

# Privacy-Security Tradeoffs in Biometric Security Systems

Lifeng Lai, Siu-Wai Ho and H. Vincent Poor

Department of Electrical Engineering,

Princeton University,

Princeton, NJ, 08544.

Email: {llai,siuho,poor}@princeton.edu

**Abstract**—Biometric security systems are studied from an information theoretic perspective. A fundamental tradeoff between privacy, measured by the normalized equivocation rate of the biometric measurements, and security, measured by the rate of the key generated from the biometric measurements, is identified. The scenario in which a potential attacker does not have side information is considered first. The privacy-security region, which characterizes the above-noted tradeoff, is derived for this case. The close relationship between common information among random variables and the biometric security system is also revealed. The scenario in which the attacker has side information is then considered. Inner and outer bounds on the privacy-security region are derived in this case.

## I. INTRODUCTION

Biometric security systems have widespread applications. One typical example is a biometric authentication system, in which one's identity is verified by his or her physical biometric characteristics. Another example is a biometric encryption system, in which secret messages are encrypted using biometric characteristics, and are decrypted by presenting the same biometric measurements. Biometric characteristics are unique and do not change dramatically over time. The employment of biometric systems relieves the burden of selecting, memorizing and protecting passwords.

There are usually two stages in biometric security systems: an enrollment stage and a release stage. In the enrollment stage, biometric characteristics, such as fingerprints, are sampled. The biometric measurements themselves or a transformation of the biometric measurements are stored in a database. In the release stage, the biometric characteristics are sampled again. Depending on the specific application, the newly sampled biometric measurements are used either to verify one's identity or used to decrypt messages. There are two main challenges in biometric security systems. Firstly, there is the issue of noise. Two different measurements of the same biometric characteristics will not produce the same result, due to measurement noise or other factors such as injuries. Hence, biometric measurements cannot be directly used for encryption or authentication. Secondly, there is a privacy issue. Biometric characteristics are stored in the database, which creates a security threat. For example, it has been shown that it is possible to recover fingerprints from minutiae points stored in

the database [1]. Unlike passwords, biometric characteristics cannot be changed. Hence, if the database is compromised, identity theft is possible.

In recent years, there has been increasing research interest in addressing these issues. A number of approaches have been proposed (see, e.g., [2]–[7]). The basic idea of these approaches is to generate a secret key and helper data during the initial enrollment stage. The key is used for encryption or authentication. The helper data is stored in the database to assist the key recovery during the release stage. More specifically, during the release stage, one will combine the noisy measurements and the helper data to recover the key. The recovered key is then used to decrypt a message or perform authentication. The existing approaches focus on maximizing the rate of the key that can be recovered successfully from the noisy measurements, under the constraint that the data stored in the database does not provide any information about the key (rigorous definitions will be given in the sequel). This approach is motivated by the fact that the security level of a security system relies on the size of the generated key. For example, in an encryption system, the equivocation of a secret message is limited by the entropy of the key [8], while in an authentication system, the success probability of the opponent's attack decreases as the key size increases [9]. From an information theoretic perspective, these existing approaches can be modelled as a problem of generating a secret key from common randomness [10]–[12], and hence the largest rate of the key can be characterized [4].

While the existing approaches solve the noise issue, they do not address the privacy issue adequately. In practice, the protection of the biometric measurements themselves is at least as important as maximizing the key rate. As noted above, to assist the recovery of the key in the release stage, the helper data must contain certain information about the biometric measurements. To increase the security level of the biometric security system, we would like to make the key rate as large as possible. On the other hand, to preserve the privacy, we need to ensure that information leakage about the biometric measurements themselves is as small as possible. One question naturally arises: can we maximize the rate of the generated key while simultaneously minimizing the information leakage about the biometric measurements? In this paper, by establishing an information theoretic foundation for biometric

security systems, we show that there exists a fundamental tradeoff between security and privacy in any biometric security system. Thus, we cannot achieve both goals simultaneously. More specifically, we first rigorously formulate the privacy-security tradeoff problem in biometric security systems. We then identify and characterize this fundamental tradeoff for several different scenarios. In the first scenario, we require perfect security of the generated key while minimizing the information leakage of the biometric measurements. In this scenario, we consider two systems corresponding to whether the user is allowed to select the key or not. In each system, we characterize the security-privacy tradeoff. Furthermore, we propose schemes that fully achieve any particular point of the tradeoff. In the second scenario, we require perfect privacy of the biometric measurements. We define the common randomness between the fingerprints obtained in the enrollment stage and the release stage. We then reveal a close relationship between this common process and the rate of a secret key that can be generated. We further study the scenario in which the attacker has side information about biometric measurements. We derive both inner and outer bounds on the privacy-security region for this scenario.

