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

# Incorporating Local Acoustic Variability Information into Short Duration Speaker Verification

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

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]DATA61, CSIRO, Sydney, Australia
[3]Institute for Infocomm Research, A*STAR, Singapore

jianbo.ma@unsw.edu.au

## Abstract

State-of-the-art speaker verification systems are based on the total variability model to compactly represent the acoustic space. However, short duration utterances only contain limited phonetic content, potentially resulting in an incomplete representation being captured by the total variability model thus leading to poor speaker verification performance. In this paper, a technique to incorporate component-wise local acoustic variability information into the speaker verification framework is proposed. Specifically, Gaussian Probabilistic Linear Discriminant Analysis (G-PLDA) of the supervector space, with a block diagonal covariance assumption, is used in conjunction with the traditional total variability model. Experimental results obtained using the NIST SRE 2010 dataset show that the incorporation of the proposed method leads to relative improvements of 20.48% and 18.99% in the 3 second condition for male and female speech respectively.

**Index Terms**: speaker verification, short duration, i-vector, probabilistic LDA, supervector

## 1. Introduction

Most state-of-the-art text-independent speaker verification systems are comprised of i-vectors, which model speaker and channel variability in a low-dimensional representation of speech utterances [1]. These are combined with Probabilistic Linear Discriminant Analysis (PLDA), which serves as a back-end to the speaker verification system [2]. Text-independent speaker verification systems conventionally require long enrolment utterances and operate on long test utterances (with typical duration ranging from 2 to 3 minutes). In practical applications, short duration speaker verification would be significantly more desirable. It should be noted that since enrolment is carried out only once and in an offline manner, it is still reasonable to assume long utterances can be used for this purpose. Discussions in this paper are therefore confined to this scenario of long enrolment and short test utterances.

In recent years there has been increasing interest in short duration text independent speaker verification systems, almost all of which focuses on the aforementioned i-vector PLDA approach. A twin model G-PLDA [3] was proposed to compensate for the duration mismatch between i-vectors of long enrolment and short test utterances. The covariance of the i-vector posterior probability was propagated to the PLDA model in [4-6]. Score domain compensation was introduced in [7] for duration mismatch using a quality measure function (QMF) that takes durations of enrolment and test utterances into account. The mismatch between long enrolment and short test duration was compensated for in [8] by adding short utterances in the training phase of a total variability matrix and the hyper-parameters of PLDA.

As described in [9], the idea behind i-vector modelling is that supervector representations of utterances can be mapped to a low dimensional space with little loss of accuracy. One of the main advantages of the i-vector framework is that channel variability can be compensated for by using techniques such as Linear Discriminant Analysis (LDA), Within Class Covariance Normalization (WCCN), and PLDA in this low-dimensional space. However, as the duration of utterance decreases, the uncertainty of i-vector representation increases and speaker verification performance degrades sharply once test utterance durations fall below 10 seconds [10]. It was also shown in [11] that phonetic mismatch has a greater influence than speaker-channel variability for short duration speaker verification. It was shown that the basic Gaussian Mixture Model-Universal Background Model (GMM-UBM) based methods are superior to subspace methods [12-14] when addressing text-dependent speaker verification with extremely short utterance (e.g. 3 seconds).

Supervectors can be regarded as representations of GMMs that differ only in their mixture means [15] and since the total variability model may only provide an incomplete representation for short durations utterances, direct modelling of the supervectors may be beneficial. Parameter tying across mixtures in the total variability model is relaxed in [16, 17] and banks of local variability vectors or concatenated local vectors are obtained. G-PLDA was then trained on top them. Our solution of modelling local acoustic variability is different. In this paper, we propose that as uncertainty of i-vector representation increases sharply for short duration utterances, we bypass the latent variable model and directly capture local acoustic variability information in the supervector space. Following this, different weighting strategies are applied in order to take the relative reliability of local acoustic information into account.

Specifically, we start by showing that as test utterances become short, the mismatch in terms of zero order statistics between test utterances and enrolment utterance become obvious. We then propose the use of modified G-PLDA in the supervector space in order to capture information about local acoustic variability in addition to the standard i-vectors, which are based on the total variability model. Here the term total variability refers to the fact that the i-vectors are lower dimensional representations of higher dimensional supervectors. The term local acoustic variability refers to the fact that the proposed method operates on sub-vectors of the supervector corresponding to each component of the UBM independently. Finally, to account for relative reliability of the



http://dx.doi.org/10.21437/Interspeech.2017-266

mean vector for each utterance, we propose likelihood, mean vector and scoring weighting to be used in hyper-parameter estimation and scoring stages.

