

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/320773488>

Cancellable Speech Template via Random Binary Orthogonal Matrices Projection Hashing

Article in Pattern Recognition · November 2017
DOI: 10.1016/j.patcog.2017.10.041

CITATIONS
0

READS
89

7 authors, including:



Kong Yik Chee
Universiti Tunku Abdul Rahman
2 PUBLICATIONS 0 CITATIONS

SEE PROFILE



Zhe Jin
Monash University (Malaysia)
29 PUBLICATIONS 215 CITATIONS

SEE PROFILE



Ming Li
Duke Kunshan University
80 PUBLICATIONS 828 CITATIONS

SEE PROFILE



Wun-She Yap
Universiti Tunku Abdul Rahman
39 PUBLICATIONS 279 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Fall reduction system [View project](#)



Cancellable speech template via random binary orthogonal matrices projection hashing



Kong-Yik Chee^a, Zhe Jin^b, Danwei Cai^{c,d}, Ming Li^{c,d}, Wun-She Yap^{a,*}, Yen-Lung Lai^a, Bok-Min Goi^a

^a Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Malaysia

^b School of Information Technology, Monash University Malaysia, Malaysia

^c SYSU-CMU Joint Institute of Engineering, School of Electronics and Information Technology, Sun Yat-Sen University, China

^d SYSU-CMU Shunde International Joint Research Institute, China

ARTICLE INFO

Article history:

Received 23 April 2017

Revised 21 September 2017

Accepted 30 October 2017

Available online 1 November 2017

Keywords:

Cancellable biometrics

Speaker recognition

RBOMP hashing

PF function

Security & privacy

ABSTRACT

The increasing advancement of mobile technology explosively popularizes the mobile devices (e.g. iPhone, iPad). A large number of mobile devices provide great convenience and cost effectiveness for the speaker recognition based applications. However, the compromise of speech template stored in mobile devices highly likely lead to the severe security and privacy breaches while the existing proposals for speech template protection do not completely guarantee the required properties such as unlinkability and non-invertibility. In this paper, we propose a cancellable transform, namely Random Binary Orthogonal Matrices Projection (RBOMP) hashing, to protect a well-known speech representation (i.e. i-vector). RBOMP hashing is inspired from Winner-Takes-All hash and further strengthened by the integration of the prime factorization (PF) function. Briefly, RBOMP hashing projects the i-vector using random binary orthogonal matrices and records the discrete value. Due to the strong non-linearity of RBOMP, the resultant hashed code withstands the template invertibility attack. Further, the experimental results suggest that the speech template generated using RBOMP hashing can still be verified with reasonable accuracy. Besides that, rigorous analysis shows that the proposed cancellable technique for speech resists several major attacks while the other criteria of biometric template protection can be justified simultaneously.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Given the advancement of technologies and the increase in the popularity of mobile devices, speaker recognition system is emerging into a rapid growing field of research. In [1], Unar et al. stated the possibilities of using voice biometric modalities in different applications involving mobile commerce and transactions. Voice, consisting of unique features of different speakers, is often used to identify and verify the legitimate user in numerous applications. Typically, speaker recognition can be categorized as speaker identification and speaker verification. Speaker identification classifies a given voice to a specific speaker, while speaker verification decides a pair of voices as from the same speaker. State-of-the-art speaker recognition systems widely use i-vector modeling as a frontend technique to jointly model speaker and channel variabilities in a speech utterance due to its favorable performance as well as its condensed representation [2]. Moreover, Probabilistic

Linear Discriminative Analysis (PLDA) is commonly adopted as a supervised backend modeling approach to strengthen speaker information while restraining channel variability and other sources of undesired variabilities [3–5]. Instances of speaker recognition systems that use both i-vector and PLDA can refer to [6,7]. It is worth mentioning that two general methods applying Deep Neural Network (DNN) to speaker recognition system brought impressive gains in performance. The first method trained a DNN acoustic model to produce frame alignments by the standard Gaussian Mixture Model (GMM) in the conventional framework [8]. The second method used the DNN acoustic model to extract phonetic features [9,10]. The phonetic features are the outputs of the bottleneck layer of a DNN or the low dimensional features after applying PCA to DNN's outputs of tied triphone state phoneme posterior probabilities. The phonetic features were then concatenated to Mel Frequency Cepstral Coefficient (MFCC) to generate tandem feature.

In i-vector/PLDA framework, a speaker recognition system can determine the authenticity of a user by matching the voice reference (i.e. i-vector) stored in the database. However, this raises the concern on the protection of the voice reference (also known as

* Corresponding author.

E-mail address: yapws@utar.edu.my (W.-S. Yap).

template) stored in the database to prevent security and privacy threats. In [11,12], it has shown that biometric template leakage is considered as one of the most harmful attacks in the biometric security system. The compromised biometric template can lead the impostor to create physical spoof from the stolen template, replace the template and gain illegitimate access to the system [12–14]. It is further complicated by the fact that biometric traits are irreplaceable once compromised. Therefore, a biometric-based application equipped with template protection capability is urgently needed.

In the literature, a number of proposals have been reported to secure the biometric templates. The existing proposals in protecting biometric template can be divided into three types: biometric cryptosystems (or helper data methods), feature transformation (or cancellable biometrics) and hybrid biometric cryptosystem [12]. Biometric cryptosystems require the usage of helper data, a biometric-dependent public information which does not reveal the original biometric template, to retrieve or generate keys. Instance of biometric cryptosystem can refer to [15]. The authentication process for this approach is to perform biometric comparison to determine the validity of the key retrieved or generated. Depending on how the helper data is derived, this approach can further be divided into key-binding or key-generation systems [16]. On the other hand, cancellable biometrics transforms the original biometric feature in such a way that it is computationally difficult to reconstruct the original biometric feature [16,17]. The advantages of using this approach is that the adversary is computationally hard to recover the original biometric feature even if the transformed feature vector had been compromised. However, the transformation of feature often leads to the loss of accuracy and this will likely degrade the performance of the biometric recognition system [17]. Instances of cancellable biometric can refer to [18,19]. Lastly, the hybrid biometric cryptosystem is the combination of biometric cryptosystems and cancellable biometrics to enjoy the strength from each type of method. An ideal template protection scheme is required and must fulfill all of the following requirements [20]:

1. Irreversibility. It should always be computationally hard for the adversary to invert the protected biometric template.
2. Unlinkability. It should always be computationally hard for the adversary to distinguish whether multiple protected biometric templates were generated using the same biometric trait of a user.
3. Revocability. The protected biometric template should be able to be revoked or renewed to replace the old template while the original template should be computationally hard to be inverted from multiple protected biometric templates derived from the same biometric trait of a user.
4. Performance. The performance of the biometric recognition rate should not be seriously degraded.

1.1. Related works

In this section, the previous works on the speech template protection are discussed and summarized. Generally, the revisit of the speech template protection schemes follows the categories of biometric template protection, i.e. cancellable biometrics, biometric cryptosystems and hybrid biometric cryptosystem [12].

1.1.1. Cancellable biometrics

Cancellable biometrics, the intentional distortion of the biometric feature, was formalized by Ratha et al. [21] to protect the privacy of the user. In the event that the cancellable feature is compromised, the same biometric feature can be mapped into another new distinct template using the pre-designed distortion characteristics. Cancellable biometrics can further be divided into biometric salting and non-invertible transformation.

Biometric salting [22] blends an auxiliary data (e.g. a user specific key or password) with the biometric feature. A concrete example of biometric salting for speech template protection is probabilistic random projection proposed by Chong and Teoh [23]. Two-dimensional principal component analysis was applied on the feature matrix before going through a random projection process via an externally derived pseudo random-number. The projected matrix was then fed into a Gaussian Mixture Model (GMM) to obtain probabilistic speaker models. The presented scheme was shown to be resisted from the stolen-token attacks where even if the token had been compromised, the recognition performance of the system was still able to retain at the feature vector level. However, the scheme was vulnerable to attack via record multiplicity (ARM) as the adversary can recover the original feature template by exploiting multiple templates generated using different random projection matrices [24].

Cancellable biometrics also often refers to the use of one-way transformation function that converts the voice feature to a protected template that is computationally hard to be inverted [22]. In 2008, Xu and Cheng [25] proposed a cancellable voice template protection method based on fuzzy vault scheme [26]. Chaff points were added to the unordered Mel-Frequency Cepstral Coefficient matrix to create a vault and a prime accumulator was used to separate the genuine points from chaff points. Besides, a non-invertible function was used to conceal the raw features while polynomial reconstruction was used for authentication. However, Chang et al. [27] revealed that the selection of the chaff points is not independent as the selection of new chaff point depends on the location of the previous selected point. It was observed that the latecomers, referring to the points added later, will likely to have more nearby points. Hence, increasing the number of chaff points will likely lead the adversary to correctly guess the genuine points. In addition to that, if the prime accumulator had been compromised, the adversary will be able to easily determine the genuine points.

Recently, Pandev et al. [28,29] proposed a new technique called deep secure encoding for protecting face template. The face features were first extracted and trained using deep convolutional neural networks to generate an unprotected binary template. The unprotected binary template was divided into n k -bit blocks. Each k -bit block was then fed as an input of a cryptographic hash function (e.g. SHA-256). Finally, the n outputs of hash function were stored in the database for matching purposes. During the matching phase, the face image is first queried. Subsequently, similar training and feature extraction processes will be carried out using the queried face image to generate an unprotected binary template. The unprotected template is then divided into n k -bit blocks as the inputs of the underlying hash function. The n outputs of the hash will then be compared with the hashed codes stored in the database. The matching is successful if i out of n outputs of the hash are matched where i must be greater than the pre-defined threshold value. The proposed scheme is interesting as a random key is chosen and is embedded during face extraction and training processes to generate an unprotected binary template while no key is needed to secure the unprotected binary template. If the template is compromised, a new key will be selected and the training process must be carried out again to re-generate a new unprotected binary template. Thus, Pandev et al. claimed that their scheme offers the property of cancellability without using key (where no key is needed after the feature extraction and training processes). This idea is different with typical template protection schemes where key is needed in securing the unprotected template to offer the property of cancellability. The size of protected template is of $n \times k$ bits. In the experiment performed by Pandev et al. using two different datasets (i.e. CMU PIE and Extended Yale B), the size of a protected template is of $64 \times 1024 = 65536$ bits.

Since a typical feature extraction method does not involve any key, we propose a template protection scheme involving a key after the feature extraction method. Our proposed scheme enjoys the benefit that one does not need to focus on the training and feature extraction processes of the underlying biometrics and our scheme can be generalised to other biometric modalities with real value representation. Other than the brute-force attack examined by Pandev et al. on their proposed method, we also provide extensive analysis on different security concerns of our proposed template protection scheme.

