Capacity and Examples of Template-Protecting Biometric Authentication Systems

Pim Tuyls and Jasper Goseling

Philips Research,
Prof. Holstlaan 4,
5656 AA Eindhoven,
The Netherlands,
pim.tuyls@philips.com, j.goseling@ieee.org

Abstract. In this paper, we formulate the requirements for privacy protecting biometric authentication systems. The secrecy capacity C_s is investigated for the discrete and the continuous case. We present, furthermore, a general algorithm that meets the requirements and achieves C_s as well as C_{id} (the identification capacity). Finally, we present some practical constructions of the general algorithm and analyze their properties.

1 Introduction

The increasing demand for more reliable and convenient security systems generates a renewed interest in human identification based on biometric identifiers such as fingerprints, iris, voice and gait. Since biometrics cannot be lost or forgotten like e.g. computer passwords, biometrics have the potential to offer higher security and more convenience for the users.

A common approach to biometric authentication is to capture the biometric templates of all users during the *enrollment phase* and to store the templates in a reference database. During the *authentication phase* new measurements are matched against the database information.

The fact that biometric templates are stored in a database introduces a number of security and privacy risks. We identify the following threats:

1. Impersonation. An attacker steals templates from a database and constructs artificial biometrics that pass authentication.
2. Irrevokability. Once compromised, biometrics cannot be updated, reissued or destroyed.
3. Exposure of sensitive personal information.

The first threat was recognized by several authors [1–3]. When an authentication system is used on a large scale, the reference database has to be made available to many different verifiers, who, in general, cannot be trusted. Especially in a networked environment, attacks on the database pose a serious threat. It was explicitly shown by Matsumoto et al. [4] that by using information stolen from a database, artificial biometrics can be constructed to impersonate people.

Construction of artificial fingerprints is possible even if only part of the template is available. Hill [5] showed that if only minutiae templates of a fingerprint are available, it is still possible to successfully construct artificial fingers that pass authentication. The second threat was first addressed by Schneier [6]. The problem is concisely paraphrased by: *"Theft of biometrics is theft of identity."* The third threat is caused by the fact that biometrics contain sensitive personal information. It is shown in [7]that fingerprints contain genetic information. From [8] on the other hand it follows that retina scans reflect information about diseases like diabetes and strokes.

We observe that a biometric authentication system does not need to store the original biometric templates. In order to protect against the threats given above, other authentication architectures are possible as well. Examples of systems that use other architectures and achieve protection of templates are *private biometrics* [9], *fuzzy commitment* [10], *cancelable biometrics* [11], *fuzzy vault* [12], *quantizing secret extraction* [13] and *secret extraction from significant components* [14]. The systems proposed in [9, 10, 12–15] are all based on an architecture that uses *helper data*. In this paper we analyze this architecture and derive performance bounds. Moreover, we propose an algorithm that implements this architecture and achieves these bounds.

This paper is organized as follows. In Section 2, we introduce our model and give definitions. In Section 3, we identify the requirements for authentication systems that protect against the threats mentioned above. We introduce the *helper data architecture*. Finally, we explain in this section the relation between the protection of biometric templates and secret extraction from common randomness. In Section 4, we derive fundamental bounds for the helper data architecture. A general algorithm that implements this helper data architecture is given in Section 5. It is shown that this algorithm satisfies the requirements and meets the performance bounds. Additionally, we show that by using the helper data architecture for template protection, the maximum achievable performance of a biometric identification system is not decreased. In Section 6, some concrete examples of the general algorithm are discussed. These examples illustrate the relation between our work and [10, 13, 14].

2 Model and Definitions

2.1 Security Assumptions

An overview of the possible attack scenarios to a biometric authentication system is given in [1–3]. In this paper, we make the following security assumptions.

- Enrollment is performed at a *trusted* Certification Authority (CA). The CA enrolls all users by capturing their biometrics, performing additional processing and adding a protected form of the user data to a database.
- The database is vulnerable to attacks from the outside as well as from the inside (malicious verifier).

- During the authentication phase an attacker is able to present artificial biometrics at the sensor.
- All capturing and processing during authentication is tamper resistant, e.g. no information about biometrics can be obtained from the sensor.
- The communication channel between the sensor and the verification authority is assumed to be public and authenticated, i.e. the line can be eavesdropped by an attacker.

