Interspeech 2018
2-6 September 2018, Hyderabad

Co-whitening of i-vectors for short and long duration speaker verification

Longting Xu[1], *Kong Aik Lee*[2], *Haizhou Li*[1], *Zhen Yang*[3]

[1] Department of Electrical and Computer Engineering, National University of Singapore
[2]Data Science Research Laboratories, NEC Corporation, Japan
[3]Broadband Wireless Communication and Sensor Network Technology Key Lab.
Nanjing University of Posts and Telecommunications, China

elexul@nus.edu.sg, kalee@ieee.org

Abstract

An i-vector is a fixed-length and low-rank representation of a speech utterance. It has been used extensively in text-independent speaker verification. Ideally, speech utterances from the same speaker would map to an unique i-vector. However, this is not the case due to some intrinsic and extrinsic factors like physical condition of the speaker, channel difference, noise and notably the duration of speech utterances. In particular, we found that i-vectors extracted from short utterances exhibit larger variance than that of long utterances. To address the problem, we propose a co-whitening approach, taking into account the duration, while maximizing the correlation between the i-vectors of short and long duration. The proposed co-whitening method was derived based on canonical correlation analysis (CCA). Experimental results on NIST SRE 2010 show that co-whitening method is effective in compensating the duration mismatch, leading to a reduction of up to 13.07% in equal error rate (EER).

Index Terms: Speaker recognition, co-whitening, short duration, i-vector, text-independent, canonical correlation analysis

1. Introduction

Text-independent speaker verification aims to verify the identity of a speaker in two different speech utterances [1]. The first forms the enrollment utterance, whereas the second utterance is provided during testing. The lexical content of two utterances are different, hence the name text-independent. It is reasonable to assume that the enrollment utterances have sufficiently long duration, since enrollment is carried out once and in an offline manner. This is usually not the case for test utterances, where the short duration is usually more desirable from the user's perspective.

There has been increasing interest in short duration text-independent speaker verification. The recognition accuracy degrades as the duration of speech utterance decreases as a result of the reduced amount of information available in the short utterance. This could partly be observed in the form of increased uncertainty (i.e., larger variance) in the distribution of the i-vectors extracted from short utterances. Paper [2] proposed to propagate the uncertainty into a PLDA classifier for unrestricted duration utterance based speaker recognition. In [3], score calibration was introduced to compensate the duration mismatch. In [4] and [5], it was shown that the relationship between i-vectors of short and long duration could be modeled by tying them to a single latent variable.

In this paper, a novel approach of simultaneously whitening i-vectors estimated from short and long utterances is proposed. This approach aims to project the i-vectors onto individual matrices simultaneously, while maximizing the relationship between them. We name this approach co-whitening, and canonical correlation analysis (CCA) is taken to measure the linear relationship. Recent paper [6] also mentioned CCA has close relationship with whitening and can be regarded as a special case of co-whitening method on two data groups. In previous works, CCA was mainly used to measure the correlation among different features [7] and fuse multimodal features in speaker recognition [8, 9].

Recent advances in deep learning have enabled a wide range of machine learning tasks [10] including automatic speech recognition, machine translation, natural language processing, just to name a few. In speaker recognition task, deep neutral network has shown to be effective for extracting i-vector like, utterance-level representation, referred to as speaker embedding [11, 12, 13, 14]. Paper [13, 14] are especially useful for short duration speaker verification. The utterance-level representations, e.g. x-vector [11, 12], d-vector [15] are processed with the same processing backend (i.e., whitening, length normalization, and probabilistic linear discriminant analysis (PLDA)) as used for i-vector. The co-whitening approach is therefore useful for i-vector and x-vector alike. We devote our attention to the former in the current paper.

In the following, we first introduce the fundamentals of i-vector paradigm in Section 2. Section 3 presents the conventional whitening method and analyzes the disadvantages using the same whitening projection for different duration utterances. We then propose a novel approach that co-whitens the short and long utterances simultaneously in Section 4, which is further validated by the experimental results in Section 5. Section 6 concludes the paper.

