], the author proposed to di-
52 agonalise the matrices involved in scoring to reduce the computational cost of
53 full matrix operations. Although this approach significantly reduces the com-
54 putational cost and does not require long enrolment utterances, it still degrades
55 the performance of PLDA-UP when test utterances are very short.

56 The utterance-dependent matrix in PLDA-UP has no speaker specific in-
57 formation. The only role it plays is to convey the reliability of i-vector. In-
58 tuitively, if two utterances are close in duration, the corresponding i-vectors
59 should have similar reliability. Based on this assumption, we have proposed in
60 [12] to group i-vectors according to their utterance durations and model the
61 reliability of i-vectors in each group by a single representative loading matrix.

62 Because these representative loading matrices can be pre-computed based on
63 development data, we can pre-compute all of the relevant terms during scoring,
64 thus saving lots of computation. In this paper, we extend our previous work in
65 the following aspects:

- 66 • We introduce a metric for measuring the distance between two covari-
67 ance matrices. Through this metric, we define a within-group distance to
68 measure the quality of the grouping schemes.
- 69 • More extensive experiments are carried out to compare the performance of
70 different grouping schemes. Also, the effectiveness of PLDA-UP and the
71 proposed fast scoring schemes on utterances with different length-ranges
72 was investigated.

73 Experimental results on the NIST 2012 SRE show that the proposed method
74 can perform as good as the PLDA-UP in all four different length-ranges inves-
75 tigated, and the scoring time can be as low as 3.18% of the PLDA-UP.

76 The organization of this paper is as follows. In Section 2 and Section 3,
77 we give a brief review of i-vector/PLDA framework and PLDA-UP. We show
78 why PLDA-UP can deal with length variability and the source of computational
79 burden is also identified. We then present the proposed fast scoring schemes
80 in Section 4. Experimental setup and results are presented in Section 5 and
81 Section 6, respectively. Finally, we conclude our findings in Section 7.

82 **2. Review of I-vector/PLDA**

83 *2.1. I-vector Extraction*

84 The i-vector approach is an extension of joint factor analysis [13, 14]. It aims
85 to extract from the acoustic vectors of an utterance a low-dimensional vector
86 that incorporates most of the speaker information. It assumes that the speaker-
87 and channel-dependent GMM-supervectors live in a low dimensional space:

$$\beta = \mathbf{m} + \mathbf{T}\eta, \quad (1)$$

88 where \mathbf{m} is the speaker- and channel-independent GMM-supervector constructed
 89 by stacking up the means of a universal background model (UBM); \mathbf{T} is a low-
 90 rank total variability matrix whose columns span the subspace where speaker-
 91 and channel-specific information varies; $\boldsymbol{\eta}$ is a latent variable which is assumed
 92 to follow a standard normal distribution. Given an utterance, its i-vector is a
 93 maximum-a-posteriori (MAP) estimate of the latent variable $\boldsymbol{\eta}$, which we de-
 94 note as $\boldsymbol{\omega}$. To estimate an i-vector of an utterance with T acoustic frames,
 95 $\mathcal{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_T\}$, the Baum-Welch statistics are used:

$$N_c = \sum_{t=1}^T \gamma_c(\mathbf{o}_t) \quad (2)$$

$$\tilde{\mathbf{f}}_c = \sum_{t=1}^T \gamma_c(\mathbf{o}_t)(\mathbf{o}_t - \mathbf{m}_c), \quad c = 1, \dots, C \quad (3)$$

96 where

$$\gamma_c(\mathbf{o}_t) = \frac{\lambda_c \mathcal{N}(\mathbf{o}_t | \mathbf{m}_c, \boldsymbol{\Sigma}_c)}{\sum_{c=1}^C \lambda_c \mathcal{N}(\mathbf{o}_t | \mathbf{m}_c, \boldsymbol{\Sigma}_c)}, \quad (4)$$

97 where \mathbf{m}_c and $\boldsymbol{\Sigma}_c$ are the mean vector and covariance matrix of the c -th mixture
 98 in the UBM. The i-vector $\boldsymbol{\omega}$ and its posterior covariance matrix $\text{cov}(\boldsymbol{\eta}, \boldsymbol{\eta})$ can
 99 be obtained by [13, 15]:

$$\boldsymbol{\omega} = \text{cov}(\boldsymbol{\eta}, \boldsymbol{\eta}) \sum_{c=1}^C \mathbf{T}_c^\top \boldsymbol{\Sigma}_c^{-1} \tilde{\mathbf{f}}_c \quad (5)$$

$$\text{cov}(\boldsymbol{\eta}, \boldsymbol{\eta}) = \mathbf{L}^{-1} = \left(\mathbf{I} + \sum_{c=1}^C N_c \mathbf{T}_c^\top \boldsymbol{\Sigma}_c^{-1} \mathbf{T}_c \right)^{-1}, \quad (6)$$

100 where \mathbf{L} is a precision matrix and \mathbf{T}_c is the c -th partition of \mathbf{T} , i.e. $\mathbf{T} =$
 101 $[\mathbf{T}_1^\top, \dots, \mathbf{T}_C^\top]^\top$.

102 2.2. Probabilistic Linear Discriminant Analysis

103 To suppress undesired intra-speaker variability in i-vectors, channel compen-
 104 sation is applied. Probabilistic linear discriminant analysis (PLDA) is found to
 105 be the most effective. Because of the heavy-tailed behaviour of i-vector distribu-
 106 tions, early PLDA is based on Students's t distribution [16]. Garcia-Romero and

107 Espy-Wilson [17] found later that by simply length-normalizing the i-vectors,
 108 Gaussian PLDA can perform equally well. Because of the nice analytical solu-
 109 tion that Gaussian PLDA can offer, it is more preferable in practice.

110 2.2.1. Pre-processing for Gaussian PLDA

111 To use Gaussian PLDA, two pre-processing steps are necessary to Gaus-
 112 sianalize i-vectors. First, a whitening transform is applied to i-vectors:

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

113 where $\bar{\boldsymbol{\omega}}$ is the global mean of i-vectors, \mathbf{W} is a transformation matrix obtained
 114 from the Cholesky decomposition of the within-class covariance matrix of i-
 115 vectors [18] and $\boldsymbol{\omega}^{\text{wht}}$ is the whitened i-vector. The second step is to apply a
 116 simple length-normalization to the whitened i-vectors:

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

117 It is customary to include linear discriminant analysis (LDA) and within-class
 118 covariance normalization (WCCN) [18] in the pre-processing steps. The whole
 119 pre-processing can be written in a more succinct fashion:

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

120 where \mathbf{P} denotes the transformation matrix that combines whitening, LDA and
 121 WCCN and \mathbf{w} is the pre-processed i-vector that is ready for PLDA modelling.

