Fundamental Tradeoff among Identification, Secrecy and Compression Rates in Biometric Identification System

Vamoua Yachongka and Hideki Yagi

Department of Computer and Network Engineering, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan
E-mail: {vajstudy, h.yagi}@uec.ac.jp

Abstract We investigate the fundamental tradeoff among the identification, secrecy, and compression rates under the condition that the prior distribution of an identified individual is unknown. We combine the techniques developed by Tuncel (2009) and Ignatenko and Willems (2010) to analyze the capacity region of the three rate tuples mentioned above. We show that when the secrecy rate is zero, the capacity region of the identification and compression rates is characterized as a form similar to the one given by Tuncel.

Keywords: biometric identification, capacity region, error probability, identification rate, secrecy rate, compression rate

1. Introduction

In recent years, biometric security has been drawing increasing attentions from the public because it uses information that is difficult to purposely impersonate or change. Some familiar technologies of this kind of security are fingerprint-based identification, iris-based identification, and voice recognition.

The discrete memoryless biometric identification system (BIS), introduced by O’Sullivan and Schmid [1] and Willems et al. [2] independently, is a system for identifying individuals from the measurable physical characteristics of each individual. In general, a BIS consists of three phases: the (I) Generation Phase, (II) Enrollment Phase, and (III) Identification Phase. In the (I) Generation Phase, biodata sequences of all individuals are generated from a stationary memoryless source. In the (II) Enrollment Phase, all the sequences in (I) are observed through a noisy discrete memoryless channel (DMC) and stored in the system database. In the (III) Identification Phase, a biodata sequence of an unknown individual is observed via another noisy DMC, and an estimated value of the unknown individual is output.

In [2], Willems et al. adopted an information-theoretic approach and clarified the identification capacity, which is the maximum achievable rate of the number of individuals when the error probability converges to zero as the length of biometric data sequences goes to infinity. In 2009, Tuncel [3] extended their model by incorporating the compression of biodata sequences stored in the system database and clarified the capacity region of the identification and compression rates. Later, Ignatenko and Willems [4] investigated the capacity region of the identification, secrecy, and compression rates under the condition that the prior distribution of an identified individual is unknown. We show that when the secrecy rate is zero, the capacity region of the identification and compression rates is characterized as a form similar to the one given in [3].

2. System Model

In this section, we define the notation used in this paper and describe the details of the system model.

2.1 Notation

$\mathcal{A}$ stands for a finite alphabet. $A$ denotes a random variable (RV) taking values in $\mathcal{A}$ and $a \in \mathcal{A}$ denotes its realization. $P_A(a) := \Pr[A = a]$ represents the probability distribution on $\mathcal{A}$, and $P_{x^a}$ represents the probability distribution of RV $A^n = (A_1, A_2, \cdots, A_n)$ on $\mathcal{A}^n$, the $n$th Cartesian product of $\mathcal{A}$. $P_{x,a}^{A^n,B^n}$ represents a joint probability distribution of a pair of RVs $(A^n, B^n)$, and its conditional probability distribution $P_{x,a}^{A^n,B^n}$ is defined as $P_{x,a}^{A^n,B^n}(a^n|b^n) = \frac{P_{x,a}^{A^n,B^n}(a^n,b^n)}{P_{x,a}^{B^n}(b^n)}$ ($\forall a^n \in \mathcal{A}^n, \forall b^n \in \mathcal{B}^n$ such that $P_{x,a}^{B^n}(b^n) > 0$).

The entropy of RV $A$ is denoted by $H(A)$, the joint entropy of RVs $A$ and $B$ is denoted by $H(A, B)$, and the mutual infor-
2.2 Model descriptions

The BIS model assumed in this study is shown in Fig. 1. As we have previously mentioned, it consists of three phases: (I) Generation Phase, (II) Enrollment Phase, and (III) Identification Phase. We explain the details of each phase.

**Generation Phase:**
Let $\mathcal{I} = \{1, 2, \cdots, M_I\}$ and $\mathcal{X}$ be the set of indexes of individual and a finite source alphabet, respectively. We assume that there are $M_I$ individuals whose index takes a value in $\mathcal{I}$. For any $i \in \mathcal{I}$, we assume that $x^n_i = (x_{i1}, x_{i2}, \cdots, x_{in}) \in \mathcal{X}^n$, an $n$-length biodata sequence of individual $i$, is generated independently and identically distributed (i.i.d.) from a stationary memoryless source $P_X$. The generating probability for each sequence $x^n_i \in \mathcal{X}^n$ is

$$P_X(x^n_i) \coloneqq \Pr[X^n = x^n_i] = \prod_{k=1}^{n} P_X(x_{ik})$$

for all $i \in \mathcal{I}$. In this paper, $x^n_i$ is called a biodata sequence of individual $i$.