## 2. Total Variability Model

Let $M$ be a supervector obtained by concatenating the mean vectors of all components of a GMM. The i-vector corresponding to $M$ is given by the well-established total variability model as follows:

$$M = M_0 + Tx \tag{1}$$

where, $M_0$ is the supervector corresponding to the universal background model (UBM), $T$ is the total variable matrix of a low rank $R$ (e.g 400), and $x$ denotes latent variables that follow a normal distribution. The i-vector is the expected value of the latent variables $x$ and is given by:

$$E(x) = \left(I + \sum_c N_c T_c^* \hat{\Sigma}_c^{-1} T_c\right)^{-1} \left(\sum_c T_c^* \hat{\Sigma}_c^{-1} F_c\right) \tag{2}$$

where $T_c$ is the $P \times R$ dimensional sub-matrix of $T$ corresponding to the $c^{th}$ Gaussian mixture component of the UBM, $C$ and $P$ are the number of components in the UBM and the dimensionality of the feature space respectively. $\hat{\Sigma}_c$ is the covariance of $c^{th}$ component of the UBM and $N_c$ and $F_c$ are the zero- and first-order statistics of $c^{th}$ component respectively.

Equation (2) suggests that if there are insufficient feature vectors under any component of the UBM, all of the latent variables will be influenced. This is not a problem in long duration utterances as every component is likely to be adequately visited and consequently the zero-order statistics are statistically stable. However, these properties are not retained in the short duration utterances because of a lack of phonetic diversity, and consequently i-vectors may not be representative of the entire acoustic landscape.

## 3. Analyses of Short Duration Utterances

To analyse the effect of short duration utterances, a preliminary experiment was conducted, comparing zero-order statistics (herein referred to as *N*-vectors) as front-ends to i-vectors with G-PLDA based speaker verification system for long and short test utterances. *N*-vectors and i-vectors were estimated from utterances in NIST SRE '04, '05, '06 and '08, Switchboard II Parts 1, 2 and 3, Switchboard Cellular Parts 1 and 2, and NIST SRE'10 [18] 8CONV-10SEC, where test utterances are around 10 seconds in length, and 8CONV-CORE, where they are around 2.5 minutes. 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 results of these are compared in Table 1, where it can be seen that the difference in accuracy between *N*-vector and i-vector based systems are more pronounced in the case of short duration utterances (10s and less). However, in the case of long duration utterances, comparable results can be obtained by using *N*-vectors or i-vectors. Since the *N*-vectors (zero-order statistics) represent only the mixture occupancy of the UBM while the i-vectors are representative of the feature space distribution corresponding to the utterance using the UBM as prior information, *N*-vectors can be expected to be much more sensitive to variations in phonetic distributions across utterances. The results in Table 1 support the idea that zero-order statistics from short utterances are much less stable compared to those from long duration utterances and that short utterances lack phonetic diversity and contain less information. This is subsequently reflected in i-vectors inferred from these statistics, and consequently plays a significant role in the degradation of the performance of short duration text independent speaker verification systems.

Table 1. *Performance* (*equal error rate (EER) %*) *using i-vector and N-vector on SRE'10 8CONV-10SEC and 8CONV-CORE conditions* (*male speakers only*).

| Condition | EER (%) | |
|---|---|---|
| | i-vector | *N*-vector |
| 8CONV-CORE | 1.51 | 2.41 |
| 8CONV-10SEC | 5.03 | 22.45 |
| 8CONV-5SEC | 10.73 | 35.57 |
| 8CONV-3SEC | 17.68 | 38.77 |

Table 2. *Trace of covariance matrix of supervectors in i-vector framework.*

| Measure | Duration | | | |
|---|---|---|---|---|
| | 2.5min | 10sec | 5sec | 3sec |
| $\bar{\sigma}$ | 3.2 | 511.7 | 1107.2 | 1974.3 |

A second observation about the effect of utterance duration can be made in terms of covariance matrices of the supervectors within the i-vector framework [9]. A basic tenet of the i-vector framework is that i-vectors are Maximum a Posterior (MAP) estimates and the covariance of the supervector, B, is related to the uncertainty of estimated mean, M (Corollary 1 of [9]). The larger the uncertainty, the less accurate the i-vector representing that supervector will be. Equation (3) shows the use of the trace of covariance matrix as a measure of this uncertainty:

$$\bar{\sigma} = \frac{1}{H} \sum_i trace(B_i) \tag{3}$$

where, $H$ is the total number of utterances. This measure of uncertainty is estimated for utterances of 2.5min, 10s, 5s and 3s durations and results are given in Table 2. It can be observed that as the duration decreases, the uncertainty increases dramatically.

