INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Twin Model G-PLDA for Duration Mismatch Compensation in Text-Independent Speaker Verification

Jianbo Ma[1], *Vidhyasaharan Sethu*[1], *Eliathamby Ambikairajah*[1,2], *Kong Aik Lee*[3]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Australia
[3]Institute for Infocomm Research, A*STAR, Singapore

jianbo.ma@student.unsw.edu.au

Abstract

Short duration speaker verification is a challenging problem partly due to utterance duration mismatch. This paper proposes a novel method that modifies the standard Gaussian probabilistic linear discriminant analysis (G-PLDA) to use two separate generative models for i-vectors from long and short utterances which are jointly trained. The proposed twin model G-PLDA employs distinct models for i-vectors corresponding to different durations from the same speaker but shares the same latent variables. Unlike the standard G-PLDA, this twin model G-PLDA takes the differences between utterances of varying durations into account. Hyper-parameter estimation and scoring formulae for the twin model G-PLDA are presented. Experimental results obtained using NIST 2010 data show that the proposed technique leads to relative improvements of 8.5% and 15.6% when tested on utterances of 5 second and 3 second durations respectively.

Index Terms: automatic speaker verification, short duration speaker verification, i-vector, G-PLDA, twin model G-PLDA

1. Introduction

Automatic speaker verification refers to technology that enables machines to verify a person's identity using their voice samples. Automatic speaker verification systems are broadly categorised into one of two types, namely, text-dependent and text-independent systems. In text-dependent system, contents of pass-phrases are fixed, which provides extra information to identify the speaker [1]. In the text-independent case, speakers are free to speak any phrases and the system cannot rely on prior knowledge of fixed pass-phrases [2]. Whilst, the development of text-independent systems is recognised as a more challenging task than that of text-dependent systems, text-independent systems are also required in a greater number of applications compared to text-dependent systems. Most state-of-the-art text-independent speaker verification systems comprise of i-vectors, which model speaker and channel variability in a low-dimensional representation of speech utterances [3]. These are combined with Probabilistic Linear Discriminant Analysis (PLDA), which serves as back-end to the speaker verification system [4].

Conventionally, text-independent speaker verification systems have required long enrolment and test utterances (e.g. 2 to 3 minute utterances). However, in real applications, it is unreasonable to expect users to speak long sentences in order to verify their identity. Short duration speaker verification would be significantly more practical. Discussions in this paper are confined to this scenario of long enrolment and short test utterances. Enrolment is carried out once in an offline manner. It is therefore reasonable to assume long utterances are available for this. Given this interest in short duration speaker-verification, 10 second test conditions were reintroduced in NIST SRE 2010 [5]. It was observed that although i-vector/G-PLDA system showed better results compared with other factor analysis systems, its performance still degrades sharply once the test utterance durations falls below 10 seconds [6]. This is partly due to the duration mismatch between enrolment and test utterances.

A number of different approaches to deal with this mismatch have been proposed. In [7-9], the covariance of the i-vector posterior probability, which describes the uncertainty, was integrated into the PLDA model. In [10], score domain compensation for duration mismatch using Quality Measure Function (QMF), which takes durations of enrolment and test utterances into account, was introduced. In [11], it was demonstrated that it is beneficial to use short utterances to train hyper-parameters of the total variability model. Statistical content matching was proposed in [12] and performed well in conditions where contents of test utterances have been spoken in enrolment utterances. However, this did not generalize to the text-independent case. Last but not least, a local variability model to capture variability in each component in the UBM was proposed in [13], although the results do not indicate that it outperforms the standard i-vector system.

In this paper, we propose a new method to handle duration mismatch when using long utterance for enrolment and short utterance for testing. Specifically, we relax the assumption that i-vectors from both long and short utterances have the same distribution and modify the standard G-PLDA model to have two sets of hyper-parameters that are jointly trained by i-vectors from both long and short utterances. In this paper we refer to this as the twin model G-PLDA. Expectation Maximization (EM) algorithm for the estimation of the proposed model as well as scoring function are presented.