2. The i-vector methodology

We use i-vector [16] as our baseline in this paper. Let $\boldsymbol{m}$ be a Gaussian Mixture Model (GMM) supervector stacked by the mean vectors of C Gaussian components, the lower dimension vector $\boldsymbol{x}$ can be computed as follows:

$$\boldsymbol{m} = \mathcal{M} + \mathbf{T}\boldsymbol{x} \tag{1}$$

where $\mathcal{M}$ is the GMM supervector of the universal background model (UBM). The low-rank matrix $\mathbf{T}$ captures total variability of each speaker's sessions. The i-vector extraction process employs EM algorithm [17].

Let $\mathcal{O} = \{o_1, o_2, \ldots, o_T\}$ represent the feature sequence of a given utterance. Given $\mathcal{O}$, the total variability matrix $\mathbf{T}$ and the covariance matrices $\boldsymbol{\Sigma}_c$ of the UBM, we infer the posterior mean of the latent variable $\boldsymbol{x}$:

$$\phi = \arg\max_{\boldsymbol{x}} \left[\prod_{c=1}^{C} \prod_{t=1}^{N_c} \mathcal{N}\left(o_t | \mathcal{M}_c + \mathbf{T}_c\boldsymbol{x}, \boldsymbol{\Sigma}_c\right) \right] \mathcal{N}\left(\boldsymbol{x}|0, \mathbf{I}\right)$$

10.21437/Interspeech.2018-1246

The solution to the above maximum a-posterior estimation is the i-vector:

$$\phi = \mathbf{L}^{-1}\left[\sum\nolimits_{c=1}^{C}\mathbf{T}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\boldsymbol{F}_c\right] \tag{2}$$

where $\mathbf{L}$ is the posterior precision given by

$$\mathbf{L} = \mathbf{I} + \sum\nolimits_{c=1}^{C} N_c\mathbf{T}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{T}_c \tag{3}$$

In the above equations, $\{N_c, \boldsymbol{F}_c\}$ are the utterance-dependent Baum-Welch statistics computed based on the UBM. In particular, N_c is the zero-order statistics computed for the c-th Gaussian by summing the frame occupancy $\gamma_c(t)$ over the entire sequence, as follows

$$N_c = \sum\nolimits_t \gamma_c(t) \tag{4}$$

while

$$\boldsymbol{F}_c = \sum\nolimits_t \gamma_c(t)\,(o_t - \boldsymbol{\mu}_c)$$

is the first-order statistics. It is common to use

$$\tilde{\boldsymbol{F}}_c = \boldsymbol{\Sigma}_c^{-1/2}\left[\sum\nolimits_t \gamma_c(t)\,(o_t - \boldsymbol{\mu}_c)\right] \tag{5}$$

representing the first-order statistics centered to the mean $\boldsymbol{\mu}_c$ and whitened with respect to covariance $\boldsymbol{\Sigma}_c$ of the UBM.

3. The whitening paradigm

As mentioned in Section 2, i-vector is assumed to be normally distributed, but practically not, thus length normalization was proposed in [18]. This technique starts with a linear whitening transformation, followed by a non-linear transformation, which is a vector unitization method.

Let $\boldsymbol{\Phi} = (\phi(1), \ldots, \phi(i), \ldots, \phi(n))$ be a matrix consisting of n i-vectors in its column. The empirical mean of this data matrix is

$$\hat{\mu}_\phi = \frac{1}{n}\sum_{i=1}^{n}\phi(i) \tag{6}$$

by which, the i-vectors are centered to zero $\bar{\boldsymbol{\Phi}} = (\phi(1) - \hat{\mu}_\phi, \ldots, \phi(i) - \hat{\mu}_\phi, \ldots, \phi(n) - \hat{\mu}_\phi)$. We then transform $\bar{\boldsymbol{\Phi}}$ with the covariance matrix $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\Phi}}}$ as follows