122 2.2.2. Gaussian PLDA as a Generative Model

123 Given R i-vectors $\{\mathbf{w}_r; r = 1, \dots, R\}$ from a speaker, PLDA assumes that
 124 they can be decomposed in the following manner:

$$\mathbf{w}_r = \boldsymbol{\mu} + \mathbf{V}\mathbf{h} + \mathbf{G}\mathbf{z}_r + \boldsymbol{\epsilon}_r. \quad (10)$$

125 This decomposition has two distinct parts: (1) the speaker-dependent part,
 126 $\boldsymbol{\mu} + \mathbf{V}\mathbf{h}$, which is the same for all i-vectors from the same speaker; (2) the
 127 utterance-dependent part, $\mathbf{G}\mathbf{z}_r + \boldsymbol{\epsilon}_r$, which varies even for the utterances from

128 the same speaker. In Eq. 10, $\boldsymbol{\mu}$ is the global mean of i-vectors and the matrix
 129 \mathbf{V} represents the speaker subspace on which the speaker factor \mathbf{h} can vary. The
 130 columns of matrix \mathbf{U} span the subspace where the channel factor \mathbf{z}_r varies. $\boldsymbol{\epsilon}_r$
 131 models the residue that is not captured by both speaker and channel subspaces
 132 and is assumed to follow a Gaussian distribution with zero mean and a diagonal
 133 covariance matrix.

134 The low dimensionality of i-vector makes it possible to conflate the channel
 135 variability and residue by using a full covariance matrix $\boldsymbol{\Sigma}$ such that:

$$\mathbf{w}_r = \boldsymbol{\mu} + \mathbf{V}\mathbf{h} + \boldsymbol{\epsilon}_r, \quad \boldsymbol{\epsilon}_r \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (11)$$

136 2.2.3. Scoring in Gaussian PLDA

137 Given a target speaker’s i-vector \mathbf{w}_s and a test i-vector \mathbf{w}_t , the log-likelihood
 138 ratio of the same-speaker hypothesis to different-speaker hypothesis can be com-
 puted by [17]:

$$\begin{aligned} S_{LR}(\mathbf{w}_s, \mathbf{w}_t) &= \log \frac{p(\mathbf{w}_s, \mathbf{w}_t | \text{same-speaker})}{p(\mathbf{w}_s, \mathbf{w}_t | \text{different-speaker})} \\ &= \frac{1}{2} \mathbf{w}_s^T \boldsymbol{\Phi} \mathbf{w}_s + \mathbf{w}_s^T \boldsymbol{\Psi} \mathbf{w}_t + \frac{1}{2} \mathbf{w}_t^T \boldsymbol{\Phi} \mathbf{w}_t + \text{const} \end{aligned} \quad (12)$$

where

$$\boldsymbol{\Phi} = \boldsymbol{\Sigma}_{tot}^{-1} - (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac} \boldsymbol{\Sigma}_{tot}^{-1} \boldsymbol{\Sigma}_{ac})^{-1} \quad (13)$$

$$\boldsymbol{\Psi} = \boldsymbol{\Sigma}_{tot}^{-1} \boldsymbol{\Sigma}_{ac} (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac} \boldsymbol{\Sigma}_{tot}^{-1} \boldsymbol{\Sigma}_{ac})^{-1} \quad (14)$$

$$\boldsymbol{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^T \quad \boldsymbol{\Sigma}_{tot} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Sigma}. \quad (15)$$

139 Note that Eqs. 13–14 can be computed beforehand. Only Eq. 12 needs to be
 140 evaluated during verification. As a result, PLDA scoring is very efficient.

141 3. Gaussian PLDA with Uncertainty Propagation

142 Despite the great success of the i-vector/PLDA framework, its performance
 143 becomes very poor if both the enrolment and test utterances have a wide range
 144 of durations. There are several reasons for this. First, in i-vector extraction,

145 the duration of utterances is totally ignored, i.e., utterances are represented by
 146 vectors of fixed dimension regardless of their duration. Recall that an i-vector is
 147 the MAP estimate of latent variable $\boldsymbol{\eta}$; the accuracy of such estimate depends on
 148 the number of acoustic vectors. By ignoring durations, all i-vectors are treated
 149 as equally reliable. Second, in PLDA modelling, it is assumed that all of the
 150 intra-speaker variabilities are represented by the covariance matrix $\boldsymbol{\Sigma}$, which is
 151 the same across all i-vectors. This is apparently not a satisfactory assumption
 152 because short utterances have more severe intra-speaker variabilities than long
 153 utterances.

154 To better accommodate utterance-length variability, a modified PLDA is
 155 proposed in [9]. The basic idea is to tightly couple i-vector extraction and PLDA
 156 modelling by propagating the uncertainty during i-vector extraction into the
 157 PLDA model. Recall that the posterior covariance matrix in Eq. 6 represents the
 158 uncertainty of the MAP point-estimate in i-vector extraction. The shorter the
 159 utterance, the larger the posterior covariances. By propagating this information
 160 into PLDA and using a loading matrix to model the variability due to duration
 161 variation, this PLDA model can better handle the length-variability than the
 162 conventional PLDA model.