2. Standard G-PLDA paradigm

I-vectors have become the de-facto technique to obtain fixed and low-dimensional representations of speech utterances for speaker verification [3]. The standard Gaussian PLDA (G-PLDA) is a generative model of the i-vectors that has been successfully applied to deal with channel variability in speaker verification system [4]. Given a set of i-vectors $\chi = \{x_{ij}; i = 1,2,\cdots,S; j = 1,2\cdots,N_i\}$, where x_{ij} denotes the i-vector corresponding to the j^{th} utterance from the i^{th} speaker, G-PLDA decomposes them as:

http://dx.doi.org/10.21437/Interspeech.2016-683

$$x_{ij} = \mu + \Phi h_i + \varepsilon_{ij} \qquad (1)$$

where Φ is a factor loading matrix, h_i is a vector of latent variables which have a standard Gaussian distribution, $N(0, I)$, and ε_{ij} is a residual term that is assumed to be Gaussian with zero mean and a full covariance matrix denoted by Σ. The latent variables (elements of h_i) are assumed to be statistically independent. By marginalizing over the latent variables, it can be shown that the i-vectors follow a normal distribution given by $\mathcal{N}(\mu, \Phi\Phi^T + \Sigma)$.

Based on this model, given an enrolment i-vector x_e and a test i-vector x_t from a trial, the log-likelihood ratio between the hypothesis that the two i-vectors are from the same speaker versus the hypothesis that they are from different speakers is calculated [14] as follows

$$Score(x_e, s_t) =$$
$$\log\left(\mathcal{N}\left(\begin{bmatrix} x_e \\ x_t \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Phi\Phi^T + \Sigma & \Phi\Phi^T \\ \Phi\Phi^T & \Phi\Phi^T + \Sigma \end{bmatrix}\right)\right) -$$
$$\log\left(\mathcal{N}\left(\begin{bmatrix} x_e \\ x_t \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Phi\Phi^T + \Sigma & 0 \\ 0 & \Phi\Phi^T + \Sigma \end{bmatrix}\right)\right) \qquad (2)$$

3. Duration Mismatch in i-vectors

As mentioned in Section 2, I-vectors are assumed to follow a standard normal distribution after whitening [14], and consequently the length of i-vector should follow a chi-square distribution. Histograms of the length (magnitude) of 200 dimensional i-vectors from long and short utterances are plotted in Figure 1. These are obtained from 9,189 i-vectors estimated from NIST SRE'04, 05, 06, 08, Switchboard II Part 1, 2, 3 and Switchboard Cellular Part 1 & 2 full conversation utterances. Correspondingly, the 9,189 i-vectors for short utterances are extracted from utterances by truncating these full utterances and using the first 10 seconds. Note that i-vectors are transformed by linear discriminative analysis (LDA) and within-class covariance normalization (WCCN) before whitening. From Figure 1 it can be seen that the histograms of the length of i-vectors from long and short segments are very distinct, suggesting that both long and short i-vectors are not identically distributed.

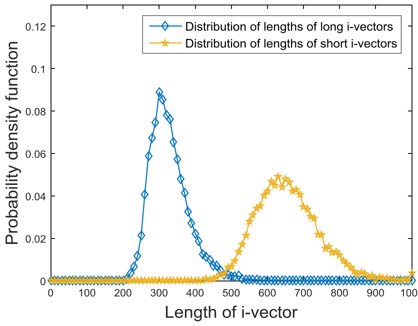

Figure 1: *Histograms of i-vector lengths (magnitudes) estimated from long and short duration utterances.*

In addition to comparing these histograms, a second measure of differences between the distribution of i-vectors from long and short utterances based on the Partition Coefficient [15, 16] is employed in this paper. The partition coefficient is an index that indicates the clustering tendency in a dataset and lies in the range $[1/k, 1]$, where k is the number of clusters. A partition coefficient value close to unity indicates that the dataset is better clustered into these k clusters. In this paper, we use the partition coefficient to test if i-vectors from different durations follow different distributions. A Gaussian mixture model with two components $(K = 2)$ was trained using length normalised i-vectors from short and long utterances. The partition coefficient (PC) is then defined as per equations (3) and (4), where μ, Σ are the mean and covariance of each Gaussian mixture component.

$$PC = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\zeta_{ik}^2 \qquad (3)$$

$$\zeta_{ik} = \frac{N(x_i|\mu_k, \Sigma_k)}{\sum_{r=1}^{K} N(x_i|\mu_r, \Sigma_r)} \qquad (4)$$

The partition coefficient estimated from the long and short utterances used to generate Figure 1 was 0.837 and given there are two clusters in this case (long and short utterance), the range of values it could have taken is [0.5,1]. This high value of partition coefficient suggests that i-vectors from long and short duration utterances have high clustering tendency and are likely to have two different distributions.