1.1.2. Biometric cryptosystems

Biometric cryptosystems [30] can broadly be divided into key-binding and key-generation. The representative instances of key-binding schemes are fuzzy commitment [31] and fuzzy vault [26]. Fuzzy commitment scheme was first proposed by Juels and Wattenburg. Fuzzy commitment is a two-steps algorithm consisting of commitment and decommitment. The fuzzy commitment scheme F commits a random codeword c using a one-way hash function h and a template x , where both c and x can be expressed as n -bit strings. Mathematically, we have $F(c, x) = (h(c), x - c)$ and the output is stored in the database. To decommit a query, x' denoted as the witness is used such that the extracted commitment $c' = f(x' - (x - c))$ where f is the decommit function. Decommitment is successful if $h(c) = h(c')$. The decommitment can always succeed if the distance between the query and the template is less than approximate half the minimum distance. In this case, the minimum distance is considered as the minimum Hamming distance between two codewords encoded by an error-correcting code.

The fuzzy commitment scheme was first realized by Inthavasis and Lopresti [32] who proposed password based cryptographic key regeneration. They utilized Dynamic Time Warping (DTW) on the extracted feature vector and mapped DTW features to a binary string called feature descriptor. Subsequently, the feature descriptor was used to define distinguishing features. The template was hardened by perturbing the template many times and one of the stable features is extracted each time. The extracted feature will be the key in DTW. The process continued until the distinguishing descriptor had less than or equal to half of the feature vector length. Finally, the hardened template was fed through the transformation, permutation and key binding processes using fuzzy commitment framework. It was shown that the security of this scheme is dominated by the password instead of the biometric feature [32].

Billeb et al. [33] proposed to construct a voice protection scheme based on the Universal Background Model (UBM). The proposed scheme binarized the supervector derived from UBM and an adapted fuzzy commitment scheme was used as the basis for the template protection scheme. Even though security analysis against unlinkability and privacy protection was provided, the proposed scheme still suffers from ARM when both key and the difference vectors are compromised. The adversary can exploit the compromised information to reconstruct the template stored in the database.

Paulini et al. [34] proposed the use of multi-bit allocation instead of single bit allocation. Different to Billeb's work, the proposed scheme divided the feature space into 2^k intervals and encodes each interval with k bits. A modified fuzzy commitment scheme was then applied on the binarized features. Their work outperformed the single bit allocation approach and preserved the performance of the recognition system with lesser degradation in the recognition ability. However, similar to Billeb's work, the presented scheme was vulnerable to ARM.

On the other hand, fuzzy vault scheme was proposed by Juels and Sudan [26]. The general idea of the proposed fuzzy vault scheme is to lock the secret key k under an unordered set A . A

polynomial p was selected in such a way that it is able to encode k into variable x . Random chaff points that do not lie on p were then added to set A , creating a vault which consists of collection of points which lie on p and chaff points. To unlock the key k by the means of set B , if B overlaps substantially with A , the collection of points that lie on polynomial p can be determined. Using these points, with error correction ability, the polynomial p can be reconstructed and thereby key k .

Johnson et al. [35] proposed a vaulted verification protocol, where a challenge-respond protocol and fuzzy vault were used in their security scheme. This work used the same database as the work [32] and the results had shown that it was able to achieve a better performance as compared to [32] under the scenario that all the keys had been compromised. The user voice feature was first separated into several blocklets and a chaff/fake blocklet was added to each real blocklet, forming many pairs of real and chaff blocklets. These pairs were then encrypted by password and stored in the template. During the authentication phase, the template was first decrypted and a challenging bitstring was generated such that real block represents "0" and chaff block represents "1". The pairs were then randomly swapped. The score computation was carried out by matching the bitstring response given by the user with the template. However, limited biometric information such as limited voice samples will not be able to vary the data in the challenge-response process due to lesser pairs of real and chaff blocks and thus the adversary will have higher probability in guessing the correct response [35].

1.1.3. Hybrid biometric cryptosystem

As biometric cryptosystems have limitations such as unable to generate multiple unlinkable templates, a hybrid approach of combining cancellable biometrics with biometric cryptosystems is proposed to overcome such limitation [12]. As the name implied, hybrid biometric cryptosystem is a combination of two or more template protection schemes such as bio-hashing with fuzzy vault scheme and key-binding scheme with non-invertible transformation [36]. Hybrid biometric cryptosystem reaps the benefit of cancellable properties from cancellable biometric while providing stronger security and privacy protection inherited from biometric cryptosystem. An instance of hybrid biometric cryptosystem is the cancellable speech template based on chaff point mixture method proposed by Zhu et al. [37] where a two-step hybrid approach (i.e. random projection and fuzzy vault) was used. The voice feature matrix was first randomly projected into another feature space and chaff points were added to the projected space instead of directly to the original feature matrix. Binary indices were used to bind the points and accumulator of genuine indices (key) were calculated using OR operator. The key will be sent to the matcher to filter out the genuine points from query using AND operator. The proposed work had shown that it was able to preserve the performance of the recognition system, however the security of the proposed work is not analyzed in detail as ARM analysis and lost key scenario were not considered. In the event that the binary indices and the key are compromised, the adversary will be able to differentiate the genuine points from randomly added chaff points.

Feng et al. [38] proposed a three-step hybrid framework for face template protection. A random projection matrix was first applied to the original biometric template to provide cancellability. To strike the balance between the security and the recognition performance, a class distribution preserving transform was then used to enhance the discriminatory power of the template and at the same time convert the template from real value to binary space. A distance function and thresholding were used in such a way that if the distance measured between the distinguishing points and the template is lower than the threshold, a "0" bit is generated, otherwise "1" bit is generated. The final step of the proposed framework

was to hash the generated binary template using MD5 hashing algorithm. The proposed work had shown a significant improvement on the recognition performance; however the proposed work was vulnerable to several security attacks. Wang and Yu [39] had outlined several drawbacks of using MD5 hash and concluded that finding a collision for MD5 is feasible.

1.2. Motivation and contribution

From the existing voice template protection schemes, we have observed that there are several issues that need to be addressed as follows:

1. **Robustness to attacks:** It is observed that most of the speech template protection schemes were vulnerable to different attacks such as attack-via multiplicity (ARM) and stolen-token attacks. The vulnerability of the scheme is most likely due to the high correlation between the templates generated using the same biometric feature. Hence, the adversary is able to derive the original template by analyzing multiple compromised templates. Thus there is an urgent need to ensure that the generated templates are independent to each another, fulfilling the unlinkability and revocability criteria.
2. **Performance degradation:** It can be seen from [32] and [36] that the transformation of the biometric feature from one space to another will cause the loss of the discriminative features. Thus, it will result in the increase of the intra-class variation and eventually lead to the drop of accuracy in the performance. Therefore, the template protection scheme should be able to preserve the performance of the system as much as possible while providing sufficient security protection.

In this paper, we propose a cancellable transform named Random Binary Orthogonal Matrices Projection (RBOMP) hashing, for the well-known voice representation, namely i-vector [2] to address the aforementioned security and privacy issues. Our proposed method is inspired from a hashing method, i.e. Winner Takes All [40] which is designed for the task of fast similarity search initially. Our main contributions are listed as follow:

- We proposed a cancellable transform e.g. RBOMP hashing to project the biometric feature to ordinal space using binary orthogonal matrices which will induce a strong non-invertible property and is resilient to small intra-class variation simultaneously.
- Prime Factorization (PF) feature is proposed to further enhance the security and privacy, more specifically, a many-to-one function, namely prime factorization approach together with a user-specific key, are incorporated.
- Security and Performance analysis. Through analysis on the security and performance of the proposed method are given to justify the common tradeoff of security and performance.
- Attack-via-Multiplicity (ARM) analysis. Extensive theoretical and simulation analyses on ARM are conducted to boost confidence towards the security against this major attack.

For the rest of the paper, a brief introduction to the generation of i-vector is provided in Section 2. Section 3 presents the proposed RBOMP hashing in detail. Section 4 demonstrates the experimental results and general security analysis. Besides, Section 5 provides detailed ARM analysis. Finally, an outline of the conclusion for this work is given in Section 6.

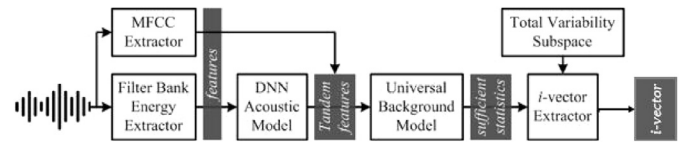


Fig. 1. The overview of the i-vector extraction.

2. Preliminaries

2.1. Generation of i-vector

The state-of-art feature extraction technique through i-vector provides a fixed-length low dimensional representation of speech utterances that preserves the speaker-specific information of each speaker. The Mel Frequency Cepstral Coefficient (MFCC), represented in a form of vectors and derived from a given utterance [4], was fed into a Universal Background Model (UBM). UBM is a K-component Gaussian Mixture Model (GMM), $\lambda = (w_k, m_k, \Sigma_k)$, where each of the symbols represents weight, mean and covariance respectively. Next, the Baum-Welch statistics is accumulated from each utterances. Hence the speaker utterances represented by a supervector (θ) that consists of additive components from speaker and channel subspace can be written in the form below,

$$\theta = m + Tx \quad (1)$$

where m represents the speaker- and channel- independent supervector (derived from the UBM), T represents the total variability matrix that spans the subspace that consist of most of the speaker-specific information and x is a standard normally distributed random vector that we refer as i-vectors [2].

The i-vector framework is greatly improved by applying senone DNNs to the speaker recognition system. In our system, tied tri-phone state phoneme posterior probabilities are first extracted by DNN acoustic model. Then, after performing logarithm to the posterior probabilities, principal component analysis is applied on top of it to get lower dimensional phonetic feature which is concatenated to MFCC to generate tandem feature. The tandem feature instead of MFCC is fed into UBM to accumulate Baum-Welch statistics in i-vector framework. Tandem feature provides discriminant phoneme information and thus achieves better performance in the speaker recognition system. Fig. 1 is the flowchart of the generation of i-vector with DNN.

2.2. Winner-Takes-All Hash

Winner-Takes-All Hash (WTA) is a method used for fast similarity search and was implemented by Google in their image search engine [40,41]. WTA used rank correlation measures and recorded the index of the maximum value of the biometric feature after applying random permutations. Different index vectors can be generated using different permutation sequences.