Due to space limitations, we provide proof Theorem 1 only and omit proof of the other theorems presented in this paper. The interested reader can refer to [13] for details.

## II. MODEL

We denote the biometric measurements sampled during the enrollment stage by  $X^n$  and the biometric measurements sampled during the verification stage by  $Y^n$ . Here, we assume that  $X^n$  and  $Y^n$  are sequences with length  $n$  taking values from  $n$ -fold product sets  $\mathcal{X}^n$  and  $\mathcal{Y}^n$ , respectively. Assume these measurements are generated according to a joint distribution

$$P_{X^n Y^n}(x^n, y^n) = \prod_{i=1}^n P_{XY}(x, y).$$

Specific models for the distribution of the biometric measurements  $X, Y$  can be found, for example, in [5].

During the enrollment stage, a key  $K$  and helper data  $V$  are generated. We assume that  $K$  takes values from  $\mathcal{K}$  with

$$\log |\mathcal{K}| = O(n). \quad (1)$$

The key  $K$  is used in the rest of the biometric security system to perform various tasks such as encrypting messages, authentication, etc. The helper data  $V$  is stored in the database, and is used to assist the recovery of the key from the noisy measurements  $Y^n$  during the release stage.

Regarding the generation of  $K$ , we consider two types of systems: namely non-randomization systems and randomization systems. In non-randomization systems, both  $K$  and  $V$  are generated from  $X^n$ , by functions  $h_n$  and  $h'_n$  respectively, so that  $V = h_n(X^n)$  and  $K = h'_n(X^n)$ . In randomization systems, the value of the key  $K$  is randomly generated during the enrollment stage. Then  $V$  is generated from the randomly chosen key  $K$  and the biometric measurements  $X^n$  by a function  $h_n^*$  so that  $V = h_n^*(X^n, K)$ .

During the release stage, by providing the noisy measurement  $Y^n$  and the helper data  $V$  stored in the database, we generate an estimate  $\hat{K}$  of the key. Let  $g_n$  be the recovery function, and thus  $\hat{K} = g_n(Y^n, V)$ . In order to perform decryption or authentication, we require an arbitrarily small error probability during the key recover stage. That is, for any  $\epsilon > 0$ , we require that  $\mathbb{P}[K \neq \hat{K}] \leq \epsilon$  for all sufficiently large  $n$ .

An attacker gaining access to the database will obtain  $V$ , which is related to both the biometric measurements  $X^n$  and the key  $K$ . In this paper, we will put various constraints on  $I(X^n; V)$  (the information leakage about the biometric measurement  $X^n$ ), and on  $I(K; V)$  (the information leakage about the generated key), and we will study the fundamental privacy-security tradeoff of biometric security systems under these constraints.

## III. PERFECT SECURITY OF THE GENERATED KEY

In this section, we require that  $V$  does not contain any information about the generated key, and hence, we require that for any  $\epsilon > 0$ ,  $\frac{1}{n}I(K; V) \leq \epsilon$  for all sufficiently large  $n$ . As mentioned before, the security level of a system is related to the rate of the generated key, and hence we measure the security level of the system by  $\frac{1}{n}H(K)$ . The privacy of the biometric measurement is defined as the normalized equivocation rate  $H(X^n|V)/H(X^n)$ . The larger this quantity, the greater the degree of privacy of the biometric measurements. If this quantity is arbitrarily close to 1, we achieve perfect privacy, which means that  $V$  does not leak any information about  $X^n$ . We have the following definition.

*Definition 1:* In a biometric security system, the privacy-security pair  $(\Delta_P, R)$  is said to be achievable, if for each  $\epsilon > 0$ , there exist an integer  $n$ , coding functions, namely  $h_n$  and  $h'_n$  in the non-randomization system and  $h_n^*$  in the randomization system, and a decoding function, namely  $g_n$ , satisfying the following conditions:

$$\frac{1}{n}H(K) \geq R, \quad (2)$$

$$H(X^n|V)/H(X^n) \geq \Delta_P, \quad (3)$$

$$\frac{1}{n}I(V; K) \leq \epsilon \quad \text{and} \quad (4)$$

$$\mathbb{P}[K \neq \hat{K}] = \mathbb{P}[K \neq g_n(Y^n, V)] \leq \epsilon. \quad (5)$$

### A. Non-randomization System

As discussed in Section II, in a non-randomization system, both the key  $K$  and data  $V$  are generated from the biometric measurement  $X^n$ . Some existing schemes, e.g. [3], [5], belong to this category. The following theorem establishes the performance limits of this biometric security system.