The mismatch of zero-order statistics between long and short duration utterances, caused by insufficient phonetic content of short utterances, is likely to be the reason for this uncertainty in the i-vector representation. The total variability matrix that maps zero and first-order statistics of each component into the total variability space is trained on long utterances and is fixed. When there is insufficient information pertaining to some components of the background model in short utterances, the distribution of all latent variables is influenced as a whole. A technique that is able to differentiate and compare component-wise information of utterances could be beneficial in this scenario and could complement the i-vector framework. Motivated by these observations, we propose the use of G-PLDA on the supervector space, with the assumption of a block-diagonal covariance matrix, to directly capture and compare local (component-wise) acoustic variability in the supervector space.

## 4. Proposed Method

### 4.1. G-PLDA in Supervector Space

In short duration speaker verification, we take into consideration both channel variability and phonetic variability

in the supervector space. Given a collection of supervectors, $\mathcal{D} = \{\mathcal{M}_{ij}; i = 1,2,\cdots,S; j = 1,2\cdots,H_i\}$, where $\mathcal{M}_{ij}$ is the supervector (after centring) corresponding to the $j^{th}$ utterance from the $i^{th}$ speaker. The generative model is prescribed as:

$$\begin{bmatrix} \mathcal{M}_{i1} \\ \vdots \\ \mathcal{M}_{iH_i} \end{bmatrix} = \begin{bmatrix} V \\ \vdots \\ V \end{bmatrix} z_i + \begin{bmatrix} \bar{\varepsilon}_{i1} \\ \vdots \\ \bar{\varepsilon}_{iH_i} \end{bmatrix} \quad (4)$$

where $V$ is a factor loading matrix and it is of $CP \times D$ dimension ($D \leq CP$), $z_i$ is a vector of latent variables which have a standard Gaussian distribution, $N(0, I)$, and $\bar{\varepsilon}_{i,j}$ are residual terms, assumed to be Gaussian distributed with zero mean and a $CP \times CP$ covariance matrix denoted by $\bar{\Sigma}$. The first two moments of the latent variables are then calculated as follows:

$$E(z_i) = (I + H_i V^* \bar{\Sigma}^{-1} V)^{-1} \sum_j V^* \bar{\Sigma}^{-1} \mathcal{M}_{ij} \quad (5)$$

$$E(z_i z_i^*) = (I + H_i V^* \bar{\Sigma}^{-1} V)^{-1} + E(z_i) E(z_i^*) \quad (6)$$

Here we assume that the covariance is block diagonal, which imposes the underlying assumption that the mixture components are independent while preserving covariance information within components (local covariance) in addition to reducing the computational burden. Using the matrix inverse lemma [19], the estimation of the moments of the latent variables can then be broken down as follows:

$$E(z_{ic}) = (I + H_i V_c^* \bar{\Sigma}_c^{-1} V_c)^{-1} \sum_j V_c^* \bar{\Sigma}_c^{-1} \mathcal{M}_{ijc} \quad (7)$$

$$E(z_{ic} z_{ic}^*) = (I + H_i V_c^* \bar{\Sigma}_c^{-1} V_c)^{-1} + E(z_{ic}) E(z_{ic}^*) \quad (8)$$

where, the subscript $c$ denotes the $c^{th}$ block of the corresponding parameter or variable.

Given an enrolment supervector $\mathcal{M}_e$ and a test supervector $\mathcal{M}_t$ from a trial, the score is calculated as follows:

$$Score(\mathcal{M}_e, \mathcal{M}_t) = \sum_c \log S_{1c} - \log S_{0c} \quad (9)$$

where

$$S_{1c} = \mathcal{N}\left(\begin{bmatrix} \mathcal{M}_{ec} \\ \mathcal{M}_{tc} \end{bmatrix}; 0, \begin{bmatrix} V_c V_c^* + \bar{\Sigma}_c & V_c V_c^* \\ V_c V_c^* & V_c V_c^* + \bar{\Sigma}_c \end{bmatrix}\right) \quad (10)$$

$$S_{0c} = \mathcal{N}\left(\begin{bmatrix} \mathcal{M}_{ec} \\ \mathcal{M}_{tc} \end{bmatrix}; 0, \begin{bmatrix} V_c V_c^* + \bar{\Sigma}_c & 0 \\ 0 & V_c V_c^* + \bar{\Sigma}_c \end{bmatrix}\right) \quad (11)$$