2.2 Biometrics

Biometric templates are processed measurement data, i.e. feature vectors. We model biometric templates as realizations of a random process. Biometrics of different individuals are independent realizations of a random process that is equal for all individuals. We assume that the processing of biometrics results in templates that can be described as a sequence of n independent identically distributed (i.i.d.) random variables with a known distribution P_X. The probability that the biometric sequence X^n of a certain individual equals x^n is defined by

$$\Pr\{X^n = x^n\} = \prod_{i=1}^{n} \mathrm{P_X}(x_i)\,, \tag{1}$$

where P_X is the probability distribution of each component, defined on an alphabet $\mathcal{X}$, which can be a discrete set or $\mathbb{R}$. [1]

Noisy measurements of biometrics are modeled as observations through a memoryless noisy channel. For a measurement Y^n of biometrics x^n we have

$$\Pr\{Y^n = y^n | X^n = x^n\} = \prod_{i=1}^{n} \mathrm{P_{Y|X}}(y_i|x_i)\,, \tag{2}$$

where $P_{Y|X}$ characterizes the memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. It is assumed that the enrollment measurements of the biometric templates are noise free.

2.3 Secret Extraction Codes (SECs)

In order to deal with noisy measurements, we introduce the notion of Secret Extraction Codes (SECs). Let $\mathcal{S}$ denote the set of secrets and let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output alphabets, respectively, of the channel representing the noisy measurements.

Definition 1 (Secret Extraction Code). *Let $n, \epsilon > 0$. An $(n, |\mathcal{S}|, \epsilon)$ Secret Extraction Code $\mathcal{C}$, defined on $\mathcal{X}^n \times \mathcal{Y}^n$, is an ordered set of pairs of encoding and decoding regions*

$$\mathcal{C} = \Big\{ (\mathcal{E}_i, \mathcal{D}_i) \,\Big|\, i = 1, 2, \ldots, |\mathcal{S}| \Big\}, \tag{3}$$

[1] For $\mathcal{X} = \mathbb{R}$ the sequence X^n is characterized by the probability density function $f_{X^n}(x^n) = \prod_i f_X(x_i)$.

where $\mathcal{E}_i \subseteq \mathcal{X}^n$ *and* $\mathcal{D}_i \subseteq \mathcal{Y}^n$*, such that*

$$\mathcal{E}_i \cap \mathcal{E}_j = \emptyset, \qquad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset, \qquad \bigcup_i \mathcal{D}_i = \mathcal{Y}^n, \tag{4}$$

for $i, j = 1, 2, \ldots, |\mathcal{S}|$*,* $i \neq j$ *and*

$$\mathrm{P}_{\mathrm{Y^n|X^n}}(\mathcal{D}_i | x_i^n) \geq 1 - \epsilon, \tag{5}$$

for all $x_i^n \in \mathcal{E}_i$ *and* $i = 1, 2, \ldots, |\mathcal{S}|$*.*

Note that a SEC provides an encoding-decoding scheme of a (possibly continuous) variable into a finite alphabet $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$ by discretization. We note that the condition in Eq. (4) guarantees that unambiguous encoding and decoding is possible and the condition of Eq. (5) implies a low False Rejection Rate (FRR).

Note that SECs are strongly related to geometric codes [16]. Furthermore, when the sets $\mathcal{C}_i$ have cardinality one, the SECs are normal error correcting codes.

3 Protection of Templates

3.1 Requirements

The requirements for an architecture that does not suffer from the threats mentioned in the introduction are:

1. The information that is stored in the database does not give sufficient information to make successful impersonation possible.
2. The information in the database provides the least possible information about the original biometrics, in particular it reveals no sensitive information.

Note that an architecture that meets those requirements, guarantees that the biometric cannot be compromised.

3.2 The Helper Data Architecture

The privacy protecting biometric authentication architecture that is proposed in [13,14] is inspired by the protection mechanism used for computer passwords. Passwords are stored in a computer system in a cryptographically hashed form. This makes it computationally infeasible to retrieve the password from the information stored in the database. The hash function is also applied to the user input that is given in the authentication phase and matching is based on the hashed values. This approach, however, cannot be used for the protection of biometric templates in a straightforward way, because the measurements in the authentication phase are inherently noisy. Since small differences at the input of one-way functions result in completely different outputs, the hashed versions of

the enrollment and the noisy authentication measurements will be different with high probability.

In order to combine biometric authentication with cryptographic techniques, we derive *helper data* during the enrollment phase. The helper data guarantees that a unique string can be derived from the biometrics of an individual during the authentication as well as during the enrollment phase. Since the helper data is stored in the database it has to be considered as public data. In order to prevent impersonation, we need to derive reference data from the biometric that is statistically independent of the helper data. In order to keep the reference data secret for somebody having access to the database, we store the reference data in hashed form. In this way impersonation becomes computationally infeasible.

