Protecting the ECG Signal in Cloud-based User Identification System
A Dissimilarity Representation Approach

Diana Batista[1,2], Helena Aidos[1], Ana Fred[1,2], Joana Santos[3], Rui Cruz Ferreira[4] and Rui César das Neves[5]

[1]*Instituto de Telecomunicações, Lisbon, Portugal*

[2]*Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal*

[3]*Escola Superior de Saúde, Cruz Vermelha Portuguesa, Lisbon, Portugal*

[4]*Hospital de Santa Marta, Lisbon, Portugal*

[5]*CAST - Cons. e Apl. em Sistemas e Tecnologias, Lda, Lisbon, Portugal*

Keywords: ECG, Biometrics, Dissimilarity Representation, Dissimilarity Increments, Cloud-based System.

Abstract: Biometric recognition has become a popular approach for user identification and authentication. However, since in ECG-based biometrics users cannot change their authentication/identification signal (unlike in password-based methods), its applicability is seriously constrained for cloud-based systems: a hacker could potentially retrieve the stored ECG signal, eternally disabling ECG-based biometrics for the attacked user. To overcome such an issue, new methodologies must be devised to enable cloud-based authentication/identification systems without requiring the transmission and storage of the user's ECG signal on remote servers. In this paper we propose an ECG biometric approach that relies on non-linear irreversible dissimilarity spaces to encode (encrypt) the user's ECG. We show how to construct the dissimilarity space, and also evaluate the system's accuracy with the dimensionality of the dissimilarity space. We show that the proposed biometric system retains similar identification errors as an equivalent system relying on the Euclidean space, while the latter can potentially be broken by using triangulation techniques to uncover the users original ECG signal.

1 INTRODUCTION

In the last years, electrocardiographic (ECG) signals have demonstrated their potential in biometrics applications (Fratini et al., 2015; Islam and Alajlan, 2017; Hejazi et al., 2016), due to its inherent characteristics. The ECG has essential properties in the context of biometrics (Odinaka et al., 2012), including: universality (it is found in all living beings), performance (performs accurately for subsets of the population), measurability (can be measured with appropriate sensors), acceptability (the sensors can be designed in a non-intrusive way), and circumvention (it is not easily spoofed, since it does not depend on any external body traits). Besides, the ECG provides intrinsic aliveness detection and is continuously available. These ECGs' properties allow the development of exciting applications, where continuous and non-intrusive authentication are demanding factors, such as electronic trading platforms, where high-security, continuous authentication is essential.

Nowadays, an enormous amount of sensitive data has been generated, containing personal and confidential information about a subject (e.g., financial status or medical records). Thus, the privacy of an individual may be compromised with the release of such sensitive information, e.g., to cloud servers. Those are susceptible to hacker attacks (see the example biometric application in figure 1), and the sensitive information released and sold to third parties. All over the web, news can be found reporting examples of hacker attacks on servers with sensitive and confidential user information[1], despite the multiple security levels that are usually employed at the communication networks and cloud systems. Hence, it is crucial to design privacy-preserving techniques to ensure the confidentiality of the users data even after security breaches, especially when dealing with sensitive data that can not be changed or replaced (e.g., the ECG signal). Data privacy-preserving techniques tend to transform the data by distortion, approximation, or even by sup-

[1]https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html
http://www.telegraph.co.uk/technology/2017/02/01/hackers-steal-25-million-playstation-xbox-players-details-major/

Batista, D., Aidos, H., Fred, A., Santos, J., Ferreira, R. and Neves, R.
Protecting the ECG Signal in Cloud-based User Identification System - A Dissimilarity Representation Approach.
DOI: 10.5220/0006723900780086
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 4: BIOSIGNALS*, pages 78-86
ISBN: 978-989-758-279-0

pression or aggregation, such that the data is not the same but a sort of approximation. However, this often leads to a deterioration of the quality of data mining results (Aggarwal and Yu, 2008). In particular, a commonly used privacy-preserving technique consists of adding noise to the data. Yet, adding a large amount of noise compromises the utility of the data, whereas a small amount allows for an easy estimation of the original signal.

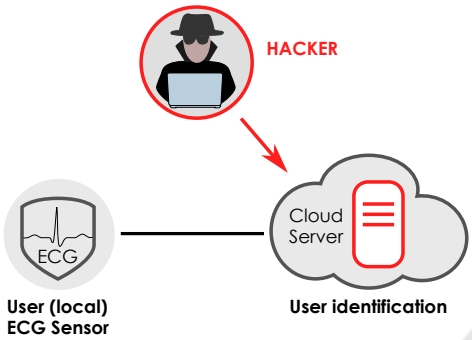

Figure 1: Cloud-based biometric system. If the users ECG signal is stored on the server, an attacker could potentially retrieve it, permanently disabling ECG-based biometric systems for the affected users.