Parameter estimation and scoring can be processed in a component-wise manner. The expectation maximization (EM) algorithm is used to estimate the hyper-parameters. A comparison between an i-vector/G-PLDA system and the proposed method, G-PLDA in supervector space, has been presented in Figure 1.

### 4.2. Likelihood weighting

In the framework proposed in Section 4.1, during the hyper-parameter training stage, the mean vector $\mathcal{M}_{ijc}$ of $c^{th}$ component of the supervectors corresponding to each session is assumed to be the same across the training data. This is not a satisfactory assumption as the number of frames aligned to each component of UBM is not equal for each utterance. In order to remedy this, the EM algorithm's M-step is modified to take into account the relative reliability of the mean vectors.

Let $\theta_c = \{V_c, \bar{\Sigma}_c\}$ denote the $c^{th}$ block of the parameters to be estimated. The M-step maximises the log likelihood. The M-step maximises the log likelihood. The auxiliary function is

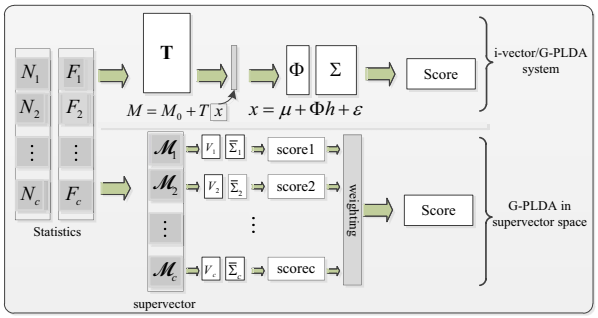


Figure 1. *Comparison between i-vector/G-PLDA system (upper panel) and G-PLDA in supervector space system (lower panel).*

$$Q(\theta_c'|\theta_c) = \mathbb{E}_z\{\log P(\mathcal{D}_c, Z_c|\theta')|X, \theta\} = \sum_i \sum_j \mathcal{L}_{ijc} \quad (12)$$

where,

$$\mathcal{L}_{ijc} = \mathbb{E}_z\left\{\left[log \frac{1}{(2\pi)^{\frac{d}{2}} |\bar{\Sigma}_c|^{\frac{1}{2}}} - \frac{1}{2}(\mathcal{M}_{ijc} - V_c z_{ic})^T \bar{\Sigma}_c^{-1} (\mathcal{M}_{ijc} - V_c z_{ic}) - \frac{1}{2} z_{ic}^* z_{ic}\right]\right\} \quad (13)$$

For each speaker, we weight the likelihood of each utterance from the speaker as

$$\mathcal{L}_{i.c} = \frac{H_i}{\sum_j N_{ijc}} \sum_j N_{ijc} \mathcal{L}_{ijc} \quad (14)$$

where $N_{ijc}$ is the zero-order statistics aligned to the $c^{th}$ component of UBM from $j^{th}$ utterance of $i^{th}$ speaker.

The interpretation of this weighting is that the relative importance of $i^{th}$ speaker to the total log likelihood is valued by the session number $H_i$, while the relative importance of each session is proportional to the factor $N_{ijc}/\sum_i N_{ijc}$, which reflects the reliability of the mean vector of this session. The updated parameters will favour those sessions which have more frames aligned to $c^{th}$ component as they should be more robust. Thus, the solution to maximize the auxiliary function is listed below:

$$V_c = \left(\sum_i \frac{H_i}{\sum_j N_{ijc}} \sum_j N_{ijc} \mathcal{M}_{ijc} E(z_i)^*\right) \left(\sum_i H_i E(z_i z_i^*)\right)^{-1} \quad (15)$$

$$\bar{\Sigma}_c = \frac{1}{\sum_i H_i} \left(\sum_i \frac{H_i}{\sum_j N_{ijc}} \sum_j N_{ijc} (\mathcal{M}_{ijc} \mathcal{M}_{ijc}^* - V_c E(z_i) \mathcal{M}_{ijc}^*)\right) \quad (16)$$