*Theorem 1:* Let  $\mathcal{C}_N$  be the set of all privacy-security pairs  $(\Delta_P, R)$  satisfying the following conditions

$$\Delta_P \leq 1 - \frac{I(U; X) - I(U; Y)}{H(X)} \quad \text{and} \quad (6)$$

$$R \leq I(U; Y), \quad (7)$$

for some (auxiliary) random variable  $U$  such that  $(U, X, Y)$  satisfies the Markov chain condition  $U \rightarrow X \rightarrow Y$ . Then any privacy-security pair  $(\Delta_P, R)$  is achievable if and only if  $(\Delta_P, R) \in \mathcal{C}_N$ .

*Proof:* Please refer to Appendix I. ■

*Remark 1:* If we set  $U = X$ , we achieve the largest rate of the key, which is  $I(X; Y)$ . Correspondingly, the privacy level is  $1 - H(X|Y)/H(X)$ . This recovers the existing results of [6], [7].

*Remark 2:* If there is an auxiliary random variable  $U$  such that  $U \rightarrow X \rightarrow Y$  and  $I(U; X) = I(U; Y)$ , then we can achieve perfect privacy, i.e.  $H(X^n|V) = H(X^n)$ . The rate of the key corresponds to the rate of common information between  $X$  and  $Y$  in the sense of [14], which can be far less than the mutual information  $I(X; Y)$  between the two random variables  $X$  and  $Y$ .

### B. Randomization Approach

During the enrollment stage, users have the freedom to choose the values of the keys but they are not required to remember them. For example, the fuzzy vault scheme studied in [2] belongs to this category. Here, the key  $K$  can be viewed as a source of additional randomness. One may conjecture that this additional randomness could enhance the performance of the biometric security system, at least for the privacy of the biometric measurements. The following theorem disproves this conjecture.

*Theorem 2:* Let  $\mathcal{C}_R$  be the set of all privacy-security pairs  $(\Delta_P, R)$  satisfying the following conditions

$$\Delta_P \leq 1 - \frac{I(U; X) - I(U; Y)}{H(X)} \quad \text{and} \quad (8)$$

$$R \leq I(U; Y), \quad (9)$$

for some (auxiliary) random variable  $U$  such that  $(U, X, Y)$  satisfies the following Markov chain condition  $U \rightarrow X \rightarrow Y$ . Then any privacy-security pair  $(\Delta_P, R)$  is achievable if and only if  $(\Delta_P, R) \in \mathcal{C}_R$ .

*Proof:* Please refer to [13] for details. ■

*Remark 3:* From here, we can see that  $\mathcal{C}_N = \mathcal{C}_R$ , and hence, randomization does not increase the region.

## IV. PERFECT PRIVACY FOR BIOMETRIC MEASUREMENTS

Instead of considering perfect security of the key, it is also important to consider the perfect privacy of the biometric measurements. To be specific, for any  $\epsilon > 0$ , we require  $I(X^n; V) \leq \epsilon$  for all sufficiently large  $n$  in this section. As discussed in Remark 2 in Section III, if we consider both perfect privacy and perfect secrecy for the generated key, i.e. both  $I(X^n; V)$  and  $\frac{1}{n}I(V; K)$  are arbitrarily small, the problem can be solved by looking at the common information between  $X$  and  $Y$ . In this section, we generalize the results by relaxing the constraints on  $I(V; K)$ . More specifically, we allow  $I(V; K)$  to range from 0 to  $H(K)$ . The goal is to understand the relationship between the rate of the generated key  $\frac{1}{n}H(K)$  and the secrecy of the key measured by  $H(K|V)/H(K)$ .

In non-randomization systems,  $H(K|X^n) = 0$ , and thus  $I(X^n; V) \leq \epsilon$  implies that  $I(K; V) \leq \epsilon$ . This case has been considered in Remark 2 in Section III. Therefore, it is sufficient to discuss only randomization systems in the remainder of this section.