Another privacy preserving technique consists of transforming the data to a new space, such as by describing sensitive data through a dissimilarity representation (Marques et al., 2015). Dissimilarity measures can be used to describe objects, by comparing pairs of objects, and, consequently, building representations of data that preserve the information therein. In this work we follow such a concept and adopt a dissimilarity representation in order to build a cloud-based biometric system that avoids storing, or otherwise transmitting, the users ECG signal. For such purpose, the system uses a public key (collection of prototypes) to locally encrypt the users ECG signal through a dissimilarity representation, before transmitting it to the server (where biometry is actually performed). Since different prototypes lead to different space representations, different public keys may be generated at any time, increasing the protection of sensitive information.

From different dissimilarity representations, we avoid the usage of the Euclidean distance to compare objects, as such a metric allows the uncovering of the original data through triangulation approaches (as in GSM navigation or surveillance applications). In contrast, we rely on a nonlinear second-order dissimilarity measure to build the dissimilarity representation and therefore obtain the encrypted signal.

The main contributions of this paper are:

- A novel remote biometric system for user identification that prevents hackers to obtain sensitive user's information. This is achieved by only storing a transformed (non-invertible) key in the cloud server, and not the user's original ECG signal.
- A method for the generation of public keys, achieved through a clustering algorithm and using a reference set of ECGs to produce it.
- We show that this key can be easily changed to ensure the privacy of the data, by changing the parameters of the algorithm, the algorithm itself, the reference set of ECGs or by a simple permutation of the current key.
- We analyse the accuracy of the proposed scheme as well as the required size of the public key in order to create the cloud-based biometric system.

The remainder of this paper is organized as follows: Section 2 presents the proposed remote ECG biometric system relying on a cloud server. Section 3 presents the concepts for the dissimilarity representation of the signals. Sections 4 describes the dataset used in the experiments, while section 5 presents the experimental setup and results. Conclusions are drawn in section 6.

2 ECG-BASED BIOMETRY

Despite the multiple security levels that are usually employed at the communication networks and cloud and remote systems, the storing or transmission of raw user's data may still compromise the safety of many modern systems. This is particularly problematic as hackers shift their modus operandi to specifically target administrators accounts, hence getting access to users' accounts and passwords. In the case of ECG-based biometric systems, this is especially distressing as a user's ECG signal cannot be modified. Hence, once a hacker acquires the user's ECG signal, he is permanently able to gain access to any ECG-based biometric system. To avoid such a problem, a new methodology is herein proposed that is built upon the concept of privacy-preserving transformations for sensitive user data. To create such transformations we rely on an non-invertible data transformation technique, using a dissimilarity data representation between user's data and a public key, which can be freely transmitted or otherwise stored on a cloud server, and that may change at any time (e.g., in the event of a hacker attack to the server).

The proposed remote biometric system works as depicted in Figure 2, and comprises two phases: *user*

enrollment, where a user ECG signal is recorded for later identification; and *user identification*, where the system matches the observed ECG signal against those of enrolled users. The following subsections describe how each of these steps work, namely enrollment (subsection 2.1), identification (subsection 2.2) and public-key generation (subsection 2.3).

2.1 User Enrollment

A user can enroll in the system by using an off-the-person sensor to acquire its ECG signal, e.g., BITalino (Alves et al., 2013). At that point, the local system requests the public key to the cloud server so it can locally encrypt the user's ECG. Hence, the real signal is never sent to the cloud server, but only an encrypted version of it, which is used whenever an identification query is requested.

The local encryption of the ECG signal of the user is made by representing its acquired heartbeats in a dissimilarity representation. This kind of representation is an attractive way to preserve the privacy of sensitive data since it is non-invertible. In this system, the dissimilarity representation is obtained by computing a dissimilarity measure between the enrolled heartbeats of the user and a public key (generated as described in subsection 2.3) received from the cloud server. After that, instead of directly storing the enrolled heartbeats, the dissimilarity representation is sent to the cloud server and stored until the subject needs to test his/her identity.

2.2 User Identification

After a user enrolls into the biometric system, such data can be used for identification purposes. For that, the user must acquire a new ECG signal from the local sensor and, using the received public key, generate the encrypted signal. As in the previous case, this encryption is performed by computing a dissimilarity representation between the new acquisition and the public key.