163 *3.1. Preprocessing for Gaussian PLDA with UP*

164 The pre-processing steps in Section. 2.2.1 also need to be applied to the
 165 posterior covariance matrices. If only linear transform \mathbf{P} is applied to an i-
 166 vector, the corresponding pre-processed covariance matrix can be obtained by:

$$167 \quad \text{cov}(\mathbf{P}\boldsymbol{\eta}, \mathbf{P}\boldsymbol{\eta}) = \mathbf{P}\mathbf{L}^{-1}\mathbf{P}^{\text{T}}, \quad (16)$$

168 which we denote as $\boldsymbol{\Lambda}$. When length-normalization is applied to an i-vector, the
 169 pre-processed covariance matrix can be approximated by [9]:

$$170 \quad \boldsymbol{\Lambda} \leftarrow \frac{\mathbf{P}\mathbf{L}^{-1}\mathbf{P}^{\text{T}}}{\|\boldsymbol{\omega}^{\text{wht}}\|}. \quad (17)$$

170 Other methods to deal with this non-linear transform on posterior matrix can
 171 be found in [9, 19].

172 *3.2. Generative Model for Gaussian PLDA with UP*

173 To propagate the uncertainty of an i-vector into the PLDA model, an utterance-
174 dependent loading matrix is added to the factor analysis model:

$$\mathbf{w}_r = \boldsymbol{\mu} + \mathbf{V}\mathbf{h} + \mathbf{U}_r\mathbf{z}_r + \boldsymbol{\epsilon}_r, \quad (18)$$

175 where \mathbf{U}_r is the Cholesky decomposition of the posterior covariance matrix $\boldsymbol{\Lambda}_r$,
176 and \mathbf{z}_r is a latent variable assumed to follow a standard normal distribution.
177 The intra-speaker variability of \mathbf{w}_r in Eq. 18 is:

$$\text{cov}(\mathbf{w}_r, \mathbf{w}_r | \mathbf{h}) = \boldsymbol{\Lambda}_r + \boldsymbol{\Sigma}, \quad (19)$$

178 where $\boldsymbol{\Lambda}_r$ varies from utterances to utterances, thus reflecting the reliability of
179 i-vector \mathbf{w}_r .

180 *3.3. Scoring in Gaussian PLDA with UP*

181 Given a target speaker's i-vector \mathbf{w}_s together with its posterior covariance
182 matrix $\boldsymbol{\Lambda}_s$ and a test i-vector \mathbf{w}_t together with its posterior covariance matrix
 $\boldsymbol{\Lambda}_t$, the log-likelihood ratio can be written as:

$$\begin{aligned} S_{LR}(\mathbf{w}_s, \mathbf{w}_t; \boldsymbol{\Lambda}_s, \boldsymbol{\Lambda}_t) &= \log \frac{p(\mathbf{w}_s, \mathbf{w}_t; \boldsymbol{\Lambda}_s, \boldsymbol{\Lambda}_t | \text{same-speaker})}{p(\mathbf{w}_s, \mathbf{w}_t; \boldsymbol{\Lambda}_s, \boldsymbol{\Lambda}_t | \text{different-speaker})} \\ &= \log p \left(\begin{bmatrix} \mathbf{w}_s \\ \mathbf{w}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_s & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Sigma}_t \end{bmatrix} \right) \\ &\quad - \log p \left(\begin{bmatrix} \mathbf{w}_s \\ \mathbf{w}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_s & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_t \end{bmatrix} \right) \\ &= \frac{1}{2} \mathbf{w}_s^\top \mathbf{A}_{s,t} \mathbf{w}_s + \mathbf{w}_s^\top \mathbf{B}_{s,t} \mathbf{w}_t + \frac{1}{2} \mathbf{w}_t^\top \mathbf{C}_{s,t} \mathbf{w}_t + D_{s,t} \quad (20) \end{aligned}$$

where

$$\mathbf{A}_{s,t} = \Sigma_s^{-1} - (\Sigma_s - \Sigma_t^{-1} \Sigma_{ac})^{-1} \quad (21)$$

$$\mathbf{B}_{s,t} = \Sigma_s^{-1} \Sigma_{ac} (\Sigma_t - \Sigma_{ac} \Sigma_s^{-1} \Sigma_{ac})^{-1} \quad (22)$$

$$\mathbf{C}_{s,t} = \Sigma_t^{-1} - (\Sigma_t - \Sigma_s^{-1} \Sigma_{ac})^{-1} \quad (23)$$

$$D_{s,t} = -\frac{1}{2} \log \begin{vmatrix} \Sigma_s & \Sigma_{ac} \\ \Sigma_{ac} & \Sigma_t \end{vmatrix} + \frac{1}{2} \log \begin{vmatrix} \Sigma_s & \mathbf{0} \\ \mathbf{0} & \Sigma_t \end{vmatrix} \quad (24)$$

$$\Sigma_t = \mathbf{V} \mathbf{V}^\top + \Lambda_t + \Sigma \quad (25)$$

$$\Sigma_s = \mathbf{V} \mathbf{V}^\top + \Lambda_s + \Sigma \quad (26)$$

$$\Sigma_{ac} = \mathbf{V} \mathbf{V}^\top. \quad (27)$$

183 It is worth to notice that Eqs. 21–24 involve terms dependent on both the target
 184 speaker’s utterance and the test utterance, which means that these terms need
 185 to be evaluated during scoring.

186 4. Fast Scoring via I-vector Grouping

187 The computational burden of PLDA-UP comes from the utterance-dependent
 188 loading matrix \mathbf{U}_r in Eq. 18, where the uncertainty is represented by $\mathbf{U}_r \mathbf{U}_r^\top$.
 189 If we have a group of i-vectors with similar reliability, one prescribed loading
 190 matrix should be sufficient to model the reliability of all of the i-vectors in the
 191 group. Furthermore, if the prescribed loading matrix can be estimated from
 192 development data, the utterance-dependent terms in Eqs. 21–24 can be pre-
 193 computed, which would greatly speed up the scoring process.

194 Suppose we have a collection of i-vectors from a speaker and they are dis-
 195 tributed into K groups indexed by k , with the members within the k -th group
 196 indexed by (k, i) . Then, the factor analysis model can be written as:

$$\mathbf{w}_{k,i} = \boldsymbol{\mu} + \mathbf{V} \mathbf{h} + \mathbf{U}_k \mathbf{z}_{k,i} + \boldsymbol{\epsilon}_{k,i}, \quad (28)$$

197 where the loading matrices $\{\mathbf{U}_k\}_{k=1}^K$ are obtained from development data. Dif-
 198 ferent grouping schemes [12] will be explored in this paper:

- 199 • Grouping i-vectors by utterance durations.
- 200 • Grouping i-vectors by the characteristics of the posterior covariance ma-
- 201 trices.