A schematic representation of the architecture described in this section is presented in Fig. 1.

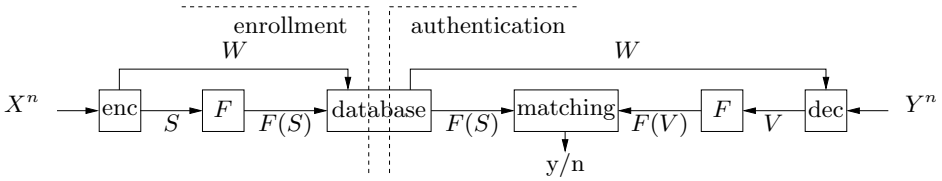

Fig. 1. The proposed authentication architecture. In the enrollment phase, the biometric template X^n is used for the derivation of a secret S and helper data W. A hashed version $F(S)$ of the secret and the helper data W are stored in a database. In the authentication phase a noisy version Y^n of the biometric template is captured. The helper data W is used to derive a secret V from Y^n. If $F(S) = F(V)$, the authentication is successful.

During the enrollment phase a secret S, belonging to an alphabet $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$, is extracted from a sequence X^n. In order to guarantee robustness to noise, the CA derives helper data W that will be used during the authentication phase to achieve noise robustness.

During the authentication phase, a noisy version Y^n of the enrollment sequence X^n is obtained. Using the helper data W, which is provided to the verifier, a secret $V \in \mathcal{S}$ is derived. The scheme is designed such that V equals S with high probability. Note that in contrast to usual biometric systems, we perform an exact match on $F(S)$ and $F(V)$.

The first requirement given in Section 3.1, is to prevent abuse of database information for impersonation. To this end the system is designed such that the mutual information between the helper data and the secret is sufficiently small and the secrets are uniformly distributed. Furthermore the set of secrets has to be sufficiently large to exclude an attack by exhaustive trial.

The helper data architecture was introduced as an architecture for verification, i.e. a situation in which an identity claim is verified. The helper data architecture can, however, also be used in an identification setting. In that case

a biometric measurement is matched against the database information of all enrolled users. In the remaining part of this paper, algorithms will be proposed in a verification setting. The extension to the identification setting is left implicit.

3.3 Relation with Secret Extraction from Common Randomness

There is a strong relation between the protection of biometric templates and *secret extraction from common randomness* [17,18]. The term common randomness is used for the situation that two parties possess sequences of correlated random variables. In the case of biometrics, the biometric is the source of common randomness. Another well-known example of this is quantum key exchange [19]. The secret extraction problem arises if the parties want to extract a common secret from the correlated data by communicating over a public channel. As the communication channel is public, the secret extraction protocol has to be designed such that no information about the secret is revealed to an eavesdropper.

Fig. 2 gives an alternative representation of the situation that was already visualized in Fig. 1. In the case of biometrics a secret S and helper data W

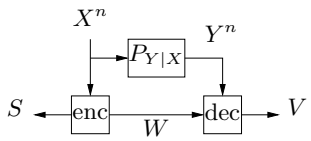

Fig. 2. The sequences X^n and Y^n are correlated random variables. The goal is to use helper data W to derive secrets S and V in such a way that W does not give much information about these secrets.

are derived from X^n during enrollment. During the authentication phase Y^n, a noisy version of X^n (and hence correlated with X^n) is obtained and a secret V is computed using the public helper data W. The helper data W is designed such that V equals S with very high probability and such that no information about S is revealed.

The main difference between secret extraction from biometrics and for instance quantum key exchange is that in the case of biometrics all helper data are be derived during enrollment. In quantum key exchange on the other hand, communications can only start after random variables are obtained by both parties. In general multiple rounds of communications are required.

4 Bounds: Secrecy and Identification Capacity

4.1 Secret Extraction

We express the size of the secrets in the rate $\mathrm{R_s}$. The maximum achievable rate is defined accordingly by the secrecy capacity $\mathrm{C_s}$.

Definition 2 (Secrecy Capacity). *The secrecy capacity* $\mathrm{C_s}$ *is the maximal rate* $\mathrm{R_s}$*, such that for all* $\epsilon > 0$*, there exist encoders and decoders that, for sufficiently large* n*, achieve*

$$\Pr\{V \neq S\} \leq \epsilon, \tag{6}$$

$$\mathrm{I}(W; S) \leq \epsilon, \tag{7}$$

$$\frac{1}{n}\mathrm{H}(S) = \frac{1}{n}\log|\mathcal{S}| \geq (\mathrm{R_s} - \epsilon). \tag{8}$$

Eq. (6) ensures correctness of the secret, Eq. (7) ensures secrecy with respect to eavesdropping of the communication line and Eq. (8) guarantees a uniform distribution of the secrets.