The newly encrypted signal is then sent to the remote server, where a proper classification algorithm will try to match it with the encrypted ECG signal that was previously stored on the server during enrollment. It should be noticed that this identification does not require the ECG to be decrypted, since it is performed over the same dissimilarity representations as used in the enrollment (and stored in the cloud server). Afterwards, the server returns the identification results to the local system.

The proposed methodology has several advantages over traditional solutions, namely because sensitive data is never transmitted nor stored in a cloud server, which can potentially be attacked by hackers. The only information stored in the cloud server is the public key and the encrypted user enrollment data. If an attack occurs, the public key can always be modified, therefore resulting in the encoding of the users data in a different dissimilarity space, which ensures the privacy of this sensitive data. Finally, new classifiers can be developed and updated directly on the server, which means that each user does not need to be concerned in updating its local system, ensuring the reliability of the entire system.

2.3 Public-key Generation

Naturally, the generation of the public key represents a critical step, since it must not contain any direct information regarding any of the enrolled users, but must still ensure that an accurate identification is attained. In other words, while the dissimilarity space cannot be constructed using the users ECG signal, it must still contain enough information about the morphology of an ECG to ensure a proper operation.

With this goal in mind, the public key is comprised of a set of carefully selected prototypes, which are obtained from a reference ECG database (i.e., an independent dataset composed by several heartbeats from different subjects (but not from any of the users), or derived synthetically). To achieve this, multiple approaches can be adopted, such as by applying a clustering algorithm over the set of heartbeats from the reference database (e.g., k-means), or by devising other prototype selection methods (García et al., 2012). As a consequence, the generation of a new public key can be performed by simply devising new data clusterings (e.g., running k-means with the same or with different k values) while using different initialization parameters. In this context, it should be highlighted that a simple random permutation of the public key (i.e., of the order of the selected set of prototypes) leads to a different space representation, therefore enabling the system to remain resilient after a hacker attack (see also section 5).

3 DISSIMILARITY REPRESENTATION

To build the dissimilarity representation used for data encryption, let us assume that we have acquired the ECG signal from a set of N subjects, $\mathcal{S} = \{S_i\}_{i=1}^{N}$, resulting in n_i heartbeats for subject S_i. This means that we have a set of heartbeats $\mathcal{H} = \{\mathbf{h}_i\}_{i=1}^{M}$, such

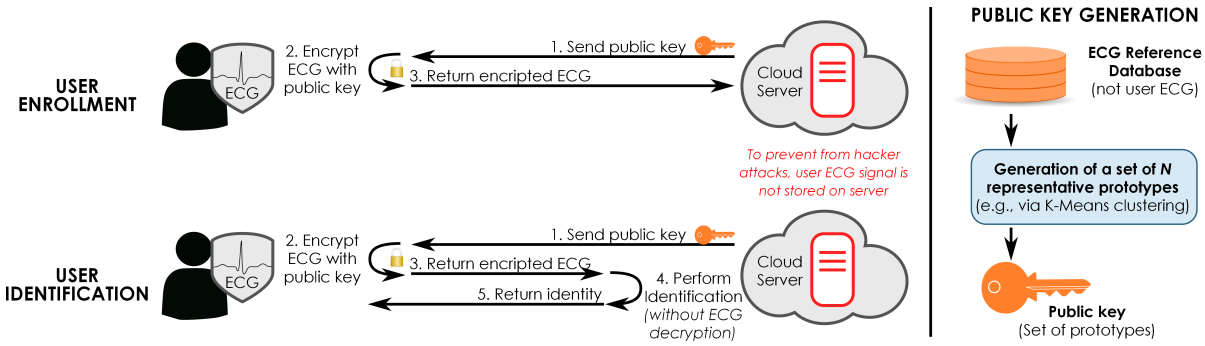

Figure 2: Proposed remote biometric system relying on a cloud server and encrypted ECG signals.

that $\sum_{i=1}^{N} n_i = M$, with n_i the number of heartbeats for subject S_i.

