INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Gain Compensation for Fast I-Vector Extraction over Short Duration

Kong Aik Lee[1] *and Haizhou Li*[2]

[1]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2]Department of Electrical and Computer Engineering, National University of Singapore
kalee@i2r.a-star.edu.sg, haizhou.li@nus.edu.sg

Abstract

I-vector is widely described as a compact and effective representation of speech utterances for speaker recognition. Standard i-vector extraction could be an expensive task for applications where computing resource is limited, for instance, on handheld devices. Fast approximate inference of i-vector aims to reduce the computational cost required in i-vector extraction where run-time requirement is critical. Most fast approaches hinge on certain assumptions to approximate the i-vector inference formulae with little loss of accuracy. In this paper, we analyze the *uniform* assumption that we had proposed earlier. We show that the assumption generally hold for long utterances but inadequate for utterances of short duration. We then propose to compensate for the negative effects by applying a simple gain factor on the i-vectors estimated from short utterances. The assertion is confirmed through analysis and experiments conducted on NIST SRE'08 and SRE'10 datasets.

Index Terms: speaker recognition, factor analysis

1. Introduction

Automatic speaker recognition refers to the use of computing algorithm to recognize a person from spoken utterances. It has enjoyed significant progress in the past decades [1, 2, 3, 4, 5, 6]. For the most part, speaker recognition is about the extraction and modeling of speaker characteristics underlying the spoken words. In the classical *Gaussian mixture model–Universal background model* (GMM–UBM) framework [7], the vocal patterns are characterized with multivariate probability densities, i.e., the GMMs, in the acoustic-spectral vector space. Many approaches based on the GMM-UBM framework have then been proposed. It was shown through the use of *joint factor analysis* [8], and more recently the *i-vector* [9], mismatch between enrollment and test utterances could be compensated by confining the variabilities within lower dimensional subspace in the parameter space of the GMM. We focus on the use of i-vector for text-independent speaker recognition in this paper.

An i-vector is a compressed representation of speaker characteristics rendered in a speech utterance, which includes as well session variabilities originated from its phonetic content and other extrinsic factors like channel, recording device, and acoustic environment [9]. Probabilistic linear discriminant analysis [10], or PLDA, is commonly used as a session compensation and scoring back-end for i-vectors [11, 12]. Figure 1 shows a pictorial depiction of the processing pipeline in the state-of-the-art i-vector PLDA system. The process begins with sufficient statistics extraction, the central component of which is the UBM. The UBM could be trained in a unsupervised manner [7] or trained to predict senone posterior via supervised training [13, 14, 15]. The training of UBM is usually performed off-line, similarly for the *total variability* and *PLDA* model training. Their high computational demand is not generally seen as a bottleneck as computing resource could always be sorted out for off-line and one-off training.

Though high computation demand could be resolved easily for off-line hyper-parameter training, run-time resource requirement remains an issue, for instance, on handheld devices [16, 17], or in face of a large volume of concurrent requests, e.g., in large-scale cloud based services [18]. To answer this challenge, approximate inference has been proposed for rapid i-vector extraction [19, 20, 21, 22, 23, 24]. In [24], we proposed a fast computation method for i-vector extraction and showed its effectively on long utterances with typical duration of two and half minutes. In a recent study, we found an inadequacy of the *uniform assumption* proposed in [24] for the case of short duration utterances. In particular, our study shows that eigenvalues of the total variability matrix play an important role when extracting i-vector from short utterances. Based on this analysis, we propose a simple *gain compensation* method to mitigate the negative effects of the uniform assumption while preserving the fast computation.

The paper is organized as follows. Section 2 reviews the fundamentals of i-vector PLDA paradigm. We touch upon those points that are required in subsequent sections. Section 3 presents the idea of orthogonal variability matrix, which is then used in Section 4 to derive the fast inference and gain compensation methods proposed in this paper. Section 5 presents the experiment results on SRE'10. Section 6 concludes the paper.

2. The i-vector PLDA pipeline

We summarize below the key features in the state-of-the-art i-vector PLDA processing pipeline from i-vector extraction to session compensation and scoring with PLDA.