The comparison of the histograms of lengths of i-vectors estimated form long and short utterances, and the partition coefficient estimated from normalised i-vectors corresponding to long and short utterances both suggest that modelling long and short duration i-vectors with the same Gaussian distribution, as is the case of the standard G-PLDA, may be inaccurate. Motivated by these limitations, we propose the twin model G-PLDA to address this problem.

4. Duration Mismatch Compensation

4.1. Proposed Twin model G-PLDA

In the proposed twin model G-PLDA, our assumption is that i-vectors from the same speaker still share identical normally distributed latent variables. However, two independent sets of factor analysis hyper-parameters are utilised to account for the mismatch between long and short duration utterances. i.e., instead of one unified 'path', we revise the standard G-PLDA model as indicated in Figure 2.

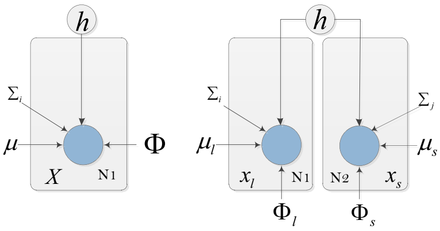

Figure 2: (a) *standard G-PLDA;* (b) *Twin Model G-PLDA*

The twin model G-PLDA can be written as:

$$x = \begin{cases} \mu_L + \Phi_L h + \varepsilon_L, & \text{for long utterances} \\ \mu_s + \Phi_s h + \varepsilon_s, & \text{for short utterances} \end{cases} \qquad (5)$$

where x denotes the i-vector; μ_L and μ_s are mean vectors for i-vector correspond to long and short utterances, respectively; Φ_L and Φ_s are the corresponding factor loading matrices; h is the vector of normally distributed latent variables and is shared by all the utterances from the same speaker. ε_l and ε_s are residuals which are different for different utterances and are assumed to be normally distributed with zero mean and covariances given by the matrices Σ_L and Σ_s for long and short utterances respectively. Thus, the hyper-parameters, $\theta = \{\mu_L, \Phi_\text{L}, \Sigma_{\text{L},}\mu_s, \Phi_\text{s}, \Sigma_\text{s}\}$, completely describe the Twin Model G-PLDA and will model the differences between long and short durations as well as within-speaker similarities of i-vectors.

As with the standard G-PLDA, the likelihood-ratio score for speaker verification is obtained by calculating the likelihood of two hypotheses. Specifically, given an enrolment i-vector x_e from a long utterance and a test i-vector x_t from a short utterance, the two hypotheses of interest are: H_s that x_e and x_t share the same latent variable h; and H_d that x_e and x_t are generated by different latent variables. Figure 3 shows the graphical models corresponding to the two hypotheses.

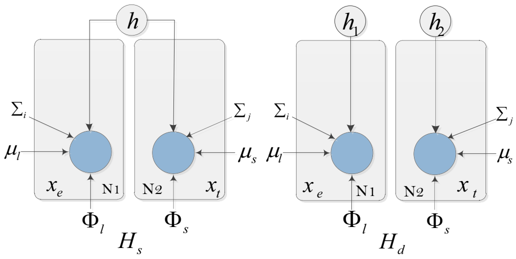

Figure 3: (a) *Hypothesis that test and enrolment i-vectors are from same speaker* (*share latent variables - h*)*;* (b) *Hypothesis that test and enrolment i-vectors are from different speakers* (*distinct latent variables -* h_1 *and* h_2)

In the twin model G-PLDA equations, i-vectors are assumed to be conditional independent. If two i-vectors share the same latent variables (H_s), factor loading matrices and other parameters can be concatenated to share the same latent variable. Similarly, if two i-vectors are generated by different latent variables (H_d), we can augment the two latent variables as they are assumed to be independent. Factor loading matrices are concatenated into a block diagonal matrix. Thereby we have developed the following equations for the two hypotheses:

$$H_s: x' = \mu' + Ah_s + \varepsilon' \tag{6}$$

$$H_d: x' = \mu' + Bh_d + \varepsilon' \tag{7}$$

where,

$$x' = \begin{bmatrix} x_e \\ x_t \end{bmatrix}, \mu' = \begin{bmatrix} \mu_L \\ \mu_s \end{bmatrix},\ A = \begin{bmatrix} \Phi_L \\ \Phi_s \end{bmatrix}, B = \begin{bmatrix} \Phi_L & 0 \\ 0 & \Phi_s \end{bmatrix}, \varepsilon' = \begin{bmatrix} \varepsilon_L \\ \varepsilon_s \end{bmatrix}.$$