Let $\mathcal{P} = \{\mathbf{h}_i^p\}_{i=1}^{T}$ be the set of heartbeats representing the selected set of prototypes (public key in figure 2), such that $\text{card}(\mathcal{P}) \leq \text{card}(\mathcal{H})$. A dissimilarity space (Pekalska and Duin, 2005) is defined as the data-dependent mapping $D(\cdot,\mathcal{P}) : \mathcal{H} \rightarrow \mathbb{R}^T$. Accordingly, each heartbeat $\mathbf{h}_i$ is described by a T-dimensional dissimilarity vector

$$D(\mathbf{h}_i,\mathcal{P}) = [d(\mathbf{h}_i,\mathbf{h}_1^p)\dots d(\mathbf{h}_i,\mathbf{h}_T^p)], \qquad (1)$$

where $d(\cdot,\cdot)$ represents a dissimilarity measure. Thus, the *dissimilarity space* is characterized by the $M \times T$ dissimilarity matrix D, where $D(\mathbf{h}_i,\mathcal{P})$ is the i-th row of D.

Three different dissimilarity representations are addressed in this manuscript, two first-order spaces, namely the Euclidean and Cosine spaces, and one second-order space, Dinc, as detailed next. In particular, the usage of a second-order dissimilarity measure provides interesting security improvements, since traditional triangulation approaches cannot be used to capture the original signal. Accordingly, as long as no substantial accuracy degradation is observed in the identification process of a subject (evaluated in section 5), second-order spaces are preferable and should be used instead.

3.1 First-order Spaces

The Euclidean space is defined by replacing the dissimilarity measure $d(\cdot,\cdot)$ in (1) by the Euclidean distance, as follows:

$$d_{\text{Euclidean}}(\mathbf{h}_i,\mathbf{h}_j^p) = \left(\sum_{l=1}^{d}(h_{il}-h_{jl}^p)^2\right)^{1/2}. \qquad (2)$$

An alternative solution for the construction of a first-order space consists in using the cosine dissimilarity, as follows:

$$d_{\text{Cosine}}(\mathbf{h}_i,\mathbf{h}_j^p) = 1 - \frac{\mathbf{h}_i \cdot \mathbf{h}_j^p}{\|\mathbf{h}_i\|\|\mathbf{h}_j^p\|}. \qquad (3)$$

3.2 Second-order Space

The dissimilarity increments (Fred and Leitão, 2003) is a second-order dissimilarity measure that can be considered for constructing a dissimilarity space (Aidos and Fred, 2015). This measure is built upon the concept of triplets of points (heartbeats), $(\mathbf{h}_i,\mathbf{h}_j,\mathbf{h}_k)$, obtained as follows: $\mathbf{h}_j$ is the nearest neighbor of $\mathbf{h}_i$ and $\mathbf{h}_k$ is the nearest neighbor of $\mathbf{h}_j$, but different from $\mathbf{h}_i$. Therefore, the *dissimilarity increments* between neighboring heartbeats is defined as

$$d_{\text{inc}}(\mathbf{h}_i,\mathbf{h}_j,\mathbf{h}_k) = \left|d(\mathbf{h}_i,\mathbf{h}_j) - d(\mathbf{h}_j,\mathbf{h}_k)\right|, \qquad (4)$$

where $d(\cdot,\cdot)$ represents the pairwise dissimilarity between two heartbeats, which can be obtained by applying any first-order dissimilarity measure (e.g., the Euclidean distance).

Dinc space: Based on the definition of dissimilarity increment, it is possible to build a dissimilarity space, where each sample of this space is described by a T-dimensional dissimilarity vector $D(\mathbf{h}_i,\mathcal{P})$. $D(\mathbf{h}_i,\mathcal{P})$ is computed by evaluating the dissimilarity increment between each heartbeat $\mathbf{h}_i$ and the public key, $\{\mathbf{h}_1^p,\cdots,\mathbf{h}_T^p\} \in \mathcal{P}$. For the dissimilarity increments space (or Dinc space), each new prototype $\mathbf{h}_j^*$ is constructed by considering the edge between an element of the public key $\mathbf{h}_j^p$ and its nearest neighbor $\mathbf{h}_{\mathbf{h}_j^p}$ in the heartbeats set $\mathcal{H}$ (obtained from the reference database). Therefore, the distance between any heartbeat $\mathbf{h}_i$ from the dataset $\mathcal{H}$ and the prototype $\mathbf{h}_j^*$ is given by

$$d(\mathbf{h}_i,\mathbf{h}_j^*) = \min\{d(\mathbf{h}_i,\mathbf{h}_j^p), d(\mathbf{h}_i,\mathbf{h}_{\mathbf{h}_j^p})\}, \qquad (5)$$

and the (i,j)-th element of the Dinc space is given by

$$D(\mathbf{h}_i,\mathbf{h}_j^*) = |d(\mathbf{h}_i,\mathbf{h}_j^*) - d(\mathbf{h}_j^*)|. \qquad (6)$$

This dissimilarity measure ensures that the matrix D is non-negative (from (6)) and asymmetric (Aidos and Fred, 2015).