2.1. I-vector extraction

Let $\mathcal{O} = \{o_1, o_2, \cdots, o_n\}$ denotes the feature vector sequence extracted from a speech utterance with n number of frames. An i-vector ϕ_r representing the observed sequence $\mathcal{O}_r$ is given by

$$\phi_r = \left(\mathbf{I} + \mathbf{T}^\mathsf{T}\mathbf{\Sigma}^{-1}\mathbf{N}_r\mathbf{T}\right)^{-1} \cdot \mathbf{T}^\mathsf{T}\mathbf{\Sigma}^{-1}\left(\mathbf{F}_r - \mathbf{N}_r\mathcal{M}\right) \quad (1)$$

where $\mathbf{T}$ is referred to as the *total variability matrix*. Apart from $\mathbf{T}$, an i-vector extractor consists as well the parameters $\mathcal{M}$ and $\mathbf{\Sigma}$ copied from the UBM. Let M be the size of the i-vector ϕ_r and F be the dimensionality of the acoustic feature vector, we see that $\mathbf{T}$ is a $CF \times M$ rectangular matrix, $\mathcal{M}$ is a $CF \times 1$ supervector obtained by concatenating the mean vectors $\mathcal{M}_c$ of the UBM. The $CF \times CF$ supervector-size matrix $\mathbf{\Sigma}$ is constructed by having its diagonal blocks made up by covariance matrices $\mathbf{\Sigma}_c$ of the UBM.

http://dx.doi.org/10.21437/Interspeech.2017-108

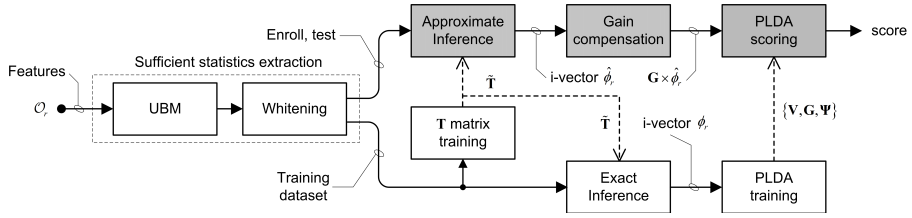

Figure 1: *Key features and the processing pipeline in a i-vector PLDA setup for a typical speaker recognition task.*

Two other quantities required in (1) are the zero and first order Baum-Welch statistics, $\mathbf{N}_r$ and $\mathbf{F}_r$, computed for a given utterance r. In particular, the $CF \times 1$ vector $\mathbf{F}_r$ is obtained by stacking the first-order statistics from all C components:

$$\mathbf{F}(c) = \sum\nolimits_t \gamma_t(c) o_t \tag{2}$$

In a similar manner, $\mathbf{N}_r$ is a $CF \times CF$ diagonal matrix, consisting of diagonal blocks of $n(c)\mathbf{I}_{F\times F}$, where

$$n(c) = \sum\nolimits_t \gamma_t(c) \tag{3}$$

is the soft count of frames aligned to the c-th Gaussian by summing the occupation probability $\gamma_t(c)$ over the entire sequence. The frame alignment could be computed using a GMM trained in an unsupervised manner [7] or a DNN trained to model senones via supervised training [13, 14, 15].

In state-of-the-art implementations [19, 25, 26], as shown in Figure 1, it is common to whiten the first-order statistics in (2) by subjecting them to the following transformation

$$\widetilde{\mathbf{F}}_r = \mathbf{\Sigma}^{-1/2} \left(\mathbf{F}_r - \mathbf{N}_r \mathcal{M} \right) \tag{4}$$

where $\mathbf{\Sigma}^{-1/2}$ is obtained by decomposing $\mathbf{\Sigma}$, for instance, with Cholesky or eigenvalue decompositions and taking the inverse. Transforming Baum-Welch statistics in this manner has similar effects on the loading matrix such that $\widetilde{\mathbf{T}} = \mathbf{\Sigma}^{-1/2}\mathbf{T}$. I-vector computation in (1) could then be expressed in term of the *normalized* statistics and parameters, as follows

$$\phi_r = \left(\mathbf{I} + \widetilde{\mathbf{T}}^{\mathsf{T}} \widetilde{\mathbf{N}}_r \widetilde{\mathbf{T}} \right)^{-1} \cdot \widetilde{\mathbf{T}}^{\mathsf{T}} \widetilde{\mathbf{F}}_r \tag{5}$$

The *normalized* form in (5) produces the same i-vector as the *basic* form in (1). We use (5) in the following sections.