202 4.1. Three Approaches to Grouping I-vectors

203 In this section, we describe and assess the quality of proposed grouping
 204 schemes. The first scheme is based on utterance durations and the last two are
 205 based on the characteristics of posterior covariance matrices.

206 One intuitive way to group i-vectors with similar reliability is to group them
 207 according to the durations of their utterances. This can be easily done by
 208 dividing the time axis (starting from the shortest duration) into a number of
 209 equal-length intervals. Then, for each interval, the uncertainties of i-vectors are
 210 represented by the posterior covariance matrix of the i-vector whose utterance
 211 duration falls on or nearest to the middle of that interval. For example, if the
 212 interval is between 10 to 20 seconds, we select the covariance matrix whose
 213 corresponding utterance duration is closest to 15 seconds.

214 Suppose the time axis is divided into K equal-length time intervals indexed
 215 by k . Then, the i -th i-vector in the k -th interval is denoted as $\mathbf{w}_{k,i}$ and its pre-
 216 processed posterior covariance matrix is denoted as $\mathbf{\Lambda}_{k,i}$ ¹ where $i = 1, 2, \dots, I_k$.
 217 Among the I_k posterior covariance matrices in the k -th interval, the one with
 218 utterance-length closest to the middle of the k -th interval is selected to represent
 219 the uncertainty of all the i-vectors insider the interval. We denote the selected
 220 matrix as $\mathbf{\Lambda}_{k,r}$. As $\mathbf{\Lambda}_{k,r}$ represents the uncertainty of all of the i-vectors insider
 221 the k -th interval, we need to assume:

$$\mathbf{\Lambda}_{k,i} \approx \mathbf{\Lambda}_{k,r} \quad \forall i \neq r. \quad (29)$$

222 To see if the above assumption holds, we introduce a within-group distance
 223 $d(\mathbf{\Lambda}_{k,i}, \mathbf{\Lambda}_{k,r})$ to measure the distances between the selected matrix and other

¹For simplicity, in the sequel we will refer the pre-processed posterior covariance matrix in Eq. 17 as the posterior covariance matrix $\mathbf{\Lambda}$ when the context is clear.

224 matrices in the k -th group [20]:

$$d(\mathbf{\Lambda}_{k,i}, \mathbf{\Lambda}_{k,r}) = \sqrt{\frac{\text{trace}\{(\mathbf{\Lambda}_{k,i} - \mathbf{\Lambda}_{k,r})^\top (\mathbf{\Lambda}_{k,i} - \mathbf{\Lambda}_{k,r})\}}{\text{trace}\{(\mathbf{\Lambda}_{k,i}^\top \mathbf{\Lambda}_{k,i}) + (\mathbf{\Lambda}_{k,r}^\top \mathbf{\Lambda}_{k,r})\}}} \quad i \neq r. \quad (30)$$

225 Note that the distance has a range between 0.0 and 1.0 such that the smaller the
 226 distance the more similar are the two matrices. We truncated 7,156 telephone
 227 conversations from NIST 2008–2010 SRE (see Section 6) into short segments so
 228 that their durations are uniformly distributed between 3 and 60 seconds. After i-
 229 vector extraction and pre-processing, we applied the above mentioned procedure
 230 to group i-vectors, i.e., the time axis was divided into five 11.4-second intervals
 231 starting from 3 seconds and ending at 60 seconds. $\mathbf{\Lambda}_{k,i}$, $i = 1, 2, \dots, I_k$, represent
 232 the posterior covariance matrices inside the k -th interval, among which $\mathbf{\Lambda}_{k,r}$
 233 was selected as the representative of the interval. The within-group distances
 234 are computed for $I_k - 1$ pairs of $\mathbf{\Lambda}_{k,r}$ and $\mathbf{\Lambda}_{k,i}$, where $i \neq r$, for a total of
 235 5 groups. The results are presented in Fig. 1(a). Each box together with its
 236 whiskers represent the variability of the within-group distances of that group.
 237 The central mark inside each box indicates the median within-group distance,
 238 and the bottom and top edges of each box indicate the 25th and 75th percentiles,
 239 respectively. The whiskers extend to the most extreme non-outliers, and the
 240 outliers are represented by the ‘+’ symbol [21].

241 We can see from Fig. 1(a) that the majority of the distances are quite small
 242 (75% of the distances are smaller than the value indicated by the upper edge
 243 of each box). As small distance means high similarity between representative
 244 matrix and the other matrices in the group, we conclude that selecting rep-
 245 resentative matrices based on utterance durations is a reasonable approach.
 246 Nevertheless, there are still some outliers in the five groups. The reason for the
 247 outliers is that utterance duration does not totally capture the information in
 248 the posterior covariance matrix. Even for utterances of exactly the same du-
 249 ration, their zero-th order statistics (N_c in Eq. 2) can be quite different, which
 250 could result in different posterior covariance matrices. Even if the posterior
 251 covariance matrices of two i-vectors are exactly the same, i.e., $\mathbf{L}_1^{-1} = \mathbf{L}_2^{-1}$ in
 252 Eq. 6, their post-processed covariance matrices ($\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ in Eq. 17) could

