The evaluation of compensation techniques in i-vector modelling for sound-based biometric recognition system

N S Ibrahim and D A Ramli
School of Electrical and Electronic, Universiti Sains Malaysia Engineering Campus, Nibong Tebal, Pulau Pinang 14300, Malaysia
dzati@usm.my

Abstract. i-vector subspace modelling is one of the recent methods that has become attractive to sound-based biometric recognition domain. This method provides a benefit of modelling both intra-domain and inter-domain variability into one low dimensional space. This paper focuses on the analysis of i-vector channel compensation techniques for the purpose of improving the i-vector sound-based biometric recognition performance. This work was mainly motivated by the need to quantify the impact of different compensation techniques to the i-vector performance specifically towards the fusion compensation approach. The performances of six channel compensation techniques: (a) whitening, (b) Within Class Covariance Normalization (WCCN), (c) Linear Discriminant Analysis (LDA), (d) whitening and WCCN, (e) whitening and LDA and (f) WCCN and LDA have been investigated in this study. 2656 syllables of bio-acoustic sounds are used as experimental data and parameters of the system are initially tuned with different GMM component sizes i.e. 16, 32, 64 and 128 number of Gaussians. To the end, we assess the effect of the tuned parameter and observe the recognition rate. Experimental results reveal that the accuracy of i-vector with the fusion of WCCN and LDA compensation outperforms other compensation approaches with result of 92.00%. Consequently, these findings allow a better understanding of the compensation approaches, in particular, the fundamental concept of the compensation procedure that leads to the success of the i-vector paradigm.

1. Introduction
Sound based biometric recognition system is a system that identify or verify a subject (speaker) either person or animal species from the characteristics of their utterance sounds [1]. Recent advances in speaker recognition have revealed the discriminant power of a new representation of spoken utterances, referred as identity vector (i-vector) [2]. i-vectors are largely used in the most recent speaker recognition systems as it is easy to compute and it bring back to a more traditional biometric pattern recognition solution. A simpler model for speaker recognition to eliminate the distinction between speaker and channel variability subspaces and to model both speaker and channel in a common constrained low dimensional space, referred as total variability space has been introduced in [2]. Channel or session variability is the variability exhibited by a given speaker from one recording session to another. This type of variability is usually attributed to channel effects although this is not strictly accurate since intra-speaker variation and phonetic variation are also involved [3]. In this approach, a voice segment is represented by a low-dimensional i-vector extracted by factor analysis.

i-vector was initially proposed by Dehak et al. [2] to provide an intermediate speaker representation between the high-dimensional Gaussian Mixture Model (GMM) supervectors and traditional low-dimensional Mel-Frequency Cepstral Coefficients (MFCC) feature representation. The extraction of
these intermediate-sized vectors was motivated by the existing super-vector-based Joint Factor Analysis (JFA) approach [3, 4]. While the JFA approach models the speaker and channel variability spaces separately, i-vectors are formed by modelling a single low-dimensional total-variability space that covers both the speaker and channel variability [5]. This approach reported that i-vector will not lose any speaker discriminant information, unlike the JFA approach, where some speaker discriminant information is lost in the channel space [6]. In fact, this approach provides an elegant way in reducing high-dimensional sequential input data to a low-dimensional fixed-length feature vector while retaining most of the relevant information [7, 8].

As the channel variability is included within the total-variability space, Dehak et al. [9] had investigated a number of standard channel compensation techniques, including Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN) to lessen the channel variability in the i-vector space. Channel compensation approaches are used to remove the nuisance effects included by the i-vector processing [10]. The advantage of applying channel compensation in the total factor space is due to the advantage of low dimension vectors as compared to GMM super-vectors hence giving a less expensive computation [11][12][13][14]. In this study, we tested six techniques of channel compensation in the total variability space for removing the nuisance effects. They are (a) whitening, (b) Within Class Covariance Normalization (WCCN), (c) Linear Discriminant Analysis (LDA), (d) WCCN followed by whitening (whiteWCCN) (e) LDA followed by whitening (whiteLDA) and (f) LDA followed by WCCN (WCCN-LDA).

The objective of this paper is to evaluate the impact of intersession compensation stages in terms of global performance including the effect after combining two compensation techniques. More precisely, we wish to measure the role of the optional compensation procedure as we suspect that this module plays a more important role in the performance of i-vector systems.