2.2. Probabilistic linear discriminant analysis

Probabilistic linear discriminant analysis [10, 11, 12], or PLDA, is a latent variable model formulated in a much similar fashion as the classical *factor analysis* [27]. In its marginalized form, PLDA is essentially a Gaussian distribution:

$$p\left(\phi_r\right) = \mathcal{N}\left(\phi_r | \mu, \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{G}\mathbf{G}^{\mathsf{T}} + \mathbf{\Psi} \right) \tag{6}$$

where the vector μ represents the global mean of all i-vectors. With such Gaussian assumption, whitening followed by unit-length normalization are usually performed on i-vector prior to PLDA modeling. The procedure was dubbed as length normalization in [28]. At the core of PLDA is the loading matrices $\mathbf{V}$ and $\mathbf{G}$, the column spaces of which model the speaker and session variabilities, while $\mathbf{\Psi}$ is a diagonal matrix of residual covariance. In some implementations, the channel and residual covariance matrices are merged into single term leading to the simplified PLDA [29]. Given a test i-vector, it could be scored against an enrollment i-vector with repeated uses of (6). See [30] for details on the score computation and [10, 29] for iterative training mechanism with expectation-maximization (EM).

3. Orthogonal variability space

Given $\widetilde{\mathbf{T}}$ a tall rectangular matrix, we consider the following transformation

$$\widetilde{\mathbf{T}}\mathbf{Q}\mathbf{\Lambda}^{-1/2} = \mathbf{U} \tag{7}$$

such that the columns of the matrix $\mathbf{U}$ are *orthonormal* (i.e., *orthogonal* and all of unit length) and span the same subspace as the column space of the original total variability matrix $\widetilde{\mathbf{T}}$. This is achieved by having $\mathbf{Q}$ and $\mathbf{\Lambda}$ be the matrices of eigenvectors and eigenvalues, respectively, from Eigen decomposition of the following

$$\widetilde{\mathbf{T}}^{\mathsf{T}}\widetilde{\mathbf{T}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{T}} \tag{8}$$

We could preserve the eigenvalues by moving the diagonal matrix $\mathbf{\Lambda}^{-1/2}$ to the right-hand-side of (7). The resulting matrix

$$\widetilde{\mathbf{T}}\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}^{1/2} \tag{9}$$

consists of orthogonal basis with their length correspond to the square-root of the eigenvalues. The matrix $\mathbf{Q}$ merely imposes a rotation on the total variability matrix.

Next, we move the matrix $\mathbf{Q}$ to the right-hand-side of (9) by post-multiplying both sides of the equations with $\mathbf{Q}^{-1}$. Substituting the result into (5), after some algebraic manipulations, we arrive at the following expression for i-vector extraction:

$$\phi_r = \mathbf{Q} \cdot \mathbf{D}_0 \left(\mathbf{D}_1 + \mathbf{U}^{\mathsf{T}} \widetilde{\mathbf{N}}_r \, \mathbf{U} \right)^{-1} \mathbf{U}^{\mathsf{T}} \widetilde{\mathbf{F}}_r \tag{10}$$

where $\mathbf{D}_0 = \mathbf{\Lambda}^{-1/2}$ and $\mathbf{D}_1 = \mathbf{\Lambda}^{-1}$. To arrive at (10) we have used the properties that $\mathbf{Q}^{-1} = \mathbf{Q}^{\mathsf{T}}$ and $\mathbf{Q}\mathbf{Q}^{\mathsf{T}} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}$ since $\mathbf{Q}$ is a square orthogonal matrix.

Comparing the i-vector extraction equation in (10) to that in (5), we see that the total variability matrix has been decomposed into components matrices consisting of the *orthogonal subspace* $\mathbf{U}$, the rotation matrix $\mathbf{Q}$ and the diagonal matrix $\mathbf{\Lambda}$ of eigenvalues. In [24], our approach to subspace orthonomalization was based on the use of non-standard Gaussian prior. In this paper, we achieve similar goal using a different approach by factoring the total variability matrix into component matrices. In particular, current formulation allows us to look into the role of eigenvalues in relation to the duration as encoded in the zero-order statistics of the input utterances.