In order to get the likelihood of each hypothesis, we evaluate the following likelihood function:

$$P(x'|h_s) = \mathcal{N}(\mu' + Ah_s, \begin{bmatrix} \Sigma_L & 0 \\ 0 & \Sigma_s \end{bmatrix}) \tag{8}$$

$$\mathrm{P(h_s)} = \mathcal{N}(0, \mathrm{I}) \tag{9}$$

Thus, the marginal likelihood for the hypothesis, H_s is:

$$P(x_e, x_t|H_s) = \mathcal{N}(\mu', AA^T + \begin{bmatrix} \Sigma_L & 0 \\ 0 & \Sigma_s \end{bmatrix}) \tag{10}$$

Similarly, the marginal likelihood for H_d is:

$$P(x_e, x_t|H_d) = \mathcal{N}(\mu', BB^T + \begin{bmatrix} \Sigma_L & 0 \\ 0 & \Sigma_s \end{bmatrix}) \tag{11}$$

The log likelihood ratio is now given by the difference between the logarithms of these two probabilities as:

$$\begin{aligned} &Score(x_e, s_t) \\ &= \log\left(\mathcal{N}\left(\begin{bmatrix} x_e \\ x_t \end{bmatrix}; \begin{bmatrix} \mu_L \\ \mu_s \end{bmatrix}, \begin{bmatrix} \Phi_L{\Phi_L}^T + \Sigma_L & \Phi_L{\Phi_s}^T \\ \Phi_s{\Phi_L}^T & \Phi_s{\Phi_s}^T + \Sigma_s \end{bmatrix}\right)\right) \\ &- \log\left(\mathcal{N}\left(\begin{bmatrix} x_e \\ x_t \end{bmatrix}; \begin{bmatrix} \mu_L \\ \mu_s \end{bmatrix}, \begin{bmatrix} \Phi_L{\Phi_L}^T + \Sigma_L & 0 \\ 0 & \Phi_s{\Phi_s}^T + \Sigma_s \end{bmatrix}\right)\right) \end{aligned} \tag{12}$$

We can see that the scoring equation given by (12) has the same structure as that given by (2). Also, note that if we set $\Phi_\mathrm{L} = \Phi_\mathrm{s}$, $\mu_L = \mu_s$ and $\Sigma_\mathrm{L} = \Sigma_\mathrm{s}$ in (12), we return to the standard G-PLDA scoring.

4.2. Twin Model G-PLDA Parameter Estimation

Hyper-parameters of standard G-PLDA are estimated using the EM algorithm from background i-vectors. In the twin model G-PLDA, two sets of hyper-parameters associated with both long and short i-vectors from the same speaker should be tied to one unique set of speaker latent variables. This is different to standard G-PLDA parameter estimation. We will show the derivation of the EM algorithm for this particular G-PLDA below.

Let $\theta = \{\mu_L, \Phi_\mathrm{L}, \Sigma_\mathrm{L}, \mu_s, \Phi_\mathrm{s}, \Sigma_\mathrm{s}\}$ denote the parameters that need to be estimated. Let x_L and x_s represent i-vectors from long and short utterances, respectively, and let h represent the latent variable. In the standard G-PLDA, i-vectors from one speaker will form a class and share one latent variable. The posterior expectation $\mathrm{E}[h]$ is then obtained by using the factor analysis model as,

$$E[h] = (\Phi^T\Sigma^{-1}\Phi + I)^{-1}\Phi^T\Sigma^{-1}(x - \mu) \tag{13}$$

In the proposed twin model G-PLDA, there are both long and short duration i-vectors from the same speaker that share the same latent variables. To estimate the model parameters, merged i-vectors will be created by concatenating one i-vector from long utterance with one from a short utterance from the same speaker. The E-step is then formulated as:

$$E[h] = (\mathrm{A}^T\Sigma'^{-1}\mathrm{A} + I)^{-1}\mathrm{A}^T\Sigma'^{-1}(x_m - \mu') \tag{14}$$

$$E[hh^T] = (\mathrm{A}^T\Sigma'^{-1}\mathrm{A} + I)^{-1} + E[h]E[h]^T \tag{15}$$

where, $\Sigma' = \begin{bmatrix} \Sigma_L & 0 \\ 0 & \Sigma_s \end{bmatrix}$, and $x_m = \begin{bmatrix} x_L \\ x_s \end{bmatrix}$.