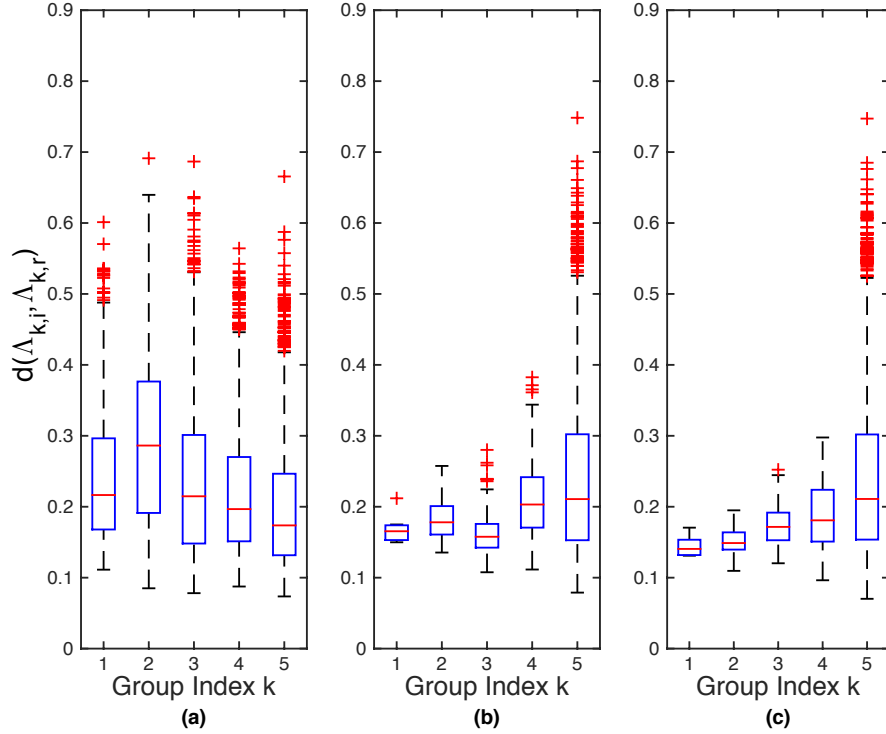


Figure 1: Distances between the representative matrix $\Lambda_{k,r}$ of the k -th group and all of the other matrices in the group. I-vector grouping schemes based on (a) utterance duration, (b) the largest eigenvalue of $\mathbf{U}\mathbf{U}^T$ and (c) the trace of $\mathbf{U}\mathbf{U}^T$.

253 be different. This is because the whitened i-vectors (ω_1^{wht} and ω_2^{wht}) are not
 254 identical in general.

255 To solve these problems, we propose two alternative approaches to grouping
 256 i-vectors using the characteristics of the posterior covariance matrices [12]. To
 257 this end, we define a scalar α , which is a function of the posterior covariance
 258 matrix:

$$\alpha = f(\mathbf{\Lambda}). \quad (31)$$

259 In Eq. 31, α could be:

- 260 1. the largest eigenvalue of $\mathbf{\Lambda}$, because the largest eigenvalue could dominate
- 261 the variances of all components; and

262 2. the trace of $\mathbf{\Lambda}$, because the trace of a covariance matrix is the sum of
 263 its eigenvalues, which summarizes the variability of all components in the
 264 corresponding i-vector.

265 Specifically, we computed α for every posterior covariance matrix after prepro-
 266 cessing. Then we divided the α -axis into K equal-spaced intervals indexed by
 267 k . The i-vectors associated with the k -th interval are denoted as $\mathbf{w}_{k,i}$ and their
 268 posterior covariance matrices are denoted as $\mathbf{\Lambda}_{k,i}$, where $i = 1, 2, \dots, I_k$. The
 269 posterior covariance matrix whose value of α is closest to the middle of the
 270 k -th interval is selected to represent the uncertainty of i-vectors in this inter-
 271 val and denoted as $\mathbf{\Lambda}_{k,r}$. Following this procedure, we divided the i-vectors
 272 extracted from the above mentioned 3–60 seconds utterances into 5 groups us-
 273 ing the largest eigenvalues and matrix traces, respectively. To evaluate the
 274 quality of these two grouping schemes, we compute the within-group distances
 275 $d(\mathbf{\Lambda}_{k,i}, \mathbf{\Lambda}_{k,r})$ for $I_k - 1$ pairs of $\mathbf{\Lambda}_{k,i}$ and $\mathbf{\Lambda}_{k,r}$, where $i \neq r$, for a total of 5 groups.
 276 The results are shown in Fig. 1(b) and Fig. 1(c) for using the largest eigenvalues
 277 and matrix traces, respectively. When compared with Fig. 1(a), there are con-
 278 siderably less outliers in Groups 1–4 in both Fig. 1(b) and Fig. 1(c), although
 279 Group 5 still has a large number of outliers.

280 4.2. Fast Scoring Procedure

281 Given a target speaker’s i-vector \mathbf{w}_s and a test i-vector \mathbf{w}_t , we need to
 282 determine their group index first, which we denoted as m and n , respectively.
 283 For the grouping scheme based on utterance duration, this can be achieved by
 284 comparing their utterance duration, denoted as $l^{(s)}$ and $l^{(t)}$, with the durations
 of the representative matrices, $\{l_k; k = 1, \dots, K\}$:

$$m = \arg \min_{k \in \{1, \dots, K\}} |l_k - l^{(s)}| \quad (32)$$

$$n = \arg \min_{k \in \{1, \dots, K\}} |l_k - l^{(t)}|. \quad (33)$$

285 For the grouping schemes based on the characteristics of the posterior covariance
 286 matrices, we need to evaluate the α -value of target speaker’s posterior covari-
 287 ance matrix $\mathbf{\Lambda}_s$, which we denoted as $\alpha^{(s)}$, and the α -value of test utterance’s

288 posterior covariance matrix $\mathbf{\Lambda}_t$, which we denoted as $\alpha^{(t)}$. Then we compared
 289 $\alpha^{(s)}$ and $\alpha^{(t)}$ with the α -value of the representative matrices, $\{\alpha_k; k = 1, \dots, K\}$,
 to determine the group identities of target speaker and test utterances:

$$m = \arg \min_{k \in \{1, \dots, K\}} |\alpha_k - \alpha^{(s)}| \quad (34)$$

$$n = \arg \min_{k \in \{1, \dots, K\}} |\alpha_k - \alpha^{(t)}|. \quad (35)$$

290 Then the log-likelihood ratio can be written as:

$$S_{LR}(\mathbf{w}_s, \mathbf{w}_t; m, n) = \frac{1}{2} \mathbf{w}_s^T \mathbf{A}_{m,n} \mathbf{w}_s + \mathbf{w}_s^T \mathbf{B}_{m,n} \mathbf{w}_t + \frac{1}{2} \mathbf{w}_t^T \mathbf{C}_{m,n} \mathbf{w}_t + D_{m,n}, \quad (36)$$

where

$$\mathbf{A}_{m,n} = \mathbf{\Sigma}_m^{-1} - (\mathbf{\Sigma}_m - \mathbf{\Sigma}_n^{-1} \mathbf{\Sigma}_{ac})^{-1} \quad (37)$$

$$\mathbf{B}_{m,n} = \mathbf{\Sigma}_m^{-1} \mathbf{\Sigma}_{ac} (\mathbf{\Sigma}_n - \mathbf{\Sigma}_{ac} \mathbf{\Sigma}_m^{-1} \mathbf{\Sigma}_{ac})^{-1} \quad (38)$$