$$\tilde{\bar{\boldsymbol{\Phi}}} = \mathbf{W}\bar{\boldsymbol{\Phi}} \tag{7}$$

where the matrix $\mathbf{W}$ is subject to the condition $\mathbf{W}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\Phi}}}\mathbf{W}^{\mathrm{T}} = \mathbf{I}$. We choose one from infinite choices of matrix $\mathbf{W}$, which is based on eigenvalues and eigenvectors. This method produces a diagonal matrix $\mathbf{D}_{\bar{\boldsymbol{\Phi}}}$ of eigenvalues and a full matrix $\mathbf{V}_{\bar{\boldsymbol{\Phi}}}$ whose columns are the corresponding eigenvectors so that

$$\boldsymbol{\Sigma}_{\bar{\boldsymbol{\Phi}}}\mathbf{V}_{\bar{\boldsymbol{\Phi}}} = \mathbf{V}_{\bar{\boldsymbol{\Phi}}}\mathbf{D}_{\bar{\boldsymbol{\Phi}}} \tag{8}$$

from which we estimate the loading matrix, as follows

$$\mathbf{W} = \mathbf{D}_{\bar{\boldsymbol{\Phi}}}^{-1/2}\mathbf{V}_{\bar{\boldsymbol{\Phi}}}^{\mathrm{T}} \tag{9}$$

I-vectors of different durations have different distributions, therefore, taking same whitening matrix trained from long duration speech for short utterance is not appropriate [5]. In this paper, we propose a novel whitening transformation, aiming to whitening the short and long utterance i-vectors simultaneously over separated matrices. Along with this, the converted short and long vectors remain maximally correlated. Mathematically, we aim to seek a pair of matrices $\mathbf{W}_s$ and $\mathbf{W}_l$, which are confined in a way that

$$\mathbf{W}_s\boldsymbol{\Sigma}_s\mathbf{W}_s^{\mathrm{T}} = \mathbf{I} \tag{10}$$

and

$$\mathbf{W}_l\boldsymbol{\Sigma}_l\mathbf{W}_l^{\mathrm{T}} = \mathbf{I} \tag{11}$$

with the constraint

$$\max_{\mathbf{W}_s,\mathbf{W}_l} cor(\mathbf{W}_s\boldsymbol{\Phi}_s, \mathbf{W}_l\boldsymbol{\Phi}_l) \tag{12}$$

Here, $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_l$ contain the corresponding short and long utterance i-vectors from same speaker, therefore having the same size.

4. Co-whitening for short and long duration utterances

As analyzed in the aforementioned section, we aim to seek matrices $\mathbf{W}_s$ and $\mathbf{W}_l$ for short and long utterances. Canonical Correlation Analysis method could be used for this purpose.

4.1. Canonical Correlation Analysis

Given two random vectors $\boldsymbol{X} = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $\boldsymbol{Y} = (y_1, \ldots, y_m)^{\mathrm{T}}$, canonical correlation analysis seeks vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that the random variables $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}$ maximize the correlation $\rho = corr\langle\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}\rangle$. The random variables $U = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$ and $V = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}$ are the first pair of canonical variables. Then one seeks to maximize the correlation ρ subject to the constraint that they are to be uncorrelated with the first pair of canonical variables; this gives the second pair of canonical variables. This procedure is iterated to $\min\{m, n\}$ times.

The parameter to maximize is

$$\rho = \frac{E(UV^{\mathrm{T}})}{\sqrt{E(UU^{\mathrm{T}})}\sqrt{E(VV^{\mathrm{T}})}} \tag{13}$$

which could be rewritten as

$$\rho = \frac{E(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{b})}{\sqrt{E(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a})}\sqrt{E(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{b})}} \tag{14}$$

The cross-covariance is an $n \times m$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = E(\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}})$ where the (i, j) entry is the covariance $E(x_i, y_j)$. Similarly, let $\boldsymbol{\Sigma}_{\boldsymbol{X}} = E(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ and $\boldsymbol{\Sigma}_{\boldsymbol{Y}} = E(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})$. The correlation parameter to be maximized becomes

$$\rho = \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{b}}} \tag{15}$$

with the constraints of $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{a} = 1$ and $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{b} = 1$.