The procedure for deriving the index vector for Winner-Takes-All Hash is described as follows:

1. **Random Permutation.** Randomly permute the feature vector, X to generate X^p where X^p denotes the permuted feature vector.
2. **Select the first K-items.** The first K-items are selected from X^p for $2 \leq K \leq n - 1$. This step reduces the length of the feature vector and hence there will be information loss during this stage.
3. **Record index of the highest value.** The index of the highest value from the first K-items is recorded and denoted as C .
4. **Repeat Step 1 to Step 3 using H different permutation sequences.** A series of indexes, C_i will be generated, where

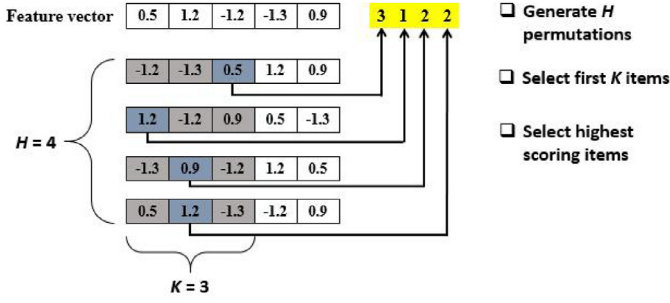


Fig. 2. An Example of the WTA Computation.

$i \in [1, H]$ and let S be the set consisting the generated C_i (i.e. $S = \{C_1, C_2, \dots, C_H\}$).

Fig. 2 shows an example of the WTA computation.

3. RBOMP hash

3.1. Baseline system

In this section, Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) is used as the baseline system. The detailed explanation and the matching protocol are discussed in Sections 3.1.1 and 3.1.2 respectively.

3.1.1. Gaussian Probabilistic Linear Discriminant Analysis (GPLDA)

Probabilistic linear discriminant analysis is widely adopted and considered as the state-of-the-art back-end modeling approach. Generally, we model the i-vectors with a Gaussian distribution assumption (GPLDA). We assume that the training data consists of j utterances from i speakers and denote the j th i-vector of the i th speaker by η_{ij} . We assume that the data are generated in the following way [42]:

$$\eta_{ij} = \phi\beta_i + \epsilon_{ij} \quad (2)$$

The speaker term $\phi\beta_i$ is dependent on the speaker. The noise term ϵ_{ij} is used to model the within-speaker variabilities and assumed to be Gaussian distributed with zero mean and diagonal covariance Σ . Suppose there are M_i i-vectors from the i th speaker, we have

$$F_i = \frac{1}{M_i} \sum_{j=1}^{M_i} \eta_{ij} \quad (3)$$

For the i th speaker, the prior and conditional distribution is defined as following multivariate Gaussian distributions:

$$P(F_i|\beta_i) = \mathcal{N}\left(\phi\beta_i, \frac{\Sigma}{M_i}\right), \quad P(\beta_i) = \mathcal{N}(0, 1) \quad (4)$$

The Expectation Maximization (EM) algorithm is employed in the modeling training. In the E -step, the posterior distribution of the hidden variable β_i given the observed F_i is:

$$P(\beta_i|F_i) = \mathcal{N}\left((I + \phi^T M_i \Sigma^{-1} \phi)^{-1} \phi^T M_i \Sigma^{-1} F_i, I + \phi^T M_i \Sigma^{-1} \phi\right) \quad (5)$$

In M -step, to maximize the conditional expectation of the log-likelihood

$$\log \left\{ \prod_{i=1}^M \sum_{j=1}^{M_i} P(\eta_{ij}, \beta_i) \right\}, \quad (6)$$

the updated ϕ and Σ are calculated as follows:

$$\phi = \left(\sum_i M_i F_i E(\beta_i^T) \right) \left(\sum_i M_i E(\beta_i \beta_i^T) \right)^{-1} \quad (7)$$

$$\Sigma = \frac{\sum_i \sum_j \eta_{ij} [\eta_{ij}^T - E(\beta_i)^T \phi^T]}{\sum_i M_i} \quad (8)$$

3.1.2. Verification score

In the speaker verification task, given a trial with two i-vectors η_i and η_j , we are interested in testing two alternative hypotheses, i.e. H_1 : both η_i and η_j are from the same speaker and they share the same speaker identity latent variable $\beta_i = \beta_j$; H_0 : they come from different speakers and the underlying hidden variables β_i and β_j are different [4,42]. The verification score can now be computed as the log likelihood ratio of these two hypotheses.

$$\text{score} = \log \frac{P(\eta_i, \eta_j | H_1)}{P(\eta_i | H_0) P(\eta_j | H_0)} \quad (9)$$

Since the corresponding distribution is all multivariate Gaussians, the score can be denoted in quadratic terms [8] as follows:

$$\begin{aligned} \text{score} &= \log \mathcal{N} \left(\begin{bmatrix} \eta_i \\ \eta_j \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{tot} & \Sigma_{ac} \\ \Sigma_{ac} & \Sigma_{tot} \end{bmatrix} \right) \\ &\quad - \log \mathcal{N} \left(\begin{bmatrix} \eta_i \\ \eta_j \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{tot} & 0 \\ 0 & \Sigma_{tot} \end{bmatrix} \right) \\ &= \eta_i^T Q \eta_i + \eta_j^T Q \eta_j + 2 \eta_i^T P \eta_j + c \end{aligned} \quad (10)$$

where c is a constant and Σ_{tot} , Σ_{ac} , Q and P are denoted as follows:

$$\Sigma_{tot} = \phi\phi^T + \Sigma \quad (11)$$

$$\Sigma_{ac} = \phi\phi^T \quad (12)$$

$$Q = \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac} \Sigma_{tot}^{-1} \Sigma_{ac})^{-1} \quad (13)$$

$$P = \Sigma_{tot}^{-1} \Sigma_{ac} - (\Sigma_{tot} - \Sigma_{ac} \Sigma_{tot}^{-1} \Sigma_{ac})^{-1} \quad (14)$$

3.2. Random Binary Orthogonal Matrices Projection (RBOMP) hashing

Inspired from WTA, we propose a new speech template protection scheme, coined as RBOMP hashing. As projection of the features from linear space to ordinal space yields a strong non-invertible property, it is computationally hard for the adversary to recover the original feature value from the protected template. However, as WTA only focuses on the rank of the features instead of the value of the features itself, the adversary may obtain the order of the features through ARM and reconstruct the original template. Hence, motivated by the fact that the returned index may be exploited by the adversary, a non-invertible function namely prime factorization is used to conceal the returned index with the help of a user-specific random token.

RBOMP is a hashing scheme consisting of k rounds of function h for $k > 1$. For ease of understanding, let i denotes the round number for $i = 1$ to k . Each round function h_i takes an i-vector X that consists of n real numbers and a random positive integer Z_i , where $1 \leq Z_i \leq 10000$, as input and generates an index S_i as output. The concatenation of indexes S_i generated in each round function h_i is denoted as the hashed code $S = S_1 || S_2 || \dots || S_k$ where $||$ denotes the concatenation. Mathematically, we have $S = \text{RBOMP}(X, Z)$ where $Z = \{Z_1, Z_2, \dots, Z_k\}$ and $S_i = h_i(X, Z_i)$. More precisely, h_i consists of the following steps:

1. *Projection, P*: Given a random binary orthogonal matrix M_i with a dimension of $n \times n$, where n is the length of the i-vector, compute the feature vector $X^F = P(X, M_i) = X \cdot M_i$ where \cdot is the matrix multiplication.

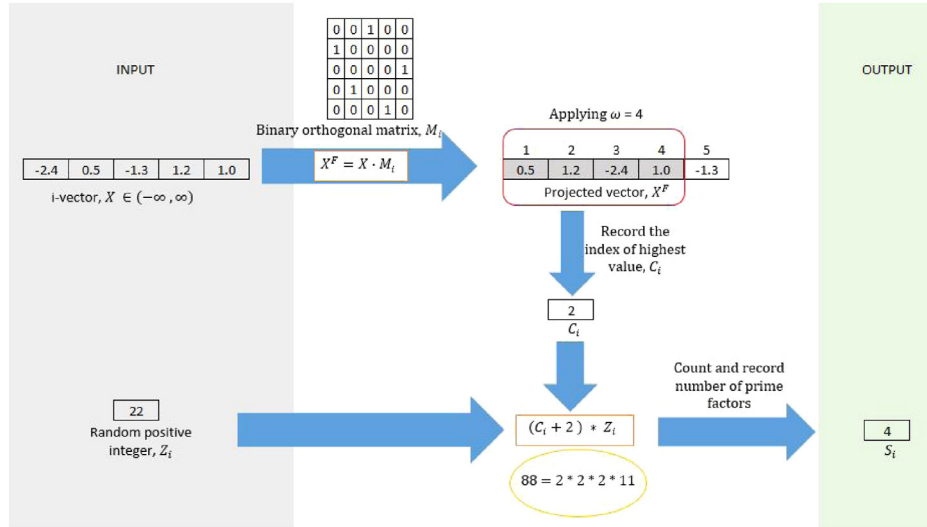


Fig. 3. Example of one-round RBOMP hash.

2. **Window, W :** Given the feature vector X^F and the window ω (the exact range of ω will be determined through experiments later), compute the windowed feature $X^W = W(X^F)$ by taking first ω real numbers of X^F . Since the length of the feature vector is reduced, certain information of the feature vector is lost.
3. **Find Intermediate Index, FI :** Given the windowed feature X^W , compute the intermediate index $C_i = FI(X^F)$ as the index/position of the highest value of the windowed feature X^W .
4. **Prime Factorisation, PF :** Given the intermediate index C_i and a positive integer Z_i , compute the index S_i as the number of prime numbers of $(C_i + 2) * Z_i$ where $*$ is the integer multiplication. The addition of 2 with C_i is performed as 1 is not a prime number and to lower the false acceptance rate (due to the fact that 2 and 3 are prime numbers).

Notice that different random binary orthogonal matrices M_i will be selected in different rounds of h function. In real world scenario, the selection of binary orthogonal matrices and random token (i.e. Z) is user-specific. In the event of the template being compromised, the user can revoke and reissue a new template by generating different binary orthogonal matrices and/or token to replace the compromised template. In our work, we will focus on lost-token scenario to evaluate the recognition performance and perform the security analysis for RBOMP hashing. In the lost-token scenario, the binary orthogonal matrices as well as the random token are assumed to be known to the adversary, therefore in the experiments, all the users are assumed to share the same binary orthogonal matrix and the random token. The pseudocode of the proposed scheme is shown in Algorithm 1 while the one-round graphical implementation of RBOMP hash is shown in Fig. 3 for illustration purposes.

3.2.1. Determining the range of ω

As the range of value of the intermediate index, C_i , is closely related to the value of ω , where $1 \leq C_i \leq \omega$. The range of ω is set in such a way that there will be at least two mappings of distinct value C_i to $PF(C_i + 2)$, where $PF(x)$ is a function that denotes the number of prime factors of x . Hence, the range of ω is set to be $(2^{q-1} * 3) - 2 \leq \omega < 2^{q+1} - 1 < 500$, where q is an integer in the range of $[2, 8]$. The motivation is to prevent the adversary from reconstructing the order of the i-vector through ARM practically (more details are discussed in Section 5).

Algorithm 1: RBOMP hashing.

Input: Window length ω , number of binary orthogonal matrices k , feature vector $X \in \mathbb{R}$, random token $Z_i \in \{1, 10000\}$

for $i = 1 : k$ **do**

Step 1: Compute $X^F = P(X, M_i) = X \cdot M_i$.