According to requirement 2 from Section 3.1, $\mathrm{I}(W; X^n)$ should be small. It was proven in [14], that in order to guarantee correctness and a large number of secrets, the helper data W should depend on the biometric template X^n. Hence $\mathrm{I}(W; X^n)$ cannot be zero. In Section 6 we show, however, that it is possible to keep $\mathrm{I}(W; X^n)$ small. More in particular, it will not be possible to derive from the helper data W a good (in the least squared sense) estimate of X^n. Finally, if the requirement of Eq. (7) is satisfied and the number of secrets is large, an impersonation attack based on artificial biometrics $\hat{X}^n$ (which is an estimate of X^n based on helper data W) is infeasible, hence satisfying requirement 1 of Section 3.1. We remark that in general $\mathrm{I}(F(S), W; X^n)$ is large in the strict information-theoretic sense. In the computational sense, however, it is infeasible to derive information about S from $F(S)$. Hence from a computational point of view $F(S)$ does not reveal information about X^n and the only information revealed about X^n is $\mathrm{I}(W; X^n)$.

The uncertainty expressed by $\mathrm{H}(S|W) = \mathrm{H}(S) - \mathrm{I}(W; S)$, defines a security parameter κ for impersonation. The number $2^{\kappa-2} + 1$ is a lower bound to the average number of attempts an attacker needs to achieve successful impersonation.

The following is a technical lemma whose proof follows from measure-theoretic entropy considerations.

Lemma 1. *For continuous random variables* X*,* Y *and* $\epsilon > 0$*, there exists a sequence of discretized random variables* X_d*,* Y_d *that converge pointwise to* X*,* Y *(when* $d \to \infty$*) such that for sufficiently large* d*,*

$$\mathrm{I}(X; Y) \geq \mathrm{I}(X_d; Y_d) \geq \mathrm{I}(X; Y) - \epsilon. \tag{9}$$

With some modifications to the results from [17, 18], the following theorem can be proven.

Theorem 1. *The secrecy capacity of a biometric system equals*

$$\mathrm{C_s} = \mathrm{I}(X; Y). \tag{10}$$

Proof. We start with the achievability argument. The proof that $\mathrm{I}(X;Y)$ can be achieved if X^n and Y^n are discrete variables, is analogous to the proof in [17]. In order to prove achievability in the continuous case, we choose $\epsilon \geq 0$, and approximate the random variables X^n, Y^n by discretized (quantized) versions, X_d^n, Y_d^n such that $\mathrm{I}(X;Y) - \mathrm{I}(X_d^n; Y_d^n) \leq \epsilon$. (The fact that such a quantization exists follows from lemma 1). Then, taking the encoder that achieves the capacity for the discrete case, it follows that we can achieve $\mathrm{I}(X_d^n; Y_d^n)$. Since this can be done for any $\epsilon \geq 0$ the proof follows.

The fact that $\mathrm{I}(X;Y)$ is an upper bound for $\mathrm{C_s}$ for discrete random variables, follows from the Fano inequality and some entropy inequalities. For the continuous case this follows again by an approximation argument. □

4.2 Biometric Identification

In the enrollment phase of an identification setting, a database is created with data from a set of $|\mathcal{M}|$ enrolled users, each identified with an index $m \in \{1, 2, \ldots, |\mathcal{M}|\}$. In the identification phase, a measurement Y^n and the information in the database are used to find the identity of an unknown (but properly enrolled) individual M. The identifier output is denoted by $\hat{M}$.

Reliability of the identification is expressed by the average error probability, assuming that the individual is chosen at random. Performance in terms of the number of users in the system is expressed by the rate R. The maximum rate at which reliable identification is possible is given by the identification capacity $\mathrm{C_{id}}$.

Definition 3 (Identification Capacity). *The identification capacity* $\mathrm{C_{id}}$ *is the maximal rate* $\mathrm{R_{id}}$*, such that for every* $\epsilon > 0$*, for sufficiently large* n*, there exists an identification strategy that achieves*

$$\text{avg } \Pr\left\{\hat{M} \neq M\right\} \leq \epsilon, \quad \textit{and } \frac{1}{n} \log |\mathcal{M}| \geq \mathrm{R_{id}} - \epsilon, \tag{11}$$

where the average is over all individuals and over all random realizations of all biometrics.

It was proven in [20, 21] that all biometric identification systems, including template protecting systems, satisfy $\mathrm{C_{id}} = \mathrm{I}(X;Y)$.