Finally, we obtain $\boldsymbol{a}$ as an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{YX}}$. Similarly, $\boldsymbol{b}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{YX}}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XY}}$. For more specific derivation or other solutions to CCA, readers can refer to [19].

4.2. Co-whitening based on CCA

In canonical correlation analysis we aim to find mutually orthogonal pairs of maximally correlated linear combinations of the variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this way, the objective of CCA can be seen to be equivalent to simultaneous whitening of both $\boldsymbol{X}$ and $\boldsymbol{Y}$, and therefore we could use it to co-whiten short

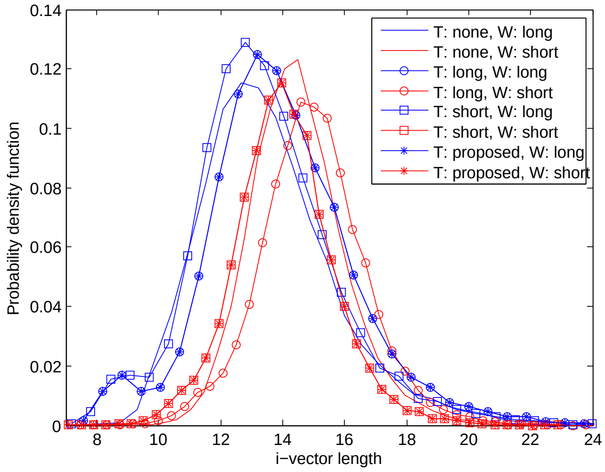

Figure 1: *Histograms of the i-vector length distribution estimated from short and long duration utterances. Notice that T is short for Train, and W is short for Whiten. Specifically, T refers to the duration of i-vectors that we use to train the whitening model, and W indicates the duration of i-vectors we whiten.*

and long utterance i-vectors. Specifically, the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$in Section 4.1 are indexed by $\phi_s(i)$ and $\phi_l(i)$ for short and long utterance i-vector pairs, respectively. The data matrix $\mathbf{\Phi}_s$ containing the observed samples in n rows is $(\phi_s(1), \ldots, \phi_s(i), \ldots, \phi_s(n))^{\mathrm{T}}$, where the empirical mean is

$$\hat{\mu}_{\phi_s} = \frac{1}{n}\sum_{i=1}^{n}\phi_s(i) \tag{16}$$

and the standard unbiased covariance estimate is

$$\hat{\mathbf{\Sigma}}_{\phi_s} = \frac{1}{n-1}\sum_{i=1}^{n}(\phi_s(i) - \hat{\mu}_{\phi_s})(\phi_s(i) - \hat{\mu}_{\phi_s})^{\mathrm{T}} \tag{17}$$

Similarly,

$$\hat{\mathbf{\Sigma}}_{\phi_s\phi_l} = \frac{1}{n-1}\sum_{i=1}^{n}(\phi_s(i) - \hat{\mu}_{\phi_s})(\phi_l(i) - \hat{\mu}_{\phi_l})^{\mathrm{T}} \tag{18}$$

$$\hat{\mathbf{\Sigma}}_{\phi_l} = \frac{1}{n-1}\sum_{i=1}^{n}(\phi_l(i) - \hat{\mu}_{\phi_l})(\phi_l(i) - \hat{\mu}_{\phi_l})^{\mathrm{T}} \tag{19}$$

where

$$\hat{\mu}_{\phi_l} = \frac{1}{n}\sum_{i=1}^{n}\phi_l(i) \tag{20}$$

The eigenvectors form the basis of two whitening matrices, $\mathbf{W}_s = (\boldsymbol{a}(1), \ldots, \boldsymbol{a}(n))^{\mathrm{T}}$ and $\mathbf{W}_l = (\boldsymbol{b}(1), \ldots, \boldsymbol{b}(m))^{\mathrm{T}}$. The whitened i-vectors can be extracted as follows

$$\begin{aligned}\widetilde{\mathbf{\Phi}}_s &= \mathbf{W}_s\mathbf{\Phi}_s \\ \widetilde{\mathbf{\Phi}}_l &= \mathbf{W}_l\mathbf{\Phi}_l\end{aligned} \tag{21}$$

with which, we continue the following steps.