1.1. i-vector based speaker recognition

i-vector approach is based on the idea of Joint Factor Analysis (JFA) [1]. According to Dehak(2019) [12], the channel factors in JFA also contain speaker-dependent information. This finding motivates them to model the total variability space (including channels and speakers) instead of modelling the channel- and speaker-spaces separately. The i-vector approach has become state-of-the-art in the speaker verification field [6] and this work shows that it can be successfully applied to other signal identification. The approach provides an elegant way of reducing high-dimensional sequential input data to a low-dimensional fixed-length feature vector while retaining most of the relevant information [9]. The main idea is that the speaker- and channel-dependent super-vectors of concatenated Gaussian Mixture Model (GMM) means can be modelled as

\[ s = m + Tw \]  

(1)

where \( m \) is the speaker- and channel-independent component of the mean super-vectors, \( T \) is a matrix of bases spanning the subspace covering the important variability (both speaker- and session-specific) in the super-vector space, and \( w \) is a standard normally distributed latent variable. For each observation sequence representing an utterance, i-vector is the Maximum A Posteriori (MAP) point estimate of the latent variable \( w \). Here, the i-vector extractor training procedure is based on the efficient implementation suggested in Glembek et al.(2011) [8].

1.2. Channel compensation techniques

Previously, Garcia-Romero et al. (2011) [11] have found the way to convert the i-vector feature behaviour from heavy-tailed to Gaussian. They have introduced the length-normalization approach using linear whitening. The length normalization approach is used to transform the non-Gaussian i-vector feature behaviour into Gaussian i-vector feature behaviour [15]. This technique follows two steps: 1. linear whitening and 2. length-normalization. A linear whitened i-vector, \( w_{wht} \), can be estimated as follows,

\[ w_{wht} = d^{-\frac{1}{2}}U^Tw \]  

(2)
where $w$ is a covariance matrix, which is estimated from development set. $U$ is an orthonormal matrix containing the eigenvectors of $w$ and $d$ is a diagonal matrix containing the corresponding eigenvalues.

Dehak et al. (2009) tested two intersession compensation method; Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) with i-vector. LDA seeks to find a new orthogonal basis (rotation) of the feature space to better discriminate between different classes (speakers). The new basis is sought to simultaneously maximize between class variance (inter speaker discrimination) and minimize within class variance (intra speaker variability) [16]. These axes can be defined using a LDA projection matrix composed of the best eigenvectors (those with highest eigenvalues) of the general eigenvalues equation

$$\Sigma_b U = \lambda \Sigma_w U$$

where $\lambda$ is the diagonal matrix of eigenvalues. The matrices $\Sigma_b$ and $\Sigma_w$ correspond to the between classes and within class covariance matrices, $U$ respectively. These are calculated as follows:

$$\Sigma_b = \sum_{i=1}^{S} (w_i - \bar{w})(w_i - \bar{w})^t$$

$$\Sigma_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - \bar{w}_s)(w_i^s - \bar{w}_s)^t$$

where $\bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} w_i^s$ is the mean of the i-vectors for each speaker, $S$ is the total number of speakers and $n_s$ is the number of utterances for each speaker $s$. In previous work, Dehak et al. (2009) assumed the mean vector of the entire speaker population $\bar{w}$ is equal to the null vector since the factors have a standard Normal distribution, $w \sim N (0, I)$, with zero mean. However, in more recent experiments, they used the actual computed global mean, rather than assuming it was zero, and found a slight performance improvement [17].

Dehak et al. (2009) successfully applied WCCN as compensation to the i-vector system, with the best performance obtained when it was preceded by LDA. The idea behind WCCN is to scale the i-vector space inversely proportional to an estimate of the in-class covariance matrix, so that directions of high intraspeaker variability are deemphasized in i-vector comparisons [15]. The within class covariance is estimated using i-vectors from a set of development speakers as

$$W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (A^t w_i^s - \bar{w}_s)(A^t w_i^s - \bar{w}_s)^t$$

where $A$ is the LDA projection matrix, $\bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} A^t w_i^s$ is the mean of the LDA projected i-vectors for each speaker $s$, $S$ is the total number of speakers, and $n_s$ is the number of utterances of each speaker $s$. They use the inverse of this matrix in order to normalize the direction of the projected i-vector components, which is equivalent to scaling the space by the matrix $B$, where $BB^t = W^{-1}$.

Most of the channel compensation techniques take direct advantage of the calculated between- and within-class scatter matrices [18]. Together with these three techniques, we experimented channel compensation techniques for i-vector and to find a combination that gives better performance for our experimented database. The model domain channel compensation approaches are presently the most active area of research, as most of the channel variations are captured at the model domain.

The paper is organized as follows: Section 2 describes the experimental methodology including the total variability modelling approach and scoring. Section 3 presents the experimental results obtained for the system. Section 4 states the conclusion.

2. Methodology
In this study, 2656 bio-acoustic signal syllables from 55 species of frog taken from our in-house database are used as experimental data. Figure 1 shows the block diagram of research framework of this work. The i-vector system consists of two main parts: front-end and back-end. The former consists
of cepstral feature extraction and UBM training, whereas the latter includes sufficient statistics computation, training of the T-matrix, i-vector extraction, dimensionality reduction and scoring.