4 DATASET

The biometric system will be tested in a database provided by a local hospital, Hospital de Santa Marta, that has been previously validated regarding biometric performance (Carreiras et al., 2014).

The used ECG records were acquired during normal hospital operation, encompassing scheduled appointments, emergency cases, and bedridden patients. For this study, we decided to focus on signals originating from individuals with normal rhythms. Consequently, each record had to be labeled by a specialist. All signals were acquired using Philips PageWriter Trim III devices, following the standard 12-lead placement, with a sampling rate of $500Hz$ and 16-bit resolution. Each record has a duration of 10 seconds. To date, we have 955 healthy subjects, whose real identities are obfuscated at the hospital.

4.1 Data Pre-processing

The raw ECG signals must be pre-processed to allow the feature extraction methods to capture the morphology of the signal and not the noise. Thus, three steps are considered in this work to obtain a set of heartbeats (see figure 3).

The filtering of the signal is a crucial step due to the presence of several noise sources during measurement, e.g., power line interference, electrode contact loss, baseline drift due to respiration, and motion artifacts (Friesen et al., 1990). Here, two median filters are applied to remove the baseline, with window sizes of 0.2 and 0.6 seconds. Afterwards, a finite impulse response low-pass filter with cut-off frequency of $40Hz$ is used to deal with high-frequency noise.

The identification of the R peak is needed to segment the ECG signal in heartbeats. Since the focus of this paper is not on algorithms for R peak detection (which have been intensively studied in prior works, e.g., (Canento et al., 2013; Friesen et al., 1990)), in this manuscript the annotations previously made by a specialist are used. After that, the segments are constructed by merely taking the ECG signal in the window [-200ms ; 400ms] in relation to each one of the identified R-peaks, leading to segments with a fixed length of 600ms.

Finally, abnormal heartbeats were removed using the DMEAN method proposed by (Lourenço et al., 2013), with parameters $a = 0.5$, $b = 1.5$ and using the Euclidean distance to compare heartbeats. From this procedure n_i heartbeats for a subject i ($i = 1,\ldots,N$) are obtained, resulting in a set of heartbeats $\mathcal{H} = \{\mathbf{h}_i\}_{i=1}^{M}$, such that $\sum_{i=1}^{N} n_i = M$.

5 EXPERIMENTAL RESULTS

5.1 Experimental Setup

Following the proposed biometric system presented in figure 2, it is crucial to define the prototypes generation process, the dissimilarity representation used in this manuscript to encrypt the signals, and the classifier stored in the cloud server. Figure 4 presents the methodology adopted for the experiments herein. In the remainder of the paper, it is assumed that the biometric system only uses sensors that acquired ECG signals from lead I (Barold, 2003).

From the dataset described in section 4, and after pre-processing the signals as described in section 4.1, a set of heartbeats is obtained and split into two sets. The first set, called *reference set*, is composed by the heartbeats of 50% of randomly chosen subjects from the original dataset. This reference set is used to produce prototypes, generating the public key that encodes the heartbeats from any given subject. The remaining 50% of subjects are the ones used to train and test the proposed biometric system. Hence, this set of heartbeats is then further split in 80% for training the model and 20% for validation.

We are assuming that a close world setup is in place: all users have gone through the enrollment phase and the system always returns an identity. Therefore, an error is accounted when the returned identity is incorrect.

To evaluate the whole system a nested cross-validation is performed, where the creation of the reference set is repeated ten times, and for each one of these reference sets, the training and validation of the model is run another ten times. The results presented here are average error rates.

Public key generation (selection of prototypes): To generate the public key, two clustering algorithms were applied to the reference set: k-means and k-medoids. The resulting centroids (or medoids) are then set as the desired prototypes, i.e., the public key being stored in the cloud server. The use of k-means to generate the set of prototypes from the reference set provides a generic template representing different morphologies of heartbeats. Consequently, it might be a good choice for a public key, since no information about a specific subject is disclosed. To analyze the influence of the size of the key in the identification results, the value of k was chosen from the set $\{2^3, 2^4, \cdots, 2^{10}\}$.

Dissimilarity representation: Each ECG signal is transformed by computing the dissimilarity represen-

Figure 3: Pre-processing steps of a raw ECG signal.