In the M-step, we optimize the auxiliary function

$$Q(\theta, \theta_{old}) = \sum_i \sum_j \int p(h_i|X, \theta_{old}) log[p(x_{ij}|h_i)p(h_i)]dh_i \tag{16}$$

where, h_i denotes the latent variables corresponding to the i^{th} speaker, x_{ij} denotes the j^{th} i-vector (both long and short utterance i-vector) from the i^{th} speaker, X denotes i-vectors from all training speakers, and θ_{old} denotes the model hyper-parameters from the previous iteration of the EM algorithm.

By taking the derivatives with respect to θ and set it to zero, we obtain the following update equations:

$$\mu_L = \frac{1}{N_L}\sum_{i,j} x_{L_{ij}} \tag{17}$$

$$\Phi_L = \left(\sum_{i,j}\left(x_{L_{ij}} - \mu_L\right)E[h_i]^T\right)\left(\sum_{i,j}E\left[h_i{h_i}^T\right]\right)^{-1} \tag{18}$$

$$\begin{aligned} \Sigma_L = \frac{1}{N_L}\sum_{i,j}\Big[&\left(x_{L_{ij}} - \mu_L\right)\left(x_{L_{ij}} - \mu_L\right)^T \\ &- \Phi_L E[h_i]\left(x_{L_{ij}} - \mu_L\right)\Big] \end{aligned} \tag{19}$$

where $x_{L_{ij}}$ denotes the i-vector corresponding to the j^{th} long utterance from the i^{th} speaker, and

$$\mu_s = \frac{1}{N_s}\sum_{i,j} x_{s_{ij}} \tag{20}$$

$$\Phi_s = \left(\sum_{i,j} \left(x_{s_{ij}} - \mu_s \right) E[h_i]^T \right) \left(\sum_{i,j} E\left[h_i {h_i}^T\right] \right)^{-1} \tag{21}$$

$$\Sigma_s = \frac{1}{N_s} \sum_{i,j} \left[\left(x_{s_{ij}} - \mu_s \right) \left(x_{s_{ij}} - \mu_s \right)^T - \Phi_s E[h_i] \left(x_{s_{ij}} - \mu_s \right) \right] \tag{22}$$

where $x_{s_{ij}}$ denotes the i-vector corresponding to the j^{th} short utterance from the i^{th} speaker.

5. Experiments and Discussion

A number of experiments were carried out to analyse the effectiveness of the proposed twin model G-PLDA. The 8CONV-10SEC condition (condition 5) of the NIST SRE'10 [5] was chosen for these experiments. Two additional conditions were created by truncating the 10 seconds test utterances to 5 and 3 seconds (using the first 5 seconds and 3 seconds of each utterance). We name these conditions 8CONV-5SEC and 8CONV-3SEC, respectively.

The baseline system is an i-vector/G-PLDA system. Standard MFCC features of 13 dimensions with their first and second derivatives were used in conjunction with a vector quantization model based voice activity detector [17] prior to feature warping [18]. Only male speakers were considered in the experiments reported in this paper and a gender-dependent universal background model (UBM) of 1024 Gaussian mixtures was created using 2040 utterances from male speakers from NIST SRE'04, 05, 06, 08, Switchboard II Part 1, 2, 3 and Switchboard Cellular Part 1 and 2. In selecting the data for training the UBM, one utterance was chosen from each speaker's available data to retain speaker diversity while reducing the overall amount of data [19]. A T matrix of rank 400 was estimated using 29,000 utterances from the 2040 speakers. I-vectors were computed for each of the development, training and test utterances using the estimated T-matrix. LDA was then applied to further reduce the dimension to 200 and followed by WCCN. I-vectors were then radial Gaussianised followed by length normalization as described in [14].

Table 1 summarises the results of the standard G-PLDA system trained on utterances of varying durations. In particular, the parameters of the G-PLDA model were trained on utterances with duration varied from 3 seconds to 2.5 minutes. For all results presented in the table, speakers were enrolled using 8 utterances of about 2.5 minutes, while the test segments were set as 10s, 5s and 3s.

The results are consistent with those in [11], which suggest that for short duration speaker verification, it is not optimal to use full utterances in the G-PLDA hyper-parameters training phase. We observed that using short development utterances (e.g. 15 seconds) benefits the shorter test scenarios of 5 seconds and 3 seconds. This may be because speaker factors estimated from long utterances do not adequately characterise the short duration utterances. By making a compromise between long and short durations by using relatively short (15sec) development utterances, a better model that characterises both long and short utterances might be obtained. These results reinforce the clear need to have a model that can take into account differences between short and long utterances.