4. Approximate i-vector extraction

4.1. Fast approximate inference

The major source of computational load in estimating an i-vector using either (5) or (10) is the computation of the $M \times M$

matrix inverse in the equation. We introduce below two approximations to speed up the computation.

i. ***Uniform regularization***: We assume that $\mathbf{D}_1 = \kappa\mathbf{I}$ in (10) such that regularization is applied uniformly across the diagonal elements of the matrix $\mathbf{U}^{\mathsf{T}}\widetilde{\mathbf{N}}_r\ \mathbf{U}$ with an average regularization factor

$$\kappa = \frac{1}{M}\sum_{l=1}^{M}\lambda_l^{-1} \tag{11}$$

where λ_l are the eigenvalues in $\mathbf{\Lambda}$.

ii. ***Uniform scaling***: Let

$$\alpha_c = \frac{n(c)}{\kappa + n(c)} \simeq \alpha \quad \text{for} \quad c = 1, 2, ..., C \tag{12}$$

where $n(c)$ are the occupancy counts defined in (3) and κ is a positive scalar. In the current paper, the value of κ is given by (11) since the *uniform regularization* assumption is applied first. We then assume a uniform scaling factor $\alpha \simeq \alpha_c$, for $0 \leq \alpha < 1$, such that α is applied uniformly across all C Gaussians of the total variability model.

To derive the fast estimation method, we start from (10) by invoking the *uniform regularization* assumption with $\mathbf{D}_1 = \kappa\mathbf{I}$. For κ a scalar, it can be shown that the following matrix inversion [31] holds:

$$(\kappa\mathbf{I} + \mathbf{U}^{\mathsf{T}}\widetilde{\mathbf{N}}_r\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}} = \mathbf{U}^{\mathsf{T}}\left(\kappa\mathbf{I} + \widetilde{\mathbf{N}}_r\ \mathbf{U}\mathbf{U}^{\mathsf{T}}\right)^{-1} \tag{13}$$

Next, we define

$$\mathbf{K} = \left(\kappa\mathbf{I} + \widetilde{\mathbf{N}}_r\ \mathbf{U}\mathbf{U}^{\mathsf{T}}\right)^{-1} \tag{14}$$

and let $\bar{\mathbf{U}}$ be a $CF \times (CF - M)$ rectangular matrix with all its columns orthogonal to the column space of $\mathbf{U}$ such that $\mathbf{U} \perp \bar{\mathbf{U}}$ and therefore

$$\mathbf{U}\mathbf{U}^{\mathsf{T}} = \mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^{\mathsf{T}} \tag{15}$$

Using (15) in (14), and let $\mathbf{A} = (\kappa\mathbf{I} + \widetilde{\mathbf{N}}_r)$, we show that

$$\mathbf{K} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\widetilde{\mathbf{N}}_r \cdot \bar{\mathbf{U}}\bar{\mathbf{U}}^{\mathsf{T}}\left(\mathbf{I} - \mathbf{A}^{-1}\widetilde{\mathbf{N}}_r\bar{\mathbf{U}}\bar{\mathbf{U}}^{\mathsf{T}}\right)^{-1}\mathbf{A}^{-1} \tag{16}$$

Combining (13), (14), (16), and substituting the result back into (10), we obtain

$$\phi_r \approx \mathbf{Q} \cdot \mathbf{D}_0\mathbf{U}^{\mathsf{T}}\left(\kappa\mathbf{I} + \widetilde{\mathbf{N}}_r\right)^{-1}\widetilde{\mathbf{F}}_r \tag{17}$$

To arrive at the above, we have invoked the *uniform scaling* assumption to eliminate the second term on the right-hand-side of (16) which is orthogonal to $\mathbf{U}$. Next, we pre-multiply both sides of (17) with $\mathbf{Q}^{\mathsf{T}}$ such that

$$\hat{\phi}_r = \mathbf{D}_0\mathbf{U}^{\mathsf{T}}\left(\kappa\mathbf{I} + \widetilde{\mathbf{N}}_r\right)^{-1}\widetilde{\mathbf{F}}_r \tag{18}$$

The resulting i-vector $\hat{\phi}_r$ contains features which are less correlated.