4.3. Comparison of different whitening methods

I-vector is assumed to be normally distributed, which is equivalent to the length of i-vector follow the chi-square distribution. In our experiments, the short and long i-vectors form an one-to-one pairs, that is, the speech for short i-vector is the first 10 seconds truncated from the long duration speech. Fig.1 illustrates the different distributions using different whitening methods. Notice that T is short for Train, and W is short for Whiten. Specifically, T:none indicates no whitening method is considered, T:long, T:short indicates the we train whitening model with long, short duration i-vectors, and T:proposed refers to the whitening model that is trained with the proposed co-whitening approach. W:short means that we plot the whitened short duration i-vectors distribution with the corresponding whitening model, and similarly for W:long. Notice that all the distributions in Fig.1 are all based on i-vectors after LDA with dimension 200, which means that the ideal Chi distribution after whitening is with 200 degrees of freedom.

Four whitening methods are compared in Fig.1. We first analyze the condition without whitening, it is obvious that short and long i-vectors are not identically distributed. Moving to the second whitening approach, where the whitening model is trained with long duration i-vectors. We whiten the short and long duration i-vectors respectively with the trained model, it is clear that the whitened long duration i-vectors are normally distributed, due to the matched duration. Notice that the whitened long duration i-vectors labeled with T:proposed, W:long are also normally distributed as the constraint by (11). Therefore, these two curves overlap with each other. We move on to the third and fourth curves, the distance of these two keep almost the same as the first two curves. From this, we see that using the same whitening model for both short and long duration speech, the distribution could not be normalized. The next two curves further verify this conclusion, where the whitening model is trained with short i-vectors. Finally, we observe the last two curves, which denote the proposed CCA based co-whitening method, it is obvious that the distance between the short and long i-vector length distribution is the closest one as compared to the other three whitening methods. The comparison of these whitening approaches suggest that co-whitening is the best choice. The experimental results will be shown in the following section to confirm the aforementioned theoretical analysis.

5. Experiments

We carried out different tasks to validate the effectiveness of proposed co-whitening method. We choose NIST SRE'10 core-10sec, 8conv-10sec under common condition 5 (CC'5), which are long enrollment vs short test scenarios. Additionally, experiment is conducted on a short vs short task (10sec-10sec). The acoustic features used in the experiments consists of 19-dimensional mel frequency cepstral coefficients (MFCC). Delta and double 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 $\mathbf{T}$ was estimated using NIST SRE'04, 05, and 06 data. The rank of the matrix $\mathbf{T}$ is set to M = 400. The dimensionality of the i-vectors was reduced to 200 with LDA. Length normalization was applied before PLDA [20]. The PLDA model was trained to have 200 speaker factors with a full covariance, while the channel factor is ignored.

Table 1 shows the speaker verification performance of different whitening approaches. The Baseline column contains three methods, where None represents whitening is not considered. Short and Long refer to the techniques that train whitening model using i-vectors estimated from short and long utterances,

Table 1: *Performance comparison under CC'5 of NIST SRE'10 core-10sec, 8conv-10sec and 10sec-10sec tasks. Left: male trials, Right: female trials. Each entry shows the equal error rate (EER) (%) and minimum detection cost function (MinDCF) at the top and bottom rows, respectively.*

Tasks		Male				Female			
		Baseline			Co-whiten	Baseline			Co-whiten
		none	short	long		none	short	long	
core-10sec	EER(%)	6.85	6.83	6.82	**6.63**	**8.23**	8.35	8.26	8.55
	MinDCF	0.91	0.92	0.92	**0.89**	0.84	0.84	0.85	**0.81**
8conv-10sec	EER(%)	2.87	2.60	2.61	**2.50**	4.99	4.81	4.91	**4.76**
	MinDCF	0.60	0.60	**0.60**	0.61	**0.59**	0.69	0.64	0.66
10sec-10sec	EER(%)	14.25	14.37	**13.53**	13.99	13.44	13.53	13.47	**12.87**
	MinDCF	**0.94**	0.96	0.96	0.96	0.97	0.95	**0.95**	0.95

respectively. Co-whiten refers to the proposed co-whitening method. Notice that we use the first 10 seconds of speech truncated from the corresponding long duration speech sample for short i-vectors.