Step 2: Compute $X^W = W(X^F)$ by constructing ω -window.

Step 3: Compute $C_i = FI(X^F)$ as the index/position of the highest value of the windowed feature X^W .

Step 4: Compute S_i as the number of prime factors of $(C_i + 2) * Z_i$.

end

Output: Hashed Code, $S = \{S_i | i = 1, \dots, k\}$ and $S \in \mathbb{Z}^+$.

3.2.2. Matching

Transforming the feature to ordinal space which is not sensitive to the value of the feature dimension shifts the focus to the implicit ordering implied by the values [40]. As rank correlation refers to the measure of the degree of correlation between the ranks of the members within a set, the similarity measurement of the feature representation can be defined as the degree to which the rank of their feature dimension agrees [40].

Let c refers to the maximum value of a given ω -sized window. The similarity score is defined as the probability of both hashed code S and S' having c at the same position (i.e. $S_x = S'_x$ for $x = 1, \dots, k$). The higher the probability implies that the hashed code S and S' have a high similarity. In our experiments, the number of collisions will be calculated by counting the number of zeros after performing element-wise subtraction between two hashed codes.

The procedure of the similarity score calculation is described as follows. Besides, Fig. 4 shows an example of similarity score computation.

1. **Taking the difference of two hashed codes.** Given an enrolled hashed code, S_x and a query hashed code, S'_x , the difference of S_x and S'_x is computed by taking $S_x - S'_x$.
2. **Count the number of zeros.** The number of "0" is counted after taking the difference of S_x and S'_x . The "0" in this case indicates a match between the hashed codes and by counting the number of "0", the total matches of two hashed codes can be determined.

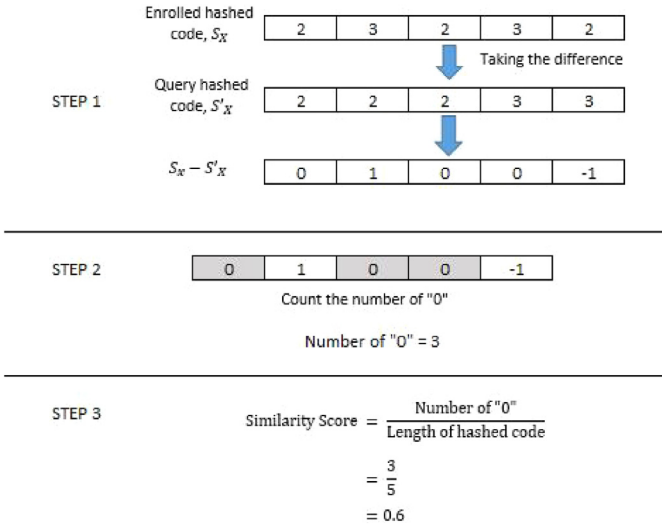


Fig. 4. Example of Similarity Score computation.

3. *Compute Similarity Score.* Similarity Score is computed by taking the total number of "0" over the length of the hashed code.

4. Experiment and analysis

The training set in this experiment is the database released through the Linguistic Data Consortium (LDC) for the NIST Speaker Recognition Evaluation (SRE) 2004–2010, as well as Switchboard-2 Phase II corpora. The Gaussian PLDA model with a full covariance residual noise term is trained on i-vectors extracted from all training data which amounted to 2790 speakers and 30,600 speech files. The eigenvoice subspace in the PLDA model is assumed to be full-rank. Besides that, there are 2391 enrolled models and 379 test segments in the evaluation set which is for NIST SRE 2010 extended condition 5 (tel-to-tel) female part task. The speakers in training set are different from the evaluation set to avoid correlation between them. The accuracy performance of the system is evaluated based on equal error rate (EER). The EER is the rate when false acceptance rate (FAR) equals to the false rejection rate (FRR).

Tandem features used in this experiment is extracted as follows. First, each utterance is processed by a Czech phoneme recognizer to perform the voice activity detection (VAD) and then converted into a sequence of 36-dimensional MFCC features consisting of 18 Mel frequency cepstral coefficients and their derivatives. Next, to extract high dimensional phonetic features, each utterance with high resolution MFCC is fed into an DNN acoustic model which is trained by Kaldi's time delay deep neural network (TDNN) [43] using 1800 h of the English portion of Fisher corpora [44]. After applying PCA, 52-dimensional features as the result are concatenated to MFCC at feature level to generate 88-dimensional hybrid tandem feature.

A gender-dependent 1024-component UBM is trained with the extracted 88-dimensional tandem features using NIST SRE 2004 and 2005 corpora. For each utterance, the corresponding tandem features are fed into UBM to compute zero-order and first-order Baum-Welch statistics. The resulting high dimensional Baum-Welch statistics are then projected on the 500-dimensional total variability subspace, which is trained with Switchboard-2 Phase II and NIST SRE 2004, 2005, 2006 and 2008 corpora, to extract i-vectors.

Since our proposed method is training-free (the training set mentioned earlier is for i-vector generation purpose), only the

speakers with five or more speech utterances are selected. After the selection process, the first five samples from each 2001 speakers are selected and used in the experiments. Such selection ensures that the total number of samples and number of impostors are balanced for the computation of genuine scores during the matching phase.

For intra-class comparison, there are a total of 20,010 genuine matches whereas for inter-class comparison, there are 2,001,000 impostor matches. To avoid biasness of the results obtained from a single random binary orthogonal matrix and random token,¹ the experiment for each parameter is repeated for five times and the average EER is obtained.

4.1. Effect of ω and k on the recognition performance of the proposed method

In this section, the effect of ω and k on the EER is investigated. The number of binary orthogonal matrices, k is set to vary from 1000, 2000, 5000 and 10,000 for different ω settings (i.e. $\omega = 4, 5, 10, 11, 12$). Fig. 5 shows the effect of different numbers of random binary orthogonal matrices and different lengths of window on the EER. The recognition performance of the system improves (indicated by lower EER) with the decrease in the value of ω and the increase in the value of k due to more information available for the verification process to distinguish the speakers. Smaller values of ω (i.e. lesser than 4) are not considered for security reasons as we need to ensure that the adversary will not be able to reconstruct the order of the i-vector through attack-via-multiplicity (ARM) practically (more details can be found in Section 3.2.1). As k increases, the EER slowly converges to a certain point as there will be no significant changes in the value of EER thereafter. More details of the security analysis will be discussed in Section 5.

4.2. Comparison of the recognition performances for different methods

Using the baseline system as the benchmark for fair comparison, the recognition performance of the proposed method is compared with other methods by selecting the lowest EER achieved from the experiments. The comparison results are summarized in Tables 1 and 2.

Compared with other methods, the proposed method can offer strong security while preserving the recognition performance of the system with acceptable degradation (approximately 1.76% more in EER as compared to the baseline system). The loss in accuracy is mainly due to the fact that there is less information used for verification as compared to the baseline system since additional user-specific helper data is used in the baseline system. The lack of information has caused the loss of some discriminatory properties of the voice feature, hence degrades the recognition performance of the proposed method.

From Table 2, it shows that the proposed method is able to produce satisfactory recognition results. It is worth mentioning that the database used in this work consists of larger number of people, i.e. 2001 people, as compared to other work. More importantly, NIST SRE series are widely used to measure the state-of-the-art speaker recognition systems. Different with other work, extensive analysis is conducted on the security of the proposed method against ARM (refer Section 5 for the details). Furthermore, inspired

¹ In real-world and ideal scenario, all users will have different sets of binary orthogonal matrix and token. Throughout our experiments, we assume the worst case scenario where all users use the same set of binary orthogonal matrix and token to evaluate the recognition and security performances of our proposed scheme under worst case scenario.

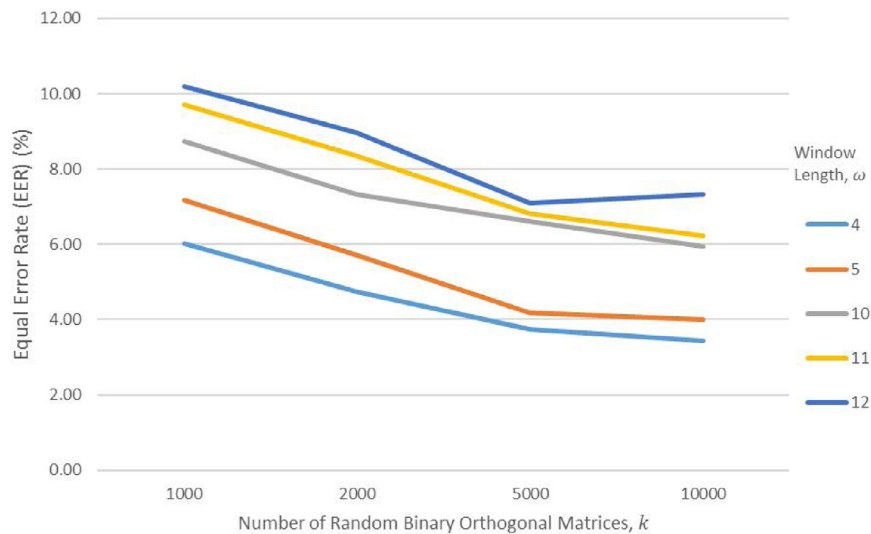


Fig. 5. EER versus number of random binary orthogonal matrices for different lengths of window.

Table 1
Comparison of the recognition performance without template protection methods.

Method	Lowest EER (%)
Baseline: GPLDA [4]	1.67
Cosine Similarity	2.41
Euclidean Distance	10.14

from the current trend of “parallelizable” (multi-core, pipeline, superscalar and vector), the proposed method can be parallelized given that each round function h is independent.

4.3. Comparison of the recognition performance for different databases

To further justify the recognition ability of the proposed scheme, the experiment is carried out on two other databases, namely Chinese Mandarin Speech Recognition Corpus – Digital String and Chinese Mandarin Speech Recognition Corpus – Conversation. To extract the speech feature in the form of i-vectors,

similar training and testing procedures are applied on both of the datasets.

The training set of Chinese Mandarin Speech Recognition Corpus – Digital String comprised of Chinese Mandarin Speech Recognition Corpus – Digital String (Mobile), Chinese Mandarin Speech Recognition Corpus – Digital String (Telephone) and Chinese Dialectal Mandarin Speech Recognition Corpus – Digital String (Telephone) which consists of 647 different speakers with a total of 19,478 utterances. For testing set, Chinese Mandarin Speech Recognition Corpus – Digital String (Desktop) with 120 speakers with a total of 3600 utterances are used.

On the other hand, Chinese Mandarin Speech Recognition Corpus – Conversation consists of 1000 speakers with 7130 speech utterances. After Voice Activity Detection (VAD) procedure, only 943 speakers remain where out of the remaining speakers, 200 speakers with 1380 utterances are selected as the testing set while the rest 743 speakers with 4400 utterances are selected as the training set.