The run-time computational cost of i-vector extraction in (5) is dominated by matrix inversion. One common practice is to pre-compute and store sub-matrices that involve [19]. The total computational cost becomes $O\left(M^3 + CFM + CM^2\right)$. With the proposed fast estimation in (18), the computation cost is greatly reduced to $O\left(CFM\right)$. Compared to [24], our current result in (18) contains an additional regularization factor κ which is absent in our previous formulation.

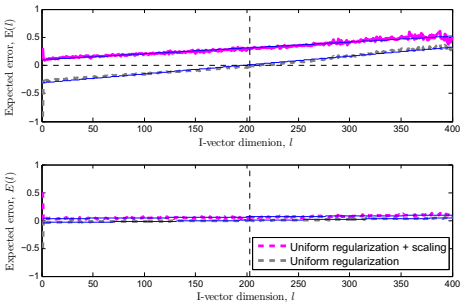

Figure 2: *Expected error estimated for (a) short utterances (upper panel) and (b) long utterance (lower panel).*

4.2. Gain compensation for short duration

Given a fixed total variability matrix, the *uniform regularization* assumption holds better for longer utterances given that the matrix $\widetilde{\mathbf{N}}_r$ has comparatively larger value on its diagonal. Similar argument is also true for the *uniform scaling* assumption whereby α_c in (12) approaches unity for larger value of $n(c)$ given by longer utterances.

To study the artifacts resulting from short-duration utterances, we define in the following two error measures between the actual i-vector ϕ_r and its fast approximation $\hat{\phi}_r$:

i. *Expected error*:

$$\mathrm{E}(l) = E\left\{|\phi_r(l)| - |\hat{\phi}_r(l)|\right\} \tag{19}$$

ii. *Gain factor* (or *ratio*):

$$\mathrm{g}(l) = \frac{E\left\{|\phi_r(l)|\right\}}{E\left\{|\hat{\phi}_r(l)|\right\}} \tag{20}$$

Here, $E(.)$ denotes the expectation with respect to the session index r. More specifically, the expected value is computed by taking the ensemble average across large number of sessions for individual dimensions l of the i-vectors, where $l = 1, 2, ..., M$.

Figure 2 shows the expected error curves evaluated for the case of short and long utterances in the upper and lower panels, respectively. I-vectors were estimated from short and long utterances with net speech duration of approximately 10 seconds (short) and 2.5 minutes (long) drawn from NIST SRE'08 dataset. For each utterance, two i-vectors were computed. One with exact inference using (10) and the other with approximate inference using (18), both taken as parallel inputs to the error measures. It is worth mentioning that we pre-multiply the i-vectors ϕ_r with $\mathbf{Q}^{\mathsf{T}}$. This decorrelate the i-vector as we have done in (18). Ideally, the expected error would tend to be zero for $\hat{\phi}_r$ a close approximate of ϕ_r. This is generally true for long long segments as shown in the lower panel of Figure 2.

To tell apart the errors introduced by the two assumptions, we show the error measures for the cases when (i) *uniform regularization* only and (ii) *uniform regularization + scaling* were applied. Also shown in the plots are the linear fits overlay-ed on top of the error curves. In the upper panel of Figure 3, the gain factor curve crosses the point where $g(l_0) = 1$ at $l_0 \approx 202$ for the current case when only uniform regularization assumption was applied. As we replaced the regularization matrix $\mathbf{\Lambda}$ with a constant regularization $\kappa\mathbf{I}$, the front elements of of the i-vectors, for $l < l_0$, are under-regularized while elements at the back,

Table 1: *Performance comparison under CC'5 of extended core task and 10sec-10sec task of NIST SRE'10 dataset.*