*Definition 2:* In a perfect privacy biometric security system, a rate-equivocation pair  $(R, \Delta_s)$  is achievable, if for any  $\epsilon > 0$ , there exist an integer  $n$ , and functions  $h_n^*$  and  $g_n$  satisfying the following conditions:

$$\frac{1}{n}H(K) \geq R, \quad (10)$$

$$I(X^n; V) \leq \epsilon, \quad (11)$$

$$H(K|V)/H(K) \geq \Delta_s \quad \text{and} \quad (12)$$

$$\mathbb{P}[K \neq \hat{K}] \leq \epsilon. \quad (13)$$

We now give a definition of the common random process among two random processes, which plays an instrumental role in our study.

*Definition 3:* For two random processes  $\mathbf{X} = (X_1, X_2, \dots)$  and  $\mathbf{Y} = (Y_1, Y_2, \dots)$ , there exists a common random process between them with entropy rate not less than  $\alpha$  if for any  $\eta > 0$ , there exist  $n$  and functions  $\psi_{X^n}$  and  $\psi_{Y^n}$  such that

$$\mathbb{P}[\psi_{X^n}(X^n) \neq \psi_{Y^n}(Y^n)] \leq \eta, \quad (14)$$

and  $n^{-1}H(\psi_{X^n}(X^n)) \geq \alpha - \eta$ .

The following result reveals a close relationship between common random processes and the requirements of the biometric security system under study.

*Theorem 3:* A privacy-rate pair  $(R, \Delta_s)$  is achievable if and only if there exists a common random process between the random processes  $\mathbf{X}$  and  $\mathbf{Y}$  with entropy rate not less than  $R\Delta_s$ .

*Proof:* Please refer to [13] for details. ■

## V. SIDE-INFORMATION AT THE ATTACKER

In this section, we consider the situation in which, besides the data  $V$  stored in the database, the attacker has side-information about the biometric characteristics. This models the situation in which the attacker obtains side-information from other sources, such as biometric characteristics stored in other databases or biometric characteristics from the relatives of the user. We denote the side observation at the attacker by  $Z^n$ , ranging in  $\mathcal{Z}^n$ , and assume that it is correlated with  $(X^n, Y^n)$ . Furthermore, we assume

$$P_{X^n Y^n Z^n}(x^n, y^n, z^n) = \prod_{i=1}^n P_{XYZ}(x, y, z).$$

Since the attacker knows both  $V$  and  $Z^n$ , the privacy level is now measured as  $H(X^n|VZ^n)/H(X^n)$ . Furthermore, to guarantee the security level of the system, the mutual information between the generated key  $K$  and  $(V, Z^n)$  should be small. Here, we consider only the non-randomization approach and have the following definition.

*Definition 4:* In a biometric system with side-information  $Z^n$  available to the attacker, the privacy-security pair  $(\Delta_P, R)$  is said to be achievable if, for any  $\epsilon > 0$ , there exist an integer

$n$ , coding functions  $h_n$  and  $h'_n$ , and a decoding function  $g_n$ , satisfying the following conditions:

$$\frac{1}{n}H(K) \geq R, \quad (15)$$

$$H(X^n|VZ^n)/H(X^n) \geq \Delta_P, \quad (16)$$

$$\frac{1}{n}I(VZ^n; K) \leq \epsilon \quad \text{and} \quad (17)$$

$$\mathbb{P}[K \neq \hat{K}] \leq \epsilon. \quad (18)$$

The following theorem characterizes an inner bound on the set of all achievable privacy-security pairs.

*Theorem 4:* Any privacy-security pairs  $(\Delta_P, R)$  satisfying the following conditions are achievable:

$$\Delta_P \leq 1 - \frac{I(X;UZ) - I(U;Y|W) + I(U;Z|W)}{H(X)} \quad \text{and}$$

$$R \leq I(U;Y|W) - I(U;Z|W), \quad (19)$$

in which  $W$  and  $U$  are (auxiliary) random variables such that  $(W, U, X, Y, Z)$  satisfy the following Markov chain condition  $W \rightarrow U \rightarrow X \rightarrow (Y, Z)$ .

*Proof:* Please refer to [13] for details. ■

The following theorem provides an outer bound for the privacy-security region achievable by any scheme.

