ECG biometric verification system: An i-vector to overcome variability factors

R X Tiong, D A Ramli and N S Ibrahim
School of Electrical & Electronic Engineering, USM Engineering Campus, Universiti Sains Malaysia, 14300, Nibong Tebal, Penang, Malaysia
dzati@usm.my

Abstract. Dealing with signal and session variability is a common problem in biometric recognition system since biometric signal is frequently inconsistent over time. Health, aging, emotional conditions and different recording settings are some of the factors that contribute to the variability issue. This cause the two samples of the same subject tends to be different from each other hence giving a mismatch effect between the enrolments and testing condition. Over the years, solving the variability problem by subspace representation concept has become prevalent. Hence, it motivates us to validate a recognition algorithm based on factor analysis perspective and we use electrocardiogram (ECG) signal for our experimental data as it is subject to change over time and sensitive to different sensors. We first model each supervectors extracted from Gaussian Mixture Model (GMM) into two different factors which are subject and session independent supervectors based on Joint Factor Analysis (JFA) algorithm. For the second model which is based on i-vector approach, the supervectors extracted from GMM is first modelled to be a single total factor and a compensation method is then employed to compensate the variability effect. Three compensation methods for the i-vector are employed which are Probabilistic Linear Discriminate Analysis (PLDA), Linear Discriminate Analysis (LDA) and Within Class Covariance Normalization (WCCN). The ECG-ID database obtained from physionet database consists of 90 subjects with a total of 310 ECG recordings; each recorded for 20 seconds are used in this study. Experimental results reveal the robustness of the i-vectors PLDA approach by giving 2.156% and 2.155% of Equal Error Rate (EER) for protocol 1 and 2, respectively.

1. Introduction
Over the decades, research direction towards a powerful subspace representation so as to solve data variability issue in speaker recognition system has become prevailing. Aronowitz et al. (2005) adapted a session based generative model into the concept of Gaussian Mixture Modeling (GMM) for the test utterance for speaker verification [1]. Similarly, the same concept have been demonstrated by Vogt et al. (2005). Then, Kenny (2007) have brought the idea of Joint Factor Analysis (JFA) for modelling speaker and session variability [3]. In another research, a comprehensive study on JFA versus Eigenchannels in speaker recognition has then been evaluated in Kenny (2007).

Subsequently, recent advances in speaker recognition have revealed the discriminant power of a new representation of spoken utterances, referred as i-vector [5][6] [7][8][9][10]. This method was initially proposed by Dehak et.al. (2009) to provide an intermediate speaker representation between the high-dimensional Gaussian Mixture Model (GMM) super-vector and traditional low-dimensional Mel-Frequency Cepstral Coefficients (MFCC) feature representation [5]. The extraction of these
intermediate-sized vectors was actually motivated by the existing supervectors from Joint Factor Analysis (JFA) approach [3][4]. However, according to Dehak et al. (2011), the channel factor estimated using JFA which are supposed to model the channel effect; holds the speaker information [7]. Thus, instead of separating the speaker and channel variability into two distinct subspaces, i-vector models a new single low-dimensional space of speaker and channel variability named total variability space. As reported in Dehak et al. (2011), i-vectors will not lose any speaker discriminant information, unlike the JFA approach, where some speaker discriminant information is lost in the channel space. The approach provides a smart way of reducing high-dimensional sequential input data to a low-dimensional fixed-length feature vector while retaining most of the relevant information.

The rest of this paper is organized as follows. Section 2 summarizes the theory of GMM, JFA and i-vector techniques. Section 3 presents the methodology. Consequently, Section 4 presents the experimental results and discussion.

2. Theory background

In this section, the fundamental theoretical concept of Gaussian mixture model-universal background model (GMM-UBM) and Joint Factor Analysis (JFA), which build up the concept of i-vector are presented.

2.1. GMM-UBM

GMM is a parametric model of the probability density function of continuous measurements or features, such as vocal-tract related spectral features [11]. It is represented as a weighted sum of Gaussian component densities and is often applied to a biometric system. GMM as a weighted sum of Gaussian component densities can be expressed as below,