5 Secure Biometric Authentication Algorithm (SBA)

We introduce a general algorithm that implements the architecture given in Fig. 1. The algorithm basically describes a class of encoders/decoders. It will be shown that the algorithm meets the requirements given by Equations (6), (7) and (8) at a maximum rate.

Initially we define a finite collection $\mathcal{C}$ of $(n, |\mathcal{S}|, \epsilon)$ SECs on $\mathcal{X}^n \times \mathcal{Y}^n$. The collection of SECs is made available in both the enrollment and the authentication phase. Furthermore, for $x^n \in \mathcal{X}^n$ we define $\Phi_{x^n} \subseteq \mathcal{C}$ as follows. A SEC $C = \{(\mathcal{E}_i, \mathcal{D}_i)\}_{i=1}^{|\mathcal{S}|} \in \Phi_{x^n}$ iff $x^n \in \bigcup_i \mathcal{E}_i$.

Enrollment

1. The biometrics x^n of the users are measured.
2. Choose a SEC C at random in Φ_{x^n}. Define w as the index of this SEC C. If $\Phi_{x^n} = \emptyset$, a SEC is selected at random from $\mathcal{C}$.
3. Given a $C = \{(\mathcal{E}_i, \mathcal{D}_i)\}_{i=1}^{|\mathcal{S}|}$, the secret s is defined as, $s = i$ if $x^n \in \mathcal{E}_i$. For $\Phi_{x^n} = \emptyset$, s is chosen at random.
4. The one-way function F is applied to s. The data $F(s)$ and w are stored in a database together with some metadata about the user identity.

Authentication

1. An individual makes an identity claim.
2. The database information $F(s)$ and w for the claimed user is retrieved.
3. A measurement y^n of the user's biometrics and the helper data w are given to the decoder.
4. The SEC $C(w)$ is used to derive the secret v as, $v = i$ if $y^n \in \mathcal{D}_i$.
5. If $F(v) = F(s)$, the user is positively authenticated.

Theorem 2. *For all $\epsilon > 0$ and sufficiently large n, a collection $\mathcal{C}$ of* SEC*s for the* SBA *algorithm can be found such that $\frac{1}{n}\mathrm{H}(S) = \frac{1}{n}\log|\mathcal{S}| \geq \mathrm{C_s} - \epsilon = \mathrm{I}(X;Y) - \epsilon$ and the requirements of Equations (6) and (7) are satisfied.*

Proof. We start with the discrete case. Fix $\epsilon > 0$. Define a set $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$, such that $|\mathcal{S}| = \exp[n(\mathrm{C_s} - \epsilon)]$. It follows from [17] that for sufficiently large n, one can find a size K collection $\mathcal{C}$ of $(n, |\mathcal{S}|, \frac{\epsilon}{2})$ SECs on $\mathcal{X}^n \times \mathcal{Y}^n$,

$$C^k = \left\{ \left(\{e_i^k\}, \mathcal{D}_i^k \right) \right\}_{i=1}^{|\mathcal{S}|}, \quad \text{for } k = 1, 2, \ldots, K, \tag{12}$$

such that $\mathrm{P_{X^n}}(e_i^k) = \mathrm{P_{X^n}}(e_j^k)$, $\bigcup_i \{e_i^k\} \cap \bigcup_i \{e_i^m\} = \emptyset$, for $k \neq m$ and $i, j = 1, 2, \ldots, |\mathcal{S}|$ and

$$\mathrm{P_{X^n}}\left(\bigcup_k \bigcup_i \{e_i^k\} \right) \geq 1 - \frac{\epsilon}{2}. \tag{13}$$

Using the collection $\mathcal{C}$ defined above for the SBA algorithm, it follows from Eq. (13) that $\Pr\{\Phi_{x^n} = \emptyset\} \leq \frac{\epsilon}{2}$. From the construction of the collection $\mathcal{C}$ and Eq. (5) we derive

$$\Pr\{S \neq V | \Phi_{x^n} \neq \emptyset\} \leq \frac{\epsilon}{2}, \quad \text{for } k = 1, 2, \ldots, K, \tag{14}$$

which combined with Eq. (13) leads to $\Pr\{S \neq V\} \leq \epsilon$.

Define W to be the index of the SEC in $\mathcal{C}$ that has to be used to extract the secret. Since for each code, all encoding sets have equal probability, the secrets are uniformly distributed and $\mathrm{I}(W;S) = 0$. Hence, requirements (6), (7) and (8) are fulfilled.