*Theorem 5:* A privacy-security pair  $(\Delta_P, R)$  is achievable (in the sense of Definition 4) only if there exist auxiliary random variables  $W$  and  $U$  such that  $W \rightarrow U \rightarrow X \rightarrow (Y, Z)$  and

$$\Delta_P \leq 1 - \frac{I(X;UZ) - I(U;Y) + I(U;Z|W)}{H(X)} \quad \text{and}$$

$$R \leq I(U;Y|W) - I(U;Z|W). \quad (20)$$

*Proof:* Please refer to [13] for details. ■

*Remark 4:* If  $\mathcal{Z} = \Phi$ , the lower-bound in Theorem 4 matches with the upper-bound in Theorem 5. Furthermore, the result recovers that of Theorem 1.

## VI. CONCLUSIONS

Biometric security systems have been studied under a privacy-security tradeoff framework. Two different scenarios, in which the attacker either has side-information about the biometric measurements or not, have been considered. In the scenario for which the attacker does not have side-information, we have considered the two cases of perfect security and perfect privacy. In both cases, the complete privacy-security tradeoff region has been identified. More specifically, an outer bound on the privacy-security pair achievable by any scheme has been derived. Moreover, a scheme has been proposed to achieve this upper bound. In the scenario for which the attacker has side-information about the biometric measurement, inner and outer bounds on the privacy-security region have been derived.

### APPENDIX I PROOF OF THEOREM 1

#### Achievability

We first give a scheme to achieve any pair in the region  $\mathcal{C}_N$ .

- 1) Fix a joint distribution  $P_{UXY}(u, x, y) = P_{U|X}(u|x)P_{XY}(xy)$ , from which we obtain the marginal distribution  $P_U(u)$ . From this joint distribution, we see that  $U \rightarrow X \rightarrow Y$ . Fix  $\gamma > 0$  and  $\eta > 0$ , and randomly select  $M = 2^{n(I(U;X)+\gamma)}$  sequences  $U^n$  from  $T_{[U],\xi}^n$ , and divide them into  $2^{n(I(U;X)-I(U;Y)+\gamma+\eta)}$  bins so that each bin contains  $2^{n(I(U;Y)-\eta)}$  typical sequences<sup>1</sup>. Here we set  $\eta = 3\xi$ . We use  $l$  as the bin index, and  $k$  as the index of the sequence within each bin. Denote the set of these  $M$  sequences as  $\mathcal{M}$ . From the construction above, we can see that each sequence  $u^n \in \mathcal{M}$  is uniquely identified by two indices  $(l(u^n), k(u^n))$ .
- 2) Enrollment stage. For any  $x^n$ , find a sequence  $u^n \in \mathcal{M}$ , so that  $(x^n, u^n) \in T_{[XU],\xi}$ . If there are more than one such  $u^n$ , we set  $U^n$  as the one with the smallest index (first compare the bin indices; if there is a tie, then compare the indices within the bin). If no such sequence exists, we set  $U^n$  to be the first sequence with index  $(l = 1, k = 1)$ . Using this procedure, we associate every  $x^n \in \mathcal{X}^n$  with a sequence  $u^n \in \mathcal{M}$ . We then store the bin index  $l(u^n)$  in the database, and set the key value as the index  $k(u^n)$ . Hence, in our scheme,  $V = L$ .
- 3) Release stage. With the noisy measurement  $y^n$ , and the bin index  $l$ , we first look for a list of sequences  $u^n$  in bin  $l$  that are jointly typical with  $y^n$ , that is  $(u^n, y^n) \in T_{[UY],\xi}$ . Then, we obtain an estimate  $\hat{u}^n$  of  $u^n$  as follows: (1) if there is only one sequence in the list, we set  $\hat{u}^n$  as this sequence; (2) if there are more than one sequences in the list, we randomly choose one sequence from the list and set  $\hat{u}^n$  as this sequence; (3) if the list is empty, we set  $\hat{u}^n$  as the first sequence in bin  $l$ . Hence, for any  $y^n \in \mathcal{Y}^n$ , we have one  $\hat{u}^n$  associated with it. We then obtain an estimate of the key  $\hat{k}$ , by setting it equal to the index of  $\hat{u}^n$  in the bin  $l$ .
- 4) Error Probability analysis.

The event  $\hat{K} \neq K$  occurs only if one of the following events occurs. (1)  $E_1$ : during the enrollment stage, there is no  $U^n$  that is jointly typical with  $X^n$ . (2)  $E_2$ : during the release stage,  $Y^n$  is not jointly typical with  $U^n$ . (3)  $E_3$ : during the release stage, there exist another  $U^n$  in bin  $L$  that is jointly typical with  $Y^n$ .