$$ P(x|\lambda) = \sum_{k=1}^{M} w_k \times g(x|\mu_k,\Sigma_k) $$

where $x$ is the dimensional feature vector, $w_k, \mu_k$ and $\Sigma_k$ are the mixture weight, mean and covariance, respectively. $g(x|\mu_k,\Sigma_k)$ for $k = 1,2,...,M$ are the Gaussian density components. Two modelling techniques that are often used are expectation-maximization (EM) and Maximum a Posteriori (MAP) adaption [11]. The former method derives the initial solution of the problem by iterative improvement and requires a significant amount of training data which may not always be available in practice. Hence, MAP adaption approach has become dominant in GMM modelling. In a standard GMM, a subject-independent world model namely universal background model (UBM) is initially trained (typically by iterative expectation-maximization (EM) algorithm) by utilizing the data collected from a large number of subjects [12]. Feature vectors of each subject data then go through a UBM based MAP adaptation process to form a mean vector as shown in Figure 1. Subsequently, GMM supervectors are formed by simply stacking all the adaption mean vectors, which consists of all the subject-dependent information. Yet, as the number of Gaussian components increase which results to the increasing number of supervectors, the complexity of the system becomes higher [12]. In addition to this weakness, as performance degradation is also observed due to session variability between training and test condition, thus a new subspace representation concept i.e. JFA is introduced [3].
2.2. JFA
In JFA [4], a supervector ($M$) is composed as additive components i.e. subject and channel subspace as described below

$$M = m + Vy + Ux + Dz$$  \hspace{1cm} (2)

where $m$ denotes the subject and session independent components, normally referred to as the UBM model, $V$ and $D$ denotes the subject dependent projection (i.e., the eigendata matrix and diagonal residual, respectively), and $U$ denotes the session or channel projection. Vector $y$ and $z$ denote the subject dependent factor in their respective projection while $x$ denotes the session dependent factor in its projection; where vector $y$, $x$ and $z$ is assumed to be a random variable with a normal distribution $N(0,1)[4]$. The GMM supervector decomposition by JFA can be illustrated as shown in Figure 2. The GMM for a source-target speaker is realized by modifying the variables of a UBM trained utilizing a large amount of data taken from different subjects.

In JFA, the basic assumption is that a subject and channel-dependent supervector $M$ that corresponds to a given recording can be decomposed into a subject supervector $s$ and a channel supervector $c$, and can be expressed as

$$M = s + c$$  \hspace{1cm} (3)

where $s$ and $c$ are statistically independent and normally distributed such that $c$ depends solely on the channel effects in the recording. Since JFA models subject- and session-dependent information into two subspaces, it possesses eigendata information within the session factor, which leads to practical difficulties in the implementation [5]. Therefore, i-vector is inspired from the coexistence of subject and session variability in same subspace [5][6].

2.3. i-VECTOR
The first single subspace-based i-vector speaker verification is introduced, where each speech utterance is expressed by a low-dimensional fixed-length vector [5]. The i-vector model constrains the GMM supervectors $s$, made up of the accumulated GMM means, to reside in one subspace, containing both channel and speaker variability, based on:

$$M = m + Tw$$  \hspace{1cm} (4)

where $m$ is the mean super-vector of UBM, $T$ is the total variability matrix where most of the subject specific information together with channel information lay within here and lastly, $w$ is the resulting i-vector. i-vector extractor maps a sequence of vector obtained from feature extraction stage into a fixed length vector, while UBM is used to collect background or common characteristics from the signal. The first-order statistics for each mixture components adjoined together to form a supervector [5]. Since the i-vector includes both subject- and session-dependent information within the same subspace
then using i-vector algorithm solely may not perform well [5]. Hence, after the i-vector extraction, an intersession compensation method must be performed to discriminate the subject- and session-dependent information. There are three intersession compensation methods used in this study. The probabilistic linear discriminate analysis (PLDA), linear discriminate analysis (LDA) and within-class covariance normalization (WCCN). The first and second methods seek for the direction in space that has maximum discriminability by modelling both inter-class and intra-class variance as a multi-dimensional Gaussian. While the third method is to estimate of the in-class covariance matrix by scaling the i-vector space inversely proportional, so that direction of high intra-speaker variability is de-emphasized in i-vector comparison [13].

3. Methodology
In this section, first, a generic description for the i-vector system using SPEAR toolbox on an Ubuntu operating system is described, and followed by the presentation of the evaluation method [14].

3.1. Database handling
The ECG-ID database obtained from physionet is employed and consists of 90 subjects with a total of 310 ECG recordings [15]. These recordings are recorded for 20 seconds using ECG lead 1. The ECG signal is digitized using a sampling frequency of 500Hz with at least 12-bit resolutions over a nominal ±10 mV range. The source format used by this database is MIT format, which consists of 3 different files with different file type extension. They are header file (.hea), annotation file (.atr) and data file (.dat). These signals are not readable by the Bob SPEAR, though all the signal information for each signal is stored in these 3 files. Bob SPEAR only support audio file format that is supported by SoundExchange (SoX). Since most applications uses the wave file (.wav) format with 16-bit ADC resolutions as input to Bob SPEAR toolbox, therefore, the target file type for these ECG signals in this database is 16-bit .wav file, as a standard input to the Bob SPEAR. Additionally, as Bob SPEAR only accepts 1-dimensional (1D) data, the MIT format is first converted into .mat file to allow MATLAB to process the signal and apply audiowrite() function to convert the ECG signal. Figure 2 shows the block diagram of ECG file conversion.