Setting $\omega = 4$ and $k = 8000$, the EER are as shown in Table 3. The results of the datasets suggest that the accuracy performance is well preserved for Chinese Mandarin Speech Recognition Corpus

Table 2
Comparison of different speech template protection methods.

Method	Database	Number of Speakers	Baseline EER Before Protection (%)	Lowest EER After Protection (%)	Brute Force Attack Complexity (bits)	Parallelizable	ARM Analysis
Universal Background Model [33]	Text- Independent Digit Corpus	701	3.40	5.42	N.A.	×	×
Multi-bit Allocation [34]	Text- Independent Digit Corpus	701	3.40	3.56	N.A.	×	×
Vaulted Voice Verification [35]	MIT Mobile Device Speaker Verification Corpus	48	11.00	6.00	$8 \sim 12^a$	×	×
Random Projection + Fuzzy Vault [37]	Mandarin Continuous Speech Recognition	40	2.22	2.22	$O(1)^b$	×	×
Proposed Method	NIST SRE 2004–2010	2001	1.67	3.43^c	40.24	✓	✓

^a Johnson et al. [35] claimed that the attacker may gain access to the recognition system with a probability of 2^{-8} to 2^{-12} in addition to the n -bit of security offered by encryption. Under the lost key scenario, the encryption can be decrypted easily by the attacker and thus the security offered by encryption can be ignored.

^b Zhu et al. [37] claimed that the time complexity in launching a brute force attack to differentiate 32 genuine points out of total 332 points is $\binom{332}{32} \approx 2^{148.11}$ assuming binary indices and key are kept secret. However, under the lost key scenario, the attacker can obtain the binary indices and key, thus the attacker can differentiate 32 genuine points out of total 332 points easily with negligible time complexity (i.e. $O(1)$).

^c We performed an experiment using all the utterances (30,600 utterances from 2790 users) to observe the robustness of the system and we obtained an EER of 7.32% using the same parameter ($\omega = 4$ and $k = 10,000$). It can be observed that the EER increases with the increase of testing samples. This is not surprise as behavioral biometrics like voice contain more human factors, e.g. emotional states [1,45]. This leads to more noise from sample to sample as compared to physical biometrics e.g. fingerprint. It is expected that behavioral biometrics provide lower level of robustness as we can observe from the experiment.

Table 3
Comparison of recognition performance for different databases.

Database	Baseline EER (%) before template protection	Lowest EER (%) after template protection
NIST SRE 2004 ~ 2010	1.67	3.43
Chinese Mandarin Speech Recognition Corpus – Digital String	3.81	7.01
Chinese Mandarin Speech Recognition Corpus – Conversation	0.60	0.89

– Conversation and not large deterioration for Chinese Mandarin Speech Recognition Corpus – Digital String and NIST SRE 2004 ~ 2010. Hence it is evident that the recognition performance of the proposed scheme is dependent on the quality of the voice feature extracted.

4.4. Brute force attack

As RBOMP projects the voice feature from linear space to ordinal space, it imposes strong non-invertible properties to the system as it is computationally difficult for the adversary to recover the original feature value in linear space. From the distribution of the feature value ranging from a minimum of -7.0000 to a maximum of $+7.0000$, if the adversary wants to guess the correct value of an i -vector with a length of 500, the guessing complexity is of $140000^{5000} \approx 2^{8547}$ attempts. Even if the adversary wants to guess the rank of the biometric feature instead of the feature value itself, it would still require a guessing complexity of $500! \approx 2^{3768}$ attempts.

The similarity score is computed based on the number of matches between the query hashed code and the enrolled hashed code. If the similarity score exceeds the threshold, the user will be deemed as the legitimate user. The threshold needed for a user to gain access to the system is approximately 0.55. Hence, using $k = 5000$, the adversary will require a minimum of $5000 \times 0.55 = 2750$ correct matches in order to gain illegitimate access to the system. Thus, it requires an average time complexity of 2^{2750} attempts to gain access to the recognition system.

Consider the scenario that the adversary does not compromise any templates, since there are only two possible outcomes for S_i , and each binary matrix is independent and uniformly distributed, one can assume that S follows binomial distribution with probability of 0.5. Let X denotes the number of correct guesses, the probability of obtaining 2750 or more correct guesses can be computed as follows:

$$\begin{aligned}
 P(X \geq 2750) &= P\left(z \geq \frac{2750 - 2500}{\sqrt{1250}}\right) \\
 &= P(z \geq 7.071) \\
 &= 2^{-40.24}
 \end{aligned} \tag{15}$$

From Eq. (15), the number of guesses required is of $2^{40.24}$. This can be referred as the average time complexity required to gain access to the system without compromising any templates. To further improve the security of the recognition system, one can limit the number of login attempts.

4.5. Revocability analysis

The revocability is evaluated by matching a particular hashed code with the other hashed codes generated from distinct random binary orthogonal matrices. A total of 100 hashed codes is derived from an i -vector with 100 different binary orthogonal matrices and the first hashed code is matched with the other hashed codes to

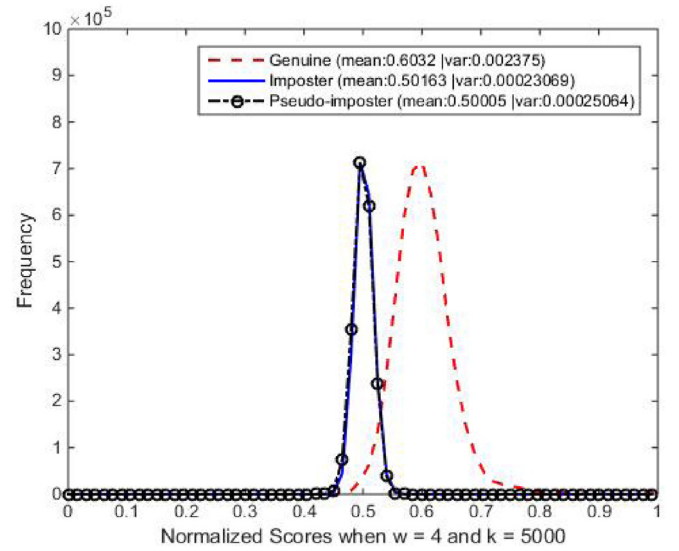


Fig. 6. Distribution of genuine, imposter and pseudo-imposter scores.

compute the pseudo-imposter scores. The process is repeated using the same random token for different users to produce a total of $99 \times 2 \times 2001 = 396,198$ scores.² The distribution of the genuine scores, imposter scores and the pseudo-imposter scores are computed using $\omega = 4$ and $k = 5000$ as shown in Fig. 6. The difference in the number of scores computed for imposter and pseudo-imposter matching is because that in pseudo-imposter matching, we only focus on matching the first generated hashed code with other generated hashed code for each i -vector. From Fig. 6, we can notice that the pseudo-imposter scores distribution resembles the imposter scores distribution. This vindicates that the newly generated hashed codes are indistinguishable to each other although they are generated from the same i -vector. Since the newly generated hashed code is uncorrelated to the old hashed code, this justifies that RBOMP hashing has fulfilled the revocability criteria.

4.6. Unlinkability analysis

The unlinkability is evaluated by introducing the pseudo-genuine scores. The pseudo-genuine score is computed by matching the hashed codes generated from different i -vector of the same user with different binary orthogonal projection matrices. Similar to the genuine matching, the pseudo-genuine match produces 20,010 scores. The overlapping of pseudo-imposter scores (from Section 4.4) and pseudo-genuine scores will indicate whether the RBOMP hashed codes generated from the same user or from another are indistinctive. The hashed codes are considered to be unlinkable when it is difficult to differentiate them. As shown in Fig. 7, there is a large overlapping between the pseudo-genuine scores distribution and the pseudo-imposter scores distribution. Hence this suggests that the RBOMP hashed is able to fulfill the unlinkability property.

5. Security analysis against Attack-Via-Record Multiplicity (ARM)

ARM refers to a privacy attack whereby the attacker uses multiple compromised templates with or without the associated information such as the parameters and algorithms to recover the original biometric template [46]. In our work, our main concern will be

² We generate 100 hashed codes for the first two samples of each user only to save the computational time of our experiments.

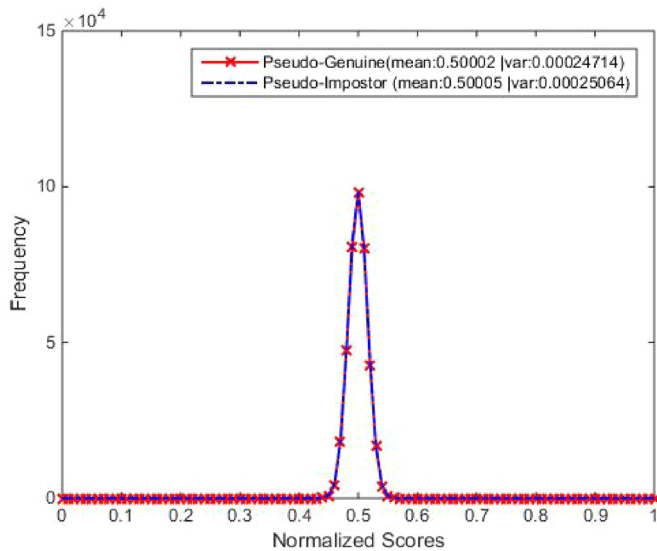


Fig. 7. Distribution of pseudo-genuine scores and pseudo-impostor scores.

on whether the adversary is able to guess the rank of the biometric feature. This is mainly due to the fact that guessing the rank of the biometric feature is relatively easier as there are lesser possibilities as compared to recovering the original feature value which

is in real number domain. If the random token Z_i and the hashed code S are compromised, the adversary might be able to obtain the intermediate index, C_i , by observing the number of prime factors in Z_i and S . This is because the value of S is computed by taking the sum of the prime factors of $Z_i * (C_i + 2)$. If one has the knowledge of C_i , he can reconstruct the order of the biometric feature. Hence for security purposes, it is important to set the range of the window length, ω , in such a way that there will be at least two possible values of intermediate indices, C mapped to each S . For instance, setting $\omega = 4$ will allow C to have 4 possible values, which are 1, 2, 3 and 4. Therefore, the number of prime factors of $C + 2$ will be 1, 2, 1 and 2 respectively. In this case, it can be seen that if S has the value of 1 then the possible value of C would be either 1 or 3. Meanwhile, if S has the value of 2 then the possible value of C would be either 2 or 4. Here the value of Z_i is not taken into account since the value of Z_i will not affect the analysis. A clear graphical representation on the mapping of C to S is shown in Fig. 8 using different setting of ω .