We compare the three methods listed in Baseline columns first. We can conclude that none of the baseline methods outperforms others. From this we gained the insight that, practically, whitening model using long or short i-vectors alone is not sufficient. However, in the 8conv-10sec task for the male trial, we observe about 10% EER reduction for whitening model using either long or short i-vectors, which is more aligned to the theoretical prediction. But it is not robust.

As mentioned, these three baseline methods show comparable results, we simply take the method without whitening to compare with our co-whiten method. We observe that co-whitening helps to improve the performance. This amounts to 3.12%, 13.07% and 1.84% for male trials in EER in the three tasks, respectively. For female trials, co-whitening causes 3.97% slight degradation in EER on core-10sec task. We observe 4.60% and 4.21% improvement on the other two tasks in EER. As for the MinDCF, the range of the relative difference between the proposed method and baseline is from -2.44% to 2.35% for the male trials. For female trials, we observe 3.49% and 1.82% improvement in the core-10sec and 10sec-10sec tasks, and 11.25% degradation in the 8conv-10sec task. Looking back at all the baselines in the 8conv-10sec task for female trial, MinDCF ranges from 0.59 to 0.69, thus the value 0.66 obtained from co-whiten method is reasonable.

Overall, both traditional whitening methods and the proposed method showed comparable results. We observe that the proposed method provides significant improvements of EER for 4 out of 6 tasks.

6. Conclusions

In this paper, we proposed a novel approach to whiten simultaneously the i-vectors extracted from both short and long duration utterances. We first analyzed the distributions of i-vectors estimated from utterances of various duration, and found that i-vectors extracted from short utterances exhibit larger variance. In this regard, we proposed a novel co-whitening method to whiten short and long i-vectors simultaneously. To experimentally validate the proposed co-whitening method, we compared the proposed CCA based co-whitening method with the baseline in different tasks. Notice that the compared baseline is conducted without whitening. We carried out the experiments using enrolled speech data with variable duration, and tested data with 10 seconds. Specifically, we conducted the experiments on core-10sec, 8conv-10sec and 10sec-10sec tasks. We observe that the proposed co-whitening method using CCA can achieve up to 13.07% improvement in EER for the male trials in 8conv-10sec task.

7. Acknowledgments

This work of Longting Xu and Haizhou Li is supported by the Neuromorphic Computing Programme under the RIE2020 Advanced Manufacturing and Engineering Programmatic Grant A1687b0033 in Singapore.

8. References

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

[2] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7649–7653.

[3] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. V. Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7663–7667.

[4] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Duration compensation of i-vectors for short duration speaker verification," *Electronics Letters*, vol. 53, no. 6, pp. 405–407, 2017.

[5] ——, "Twin model g-plda for duration mismatch compensation in text-independent speaker verification," in *INTERSPEECH*, vol. 25, 2016, pp. 1853–1857.

[6] J. Takoua and S. Korbinian, "Probabilistic canonical correlation analysis: A whitening approach," *arXiv:1802.03490 [stat.ME]*, 2018.

[7] R. K. Das and S. R. M. Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *Journal of the Acoustical Society of America*, vol. 140, no. 1, p. 184, 2016.

[8] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[9] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, 2006, pp. I–I.

[10] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[11] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, 2017, pp. 999–1003.

[12] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop*, 2017, pp. 165–170.

[13] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv:1705.02304 [cs.CL]*, 2017.

[14] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *INTERSPEECH*, 2017.

[15] E. Variani, X. Lei, E. Mcdermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4052–4056.

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

[17] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

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

[19] D. Weenink, "Canonical correlation analysis," in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 2003, pp. 81–99.

[20] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.