2.1. Feature extraction
Voice activity detection (VAD) is first executed so that accurate segmentation of utterance signal can be done. Energy based VAD technique is used where frame-level energy values are computed followed by data normalization and finally, the detection of voice activity process [19]. The class with the higher mean is considered as voice, and the corresponding voice segments are kept before being smoothed. The pre-processed data is then saved into bob library in HDF5 file format. Bob is a free signal processing and machine learning library, used in this experiment. The second step is the process of extracting features from the pre-processed segment from the previous step. The Mel-Frequency Cepstral Coefficients (MFCC) feature extraction method is executed in this procedure.

2.2. i-vector extraction
Subsequently, i-vector modelling and enrolment step. This step involves the computation of a total variability matrix. Low-dimensional i-vectors are then extracted from each of the utterances. For each utterance, the corresponding feature sequence is finally converted to i-vector based on four different UBM size components trained on pooled features from all samples of the training data. The four UBM sizes are 16, 32, 64 and 128 components as stated in Figure 1.

2.3. Channel compensation
The purpose of channel compensation is to eliminate the nuisance effect. In this step, we map extracted i-vectors into another space by executing the following compensation algorithms which are 1) whitening: to reduce non-Gaussian effects as well as mismatch between training and testing subsets, 2) within-class covariance normalization (WCCN): to normalize the within-class covariance matrix of training i-vectors 3) Linear Discriminant Analysis (LDA): to learn a linear projection that maximizes between-class variations while minimizing within-class variations [20]. In this study, the combination of two compensation techniques was also experimented which are WCCN followed by whitening (whiteWCCN), LDA followed by whitening (whiteLDA) and LDA followed by WCCN (WCCNLDA).

2.4. Classification and Evaluation
Matching between a test utterance and a target species is done using fast scoring classifier algorithm based on cosine distance. With all six compensation techniques applied, the cosine similarity is written as equation (7)

\[
\text{score}(w_{\text{target}}, w_{\text{test}}) = \frac{\langle w_{\text{target}}, w_{\text{test}} \rangle}{||w_{\text{target}}|| ||w_{\text{test}}||}
\]  (7)
where i-vector was originally considered as a feature to the classifier. It operates by comparing the angles between a test i-vector, \( w_{\text{test}} \), and a target i-vector \( w_{\text{target}} \) [9].

Evaluation measures i.e. recognition rate (%) and Cumulative Match Curve (CMC) plot are used in this experiment which are also available in Spear Toolbox.

### 3. Result and discussion

In this paper, experiments were carried out using Spear toolbox in Ubuntu operating system. Spear is an open source and extensible toolbox for state of the art speaker recognition. This toolbox is built on top of Bob, a free signal processing and machine learning library [21].

Table 1 shows the recognition rate of the development set for i-vector whitening, WCCN, LDA, whiteWCCN fusion, whiteLDA fusion and WCCNLDA fusion for 4 different GMM components. The best overall rates are highlighted in bold font. Experimental result observed that the best performances were obtained by the WCCNLDA fusion with 92.00% performance. The justification of this observation can be explained as follows. The LDA is a supervised method of dimensionality reduction. It defines new spatial axes that minimize the intra-class variance caused by channel effects and maximize the variance between speakers. So, i-vectors are subjected to the projection matrix obtained by LDA then the WCCN approach computes a cosine score, according to the metric of a within-class covariance matrix.

The simple whitening techniques also gives competitive result with 91.11% which is relatively good. Whitening technique reduces the non-Gaussian effects as well as mismatch between training and testing subsets. However, the fusion of whitening either with WCCN or LDA shows a degradation of performance. This is because the WCCN is not necessary to perform here whenever the total variability matrix has been normalized. On the other hand, the unconvincing performance of whiteLDA fusion is due to the reason that LDA attempts to define new special axes that minimize the intra-class variance caused by channel effects, and to maximize the variance between speakers. In this case, fusion is risky because whitening technique has already reduced the non-Gaussian effect and some discriminant information is lost during processing.

Based on overall result, it shows that WCCNLDA fusion compensation method gives the outstanding performances for all 128, 64, 32 and 16 GMM components when compared to the other compensation methods for our frog sound identification system as coloured in yellow in Table 1. Nevertheless, the performances of the other compensation methods are still competitive. The best performance for each method regarding the GMM component is given in Table 1.