	CC'5		CC'6		CC'8		10sec-10sec	
	EER(%)	DCF10	EER(%)	DCF10	EER(%)	DCF10	EER (%)	DCF08
Exact	2.91	0.448	5.21	0.837	2.71	0.495	14.52	0.676
Ureg	2.79	0.476	5.13	0.830	2.63	0.484	18.62	0.824
Ureg + Uscale	2.94	0.504	5.81	0.887	2.75	0.536	19.35	0.850
Ureg + G	2.87	0.481	5.19	0.839	2.68	0.492	14.08	0.688
Ureg + Uscale + G	2.97	0.504	5.88	0.891	2.71	0.534	**13.97**	**0.695**

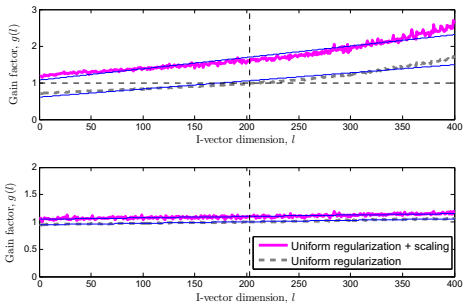

Figure 3: *Gain factor estimated for (a) short utterances (upper panel) and (b) long utterance (lower panel).*

where $l > l_0$, are over-regularization. As a result, the gain factor curve exhibits a positive slope. Now, if we introduce further the *uniform scaling* assumption, we observed a positive offset, which signifies the error being uniformly distributed across entire i-vector. For long utterances, we see that the gain factor is close to unity, where $g(l) \approx 1$ for $l = 1, 2, .., M$, which signifies that the both *uniform regularization* and *scaling* assumptions would hold better in this case.

Empirical analysis above shows that the proposed *approximate inference* results in i-vectors with amplitude deviated from exact inference, especially for the case of short utterances. We propose a simple compensation using the gain factor, as follows:

$$\hat{\phi}_r \longleftarrow \mathbf{G}\hat{\phi}_r \tag{21}$$

where $\mathbf{G}$ is a diagonal matrix with $\mathbf{G}(l,l) = g(l)$ for $l = 1, 2, .., M$. As before, the gain factors are determined empirically from data. In this paper, we estimated the gain factors from SRE'08 dataset and validate the results on SRE'10 test set. The speaker sets in these dataset are disjoint. Using (21) together with (18), we pre-compute and store the term $\mathbf{G}\mathbf{D}_0\mathbf{U}^{\mathsf{T}}$ to speed up computation.

5. Experiments

Experiments were conducted on NIST SRE'10 dataset [32]. The recognition accuracy is given in terms of *equal error rate* (EER) and *minimum decision cost function* (min DCF). We report results on *10sec* task and three *common conditions* (CCs) defined for the *extended core* task of SRE'10 – CC5, CC6 and CC8 corresponding to *normal*, *high* and *low* vocal efforts in test and *normal* vocal effort in enrollment. We emphasize on the effects and compensation of short duration in the current paper. We refer interested reader to our previous report [24] for a comparison of the proposed method to various fast i-vector extraction techniques.

The acoustic features used in the experiments consist of 19-dimensional mel frequency cepstral coefficients (MFCC). Delta and delta-delta features were appended giving rise to 57-dimensional feature vector. We used gender-dependent UBM consisting of 512 mixtures with full covariance matrices. The total variability matrix $\widetilde{\mathbf{T}}$ was trained using Switchboard, NIST SREs 04, 05, and 06 datasets. The rank M of the matrix $\widetilde{\mathbf{T}}$ was set 400. Length normalization [28] was applied prior to PLDA as described in Section 2.2. The PLDA was trained to have 200 speaker factors with a full residual covariance for channel modeling. We trained one PLDA using i-vectors extracted with exact inference and used for all experiments whereby enrollment and test i-vectors were extracted with either *exact* or *approximate* inference methods (see Fig. 1). As in Section 4.2, we pre-multiply the exact i-vector ϕ_r of (10) with $\mathbf{Q}^{\mathsf{T}}$ so that it is consistent with the approximation in (18).