$$\mathbf{C}_{m,n} = \mathbf{\Sigma}_n^{-1} - (\mathbf{\Sigma}_n - \mathbf{\Sigma}_m^{-1} \mathbf{\Sigma}_{ac})^{-1} \quad (39)$$

$$D_{m,n} = -\frac{1}{2} \log \begin{vmatrix} \mathbf{\Sigma}_m & \mathbf{\Sigma}_{ac} \\ \mathbf{\Sigma}_{ac} & \mathbf{\Sigma}_n \end{vmatrix} + \frac{1}{2} \log \begin{vmatrix} \mathbf{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_n \end{vmatrix} \quad (40)$$

$$\mathbf{\Sigma}_n = \mathbf{V} \mathbf{V}^T + \mathbf{\Lambda}_n + \mathbf{\Sigma} \quad (41)$$

$$\mathbf{\Sigma}_m = \mathbf{V} \mathbf{V}^T + \mathbf{\Lambda}_m + \mathbf{\Sigma} \quad (42)$$

$$\mathbf{\Sigma}_{ac} = \mathbf{V} \mathbf{V}^T. \quad (43)$$

291 Because Eqs. 37–40 do not depend on the test utterance, they can be pre-
 292 computed. For the grouping scheme based on utterance duration, the only extra
 293 computation is Eq. 33 during verification. For the grouping schemes based on
 294 covariance matrix’s characteristics, we need to evaluate Eq. 31 and Eq. 35.

295 5. Experimental Setup

296 5.1. Acoustic Front-End Processing

297 Speech data from NIST 2005–2010 Speaker Recognition Evaluation (SRE)
 298 were used for system development. For performance evaluation, NIST 2012 SRE

299 [22] were used. For each utterance, a two-channel voice activity detector (VAD)
300 [23] was applied to remove silent regions. Then a 25-ms Hamming window
301 was used to extract 19 mel frequency cepstral coefficients (MFCC) and log-
302 energy plus their first and second derivatives. Cepstral mean normalization and
303 feature warping [24] were applied to compensate for channel variability in the
304 MFCC vectors. In order to simulate utterances with arbitrary duration, four
305 set of utterances with duration ranging from 3–20 seconds, 3–30 seconds, 3–40
306 seconds and 3–60 seconds, respectively, were created by truncating speech files
307 from NIST 2012 SRE (core set, male speaker).

308 *5.2. Speaker Model Training*

309 Full-length microphone and telephone utterances from NIST 2005–2008 SREs
310 were used to train a gender-dependent UBM with 1024 Gaussian components
311 and an i-vector extractor with 500 total factors. Then, i-vectors were extracted
312 from the above mentioned truncated speech files. WCCN together with length-
313 normalization were applied to reduce the heavy-tailed behavior of i-vectors.
314 LDA was applied to project the i-vectors to a 200 dimensional subspace with
315 better speaker discrimination. Another WCCN was then applied to reduce
316 the undesired high within-class variability in the LDA-projected space. Then a
317 PLDA models were trained using the pre-processed i-vectors (Eq. 9). PLDA-UP
318 model was trained using the pre-processed i-vectors together with their posterior
319 covariance matrices. For fast scoring systems, we obtained the representative
320 matrices from the truncated telephone utterances in NIST 2006–2010 SRE, fol-
321 lowing the procedures described in Section 4. According to different schemes
322 specified in Table. 5.2, we have three fast scoring systems.

323 **6. Results and Analysis**

324 System performance was based on the truncated speech segments of Common
325 Conditions 2 and 4 of NIST 2012 SRE (core set, male speakers). Equal error
326 rate (EER), minimum detection cost function (minDCF) in NIST 2012 SRE
327 were used as performance metrics.

System	Criteria for Grouping i-vectors
Sys. 1	Utterance length (after VAD)
Sys. 2	The largest eigenvalue of posterior covariance matrix
Sys. 3	The trace of posterior covariance matrix

Table 1: The criteria for grouping i-vectors used by the 3 systems.

328 Fig 2 shows a bar chart of the EERs and total scoring time of PLDA, PLDA-
329 UP and the three fast scoring systems with different numbers of i-vector groups.
330 Obviously, the bar chart suggests that our fast scoring systems significantly
331 reduce the scoring time while maintaing the good performance of PLDA-UP.
332 The following sub-sections gives a detailed analysis of the results.

333 6.1. Performance of Fast Scoring Systems

334 Table 2 shows the EER and minDCF obtained by PLDA, PLDA-UP and
335 the three fast scoring systems in common conditions 2 and 4, respectively. The
336 results have two implications:

- 337 • PLDA-UP outperforms the conventional PLDA in all the four duration
338 ranges. The extent of improvement depends on the range of utterance
339 length. We can see that the performance margin is the greatest when
340 utterance-length ranges from 3–20 seconds.
- 341 • Dividing i-vectors into five groups ($K = 5$) seems to be sufficient for all
342 of the four duration ranges. Only System 1 in CC2 and System 2 in CC4
343 show noticeable improvement in both EER and minDCF when the number
344 of groups increases from 5 to 10.
- 345 • There is no clear winner among the three fast scoring systems. All three
346 perform equally well as compared to PLDA-UP. In some settings, the fast
347 scoring systems even perform better than PLDA-UP, although by a very
348 small margin only.