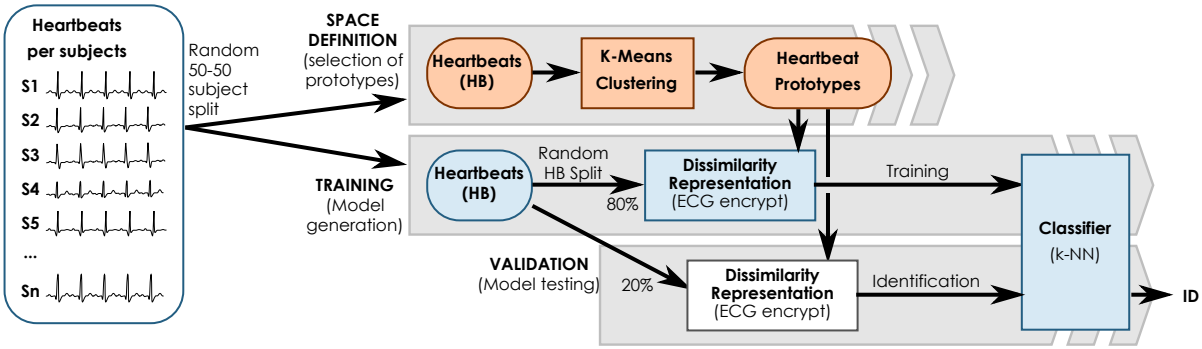

Figure 4: Experimental setup.

tation between each of its heartbeats and the public key. Three types of dissimilarity measures were applied: the Euclidean distance, the cosine distance, and the dissimilarity increments. The first two representations are based on a first-order dissimilarity measure whereas the second one is based on a second-order dissimilarity measure, providing a more difficult way to trace back the original ECG signal of a user.

Classification: A classification algorithm must be used to perform user identification on the cloud server, namely to compare the encrypted key (dissimilarity-represented heartbeat set) stored on the server during the enrollment phase, and the key used for querying user identification. In this paper, a k-nearest neighbor is considered, by setting $k = 3$ and the cosine distance, since the latter shows to provide better results than the Euclidean distance.

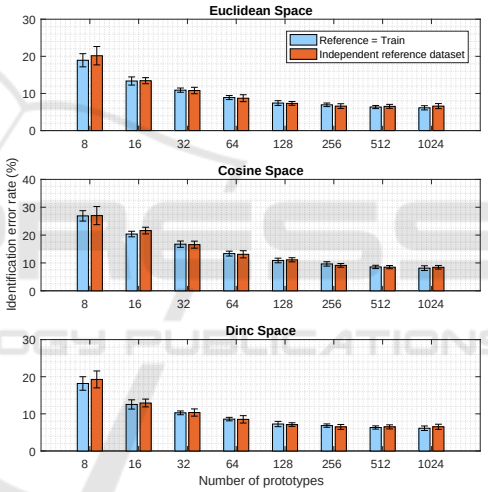

Figure 5: Evaluation of the error rates when using the original training dataset or an independent reference dataset for prototype selection with the k-means algorithm.

5.2 Results

Figure 5 presents a study of the number of prototypes that are required to generate a suitable public key, for each dissimilarity representation considered in this paper (Euclidean, Cosine, and Dinc). Moreover, it also shows the difference (in error rates) of using an independent database to generate the set of prototypes.

As can be seen, all spaces present a similar behavior: when using a reduced set of prototypes (e.g., a public key with length eight), the error rates are quite high; notwithstanding, the error significantly decreases as the number of considered prototypes increases, with the accuracy becoming stable for a large number of prototypes. Furthermore, it is quite visible from this set of experiments that the prototypes obtained from an independent set of subjects (the reference set) do not degrade the system performance. In fact, especially for larger number of prototypes, the identification error is the same.

Figure 6 presents the comparison between the three dissimilarity spaces when the reference set is used to obtain the public key. We can notice that all three spaces achieve the minimum error rate between 256 and 512 prototypes. It is not therefore useful to generate a public key larger than that, since it would only increase memory and computation requirements.