Using the union bound, we have

$$\mathbb{P}[\hat{K} \neq K] \leq \mathbb{P}[E_1] + \mathbb{P}[E_2 \cap E_1^c] + \mathbb{P}[E_3 \cap E_1^c]. \quad (21)$$

Since there are  $M = 2^{n(I(U;X)+\gamma)}$  typical sequences  $U^n$  in  $\mathcal{M}$ , for any  $\gamma > 0$ ,  $\mathbb{P}[E_1]$  goes to zero when  $n$  is sufficiently large. In the following, we can condition on the event that  $(U^n, X^n) = (u^n, x^n) \in T_{[UX],\xi}^n$ . Due to the Markov lemma [15], if given  $(u^n, x^n) \in T_{[UX],\xi}^n$ , then

$$\mathbb{P}[(u^n, x^n, Y^n) \in T_{[UXY],\xi}^n] > 1 - \xi \quad (22)$$

for  $n$  sufficiently large. Thus  $\mathbb{P}[E_2 \cap E_1^c] \leq \xi$ .

<sup>1</sup>In this paper, the notion of strong typicality follows from [15].

The probability of the third type of error can be bounded as follows:

$$\mathbb{P}[E_3 \cap E_1^c] \quad (23)$$

$$\begin{aligned} &= \mathbb{P}[\exists \tilde{u}^n \neq u^n \text{ in the bin } l \text{ and } (\tilde{u}^n, Y^n) \in T_{[UY], \xi}^n] \\ &\leq \left(2^{n(I(U;Y)-\eta)} - 1\right) \left(2^{n(H(U|Y)+\xi)} - 1\right) \\ &\quad \cdot 2^{-n(H(U)-\xi)} \end{aligned} \quad (24)$$

$$\begin{aligned} &\leq \left(2^{n(I(U;Y)-\eta)}\right) \left(\left(2^{n(H(U|Y)+\xi)}\right) \cdot 2^{-n(H(U)-\xi)}\right) \\ &= \left(2^{n(I(U;Y)-\eta)}\right) \left(2^{-n(I(U;Y)-2\xi)}\right) \end{aligned} \quad (25)$$

$$= \left(2^{-\frac{n\eta}{3}}\right), \quad (26)$$

which tends to 0 as  $n \rightarrow \infty$ .

Hence, for any  $\epsilon > 0$ ,  $\mathbb{P}[\hat{K} \neq K]$  can be made to be less than  $\epsilon$  with a sufficiently large  $n$ .

### 5) Rate analysis

For any  $u^n$  with  $l(u^n) \neq 1$  and  $k(u^n) \neq 1$ , we have

$$\mathbb{P}[U^n = u^n] \leq \sum_{x^n \in T_{[X|U], \xi}(u^n)} P_X^n(x^n) \quad (27)$$

$$\leq 2^{(-n(I(U;X)-\zeta))}, \quad (28)$$

in which  $\zeta$  is a function of  $\xi$ , and goes to zero as  $n$  increases.

Thus,

$$\begin{aligned} H(U^n) &= \sum_{u^n \in \mathcal{M}} -\mathbb{P}[U^n = u^n] \log(\mathbb{P}[U^n = u^n]) \\ &\geq \sum_{u^n \in \mathcal{M}} \mathbb{P}[U^n = u^n] (nI(U;X) - \zeta) \\ &= n(I(U;X) - \zeta) \end{aligned} \quad (29)$$

On the other hand,

$$H(V) \leq n(I(U;X) - I(U;Y) + \gamma + \eta),$$

since the value of  $l$  ranges from 1 to  $2^{n(I(U;X)-I(U;Y)+\gamma+\eta)}$ .

From

$$H(U^n) = H(K, V) = H(V) + H(K|V),$$

we have

$$\begin{aligned} H(K) &\geq H(K|V) \\ &= H(U^n) - H(V) \\ &\geq n(I(U;Y) - \zeta - \gamma - \eta). \end{aligned} \quad (30)$$

So the rate of the key is larger than  $I(U;Y) - \zeta - \gamma - \eta$ .