### 4.3. Mean vector weighting

In the E-step of the EM algorithm, all sessions from one speaker are used to estimate the posterior probability of the latent variables and a weighting similar to the proposed likelihood weighting described in Section 4.2 can be applied to the posterior probability estimation. The idea is that mean vectors, $\mathcal{M}_{ijc}$, from the same speaker should not be regarded identically. The larger the value, $N_{ijc}$, the more reliable the corresponding mean vector is and it should be assigned a higher weight when estimating the posterior probability. Specifically, each mean vector is weighted by its corresponding zero-order statistic in the E-step. The revised mean of the posterior distribution is now:

$$E(z_{ic}) = (I + H_i V_c^* \bar{\Sigma}_c^{-1} V_c)^{-1} \frac{H_i}{\sum_j N_{ijc}} \sum_j N_{ijc} V_c^* \bar{\Sigma}_c^{-1} \mathcal{M}_{ijc} \quad (17)$$

The covariance of the posterior probability should also be modified to take into account the relative reliability of the

mean vector but this is not pursued in this paper since the covariance of posterior probability is excluded from the estimation of the expectation of the latent variable. Thus, the second moment of latent variable has the same form as (8).

### 4.4. Score weighting

As per (9), the G-PLDA score in the supervector space with block diagonal covariance assumption is the summation of the sub-scores of each component of the UBM. The question can again be raised as to whether these sub-scores are equally important. It is reasonable to assume that the relative reliability of a sub-score should be taken into account by weighting them by the amount of frames that are presented in the corresponding UBM component. The sub-scores are only weighted for the short test utterances. We propose the following weights:

$$\gamma_c = \frac{N_{tc}}{\sum_c N_{tc}} \tag{18}$$

The final score is then calculated as:

$$Score(\mathcal{M}_e, \mathcal{M}_t) = \sum_c \gamma_c (\log S_{1c} - \log S_{0c}) \tag{19}$$

## 5. Experiments and discussion

A number of experiments were conducted to analyse the effectiveness of the proposed method. The 8CONV-10SEC condition of the NIST SRE '10 [18] was chosen for these experiments along with the 8CONV-5SEC and 8CONV-3SEC conditions (as described in Section 3). 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 [20] prior to feature warping [21]. Gender-dependent UBMs of 1024 Gaussian mixtures were created using utterances from NIST SRE'04, 05, 06, 08, Switchboard II Part 1, 2, 3 and Switchboard Cellular Part 1 and 2. One utterance was chosen from each speaker's available data to retain speaker diversity, while reducing the overall amount of data [22, 23]. T matrices of rank 400 were estimated by using MSR Toolbox [24]. LDA was then applied to further reduce the dimension to 200. The i-vectors were then radially Gaussianised followed by length normalization as described in [25]. The dimensionality of the speaker factors in the baseline is set as 200. For the proposed method, identical MFCC features and UBMs were used. In order to train a better model, short utterances (e.g. 10s) from NIST SRE'04, 05, 06, 08 were added to the T matrix and G-PLDA training datasets [8]. Identical development, training and test sets were employed for the baseline and proposed systems. In addition to the baseline i-vector/GPLDA system, the proposed system is also compared to a LVM system [16] utilising the same front-end and UBM as the proposed technique.

Table 3 summarises the performances of the i-vector/G-PLDA baseline system, the LVM system and the proposed system. The term S-GPLDA is used to represent the proposed G-PLDA in supervector space with block diagonal covariance assumptions without any of the weighting techniques presented in Section 4. It can be seen that there is a performance gap between the proposed system and the baseline. However, we also observe that as the duration of test utterance decrease, this gap becomes smaller.

The abbreviations S_W, SL_W and SLM_W are used to denote the proposed method with the additional score weighting (Section 4.4), score and likelihood weighting (Section 4.2), and score, likelihood and mean vector weighting (Section 4.3) respectively. Based on the results in Table 3 it can be seen that when all three weighting techniques were used (SLM_W), with 10s and 5s test conditions, the performance of the proposed approach was inferior to those of the baseline system. However, the gap decreased as the duration of test utterance decreased and a slight improvement was observed for both male and female speech under the 3s test condition. Compared with the LVM system [16], the proposed system obtained superior performances on 5s and 3s for both male and female conditions, but not in the longer 10s test condition. This may be expected since LVM is not specifically developed to deal with short utterances, while the proposed method is tailored for them. Specifically, for short duration utterances, latent variables inferred from each mixture may not be reliable.