In what respects the comparison between dissimilarity spaces, it is clear that using the Cosine space

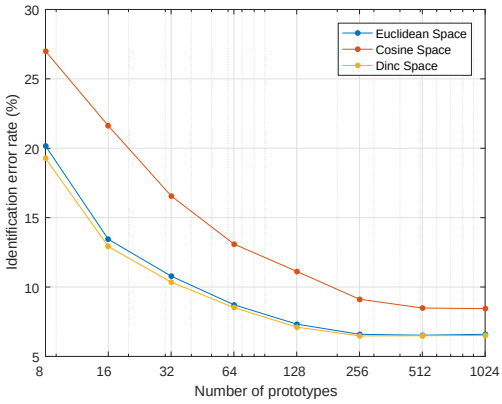

Figure 6: Evaluation of the error rates in the three dissimilarity spaces when using an independent reference dataset for prototype selection.

results in the worst error rates for any choice of number of prototypes. Euclidean and Dinc spaces achieve similar performances. However, the Dinc space has an advantage over the Euclidean dissimilarity space: since it is based on an asymmetric measure, it is equivalent to using a non-invertible transformation of the data. If the cloud server is hacked, and the public key revealed, the data encrypted with the Euclidean dissimilarity measure can potentially be broken by using triangulation techniques, while data encrypted with the dissimilarity increments measure is more difficult to decrypt.

A few modifications to the experimental setup of figure 4 can be envisioned. We explore here two of these possibilities: changing the clustering algorithm used to construct the prototypes, and altering the number of subjects used to create the public key.

Besides the *k*-means algorithm, an obvious choice to cluster the heartbeats is the *k*-medoids algorithm. Figure 7 shows the evaluation of the error rates for the three spaces when using the training dataset and an independent dataset for prototype selection.

If we compare the results obtained here with the ones from figure 5, it is clear that the same conclusions can be drawn in what regards the number of prototypes, the differences between datasets, and the performance of the three spaces. It is worth noting that there is a slight improvement overall when using *k*-medoids as opposed to *k*-means to generate the public key.

Unlike the *k*-means algorithm, when clustering with *k*-medoids, the clusters' centroids are actual samples. This has an important consequence here. The public key now contains prototypes that are not generic, they are real heartbeats from specific subjects. In order to prevent ECG data from users enrolled in the system to be gathered from an unwanted

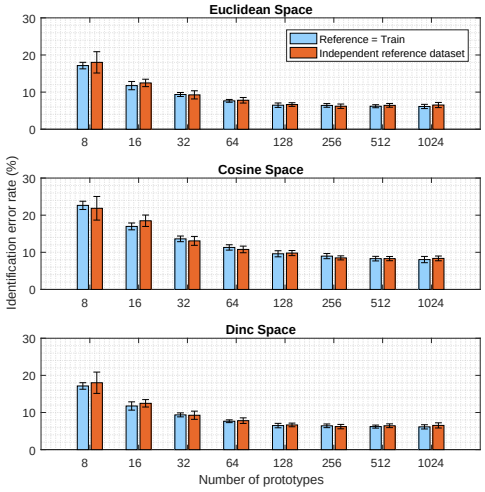

Figure 7: Evaluation of the error rates when using the original training dataset or an independent reference dataset for prototype selection with the *k*-medoids algorithm.

third party, it is therefore advisable to use an independent reference dataset to select the prototypes. Since the identification error rates are very similar for the two datasets, this should not degrade the system performance.

Figure 8 shows the influence on the identification error rate of altering the number of subjects used to construct the public key. An independent reference dataset is used with the number of subjects varying from 10 to 100% of the initial random 50-50 subject split. For both *k*-means and *k*-medoids, *k* was set to 256 and the Dinc space was used.

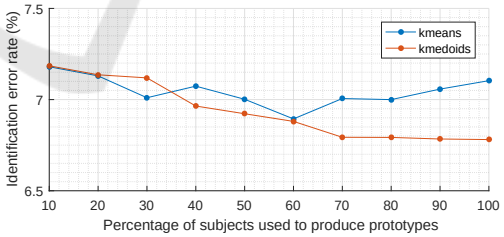

Figure 8: Evaluation of the influence of the number of subjects used to construct the prototypes.

As can be observed, the identification error rates are not significantly affected by the variation of the number of subjects used to generate the public key. In fact, when varying the number of subjects from 10% (46 subjects) to 100% (463 subjects), the error rate stays relatively close to 7% for both clustering algorithms (with standard deviations in the 0.5 - 1% range). This observation can be important when envisioning a real-world application of the proposed system: it is not necessarily important to have a massive

amount of data to generate the public key. As long as the selected prototypes are able to capture the relevant characteristics of the heartbeats, the biometric system should be able to maintain its performance.