From different perspective, the performances can also be observed by the CMC plot as in Figure 2. The figure shows the CMC plot for 6 variation of compensation methods from 4 different GMM components that give the best performance for the development and evaluation sets. The y-axis shows the probability of a species being included in the top-\( k \) rank, which is indicated by the value on x-axis. Those plots confirm the same conclusion as above.

| GMM Component | Whitening | WCCN | LDA | WhiteWCCN | WhiteLDA | WCCNLDA |
|---------------|-----------|------|-----|-----------|----------|---------|
| i-vector-128  | 73.78     | 74.67| 71.56| 70.89     | 73.33    | **77.33**|
| i-vector-64   | 80.00     | 81.33| 74.22| 77.78     | 76.44    | **84.44**|
| i-vector-32   | 85.33     | **88.89**| 75.56| **86.67**| 79.11    | **92.00**|
| i-vector-16   | **91.11**| 85.33| **76.89**| 84.44    | 78.22    | **91.56**|

*Table 1. Performance summary. Recognition rate (%) at rank-1 for i-vector extraction with 6 different compensation techniques.*
Figure 2. CMC curve for six different compensation methods on (a) the development set and (b) the corresponding evaluation set.

4. Conclusion
The total factor space designed in i-vector algorithm contains both speaker and session information. We presented in this paper a set of compensation technique experiments in order to remove the nuisance effects. Compared to our baseline single compensation techniques, the fusion of the WCCN and LDA method gives the best performance among others. We can conclude that the compensation part in i-vector modeling is essential and correct combination of the compensation techniques is crucial in order to achieve satisfying i-vector performances.
References

[1] Greenberg C S Stanford V M Martin A F Yadagiri M Doddington G R Godfrey J J and Hernandez J 2013 Proc. in 14th Annual Conference of the International Speech Communication Association (Lyon, France) (INTERSPEECH) p 1971.

[2] Dehak N Kenny P Dehak R Dumouchel P and Ouellet P 2011a IEEE Trans. on Audio, Speech, and Language Processing 19 p 788.

[3] Kenny P Boulianne G Ouellet P and Dumouchel P 2007 IEEE Trans. on Audio, Speech, and Language Processing 15 pp 1435.

[4] Kenny P Ouellet P Dehak N Gupta V and Dimouchel P 2008 IEEE Trans. Audio, Speech Lang. Process 16 p 980.

[5] Kenny P Boulianne G Ouellet P and Dumouchel P 2007 IEEE Trans. Audio, Speech Lang. Process 15 p 1448.

[6] Dehak N Torres-Carrasquillo P Reynolds D and Dehak R 2011 Proc. 12th Annual Conference of the International Speech Communication Association (Florence, Italy) (INTERSPEECH) p 857.

[7] Xu L Lee K A Li H Yang Z 2018 IEEE/ACM Transactions on Audio Speech and Language Processing 26 p 749.

[8] Glembek O Burget L Matějka P Karafiát M and Kenny P 2011 Proc. in IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP) (Prague, Czech Republic) p 4516.

[9] Dehak N Dehak R Kenny P Brummer N Ouellet P and Dumouchel P 2009 Proc. in 10th Annual Conference of the International Speech Communication Association (Brighton, United Kingdom) (INTERSPEECH) p 1559.

[10] Behravan H Hautamaki V and Kinnunen T 2015 Speech Communication 66 p 118.

[11] Garcia-Romero D and Espy-Wilson C Y 2011 Proc. of the Annual Conf. of the Int. Speech Communication Association (Florence, Italy) (INTERSPEECH) p 249.

[12] Dehak N 2009 Ph.D. thesis, Ecole de Technologie Superieure, Montreal.

[13] Ibrahim N S Ramli D A 2018 Procedia Computer Science 126 p 1534.

[14] Ibrahim N S Ramli D A 2019 Lecture Notes in Electrical Engineering 547(76) (Springer Nature).

[15] Hatch A Kajarekar S and Stolcke A 2006 Proc. in International Conference on Spoken Language Processing (Pittsburgh, PA, USA) (INTERSPEECH-ICSLP) p 1471.

[16] Vogt R J and Sridharan S 2008 Computer Speech & Language 22 p 17.

[17] Lei Y Burget L and Scheffer N 2013 Proc. in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (Vancouver, Canada) p 6788.

[18] Solomonoff A Campbell W M Boardma I 2005 Proc. in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (Pittsburgh, PA, USA) 1 p 629.

[19] Huang Z Cheng Y Li K Hautamaki V and Lee C 2013 Proc. in 14th Annual Conference of the International Speech Communication Association (Lyon, France) (INTERSPEECH) p 2282.

[20] Gahabi O and Hernando J 2017 IEEE/ACM Transactions on Audio Speech and Language Processing 25 p 807.

[21] Khoury E El Shafey L and Marcel S 2014 Proc. in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Florence, Italy) 2 p 1655.

Acknowledgments

This work was sponsored and supported by Universiti Sains Malaysia under Research University Grant (RU) 1001.PELECT.9014057.