Given that the proposed G-PLDA in supervector space was designed to capture local acoustic variability to complement the total variability framework of the baseline i-vector system, the baseline and the proposed system can be expected to be complementary and fuse well. In the experiments reported in this paper, we fuse systems at the score level. Scores from the baseline system and the proposed system with all three proposed weighting techniques (SLM_W) were fused using the BOSARIS Toolkit [26] and denoted as Fusion1. Based on the results it is clear that the two approaches are complementary and the fusion leads to substantial improvements, particularly in the 3s test condition. To the best of our knowledge, these are the best reported results on the 5s and 3s test conditions. Finally, the baseline was also fused with LVM system (denoted as Fusion2) and compared to the proposed system. We can see that the proposed system outperformed the Fusion2 system for all conditions.

Table 3. *Performance EER (%) of baseline system and proposed system on SRE'10 8CONV-10SEC and 8CONV-5SEC and 5SEC and 8CONV-3SEC conditions.*

| | EER (%) | | | | | |
|---|---|---|---|---|---|---|
| | Male | | | Female | | |
| | 10s | 5s | 3s | 10s | 5s | 3s |
| Baseline | 5.03 | 10.73 | 17.68 | 6.16 | 12.43 | 18.90 |
| S-GPLDA | 14.52 | 18.85 | 22.27 | 17.44 | 20.95 | 25.50 |
| S_W | 13.49 | 15.81 | 18.14 | 14.16 | 18.05 | 19.64 |
| SL_W | 12.33 | 15.46 | 18.47 | 11.96 | 16.19 | 19.39 |
| SLM_W | 12.34 | 14.69 | 17.27 | 12.18 | 16.00 | 18.76 |
| Fusion1 | 4.40 | 8.99 | 14.06 | 5.92 | 11.24 | 15.31 |
| LVM | 9.60 | 16.57 | 22.76 | 11.34 | 17.98 | 22.95 |
| Fusion2 | 4.65 | 10.20 | 16.28 | 6.07 | 11.53 | 17.85 |

## 6. Conclusions

This paper proposes the augmentation of the traditional speaker verification system with G-PLDA on the supervector space using a block-diagonal covariance matrix assumption to capture component wise local acoustic variability in order to improve performance on short duration utterances. Three weighting schemes were proposed to take into account the limited phonetic content of short duration utterances and the proposed approach was validated on 10s, 5s and 3s test utterances from the NIST SRE'10 dataset. The experimental results suggest that, the proposed method complements the traditional total variability space modelling approach by incorporating local acoustic variability information with greater benefits being observed for shorter test utterances.

# 7. References


[1] 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.

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

[3] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Twin Model G-PLDA for Duration Mismatch Compensation in Text-Independent Speaker Verification," *in INTERSPEECH*, 2016.

[4] 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: IEEE.

[5] 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: IEEE.

[6] 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, no. 4, pp. 846-857, 2014.

[7] 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: IEEE.

[8] [8] 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.

[9] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345-354, 2005.

[10] 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: International Speech Communication Association (ISCA).

[11] T. Stafylakis, P. Kenny, V. Gupta, J. Alam, and M. Kockmann, "Compensation for phonetic nuisance variability in speaker recognition using DNNs." *in Odyssey*, 2016

[12] A. Larcher, K.-A. Lee, B. Ma, and H. Li, "RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases," *in INTERSPEECH*, 2012, pp. 1580-1583.

[13] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1-13, 2015.

[14] J. Ma, S. Irtza, K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Parallel Speaker and Content Modelling for Text-dependent Speaker Verification," *in INTERSPEECH*, 2016.

[15] A. McCree, D. Sturim, and D. Reynolds, "A new perspective on GMM subspace compensation based on PPCA and Wiener filtering," DTIC Document2011.

[16] 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.

[17] L. Chen, K. A. Lee, L. R. Dai, and H. Li, "Quasi-factorial prior for i-vector extraction," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2484-2488, 2015.

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

[19] C. M. Bishop, "Pattern recognition," Machine Learning, vol. 128, 2006.

[20] 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: Citeseer.

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

[22] 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, no. 7, pp. 1890-1899, 2011.

[23] T. Hasan and J. H. Hansen, "A study on universal background model training in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1890-1899, 2011.

[24] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, 2013.

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

[26] N. Brümmer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.