For the proof of the continuous case, we proceed as follows. From Lemma 1, it follows that we can construct quantized variables X_d^n, Y_d^n in such a way that

$$\mathrm{I}(X_d^n; Y_d^n) \geq \mathrm{I}(X^n; Y^n) - \frac{\epsilon}{2}, \tag{15}$$

if d is sufficiently large. The quantization of X^n, Y^n to X_d^n, Y_d^n induces a partition $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_d\}$ of $\mathbb{R}^n$. The random variables X_d^n and Y_d^n defined on the set $\{1, 2, \ldots, d\}$ have probability distribution $\mathrm{P}_{\mathrm{X_d^n}}(i) = \int_{\mathcal{P}_i} f_X(x^n) dx^n$, for all $i = 1, 2, \ldots, d$. Define a size K collection $\mathcal{C}_d$ of SECs on $\{1, 2, \ldots, d\} \times \{1, 2, \ldots, d\}$ as follows,

$$C_d^k = \left\{ \left(\{e_i^k\}, \mathcal{D}_i^k \right) \right\}_{i=1}^{|\mathcal{S}|}, \quad \text{for } k = 1, 2, \ldots, K. \tag{16}$$

For each SEC C_d^k there is a corresponding SEC C^k on $\mathbb{R}^n \times \mathbb{R}^n$ given by,

$$C^k = \left\{ \left(\mathcal{P}_{e_i^k}, \bigcup_{j \in \mathcal{D}_i^k} \mathcal{P}_j \right) \right\}_{i=1}^{|\mathcal{S}|}, \tag{17}$$

resulting in a collection $\mathcal{C}$ of SECs defined on $\mathbb{R}^n \times \mathbb{R}^n$. Note that secrets derived from X^n and Y^n, by means of the collection $\mathcal{C}$, are equal to those derived from X_d^n and Y_d^n using $\mathcal{C}_d$. It was proven for the discrete case that a collection $\mathcal{C}_d$ can be found such that $\Pr\{S \neq V\} \leq \epsilon$, $\mathrm{I}(W; S) = 0$ and

$$|\mathcal{S}| = \exp\left[n\Big(\mathrm{I}(X_d^n; Y_d^n) - \frac{\epsilon}{2} \Big) \right] \geq \exp\left[n\Big(\mathrm{C_s} - \epsilon \Big) \right], \tag{18}$$

where the last inequality follows from Eq. (15) and $\mathrm{C_s} = \mathrm{I}(X; Y)$ (Theorem 1). This concludes the proof for the continuous case. □

Theorem 3. *For all $\epsilon > 0$, there exists a collection $\mathcal{C}$ of* SEC*s such that an identification scheme based on the* SBA *algorithm achieves* both $\frac{1}{n}\mathrm{H}(S) = \frac{1}{n}\log|\mathcal{S}| \geq \mathrm{C_s} - \epsilon$ and $\frac{1}{n}\log|\mathcal{M}| \geq \mathrm{C_{id}} - \epsilon$*, while satisfying the requirements of Equations (6), (7) and (11).*

Proof. (*sketch*) Given $\epsilon > 0$, choose ϵ' such that $\epsilon > \epsilon' > 0$. Theorem 2 states that for sufficiently large n, there exists a collection $\mathcal{C}$ of SECs such that, an implementation of the SBA algorithm using $\mathcal{C}$, satisfies $|\mathcal{S}| = \exp[n(\mathrm{I}(X; Y) - \epsilon')]$, $\Pr\{S \neq V\} \leq \frac{\epsilon}{2}$ and $\mathrm{I}(W; S) = 0$.

Applying the SBA algorithm results in a uniform random assignment of secrets to all users. Let the number of users in the system satisfy $|\mathcal{M}| = \exp[n(\mathrm{I}(X; Y) - \epsilon)]$. The collision probability P_{coll} that two users share the same secret is bounded as follows

$$P_{coll} = \sum_{m=2}^{|\mathcal{M}|} \frac{1}{\exp\left[n\Big(\mathrm{I}(X; Y) - \epsilon' \Big) \right]} \leq \exp\Big[-n(\epsilon - \epsilon') \Big] \leq \frac{\epsilon}{2},$$

where the last inequality holds for sufficiently large n. The overall error probability is upper bounded by

$$\text{avg } \Pr\{\hat{M} \neq M\} \leq P_{coll} + \Pr\{S \neq V\} \leq \epsilon, \tag{19}$$

which is the requirement of Eq. (11). □

6 Code Constructions for the SBA Algorithm

In this section we give two examples of SEC constructions for the SBA algorithm and show that these constructions meet the requirements.