The input to the system is the .wav file which has to be assigned into 3 different datasets, which are training, development and evaluation sets before being fed into the system [14].

![Figure 2. Block diagram of ECG file conversion](image)

Training set is used to train the Universal Background Model (UBM), development set is used to define threshold for the authentication and lastly, evaluation set is used to evaluate the system performance. i-vector requires a lot of data for the training set [14]. There are only 310 recordings in the database which is merely enough for i-vector training. Thus, all the subjects are assigned to the training set. Development and evaluation set are dependent on the protocol used.

There are 2 protocols introduced in this study; unbiased recognition protocol and all-subject recognition protocol. The former namely, protocol 1, performs biometric recognition without the presence of the subjects in the threshold decision that was computed by the development set. Hence, there is no threshold re-train required when there are new subjects come in. In the dataset assignment, recording of first 60 subjects has been assigned to development set while the remaining 30 subjects are assigned to evaluation set. For all-subject recognition protocol (namely, protocol 2), every subject
must have a pre-trained model before proceeds to evaluation. The optimal parameters are set by using the training set and developed a subject-specific model. Hence, there are 90 models dedicated to recognizing each of the 90 subjects. Bob SPEAR standardizes a file list structure to ease the development as shown below. There are probes list and model list needed for development and evaluation set each which by means, recordings within single subject needs to be further divided into these 2 lists. In this study, we divide the recordings into these lists equally as shown in Table 1 and 2 below.

| Table 1. File allocation for development and evaluation sets for protocol 1 |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| UBM | Development set | Evaluation set |
| | Enrolment | Testing | | Enrolment | Testing |
| Subject | No. of wav file | Subject | No. of wav file | Probe | Score | Subject | No. of wav file | Probe | score |
| 90 | 310 | 60 | 123 | 103 | 6180 | 30 | 47 | 37 | 1110 |

| Table 2. File allocation for deployment set for protocol 2 |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| UBM | Development set | Deployment set |
| | Enrolment | Testing | | Enrolment | Testing |
| Subject | No. of wav file | Subject | No. of wav file | Probe | Score | Subject | No. of wav file | Probe | score |
| 90 | 310 | | | | | 90 | 170 | 140 | 12600 |

In the pre-processing stage, we choose two Gaussian energy-based model as our pre-processing method since the Gaussian model is often used for signal pre-processing as it offers usability, flexibility and widely used for signal processing research. For feature extraction, Mel Frequency Cepstrum Coefficient (MFCC) is computed, which is the most commonly used for ECG signal among all other feature extraction method found in the toolbox.

The next block in the framework is supervectors extraction (as given in Figure 1). This extraction aims to compute UBM as a projector using the training set. In a subject dependent GMM, the pre-mixture posterior probability of each feature vector (GMM-posterior) is computed and used. The computation of the statistics is called as UBM projection step, which aims to project cepstral features using the output from previous MFCC extraction. In this study, different Gaussian components are examined, specifically for 8, 16, 32, 64, 128 GMM components. The GMM supervectors are now ready for i-vector and JFA.

4. Results and discussions

Performance on the protocol 1 can be described as the following. Table 3 and Table 4 show the system performance evaluated using equal error rate (EER) on the development set and the half total error rate (HTER) on the evaluation set respectively for GMM, JFA, i-vector PLDA, i-vector LDA and i-vector WCCN. From the experimental results, we observe that the i-vector PLDA achieved the best performance in most experiments. The best EER performance, 2.156%, was given by the i-vector PLDA system with 16 GMM component and for the HTER performance, the best result was also from the i-vector PLDA system with 0.746% with 16 GMM component. Experimental results also reveal that JFA methods fail to produce satisfying results compared to i-vector system. This confirms the idea of total variability space on speech utterance signals by Dehak et al. (2011) is also true for the ECG signals. The observations also highlight the unsatisfying GMM system performance which due to limited data to tune the GMM parameters correctly for optimal generalization performance. The same trend can also be observed in the detection error trade-off (DET) evaluation as given in Figure 3 and 4.
### Table 3. EER performance with different Gaussian components in protocol 1

| GMM Components | GMM | JFA | i-vector PLDA | i-vector LDA | i-vector WCCN |
|----------------|-----|-----|--------------|-------------|--------------|
| 8              | 16.000 | 20.968 | 3.172        | 7.111       | 4.913        |
| 16             | 10.762 | 17.841 | 2.156        | 4.944       | 3.652        |
| 32             | 9.778  | 14.762 | 2.825        | 5.142       | 6.743        |
| 64             | 8.889  | 18.785 | 4.913        | 5.974       | 5.147        |
| 128            | 17.778 | 21.452 | 5.674        | 6.147       | 6.603        |