Method		K	Duration Range (seconds)							
			3-20		3-30		3-40		3-60	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
C C 2	PLDA	-	7.41	0.802	6.42	0.665	5.35	0.576	4.20	0.520
	PLDA-UP	-	6.25	0.714	5.43	0.637	4.73	0.563	3.81	0.493
	Sys. 1	5	6.35	0.711	5.54	0.625	4.92	0.554	3.94	0.478
		10	6.17	0.703	5.33	0.625	4.57	0.553	3.80	0.479
		15	6.11	0.710	5.33	0.628	4.69	0.562	3.81	0.479
	Sys. 2	5	6.10	0.723	5.50	0.633	4.66	0.580	3.91	0.485
		10	6.28	0.712	5.49	0.630	4.73	0.566	3.76	0.49
		15	6.30	0.715	5.42	0.620	4.62	0.572	3.77	0.495
	Sys. 3	5	6.14	0.716	5.33	0.621	4.62	0.569	3.87	0.486
		10	6.27	0.713	5.39	0.630	4.73	0.565	3.81	0.485
		15	6.25	0.715	5.36	0.628	4.75	0.567	3.84	0.487
	C C 4	PLDA	-	14.66	0.899	12.06	0.792	10.88	0.710	9.22
PLDA-UP		-	13.28	0.878	11.34	0.809	10.23	0.731	8.71	0.665
Sys. 1		5	13.24	0.869	11.16	0.791	9.98	0.720	8.68	0.641
		10	13.25	0.871	11.06	0.795	9.69	0.712	8.86	0.649
		15	13.33	0.869	11.06	0.794	9.93	0.718	8.56	0.646
Sys. 2		5	13.14	0.878	11.63	0.813	10.44	0.734	8.82	0.662
		10	13.23	0.877	11.52	0.809	10.11	0.731	8.77	0.652
		15	13.22	0.876	11.31	0.809	10.12	0.727	8.68	0.655
Sys. 3		5	13.34	0.875	11.47	0.807	10.55	0.739	8.97	0.659
		10	13.53	0.878	11.26	0.805	10.37	0.736	9.10	0.670
		15	13.39	0.877	11.33	0.807	10.41	0.734	9.02	0.673

Table 2: The performance of PLDA, PLDA-UP and the three fast scoring systems on the truncated speech data from NIST 2012 SRE.

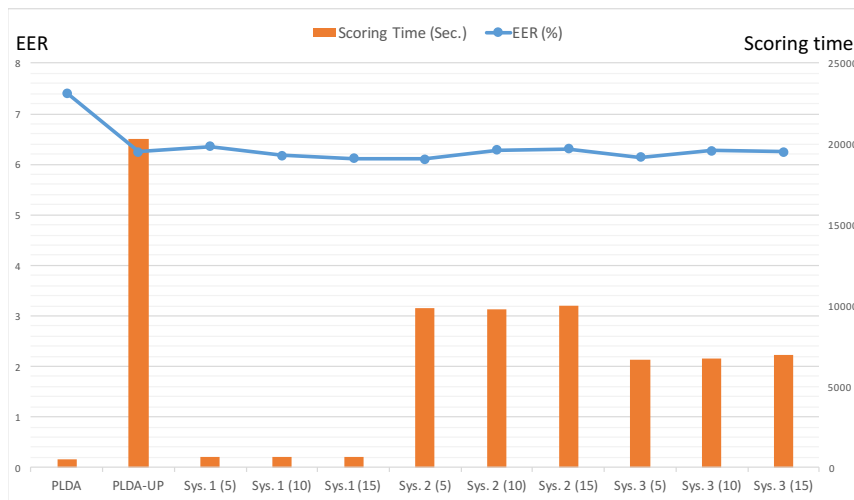


Figure 2: A bar chart showing the EERs and total scoring time of PLDA, PLDA-UP and the three fast scoring systems with different numbers of i-vector groups. For each system, the number inside the parenthesis indicates the number of i-vector groups for that fast scoring system.

349 6.2. Running time

350 The total scoring time and its breakdown for different scoring methods in
 351 CC2 of NIST 2012 SRE are shown in Table 3. Apparently, the conventional
 352 PLDA is the most economical in term of computational cost, as it only involves
 353 vector-matrix multiplications during scoring. By contrast, the PLDA-UP is the
 354 most computational expensive method, with scoring time 44 times that of the
 355 conventional PLDA. The most computational expensive part of PLDA-UP is the
 356 evaluation of Eqs. 21–24, which takes up over 60% of the scoring time. Besides
 357 Eqs. 21–24, the preprocessing of covariance matrices is also computationally
 358 expensive, taking up about 30% of the scoring time. Because our fast scoring
 359 systems do not involve utterance-dependent loading matrices, computations in
 360 Eqs. 21–24 can be done before verification, thus the scoring time is greatly
 361 reduced. However, for System 2 and System 3, we still need to preprocess
 362 the covariance matrices of test utterances, which occupies most of the scoring
 363 time of these two systems. Besides, System 2 also requires to perform eigen-

364 decomposition, which makes it the slowest one among the three systems. For
365 System 1, because the only extra computation besides the scoring function is
366 the simple scalar comparison in Eq. 33, its scoring time is very close to that of
367 the conventional PLDA.

368 **7. Conclusion**

369 In this paper, we proposed a fast scoring method for PLDA with uncertainty
370 propagation (UP). The utterance-dependent loading matrices in UP is replaced
371 by similar ones obtained from development data. The experiments in NIST 2012
372 have shown that the proposed methods have the same ability to deal with short
373 utterances as UP while the computational cost can be reduced to the one very
374 close to that of the conventional PLDA. The proposed method has important
375 implication in the real-life speaker verification, since in most applications the
376 utterance lengths are difficult to control and computation cost is one of the main
377 concerns beside performance.

378 **8. Acknowledgment**

379 This work was supported in part by The RGC of Hong Kong SAR (Grant
380 Nos. PolyU 152117/14E and PolyU 152068/15E) and in part by the Taiwan
381 MOST with Grant 105-2221-E-009-137-MY2.

382 **References**

- 383 [1] S. O. Sadjadi, J. Pelecanos, S. Ganapathy, The IBM speaker recog-
384 nition system: Recent advances and error analysis, arXiv preprint
385 arXiv:1605.01635.
- 386 [2] M. W. Mak, R. C. Hsiao, B. Mak, A comparison of various adaptation
387 methods for speaker verification with limited enrollment data, in: Proc.
388 ICASSP, Vol. 1, IEEE, 2006, pp. 929–932.