### 6) Security Analysis.

Now, we bound  $I(K;V)$ , the mutual information between the generated key and the data stored in the database.

$$\begin{aligned} I(K;V) &= H(K) - H(K|V) \\ &\leq n(I(U;Y) - \eta) - n(I(U;Y) - \zeta - \gamma - \eta) \\ &\leq n(\gamma + \zeta), \end{aligned} \quad (31)$$

since the value of  $k$  ranges from 1 to  $2^{n(I(U;Y)-\eta)}$  and (30).

From (31), we see that data  $V$  stored in the database does not contain too much information about the generated  $K$ .

### 7) Equivocation analysis.

$$\begin{aligned} H(X^n|V) &= H(X^n, U^n|V) - H(U^n|V, X^n) \quad (32) \\ &= H(U^n|V) + H(X^n|U^n, V) \\ &\quad - H(U^n|X^n, V) \end{aligned} \quad (33)$$

$$\begin{aligned} &\stackrel{(a)}{\geq} nI(U;Y) + H(X^n|U^n, V) \\ &\quad - H(U^n|X^n) - n(\zeta + \gamma + \eta) \end{aligned} \quad (34)$$

$$\begin{aligned} &\stackrel{(b)}{=} nI(U;Y) + H(X^n|U^n) \\ &\quad - H(U^n|X^n) - n(\zeta + \gamma + \eta) \end{aligned} \quad (35)$$

$$\begin{aligned} &= nI(U;Y) + H(X^n) - H(U^n) \\ &\quad - n(\zeta + \gamma + \eta) \end{aligned} \quad (36)$$

$$\begin{aligned} &\stackrel{(c)}{\geq} nI(U;Y) + nH(X) - nI(X;U) \\ &\quad - n\gamma - n(\zeta + \gamma + \eta). \end{aligned} \quad (37)$$

Here, (a) is due to (30), since  $H(U^n|V) = H(K, V|V) = H(K|V)$ . (b) is due to the fact that  $V$  is a function of  $U^n$ . (c) is due to that  $U^n$  takes at most  $2^{n(I(U;X)+\gamma)}$  different values. Hence, the privacy level of the biometric measurements is guaranteed.

### Converse

Here, we show that  $\mathcal{C}_N$  is the largest region we can achieve using any encoding functions  $h_n, \hat{h}_n$  and decoding function  $g_n$  that satisfy the conditions specified in Definition 1.

$$\begin{aligned} H(X^n|V) &= H(X^n) - I(X^n;V) \\ &= H(X^n) - H(V) + H(V|X^n) \\ &= H(X^n) - H(V) \\ &\leq H(X^n) - H(V|Y^n) \\ &= H(X^n) - H(V, K|Y^n) + H(K|V, Y^n) \\ &\leq H(X^n) - H(V, K|Y^n) + n\delta_n, \end{aligned} \quad (38)$$

where (38) is due to Fano's inequality.

By rewriting  $H(VK|Y^n)$  as  $H(Y^n|KV) + H(KV) - H(Y^n)$ , we continue

$$\begin{aligned} H(X^n|V) &\leq H(X^n) - H(Y^n|KV) - H(KV) \\ &\quad + H(Y^n) + n\delta_n \\ &\leq H(X^n) - \sum_{i=1}^n H(Y_i|KVY^{i-1}) \\ &\quad - I(KV; X^n) + H(Y^n) + n\delta_n \\ &\leq H(X^n) - \sum_{i=1}^n H(Y_i|KVY^{i-1}X^{i-1}) \\ &\quad - I(KV; X^n) + H(Y^n) + n\delta_n \\ &= \sum_{i=1}^n \{H(X_i) - H(Y_i|KVX^{i-1}) \\ &\quad - I(KV; X_i|X^{i-1}) + H(Y_i)\} + n\delta_n, \end{aligned}$$

which is due to the fact that  $Y^{i-1} \rightarrow (K, V, X^{i-1}) \rightarrow Y_i$ . To show this Markov chain relationship, we first have that  $Y^{i-1} \rightarrow X^{i-1} \rightarrow X^n Y_i$ , which leads to  $Y^{i-1} \rightarrow X^{i-1} \rightarrow X^n Y_i \rightarrow (K, V, Y_i)$ , and thus  $Y^{i-1} \rightarrow (K, V, X^{i-1}) \rightarrow Y_i$ .