Another interesting perspective of the proposed cloud-based biometric system is the change of the public key, which can be made by applying another clustering algorithm, or changing the number of prototypes in the clustering algorithms used in this paper. In this case, it should be highlighted that a simple permutation of the prototypes after training the classifier shows an error rate higher than 95% for all three spaces. This means that, when the system is attacked, a mere permutation of the public key is able to significantly change the identification process, whereas a more substantial change should make sure that an hacker obtaining the remotely-stored user key can no longer be identified by the system.

6 CONCLUSIONS

This paper proposes a new ECG-based biometric approach for cloud systems, which locally encrypts the ECG signal through a dissimilarity representation. Such representation is obtained by applying a non-linear and non-invertible transformation, the dissimilarity increments, between the public key, stored on the server, and the real-time acquired ECG signal. This provides significant advantages, as it does not require the users' ECG signals to be stored on the server, but only a transformed version of it. In traditional approaches a hacker might be able to retrieve the original ECG signal and thus forever compromise the usage of ECG biometrics for that user. However, in the proposed system, the hacker will only capture the public key and a transformed version of the signal. Accordingly, under such circumstances, a new public key can be easily generated by simply selecting a new set of prototypes and by asking the user to perform a new enrollment.

The experimental results show that the proposed methodology provides no significant degradation in the identification error rates, especially when the selected prototypes are generated from a reference dataset, independent of the users data, i.e., it is composed by the ECG signals of independent (unidentified) users.

ACKNOWLEDGEMENTS

This work was supported by the Portuguese Foundation for Science and Technology, under scholarship number SFRH/BPD/103127/2014 and grant number PTDC/EEI-SII/7092/2014.

REFERENCES

Aggarwal, C. C. and Yu, P. S. (2008). *Privacy-preserving data mining: models and algorithms*, volume 34 of *Advances in Database Systems*. Springer.

Aidos, H. and Fred, A. (2015). A novel data representation based on dissimilarity increments. In Feragen, A., Loog, M., and Pelillo, M., editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9370, pages 1–14. Springer, Copenhagen, Denmark.

Alves, A. P., Silva, H., Lourenço, A., and Fred, A. (2013). BITalino: A Biosignal Acquisition System based on the Arduino. In *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS)*, pages 261–264.

Barold, S. S. (2003). Willem einthoven and the birth of clinical electrocardiography a hundred years ago. *Cardiac electrophysiology review*, 7(1):99–104.

Canento, F., Lourenço, A., Silva, H., and Fred, A. (2013). Review and Comparison of Real Time Electrocardiogram Segmentation Algorithms for Biometric Applications. In *Proceedings of the International Conference on Health Informatics (HEALTHINF)*.

Carreiras, C., Lourenço, A., Fred, A., and Ferreira, R. (2014). ECG Signals for Biometric Applications - Are we there yet? In *Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics*, pages 765–772, Vienna, Austria. SCITEPRESS - Science and and Technology Publications.

Fratini, A., Sansone, M., Bifulco, P., and Cesarelli, M. (2015). Individual identification via electrocardiogram analysis. *Biomedical engineering online*, 14(1):78.

Fred, A. L. N. and Leitão, J. M. N. (2003). A new cluster isolation criterion based on dissimilarity increments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):944–958.

Friesen, G. M., Jannett, T. C., Jadallah, M. A., Yates, S. L., Quint, S. R., and Nagle, H. T. (1990). A Comparison of the Noise Sensitivity of Nine QRS Detection Algorithms. *IEEE Transactions on Biomedical Engineering*, 37(1):85–98.

García, S., Derrac, J., Cano, J. R., and Herrera, F. (2012). Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):417–435.

Hejazi, M., Al-Haddad, S., Singh, Y. P., Hashim, S. J., and Aziz, A. F. A. (2016). ECG biometric authentication based on non-fiducial approach using kernel methods. *Digital Signal Processing*, 52:72–86.

Islam, M. S. and Alajlan, N. (2017). Biometric template extraction from a heartbeat signal captured from fingers. *Multimedia Tools and Applications*, pages 1–25.

Lourenço, A., Silva, H., Carreiras, C., and Fred, A. (2013). Outlier Detection in Non-intrusive ECG Biometric System. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7950, pages 43–52.

Marques, F., Carreiras, C., Lourenço, A., Fred, A., and Ferreira, R. (2015). ECG Biometrics Using a Dissimilarity Space Representation. In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pages 350–359.

Odinaka, I., Lai, P. H., Kaplan, A. D., O'Sullivan, J. A., Sirevaag, E. J., and Rohrbaugh, J. W. (2012). ECG biometric recognition: A comparative analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824.

Pekalska, E. and Duin, R. P. W. (2005). *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific Pub Co Inc.