6.1 Secret Extraction from Significant Components

The biometrics are modeled as i.i.d sequences of Gaussian distributed random variables with zero mean, i.e. $X_i \sim \mathcal{N}(0, \sigma_X^2)$. We assume additive uncorrelated Gaussian noise, i.e. $Y^n = X^n + N^n$, where $N_i \sim \mathcal{N}(0, \sigma_N^2)$. The secrecy capacity of these biometrics is $\mathrm{C_s} = n \log(1 + \frac{\sigma_X^2}{\sigma_N^2})$.

The scheme extracts binary secrets of length k from biometric sequences of length n. The secrets are derived as the sign of k components that have "large" absolute value. We define a collection $\mathcal{C}$ of SECs for encoding and decoding. First we define some base sets,

$$E_0 = (-\infty, \infty), \quad E_{-1} = (-\infty, -\delta], \quad E_1 = [\delta, \infty), \tag{20}$$

$$D_0 = (-\infty, \infty), \quad D_{-1} = (-\infty, 0], \quad D_1 = (0, \infty), \tag{21}$$

where δ is chosen sufficiently large considering σ_X^2 and σ_N^2. The SECs in $\mathcal{C}$ are indexed by an ordered set w,

$$w = \{i_1, i_2, \ldots, i_k\} \subseteq \{1, 2, \ldots, n\}. \tag{22}$$

The SEC C^w extracts a secret from the components denoted by w and is defined by

$$C^w = \left\{ \left(E_0^{w^c} \times E_{-1}^{\sigma} \times E_1^{\sigma^c} \,,\, D_0^{w^c} \times D_{-1}^{\sigma} \times D_1^{\sigma^c} \right) \middle| \sigma \subseteq w \right\}, \tag{23}$$

where w^c is the complement of w relative to $\{1, 2, \ldots, n\}$ and σ^c is the complement of σ relative to w. Furthermore $E_0^{w^c} \times E_{-1}^{\sigma} \times E_1^{\sigma^c}$ is the n dimensional Cartesian product with E_0, E_{-1} and E_1 at positions w^c, σ and σ^c, respectively.

It follows from the results in [14] that $\mathrm{I}(W; S) = 0$, $|\Phi_{x^n}|$ is sufficiently large for almost all $x^n \in \mathcal{X}^n$ and $\mathrm{I}(W; X^n) < k$. The impersonation security parameter κ for this scheme is $\kappa = k$.

6.2 Secret Extraction from Discrete Biometrics

In this section we model the biometrics as binary uniform i.i.d. sequences $X^n \in \{0,1\}^n$. The authentication sequence Y^n is an observation of X^n through a binary symmetric channel with cross-over probability p. It follows from Theorem 1 that this channel results in a secrecy capacity equal to $\mathrm{C_s} = 1 - \mathrm{H}(p)$.

Take an error correcting code $C = \{c_1^n, c_2^n, \ldots, c_{|\mathcal{S}|}^n\}$ on $\{0,1\}^n$. The error correcting capability of C implies decoding sets (balls) $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{|\mathcal{S}|}$. A collection $\mathcal{C}$ of SECs is constructed as follows. For every $w \in \{0,1\}^n$,

$$C^w = \left\{ \Big(\{c_i^n + w\}, \mathcal{D}_i + w \Big) \right\}_{i=1}^{|\mathcal{S}|}, \tag{24}$$

where $\mathcal{D}_i + w = \{x^n + w | x^n \in \mathcal{D}_i\}$. Note that $C^{x^n + c_i^n}$, $i = 1, 2, \ldots, |\mathcal{S}|$, is a SEC containing x^n as one of the encoding regions. It follows that $|\Phi_{x^n}| = |\mathcal{S}|$ for all $x^n \in \mathcal{X}^n$.

Proposition 1. *For all $\epsilon > 0$ and sufficiently large n, the error correcting code C used to construct $\mathcal{C}$, can be chosen such that the scheme achieves*

$$\Pr\{S \neq V\} \leq \epsilon \tag{25}$$

$$|\mathcal{S}| = \exp[n(\mathrm{C_s} - \epsilon)] = \exp[n(1 - H(p) - \epsilon)] \tag{26}$$

The proof follows directly from the channel coding theorem.