As stated earlier, by observing the number of prime factors in Z_i and S , the adversary might be able to recover the value of $C + 2$. However, since there are two or more mappings to each value of S , (i.e. many-to-one mapping), there will be an increase in complexity for the adversary to obtain the correct value of C . Fig. 9 illustrates how the adversary might be able to recover the value of C .

In our work, using $\omega = 4$ and the minimum number of random binary matrices, k is set to 5000, the naïve approach for the adversary to recover the correct value of C would be 2^{5000} as

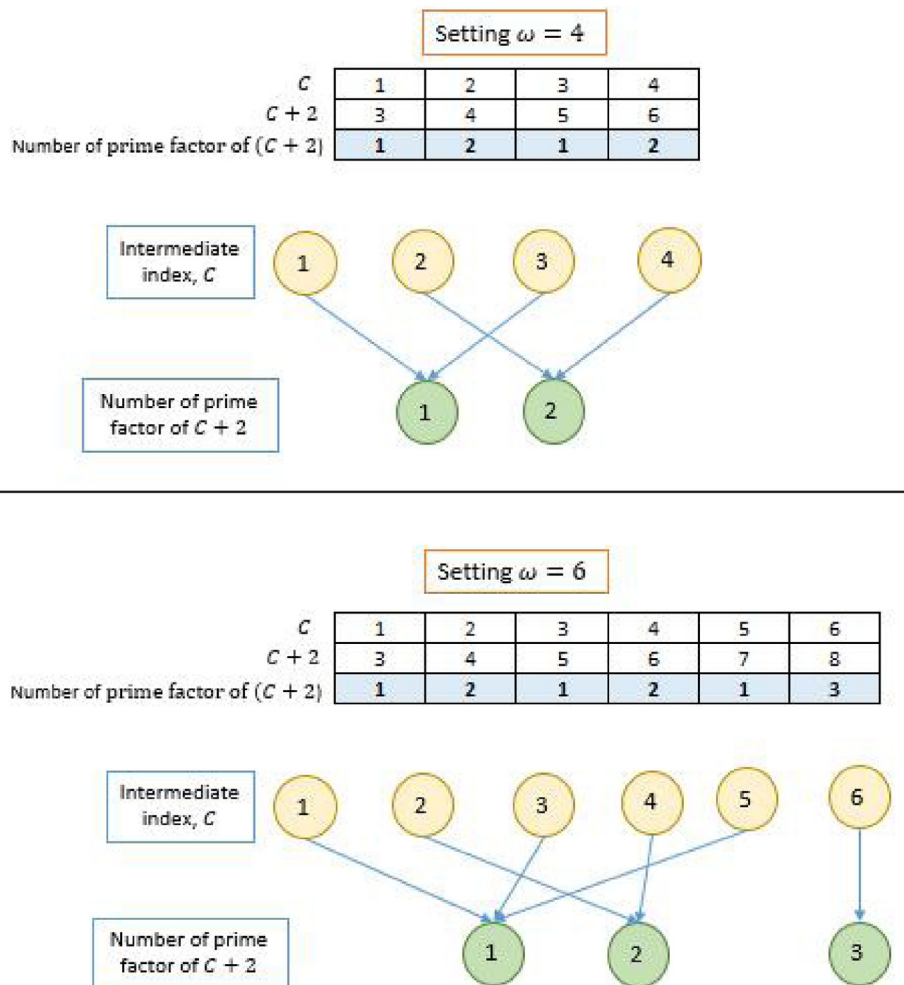


Fig. 8. Mapping of intermediate index, C , at $\omega = 4$ and $\omega = 6$ respectively.

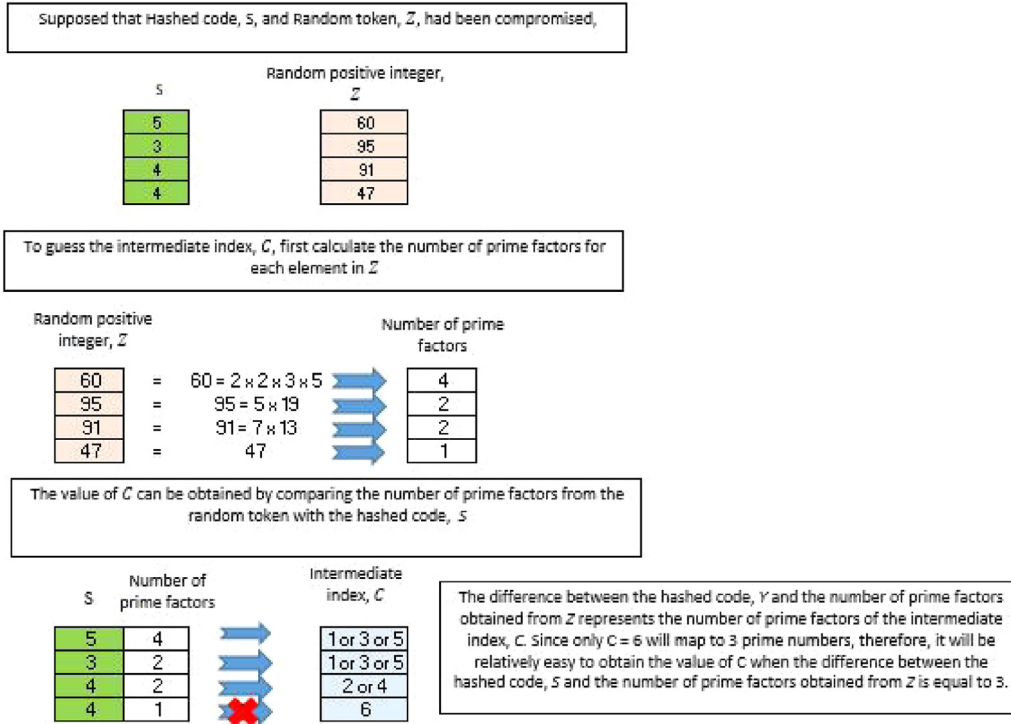


Fig. 9. Guessing the index, C , at $\omega = 6$ when both hashed code and random token are compromised.

for each round, there will be two possibilities of C . However, in real world scenario, the adversary would require much lesser than 2^{5000} attempts. Given the threshold of acceptance is 0.55 (as given in Section 4.3), using false accept attack, the adversary only needs to guess 2750 bits correctly. Since the probability of guessing a bit correctly is 0.5, one would expect the adversary to obtain 2500 correct guesses (out of 5000 guesses) on average. In other words, the adversary would only require another 250 correct guesses to gain access to the system. Given that the S_i follows binomial distribution, the average time complexity to access the system will be $2^{40.24}$ as stated in Section 4.3.

Given the worst case scenario that the adversary will always obtain the binary matrices he desired, the adversary may require lesser number of binary matrices to derive the order of the i-vector. Fig. 10 shows how the adversary is able to obtain the order of i-vector using the desired binary matrices.

Using Fig. 10 as an example, given that the adversary has compromised one binary matrix and the number of prime factor of $(C + 2)$ is 1, the adversary will be able to recover the windowed vector, X_1^w , after applying the compromised binary matrix. By using the information of X_1^w and number of prime factor of $(C + 2)$, the adversary will conclude that the possible position for the largest value among the four elements in X_1^w will be at position 1 or 3. In other words, the elements in position 1 or 3 are the candidates for the largest value of among these four elements. Next, in order to determine the correct position of the largest value, he will need to compare with another windowed vector, X_2^w , where X_2^w can be obtained after applying a different binary matrix, or here we define as the desired binary matrix. The desired binary matrix is determined in such a way that it will result in X_2^w having the same four elements as X_1^w and there will be only one common element between the candidates for the largest value of X_1^w and X_2^w (i.e. the candidates for largest value of X_1^w are {"a", "b"} and the candidates for largest value of X_2^w are {"d", "a"}). The common element between this two sets is "a" and hence "a" can be determined as having largest value among these four elements). For any

compromised binary matrix, there will be a total of 16 desired binary matrices that can be used to derive the largest value of the elements. As the length of i-vector is 500, there will be $\binom{500}{4} \approx 2^{31.26}$ unique binary matrices. Under the situation where the adversary will always obtain the desired binary matrices, it will only require him $\frac{2^{31.26}}{5000} \approx 2^{18.97}$ minimum number of templates to reconstruct the full order of the i-vector.

However, in real-world scenario the adversary will not always obtain the binary matrices that he wants. Hence if given the knowledge of all the possible binary matrices that can be generated, or here we refer as the distinct binary matrices, the adversary will be able to find a suitable pairing of the distinct binary matrices he needs and derives the order of the i-vector. Therefore, the focus of the issue will be on how many attempts are required for the adversary to obtain all the distinct binary matrices. This scenario can be reduced to the coupon collector's problem [47]. The coupon collector's problem is a probability problem where it describes the probability or the expected trials required to collect all different coupons from a finite set with replacement. In our case, the distinct binary matrices that provide useful information can be viewed as the coupon as the adversary are required to obtain all different useful binary matrices to reconstruct the ordering of i-vector. As each of the binary matrices are uniformly generated, one can say that each binary matrix is equally likely to be obtained at any time with a probability of $\frac{1}{m}$, where m is the total number of distinct matrices, $2^{31.26}$. Let X_i be the random variable for the number of trials required to complete the order of i-vector, and the probability of obtaining a new distinct binary matrix will be $\frac{m-i+1}{m}$ where m is the total number of distinct matrices and i is with the range from $[1, m]$. By the independence assumption, X_i , $i \in \{1, m\}$, is independent to each other and it follows a geometric distribution with the parameter, $p = \frac{m-i+1}{m}$. The expected number of trials of a particular distinct matrix, $E(X_i)$, can be computed using the formula $\frac{1}{p}$ and hence taking the sum of all the number of expected trials of distinct matrices, the total expected number of trials, $E(X)$,