### Table 4. HTER performance with different Gaussian components in protocol 1

| GMM Components | GMM | JFA | i-vector PLDA | i-vector LDA | i-vector WCCN |
|----------------|-----|-----|--------------|-------------|--------------|
| 8              | 19.175 | 24.206 | 0.979        | 5.947       | 5.744        |
| 16             | 14.245 | 18.667 | 0.746        | 3.728       | 5.732        |
| 32             | 10.270 | 16.508 | 1.538        | 4.124       | 6.133        |
| 64             | 9.603  | 18.444 | 5.033        | 4.874       | 7.066        |
| 128            | 20.215 | 22.841 | 0.885        | 3.814       | 5.883        |

---

**Figure 3.** DET plot for development set.

**Figure 4.** DET plot for evaluation set.
Performance for protocol 2 (as given in Table 5 and Figure 5) is expected to have better performance compared to protocol 1 as different ECG signals from the same person appear in both training and testing. Given this possibility, we would expect the performances to be significantly better. However, this only true for the i-vector PLDA systems (except for the 32 GMM component) which proves the ability of PLDA in maximizing the inter-individual variance and minimizing the intra-individual variance based on probability concept. For protocol 2, the best EER performance, 2.155%, was given by the i-vector PLDA system with 16 GMM component.

### Table 5. EER performance with different Gaussian components in protocol 2

| GMM Components | GMM | JFA | i-vector PLDA | i-vector LDA | i-vector WCCN |
|----------------|-----|-----|---------------|--------------|---------------|
| 8              | 18.477 | 25.144 | 2.183 | 8.155 | 6.411 |
| 16             | 15.895 | 19.180 | **2.155** | **6.222** | **5.825** |
| 32             | 12.322 | **17.333** | 4.944 | 7.542 | 5.121 |
| 64             | **11.111** | 21.632 | 2.869 | 8.048 | 6.833 |
| 128            | 19.177 | 24.147 | 2.857 | 9.552 | 7.019 |

![Figure 5. DET plot for deployment set.](image-url)
References

[1] Vogt R Baker B and Sridharan S 2005 Proc. Eurospeech (Lisbon) p 3117
[2] Aronowitz H Burshtein D Amir A 2005 Proc. Int. Conf. on Audio, Speech Signal Processing (Philadelphia) (IEEE) p 733
[3] Kenny P Boulianne G Ouellet P and Dumouchel 2007 IEEE Transaction Audio, Speech, Language Process 15 p 1448
[4] Kenny P Boulianne G Ouellet P and Dumouchel P 2007 IEEE Transaction Audio, Speech, Language Process 15 p 1435.
[5] Dehak N Dehak R Kenny P Brummer N Ouellet P and Dumouchel P 2009 Proc. of the Annual Conf. of the Int. Speech Communication Association (Brighton) (INTERSPEECH) p 1559
[6] Dehak N Dehak R Glass J Reynolds D and Kenny P 2010 Proc. of Odyssey The Speaker and Language Recognition Workshop (Brno) (Odyssey) p 71
[7] Dehak N Kenny P J Dehak R Dumouchel P and Ouellet P 2011 IEEE Transactions on Audio, Speech and Language Processing 19 p 788.
[8] Garcia-Romero D and Espy-Wilson C Y 2011 Proc. of the Annual Conf. of the Int. Speech Communication Association (Florence) (INTERSPEECH) p 249
[9] Ibrahim N S Ramli D A 2018 Procedia Computer Science 126 p 1534.
[10] Ibrahim N S Ramli D A 2019 Lecture Notes in Electrical Engineering 547(76) (Springer Nature).
[11] Reynolds D A Quatieri T F and Dunn R B 2000 Digital Signal Processing 10 p 19
[12] Togneri R and Pulella D 2011 IEEE Circuits and Systems Magazine 11 p 23
[13] Hatch A Kajarekar S and Stolcke A 2006 Int. Conf. on Spoken Language Processing (Pittsburgh) (INTERSPEECH)
[14] Khoury E El Shafey L and Marcel S 2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (Florence) vol 2 (IEEE) p 1655
[15] Goldberger A L Amaral L A Glass L Hausdorff J M Ivanov P C Mark R G Mietus J E Moody G B Peng C K and Stanley H E 2000 Circulation 101 p 215

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