We continue

$$\begin{aligned} H(X^n|V) &= \sum_{i=1}^n \{H(X_i) - H(Y_i|KVX^{i-1}) \\ &\quad - I(KVX^{i-1}; X_i) + H(Y_i)\} + n\delta_n \\ &= \sum_{i=1}^n \{H(X_i) + I(U_i; Y_i) - I(U_i; X_i)\} + n\delta_n, \end{aligned}$$

which is due to the fact that  $H(X_i|X^{i-1}) = H(X_i)$ . In the last equation, we set  $U_i = KVX^{i-1}$ .

On the other hand

$$\begin{aligned} H(K, V) &= H(K) + H(V) - I(K; V) \\ &\geq H(K) + H(V) - n\epsilon, \end{aligned} \quad (39)$$

due to the requirement that  $I(K; V) \leq n\epsilon$ , from definition 1.

Now,

$$\begin{aligned} H(K, V) &\stackrel{(a)}{=} I(K, V; X^n) \\ &= \sum_{i=1}^n I(K, V; X_i|X^{i-1}) \\ &= \sum_{i=1}^n I(K, V, X^{i-1}; X_i) \\ &= \sum_{i=1}^n I(U_i; X_i), \end{aligned} \quad (40)$$

in which (a) is due to the fact that  $(K, V)$  are functions of  $X^n$ .

Hence,

$$H(K) \leq \sum_{i=1}^n I(U_i; X_i) - H(V) + n\epsilon + n\delta_n. \quad (41)$$

Since  $V$  is a function of  $X^n$ , we have  $H(V, X^n) = H(X^n)$ . Together with (39), we get  $H(V) \geq \sum_{i=1}^n (I(U_i; X_i) - I(U_i; Y_i))$ . It follows from (41) that

$$H(K) \leq \sum_{i=1}^n I(U_i; Y_i) + n\epsilon + n\delta_n, \quad (42)$$

where we have substituted  $H(V)$  with equation (39).

Now, by introducing a random variable  $Q$  uniformly distributed over the set  $\{1, \dots, n\}$ , and setting  $U = (U_Q, Q)$ ,  $X = X_Q$  and  $Y = Y_Q$ , we get the desired result by following the standard single-letter characterization technique [15].

## REFERENCES

- [1] A. Ross, J. Shah, and A. Jain, "From template to image: Reconstructing fingerprints from minutiae points," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 544–560, Apr. 2007.
- [2] A. Juels and M. Sudan, "A fuzzy vault scheme," in *Proc. IEEE Intl. Symposium on Information Theory*, (Lausanne, Switzerland), pp. 293–297, June–July 2002.
- [3] Y. Sutcu, Q. Li, and N. Memon, "Protecting biometric templates with sketch: Theory and practice," *IEEE Trans. Inf. Forensics and Security*, vol. 2, pp. 503–512, Sept. 2007.
- [4] P. Tuyls and J. Goseling, *Biometric Authentication*. Berlin: Springer, 2004.
- [5] S. Draper, A. Khisti, E. Martinian, A. Vetro, and J. Yedidia, "Using distributed source coding to secure fingerprint biometrics," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (Honolulu, HI), pp. 129–132, Apr. 2007.
- [6] G. Cohen and G. Zemor, "The wire-tap channel applied to biometrics," in *Proc. IEEE Intl. Symposium on Information Theory and its Applications*, (Parma, Italy), Oct. 2004.
- [7] T. Ignatenko and F. Willems, "On privacy in secure biometrics authentication systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (Honolulu, HI), pp. 121–124, Apr. 2007.
- [8] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, pp. 656–715, Oct. 1949.
- [9] L. Lai, H. El Gamal, and H. V. Poor, "Authentication over noisy channels," *IEEE Trans. Inf. Theory*. To appear.
- [10] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE Trans. Inf. Theory*, vol. 39, pp. 733–742, May 1993.
- [11] R. Ahlswede and I. Csiszar, "Common randomness in information theory and cryptography, Part I: Secret sharing," *IEEE Trans. Inf. Theory*, vol. 39, pp. 1121–1132, July 1993.
- [12] R. Ahlswede and I. Csiszar, "Common randomness in information theory and cryptography, Part II: CR capacity," *IEEE Trans. Inf. Theory*, vol. 44, pp. 225–240, Jan. 1998.
- [13] L. Lai, S.-W. Ho, and H. V. Poor, "Privacy-security tradeoffs in biometric key generation systems," 2008. Working paper, available via llai@princeton.edu.
- [14] P. Gacs and J. Korner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, pp. 149–162, 1973.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.