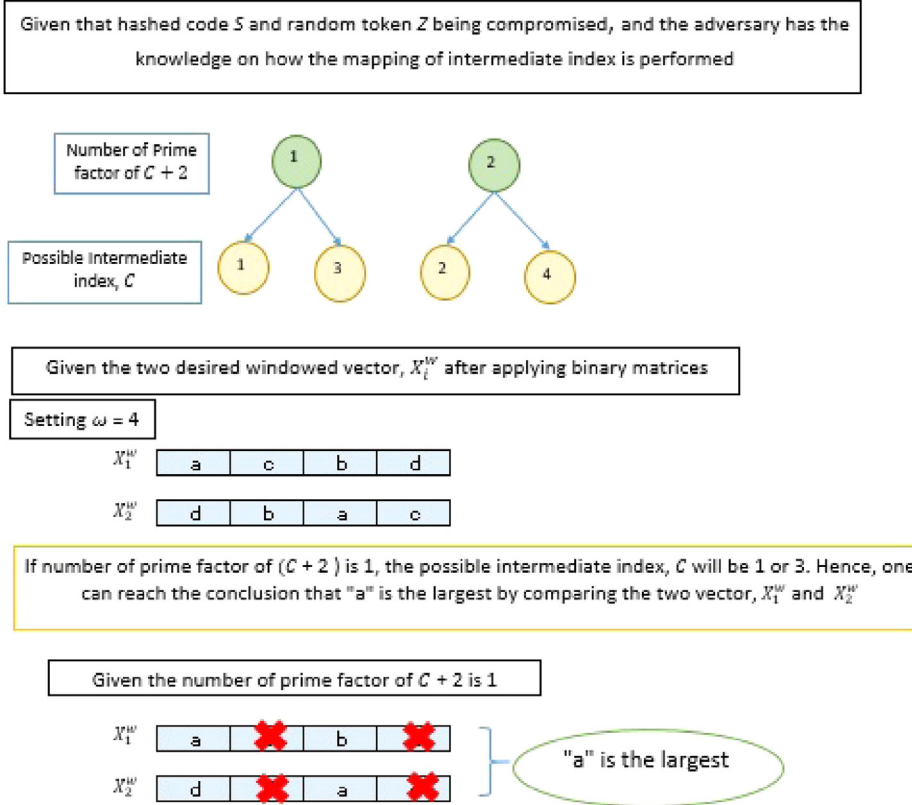


Fig. 10. Deriving the order of i-vector by comparing two desired windowed vector.

needed to obtain m distinct matrices will be as follows:

$$X = X_1 + X_2 + X_3 + \dots + X_{m-1} + X_m \quad (16)$$

$$\begin{aligned} E(X) &= E(X_1) + E(X_2) + E(X_3) + \dots + E(X_{m-1}) + E(X_m) \\ &= \frac{m}{m} + \frac{m}{m-1} + \frac{m}{m-2} + \dots + \frac{m}{2} + m \\ &= m \sum_{i=1}^m \frac{1}{i} \end{aligned} \quad (17)$$

Subsequently, using approximation formula, we obtain

$$E(X) = m(\log m + \sigma + \frac{1}{2m} + O(\frac{1}{m^2})) \quad (18)$$

where $\sigma \approx 0.5772156649$ is the Euler-Mascheroni constant. From Eq. (18), given $m = 2^{31.26}$, the expected number of trials needed to obtain m distinct binary matrices is around $2^{35.70}$. Fig. 11 shows the expected number of trials needed to collect m distinct binary matrix.

Hence, the expected number of templates required for the adversary to obtain all the distinct coupons will be $\frac{2^{35.70}}{5000} \approx 2^{23.41}$. However, it is impractical for an adversary to compromise $2^{23.41}$ templates. This can be referred as the data complexity required to gain access to the system with probability of one where the time complexity is negligible.

To relax the data complexity and the time complexity, we consider the scenario where the adversary only requires to compromise a small number of templates to gain access to the system with smaller time complexity. Given that the adversary is able to compromise some number of templates, the adversary will be able to deduce some correct number of S_i , from the compromised templates and guess the remaining number of S_i to gain access to the system. Thus, the adversary can launch the false accept attack

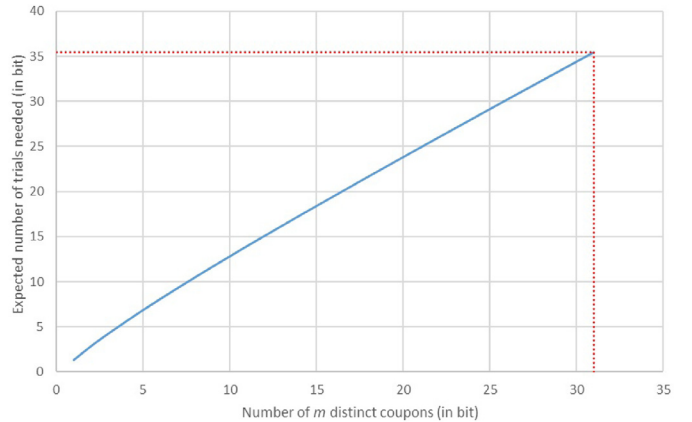


Fig. 11. Expected number of trials needed to collect m distinct coupons.

(i.e. described in Section 4.3) with lesser time complexity. Table 4 shows the relationship of the number of intermediate indices from the templates with the corresponding time complexity required to access to the system.

In our work, the intermediate indices, C , used is 5000 and the threshold to access the system is 0.55. In other words, the adversary will require 2750 correct guesses to access the system. Hence the remaining number of intermediate indices, C , required to access the system will be referring to how many guesses are needed in order to exceed the threshold (2750 correct guesses) after certain number of C s have been compromised. Assuming the remaining number of intermediate indices required to access the system follows binomial distribution with the parameter, $p = .5$, the remaining number of intermediate indices required to guess is denoted as N and the number of correct guesses is denoted with X .

Table 4

Comparison between number of intermediate indices compromised versus time complexity required to access the system.

Number of Intermediate Indices Compromised from Template	Remaining Number of Intermediate Indices Required to Guess	Remaining Number of Intermediate Indices Required to Access the System	Mean	Standard Deviation	Time Complexity (bits)
100	4900	2650	2450	35.00	27.436
200	4800	2550	2400	34.64	17.435
300	4700	2450	2350	34.28	9.146
400	4600	2350	2300	33.91	3.833
500	4500	2250	2250	33.54	1
600	4400	2150	2200	33.17	Negligible
700	4300	2050	2150	32.79	Negligible
800	4200	1950	2100	32.40	Negligible
900	4100	1850	2050	32.02	Negligible

The time complexity is calculated by taking the reciprocal of the probability of X greater than or equal to the remaining number of intermediate indices, C , required to access the system, with mean of $N \times p$ and standard deviation of $\sqrt{N \times p \times (1 - p)}$. From Table 4, it can be seen that the time complexity is negligible if the number of intermediate indices, C , compromised is at least 500. Hence one can expect the adversary will gain access to the system with probability of one once he/she has compromised at least 500 intermediate indices.

In order to obtain the minimum number of templates required to compromise at least 500 intermediate indices (as the time complexity is negligible after 500 intermediate indices are compromised), an experiment is carried out to determine the minimum number of templates required to compromise for different data size. Experiment is performed on a system consisting of 48GB RAM running on Ubuntu OS. The procedure of the experiment is as follows:

1. An empty array is created consisting of m rows and 24 columns (as there are a total of 24 possible binary matrices that will result in the windowed vector consisting the same four elements), where m is the data size.
2. Mersenne Twister pseudorandom number generator [48] is used to generate a pseudorandom number and the occurrence of the random number is marked on the empty array.
3. The 24 columns of the empty array are divided into three groups consisting of eight columns each, namely P , Q and R in such a way that a collision is only considered when there is at least a pairing of occurrences of different groups of the same row (i.e. elements in group Q and R can be referred as the desired binary matrices for the elements in group P). As explained previously, for any compromised binary matrix, there will be a total of 16 desired matrices that can be used to derive the largest value of the four elements. Here, the pseudorandom number symbolizes that a particular binary matrix is compromised and along with a desired matrix (another random number in a different group), the adversary will be able to derive the largest value (referred as collision in this experiment).
4. Step 2 to Step 3 are repeated and number of runs are recorded when the number of collisions reached 500.
5. The experiment is repeated for five times to obtain the average number of runs needed to obtain 500 collisions.
6. The minimum number of templates required to be compromised is computed by dividing the number of runs with 5000 (as a template consists of 5000 runs).
7. The experiment is repeated for different data sizes, d , and the minimum number of templates needed is recorded.

As we have $m = 2^{31.26}$ unique matrices (i.e. obtained from $\binom{500}{4}$), the minimum number of templates required to be compromised is approximately 600 as shown in Fig. 12. Thus, it is imprac-

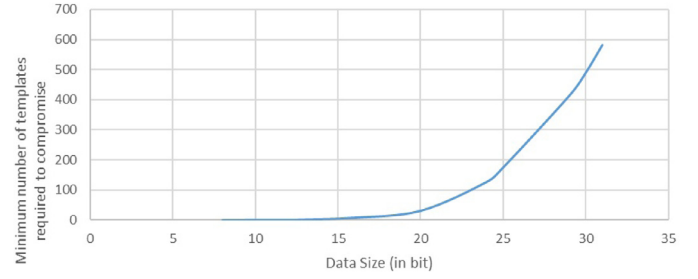


Fig. 12. Minimum number of templates required to compromise to access the system for different data sizes.

tical to compromise 600 templates from different database systems in the real world.

6. Conclusion

In this paper, we have proposed a cancellable speech template protection scheme namely RBOMP hashing. Extensive experimental results and theoretical analysis have vindicated that RBOMP is able to survive major security and privacy attacks at the same time able to preserve the verification performance. We also have demonstrated that the scheme is able to satisfy the evaluation criteria of the biometric scheme, for instance cancellability and revocability, and the user is not required to keep the binary orthogonal matrix or random token in secret. The scheme reaps the benefit of the fast similarity search from WTA and able to achieve a strong non-invertible property with the addition of the non-invertible function, prime factorization and a user-specific random token. In addition to that, a detailed theoretical and experimental analysis on the security against ARM is well-demonstrated to justify that the proposed scheme is able to resist against ARM practically. Lastly, we believe that the proposed method is not limited to voice biometric modality but other popular biometric modalities such as fingerprint and face with real-value representation.

Acknowledgements

This research was funded in part by Ministry of Science, Technology and Innovation, Malaysia under MOSTI Science Fund number 01-02-11-SF0189, National Natural Science Foundation of China (61401524, 61773413), Natural Science Foundation of Guangdong Province (2014A030313123), Natural Science Foundation of Guangzhou City (201707010363), Science and Technology Development Foundation of Guangdong Province (2017B090901045) and National Key Research and Development Program (2016YFC0103905).