Proposition 2. *The scheme achieves*

$$\mathrm{I}(W; S) = 0. \tag{27}$$

Proof. First observe that the secret extracted from x^n using the SEC $C^{x^n + c_i^n}$ is exactly i. This leads to $w = x^n + c_s^n$. Since the biometric data x^n is uniformly distributed, we have $\mathrm{H}(W|S) = \mathrm{H}(W|C_s^n) = n$. Furthermore, since there are 2^n different SECs, $\mathrm{H}(W) = n$, which leads to $\mathrm{I}(W; S) = 0$. □

It follows that the impersonation security parameter $\kappa = \log |\mathcal{S}|$. Finally we note that,

$$\mathrm{I}(W; X^n) = \mathrm{H}(X^n) - \mathrm{H}(X^n|W) = n - \log |\mathcal{S}| \tag{28}$$

and in case of $|\mathcal{S}|$ near capacity, $\mathrm{I}(W; X^n) \approx nH(p)$.

The construction presented here gives a rigorous formalism to *fuzzy commitment* [10] and *quantized secret extraction* [13]. We note that this construction can be generalized from binary to larger alphabets.

References

1. Putte, T.v.d., Keuning, J.: Biometrical fingerprint recognition: Don't get your fingers burned. In: IFIP TC8/WG8.8 Fourth Working Conference on Smart Card Research and Advanced Applications. Kluwer Academic Publishers (2000) 289–303
2. Bolle, R.M., Connell, J., Pankanti, S., Ratha, N.K., Senior, A.W.: Biometrics 101. Report RC22481, IBM Research (2002)
3. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer-Verlag New York, Inc. (2003)
4. Matsumoto, T., Matsumoto, H., Yamada, K., Hoshino, S.: Impact of artificial "gummy" fingers on fingerprint systems. In: Optical Sec. and Counterfeit Deterrence Techn. IV. Volume 4677 of Proc. of SPIE. (2002)
5. Hill, C.J.: Risk of masquerade arising from the storage of biometrics. Bachelor of science thesis, Dept. of CS, Australian National University (2002)
6. Schneier, B.: Inside risks: The uses and abuses of biometrics. Comm. of the ACM **42** (1999) 136
7. Penrose, L.: Dermatoglyphic topology. Nature **205** (1965) 545–546
8. Bolling, J.: A window to your health. Jacksonville Medicine **51** (2000) Special Issue: Retinal Diseases.
9. Davida, G., Frankel, Y., Matt, B.: On enabling secure applications through offline biometric identification. In: Proc. of the IEEE 1998 Symp. on Security and Privacy, Oakland, Ca. (1998) 148–157
10. Juels, A., Wattenberg, M.: A fuzzy commitment scheme. In: Sixth ACM Conf. on Comp. and Comm. Security, Singapore (1999) 28–36
11. Ratha, N., Connell, J., Bolle, R.: Enhancing security and privacy in biometrics-based authentication systems. IBM Systems Journal **40** (2001) 614–634
12. Juels, A., Sudan, M.: A fuzzy vault scheme. In: Proc. of the 2002 IEEE Int. Symp. on Inf. Theory, Lausanne, Switzerland (2002) 408
13. Linnartz, J.P., Tuyls, P.: New shielding functions to enhance privacy and prevent misuse of biometric templates. In: Proc. of the 4th Int. Conf. on Audio and Video Based Biometric Person Authentication, Guildford, UK (2003) 393–402
14. Verbitskiy, E., Tuyls, P., Denteneer, D., Linnartz, J.P.: Reliable biometric authentication with privacy protection. In: Proc. of the 24th Symp. on Inf. Theory in the Benelux, Veldhoven, The Netherlands (2003) 125–132
15. Dodis, Y., Reyzin, L., Smith, A.: Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. (Accepted at Eurocrypt 2004)
16. Csirmaz, L., Katona, G.: Geometrical cryptography. In: Proc. of the Int. Workshop on Coding and Cryptography, Versailles, France (2003) 101–109
17. Ahlswede, R., Csiszar, I.: Common randomness in information theory and cryptography. I. secret sharing. IEEE Trans. Inform. Theory **39** (1993) 1121–1132
18. Maurer, U., Wolf, S.: Information-theoretic key agreement: From weak to strong secrecy for free. In: Advances in Cryptology — EUROCRYPT '00. Volume 1807 of LNCS., Springer-Verlag (2000) 351–368
19. Bennett, C.: Quantum cryptography: Uncertainty in the service of privacy. Science Magazine **257** (1992) 752–753
20. O'Sullivan, J.A., Schmid, N.A.: Large deviations performance analysis for biometrics recognition. In: Proc. of the 40th Allerton Conference. (2002)
21. Willems, F.M.J., Kalker, T., Goseling, J., Linnartz, J.P.: On the capacity of a biometrical identification system. In: Proc. of the 2003 IEEE Int. Symp. on Inf. Theory, Yokohama, Japan (2003) 82