Table 1 shows the speaker verification performance for enrollment and test i-vectors extracted using *exact* inference (**Exact**), approximate inference with *uniform regularization* only (**Ureg**) and *uniform regularization + scaling* (**Ureg + Uscale**). For CC'5, where long segments were used for enrollment and test, approximate inference causes slight degradation when both uniform assumptions were applied. This amounts to 1.03% and 4.1% in EER and MinDCF, respectively. From the results on CCs 5, 6, and 8, it is interesting to note that the uniform regularization assumption alone does not affect much the speaker verification performance for long segments. This is however not the case for short utterances, which can be seen from the results in 10sec–10sec task. We observed 28% and 22% degradation on EER and MinDCF with uniform regularization assumption alone (**Ureg**), while 33% and 26% with both uniform *regularization* and *scaling* assumptions applied (**Ureg + Uscale**). This negative effects could be compensated with the used of gain factor as shown in the last two rows of the table. For 10sec–10sec, we obtained a significant improvement of 28% on EER and 18% on MinDCF when gain compensation is applied on top of the approximated i-vector. Notice that gain compensation is not required for long utterances though we do not observe much differences with and without such compensation applied.

6. Conclusions

We showed analytically and empirically that the eigenvalues of the total variability matrix play an important role when extracting i-vector from short utterances while less important for long utterances. The uniform regularization assumption speeds up the computation significantly but inadvertently removes the eigenvalues. This results in i-vector with modified amplitude compared to the exact inference, especially for the case of short utterances. We showed that the amplitude differences could be compensated with a simple gain factor estimated empirical from data. To this end, we estimated the gain factor on SRE'08 and observed that it generalizes to SRE'10 test set. We obtained a significant improvement of 28% on EER when gain compensation is applied on top of the approximated i-vector.

7. References

[1] A. E. Rosenberg, "Automatic speaker verification: a review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, Apr 1976.

[2] G. R. Doddington, "Speaker recognition – identifying people by their voices," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, Nov 1985.

[3] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, pp. 5–17, Oct 1986.

[4] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, 2004.

[5] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from vectors to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[6] J. H. L. Hansen and T. Hasan, "How humans and machines recognize voices: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[8] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[10] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Computer Vision*, 2007, pp. 1–8.

[11] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010, p. 14.

[12] K. A. Lee, A. Larcher, C. You, B. Ma, and H. Li, "Multi-session PLDA scoring of i-vector for partially open-set speaker detection," in *Proc. Interspeech*, 2013, pp. 3651–3655.

[13] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE ICASSP*, 2014, pp. 1695–1699.

[14] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 293–298.

[15] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[16] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," *IEEE Signal Processing Society Speech and language Technical Committee Newsletter*, 2013.

[17] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello, "Continuous user authentication on mobile devices: Recent progress and remaining challenges," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 49–61, July 2016.

[18] "Nuance VocalPassword," http://www.nuance.com/for-business/customer-service-solutions/voice-biometrics/vocalpassword/, accessed: 2017-03-08.

[19] O. Glembek, L. Burget, P. Matjka, M. Karafit, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proc. IEEE ICASSP*, May 2011, pp. 4516–4519.

[20] H. Aronowitz and O. Barkan, "Efficient approximated i-vector extraction," in *Proc. IEEE ICASSP*, 2012, pp. 4789–4792.

[21] W. M. Campbell, D. Sturim, B. J. Borgstrom, R. Dunn, A. McCree, T. F. Quatieri, and D. A. Reynolds, "Exploring the impart of advances front-end processing on NIST speaker recogntiion microphone tasks," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 180 –186.

[22] S. Cumani and P. Laface, "Factorized sub-space estimation for fast and memory effective i-vector extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 248–259, 2014.

[23] L. Xu, K. A. Lee, H. Li, and Z. Yang, "Sparse coding of total variability matrix," in *Proc. Interspeech*, 2015, pp. 1022–1026.

[24] ——, "Rapid computation of i-vector," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2016.

[25] P. Kenny, "A small footprint i-vector extractor," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, vol. 2012, 2012.

[26] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *Proc. IEEE ICASSP*, 2016, pp. 5095–5099.

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

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

[29] A. Sizov, K. A. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Proc. S+SSPR*, 2014.

[30] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Minimum divergence estimation of speaker prior in multi-session PLDA scoring," in *Proc. ICASSP*, 2014.

[31] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*. Technical University of Denmark, 2008.

[32] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Proc. Interspeech*, 2010, pp. 2726–2729.