Method	Task	Time (Sec.)	% of Total Time
PLDA	Preprocess i-vectors in Eq. 9	11	2.37%
	Scoring in Eq.12	179	38.66%
	Other operations	273	58.96%
	Overall	463	100.00%
PLDA-UP	Preprocess i-vectors in Eq. 9	11	0.05%
	Preprocess \mathbf{L}_t^{-1} in Eq. 17	5966	29.32%
	Evaluate $\mathbf{A}_{s,t}, \mathbf{B}_{s,t}, \mathbf{C}_{s,t}, D_{s,t}$ in Eq. 21-24	12485	61.37%
	Scoring in Eq. 20	294	1.44%
	Other operations	1585	7.79%
	Overall	20341	100.00%
Sys. 1	Preprocess i-vectors in Eq. 9	11	1.7%
	Scalar comparison in Eq. 33	11	1.7%
	Scoring in Eq.36	306	47.29%
	Other operations	319	49.3%
	Overall	647	100.00%
Sys. 2	Preprocess i-vectors in Eq. 9	11	0.11%
	Preprocess \mathbf{L}_t^{-1} in Eq. 17	5966	60.33%
	Compute eigenvalues of \mathbf{A}_t	3275	33.12%
	Scalar comparison in Eq. 35	11	0.11%
	Scoring in Eq.36	306	3.09%
	Other operations	319	3.22%
	Overall	9888	100.00%
Sys. 3	Preprocess in Eq. 9	11	0.16%
	Preprocess \mathbf{L}_t^{-1} in Eq. 17	5966	89.93%
	Compute the traces of \mathbf{A}_t	21	0.31%
	Scalar comparison in Eq. 35	11	0.16%
	Scoring in Eq.36	306	4.61%
	Other operations	319	4.8%
	Overall	6634	100.00%

Table 3: Detailed timing reports obtained by Matlab Profiler for experiments in CC2. We used five loading matrices ($K=5$) for each fast scoring system in the experiments. See Table 5.2 for the configurations of Sys. 1–3.

- 389 [3] R. Vogt, S. Sridharan, M. Mason, Making confident speaker verification
390 decisions with minimal speech, *IEEE Transactions on Audio, Speech, and*
391 *Language Processing* 18 (6) (2010) 1182–1192.
- 392 [4] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, M. W. Mason,
393 I-vector based speaker recognition on short utterances, in: *Proc. ISCA,*
394 *IEEE*, 2011, pp. 2341–2344.
- 395 [5] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, P. Dumouchel,
396 I-vector/PLDA variants for text-dependent speaker recognition, *Tech. rep.,*
397 *Centre de Recherche Informatique de Montreal (CRIM)* (2013).
- 398 [6] A. Larcher, K. A. Lee, B. Ma, H. Li, Text-dependent speaker verification:
399 Classifiers, databases and RSR2015, *Speech Communication* 60 (2014) 56–
400 77.
- 401 [7] T. Hasan, R. Saeidi, J. H. Hansen, D. A. van Leeuwen, Duration mis-
402 match compensation for i-vector based speaker recognition systems, in:
403 *Proc. ICASSP, IEEE*, 2013, pp. 7663–7667.
- 404 [8] L. Li, D. Wang, C. Zhang, T. F. Zheng, Improving short utterance speaker
405 recognition by modeling speech unit classes, *IEEE/ACM Transactions on*
406 *Audio, Speech, and Language Processing* 24 (6) (2016) 1129–1139.
- 407 [9] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, P. Dumouchel, PLDA
408 for speaker verification with utterances of arbitrary duration, in: *Proc.*
409 *ICASSP, IEEE*, 2013, pp. 7649–7653.
- 410 [10] S. Cumani, O. Plchot, P. Laface, On the use of i-vector posterior distribu-
411 tions in probabilistic linear discriminant analysis, *IEEE/ACM Transactions*
412 *on Audio, Speech, and Language Processing* 22 (4) (2014) 846–857.
- 413 [11] S. Cumani, Fast scoring of full posterior PLDA models, *IEEE/ACM Trans-*
414 *actions on Audio, Speech, and Language Processing* 23 (11) (2015) 2036–
415 2045.

- 416 [12] W. W. Lin, M. W. Mak, Fast scoring for PLDA with uncertainty propaga-
417 tion, in: Odyssey The Speaker and Language Recognition Workshop, 2016,
418 pp. 31–38.
- 419 [13] P. Kenny, Joint factor analysis of speaker and session variability: Theory
420 and algorithms, Tech. rep., Centre de Recherche Informatique de Montreal
421 (CRIM) (2005).
- 422 [14] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Joint factor analysis
423 versus eigenchannels in speaker recognition, IEEE Transactions on Audio,
424 Speech, and Language Processing 15 (4) (2007) 1435–1447.
- 425 [15] M. W. Mak, Lecture notes on uncertainty propagation for i-vector/PLDA
426 speaker verification, Tech. rep., The Hong Kong Polytechnic University
427 (2015).
- 428 [16] P. Kenny, Bayesian speaker verification with heavy-tailed priors, in:
429 Odyssey, 2010, pp. 1–14.
- 430 [17] D. Garcia-Romero, C. Y. Espy-Wilson, Analysis of i-vector length normal-
431 ization in speaker recognition systems, in: Proc. Interspeech, 2011, pp.
432 249–252.
- 433 [18] A. O. Hatch, S. S. Kajarekar, A. Stolcke, Within-class covariance normal-
434 ization for SVM-based speaker recognition, in: Proc. Interspeech, 2006.
- 435 [19] W. Rao, M. W. Mak, K. A. Lee, Normalization of total variability matrix
436 for i-vector/PLDA speaker verification, in: Proc. ICASSP, IEEE, 2015, pp.
437 4180–4184.
- 438 [20] I. L. Dryden, A. Koloydenko, D. Zhou, Non-euclidean statistics for covari-
439 ance matrices, with applications to diffusion tensor imaging, The Annals
440 of Applied Statistics (2009) 1102–1123.
- 441 [21] MathWorks, Box plot documentation, [http://www.mathworks.com/help/
442 stats/boxplot.html](http://www.mathworks.com/help/stats/boxplot.html).

- 443 [22] NIST, The NIST year 2012 speaker recognition evaluation plan, [http:](http://www.nist.gov/itl/iad/mig/sre12.cfm)
444 [//www.nist.gov/itl/iad/mig/sre12.cfm](http://www.nist.gov/itl/iad/mig/sre12.cfm).
- 445 [23] M. W. Mak, H. B. Yu, A study of voice activity detection techniques for
446 nist speaker recognition evaluations, *Computer Speech & Language* 28 (1)
447 (2014) 295–313.
- 448 [24] J. Pelecanos, S. Sridharan, Feature warping for robust speaker verifica-
449 tion, in: *Proc. Odyssey The Speaker and Language Recognition Workshop*,
450 Crete, Greece, 2001, pp. 213–218.