References

- [1] J.A. Unar, W.C. Seng, A. Abbasi, A review of biometric technology along with trends and prospects, *Pattern Recognit.* 47 (8) (2014) 2673–2688.
- [2] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, *IEEE Trans. Audio, Speech, Language Process.* 19 (4) (2011) 788–798.
- [3] P. Matějka, O. Glembek, F. Castaldo, M.J. Alam, O. Plchot, P. Kenny, J. Černocký, Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification, in: *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4828–4831.
- [4] D. Garcia-Romero, C.Y. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 256–259.
- [5] P. Kenny, Bayesian speaker verification with heavy-tailed priors, in: *The Speaker and Language Recognition Workshop (Odyssey)*, 2010, p. 14.
- [6] K. Sheng, W. Dong, W. Li, J. Razik, F. Huang, B. Hu, Centroid-aware local discriminative metric learning in speaker verification, *Pattern Recognit.* 72 (2017) 176–185.
- [7] T. Pechovsky, A. Sizov, Comparison between supervised and unsupervised learning of probabilistic linear discriminant analysis mixture models for speaker verification, *Pattern Recognit. Lett.* 34 (11) (2013) 1307–1313.
- [8] Y. Lei, N. Scheffer, L. Ferrer, M. McLaren, A novel scheme for speaker recognition using a phonetically-aware deep neural network, in: *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1695–1699.
- [9] M. Li, L. Liu, W. Cai, W. Liu, Generalized i-vector representation with phonetic tokenizations and tandem features for both text independent and text dependent speaker verification, *J. Signal Process. Syst.* 82 (2) (2016) 207–215.
- [10] P. Matějka, L. Zhang, T. Ng, S.H. Mallidi, O. Glembek, J. Ma, B. Zhang, Neural network bottleneck features for language identification, in: *The Speaker and Language Recognition Workshop (Odyssey)*, 2014, pp. 299–304.
- [11] J. Mwema, M. Kimwele, S. Kimani, A simple review of biometric template protection schemes used in preventing adversary attacks on biometric fingerprint templates, *Int. J. Comput. Trends Technol.* 20 (1) (2015) 12–18.
- [12] A.K. Jain, K. Nandakumar, Biometric authentication: system security and user privacy, *IEEE Comput.* 45 (11) (2012) 87–92.
- [13] A.K. Jain, K. Nandakumar, A. Nagar, Biometric template security, *EURASIP J. Adv. Sig. Proc.* 2008 (2008).
- [14] J. Mwema, S. Kimani, M. Kimwele, A study of approaches and measures aimed at securing biometric fingerprint templates in verification and identification systems, *Int. J. Comput. Appl. Technol. Res.* 4 (2) (2015) 108–119.
- [15] Z. Jin, A.B.J. Teoh, B.M. Goi, Y.H. Tay, Biometric cryptosystems: a new biometric key binding and its implementation for fingerprint minutiae-based representation, *Pattern Recognit.* 56 (2016) 50–62.
- [16] C. Rathgeb, A. Uhl, A survey on biometric cryptosystems and cancelable biometrics, *EURASIP J. Inf. Secur.* 2011 (2011) 3.
- [17] A.K. Jain, A. Ross, U. Uludag, Biometric template security: challenges and solutions, in: *13th European Signal Processing Conference (EUSIPCO)*, 2005, pp. 1–4.
- [18] A.B.J. Teoh, W.K. Yip, S. Lee, Cancellable biometrics and annotations on bio-hash, *Pattern Recognit.* 41 (6) (2008) 2034–2044.
- [19] Y.L. Lai, Z. Jin, A.B.J. Teoh, B.M. Goi, W.S. Yap, T.Y. Chai, C. Rathgeb, Cancellable iris template generation based on indexing-first-one hashing, *Pattern Recognit.* 64 (2017) 105–117.
- [20] K. Nandakumar, A.K. Jain, Biometric template protection: bridging the performance gap between theory and practice, *IEEE Signal Process. Mag.* 32 (5) (2015) 88–100.
- [21] N.K. Ratha, S. Chikkerur, J.H. Connell, R.M. Bolle, Generating cancelable fingerprint templates, *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (4) (2007) 561–572.
- [22] R.D. Labati, V. Piuri, F. Scotti, Biometric privacy protection: guidelines and technologies, in: *International Conference on E-Business and Telecommunications (ICETE)*, 2012, pp. 3–19.
- [23] L.Y. Chong, A.J. Teoh, Probabilistic random projections and speaker verification, in: *International Conference on Biometrics (ICB)*, 2007, pp. 445–454.
- [24] Y. Wang, K.N. Plataniotis, An analysis of random projection for changeable and privacy-preserving biometric verification, *IEEE Trans. Syst., Man, Cybern., Part B* 40 (5) (2010) 1280–1293.
- [25] W. Xu, M. Cheng, Cancelable voiceprint template based on chaff-points-mixture method, in: *2008 International Conference on Computational Intelligence and Security (CIS)*, 2008, pp. 263–266.
- [26] A. Juels, M. Sudan, A fuzzy vault scheme, in: *Proceedings IEEE International Symposium on Information Theory*, 2002.
- [27] E. Chang, R. Shen, F.W. Teo, Finding the original point set hidden among chaff, in: *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, 2006, pp. 182–188.
- [28] R. Pandey, Y. Zhou, V. Govindaraju, Deep secure encoding: an application to face recognition, 2015, arXiv:1506.043401–10.
- [29] R. Pandey, Y. Zhou, B.U. Kota, V. Govindaraju, Deep secure encoding for face template protection, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 77–83.
- [30] C. Soutar, D. Roberge, A. Stoianov, R. Gilroy, V. Kumar, Biometric encryption, in: R.K. Nichols (Ed.), *ICSA Guide to Cryptography*, McGraw-Hill, 1999.
- [31] A. Juels, M. Wattenberg, A fuzzy commitment scheme, in: *Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS)*, 1999, pp. 28–36.
- [32] K. Inthavisas, D.P. Lopresti, Speech cryptographic key regeneration based on password, in: *2011 IEEE International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–7.
- [33] S. Billeb, C. Busch, H. Reininger, K. Kasper, C. Rathgeb, Biometric template protection for speaker recognition based on universal background models, *IET Biom.* 4 (2) (2015) 116–126.
- [34] M. Paulini, C. Rathgeb, A. Nautsch, H. Reichau, H. Reininger, C. Busch, Multi-bit allocation: preparing voice biometrics for template protection, in: *The Speaker and Language Recognition Workshop (Odyssey)*, 2016, pp. 291–296.
- [35] R.C. Johnson, W.J. Scheirer, T.E. Boulton, Secure voice-based authentication for mobile devices: vaulted voice verification, in: *Proc. SPIE 8712, Biometric and Surveillance Technology for Human and Activity Identification X*, 2013.
- [36] E. Chandra, K. Kanagalakshmi, Cancelable biometric template generation and protection schemes: a review, in: *International Conference on Electronics Computer Technology (ICECT)*, 2011, pp. 15–20.
- [37] H. Zhu, Q. He, Y. Li, A two-step hybrid approach for voiceprint-biometric template protection, in: *International Conference on Machine Learning and Cybernetics (ICMLC)*, 2012, pp. 560–565.
- [38] Y.C. Feng, P.C. Yuen, A.K. Jain, A hybrid approach for face template protection, in: *Proc. SPIE 6944, Biometric and Surveillance Technology for Human and Activity Identification V*, 2008.
- [39] X. Wang, H. Yu, How to break MD5 and other hash functions, in: *Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2005, pp. 19–35.
- [40] J. Yagnik, D. Strelow, D.A. Ross, R. Lin, The power of comparative reasoning, in: *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2431–2438.
- [41] T. Dean, M.A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, J. Yagnik, Fast, accurate detection of 100,000 object classes on a single machine, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1814–1821.
- [42] S.J.D. Prince, J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: *IEEE Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.
- [43] D. Snyder, D. Garcia-Romero, D. Povey, Time delay deep neural network-based universal background models for speaker recognition, in: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 92–97.
- [44] C. Cieri, D. Miller, K. Walker, The fisher corpus: a resource for the next generations of speech-to-text, in: *International Conference on Language Resources and Evaluation (LREC)*, 2004, pp. 69–71.
- [45] A.K. Jain, K. Nandakumar, A. Ross, 50 years of biometric research: accomplishments, challenges and opportunities, *Pattern Recognit. Lett.* 79 (1) (2016) 80–105.
- [46] W.J. Scheirer, T.E. Boulton, Cracking fuzzy vaults and biometric encryption, in: *Biometrics Symposium*, 2007, pp. 1–6.
- [47] S. Ferrante, M. Saltalamacchia, The coupon collector's problem, *Mater. Math.* 2014 (2) (2014) 1–35.
- [48] M. Matsumoto, T. Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator, *ACM Trans. Model. Comput. Simul.* 8 (1) (1998) 3–30.

Kong-Yik Chee obtained his BSc degree (Hons) in Actuarial Science from Universiti Tunku Abdul Rahman (UTAR), Malaysia in 2015. Currently, he is pursuing M.Eng.Sc degree at UTAR. His research interests include biometrics security, particularly in voice template protection.

Zhe Jin obtained his BIT (Hons) in Software Engineering, MSc (I.T.) from Multimedia University (MMU), Malaysia in 2007 and 2011 respectively, and PhD degree in Engineering from University Tunku Abdul Rahman (UTAR), Malaysia in 2016. He is now a lecturer in School of Information Technology, Monash University Malaysia. His research interest is biometrics security, particularly in fingerprint template protection.

Danwei Cai received his B.S. degree in software engineering from Sun Yat-Sen University, China, in 2016. He is currently a master student at Sun Yat-Sen University.

Ming Li received his B.S. degree in communication engineering from Nanjing University, China, in 2005 and his M.S. degree in signal processing from the Institute of Acoustics, Chinese Academy of Sciences, in 2008. He joined the Signal Analysis and Interpretation Laboratory (SAIL) at USC on a Provost fellowship in 2008 and received his Ph.D. in Electrical Engineering in May 2013. He is currently an assistant professor at SYSU-CMU Joint Institute of Engineering, an associate professor at school of electronics and information technology at Sun Yat-Sen University. His research interests are in the areas of speech recognition, multimodal signal processing, multimodal human state recognition, speaker verification, language identification, multimodal biometrics, affective computing with applications to behavioural informatics notably in health and security. He has published more than 70 papers and served as scientific committee members and reviewers for multiple conferences and journals. Works co-authored with his colleagues have won awards at Body Computing Slam Contest 2009, IEEE DCOSS 2009, Interspeech2011-Speaker State Challenge, Interspeech2012-Speaker Trait Challenge, and ISCSLP 2014 best paper award. He received the IBM faculty award at 2016.

Wun-She Yap holds the Chair in Centre for Cyber Security at the Universiti Tunku Abdul Rahman (UTAR). He is now an assistant professor in Lee Kong Chian Faculty of Engineering and Science, UTAR, Malaysia. He has been invited to serve as program committees of a number of peer-reviewed security conferences. His research interests include design and analysis of both asymmetric and symmetric cryptographic primitives.

Yen-Lung Lai obtained his BSc degree (Hons) in Physics from Universiti Tunku Abdul Rahman (UTAR), Malaysia in 2015. Currently, he is pursuing M.Eng.Sc degree at UTAR. His research interests include biometrics (iris and fingerprint), information security and machine learning.

Bok-Min Goi received his B.Eng degree from University of Malaya (UM) in 1998, and the M.Eng.Sc and PhD degrees from Multimedia University (MMU), Malaysia in 2002 and 2006, respectively. He is now the Dean and a professor in Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman (UTAR), Malaysia. He was also the General Chair for ProvSec 2010 and CANS 2010, Programme Chair for IEEE-STUDENT 2012 and Cryptology 2014, and the PC members for many crypto / security conferences. His research interests include cryptology, security protocols, information security, digital watermarking, computer networking and embedded systems design. He is a senior member of the IEEE and corporate member of the IEM, Malaysia.