Table 1 also shows the accuracies of speaker verification systems employing the twin model G-PLDA. As there are two sets of hyper-parameters in the twin model G-PLDA, i-vectors from both long and short (truncated) utterances are needed to train the parameters. Thus in this experiment, i-vectors from full 2.5 minutes utterances are used along with i-vectors from truncated utterances of varying durations (given in Table 1) to estimate the two sets of hyper-parameters. Enrolment and test utterances are identical to the ones used with the standard G-PLDA. It is clear from the results that the proposed method outperforms the baseline approach for all 3 short test utterance durations. For 5 seconds and 3 seconds test condition, we obtained the best performance when using truncated utterances of 10s duration to train the twin model G-PLDA hyper-parameters. Relative improvements of 8.5% and 15.6% were observed for the 5sec and 3sec conditions respectively when comparing the proposed twin model G-PLDA to the standard G-PLDA. When compared with the best results obtained by standard G-PLDA (15s training data), 3.4% and 6.9% relative improvements are observed for the 5sec and 3sec test duration conditions respectively.

It was observed that when the duration of training utterances fall below 10 seconds, the performance of the overall system drops again. This is probably the result of not having sufficient training data frames for the estimation of the model parameters. Similarly, although the improvement in the 10 seconds condition was minor compared to the baseline, the trends remained consistent. One reason why the improvement is so small may be because the proposed method is more useful when dealing with the more severe cases such as the 5 seconds and 3 seconds tests than with the slightly longer duration 10 second tests. For longer utterances, as the mismatch between enrolment and test utterances is not as severe, the two sets of hyper-parameters of the discriminative G-PLDA will be similar and the proposed method will not be much more efficient than the standard G-PLDA.

Table 1. *Performance* (*equal error rate*) *using standard and the proposed twin model G-PLDA on SRE'10 8CONV-10SEC and additional 5sec and 3sec conditions* (*male speakers only*).

Training data	Test duration					
	Standard G-PLDA			Twin Model G-PLDA		
	10s	5s	3s	10s	5s	3s
3s	12.21	15.48	20.35	9.78	13.37	18.02
5s	11.05	13.95	18.60	8.52	12.79	16.28
10s	8.14	13.15	17.27	6.98	**11.80**	**15.70**
15s	6.93	**12.21**	**16.86**	6.85	11.86	16.28
20s	6.40	13.37	17.55	6.98	12.79	16.77
30s	5.81	12.21	18.02	6.51	12.21	17.82
2.5min	6.40	12.89	18.60	6.40	12.89	18.60

6. Conclusions

In this paper, we proposed a novel method to deal with duration mismatch in the case of long enrolment and short test speaker verification. Initial experimental results demonstrated that i-vectors from long and short utterances have distinct distributions which contradict the assumptions in standard G-PLDA models. Hence, the standard G-PLDA model was modified to have two separate generative paths and jointly trained by i-vectors from long utterances and short utterances. Scoring equations were developed as well. The efficacy of the proposed technique was validated on the NIST SRE'10 8CONV-10SEC male condition and additional shorter duration conditions using the truncated 5 and 3 seconds test data. The proposed twin model G-PLDA additionally provides a new avenue for utterance mismatch compensation using twin i-vector transformations, which will be pursued in future work.

7. References

[1] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication,* vol. 60, pp. 56-77, 2014.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication,* vol. 52, pp. 12-40, 2010.

[3] 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, pp. 788-798, 2011.

[4] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey*, 2010, p. 14.

[5] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[6] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2011, pp. 2341-2344.

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

[8] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7644-7648.

[9] S. Cumani, O. Plchot, and P. Laface, "On the use of i–vector posterior distributions in Probabilistic Linear Discriminant Analysis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on,* vol. 22, pp. 846-857, 2014.

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

[11] A. K. Sarkar, D. Matrouf, P.-M. Bousquet, and J.-F. Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification," in *INTERSPEECH*, 2012, pp. 2662-2665.

[12] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *INTERSPEECH*, 2014, pp. 1317-1321.

[13] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Local variability modeling for text-independent speaker verification," in *Proceedings of Odyssey: Speaker and Language Recognition Workshop*, 2014.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, 2011, pp. 249-252.

[15] M.-W. Mak, X. Pang, and J.-T. Chien, "Mixture of PLDA for Noise Robust I-Vector Speaker Verification," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on,* vol. 24, pp. 130-142, 2016.

[16] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems,* vol. 17, pp. 107-145, 2001.

[17] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *ICASSP*, 2013, pp. 7229-7233.

[18] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[19] T. Hasan and J. H. Hansen, "A study on universal background model training in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1890-1899, 2011.