**Enrollment Phase:**
Let $\mathcal{J} = \{1, 2, \cdots, M_J\}$ and $\mathcal{S} = \{1, 2, \cdots, M_S\}$ be the set of indexes of codewords stored in the database and individuals’ secret data, respectively. In this phase, all biodata sequences are observed via a stationary DMC $\{\mathcal{Y}, P_{Y|X}, \mathcal{X}\}$, where $\mathcal{Y}$ is a finite output alphabet of $P_{Y|X}$. The corresponding probability that a biodata sequence $x^n_i \in \mathcal{X}^n$ is observed as $y^n_i = (y_{i1}, y_{i2}, \cdots, y_{in}) \in \mathcal{Y}^n$ via the DMC $P_{Y|X}$ is

$$\Pr[Y^n = y^n_i | X^n = x^n_i] = \prod_{k=1}^{n} P_{Y|X}(y_{ik}|x_{ik})$$

for all $i \in \mathcal{I}$. Afterwards, the observed biodata sequence $y^n_i$ of individual $i \in \mathcal{I}$ is encoded into codeword $j(i) \in \mathcal{J}$ and secret data $s(i) \in \mathcal{S}$ as

$$f(y^n_i) = (j(i), s(i)) \quad (i \in \mathcal{I})$$

where $f : \mathcal{Y}^n \rightarrow \mathcal{J} \times \mathcal{S}$ is the encoder function. Codeword $j(i)$ is a compressed version of sequence $y^n_i$ and it is stored at position $i$ in the database, which can be accessed by the system decoder. The secret data $s(i)$ is returned to individual $i$ and kept confidential.

**Identification Phase:**
Biodata sequence $x^n_w \in \mathcal{X}^n$ of an unknown $w \in \mathcal{I}$ is observed via a DMC $\{\mathcal{Z}, P_{Z|X}, \mathcal{X}\}$, where $\mathcal{Z}$ is a finite output alphabet of $P_{Z|X}$.

The generating probability for each $x^n_w \in \mathcal{X}^n$ is

$$P_Z(z^n) = \Pr[Z^n = z^n] = \prod_{k=1}^{n} P_{Z|X}(z_{ik}|x_{ik})$$

for all $w \in \mathcal{I}$. In this paper, logarithms are of base two.

**Remark 1** RV $W$ is independent of $X^n(i), Y^n(i), J(i), S(i)$, and $Z^n$ for arbitrary $i$. In this paper, we assume neither that the identified individual index $W$ is uniformly distributed over $\mathcal{I}$ nor that there is a prior distribution of $W$.

Tuncel [3] also considered the compression of $y^n_i$ before it is placed into storage, but there is no secret data generation in his model.

3. Problem Formulation

3.1 Definitions and main results

Now we are in a position to introduce the main result of this study. We denote by $W$ and $\hat{W}$ the RVs corresponding to the unknown individual $w$ and its estimate $\hat{w}$, respectively.

**Definition 1** The tuple of identification, secrecy, and compression rates $(R_I, R_S, R_J)$ is said to be achievable if for all $\delta > 0$ and large enough $n$ there exist pairs of encoders and decoders that satisfy

$$\Pr[W \neq \hat{W} | W = i] \leq \delta$$

$$\frac{1}{n} \log M_I \geq R_I - \delta$$

$$\frac{1}{n} \log M_J \geq R_J + \delta$$

$$\frac{1}{n} H(S(i)) + \delta \geq \frac{1}{n} \log M_S \geq R_S - \delta$$

$$\frac{1}{n} I(S(i); J(i)) \leq \delta$$

Moreover, the capacity region $\mathcal{R}$ is defined as the closure of the set of all achievable rate tuples.
In Definition 1, Eq. (5) is the reliability condition of the system, which means that the identification error probability (IEP) can be made arbitrarily small. Unlike the BIS in [2], the decoder of our model outputs only an individual index but not its secret data in view of an application in which the secret data of an individual is kept confidential after it is generated. Equation (6) is the condition of the identification rate, which requires that the rate is greater than or equal to $R_i$. Equation (7) is the condition of the coding rate of codewords stored in the system database, which should be less than or equal to $R_S$. Finally, from the biodata protection perspective, Eq. (9) gives the amount is less than or equal to $\delta$. This is an arbitrary small quantity, which indicates that codeword $J(i)$ and secret data $S(i)$ are almost independent of each other.

Indeed, the secrecy leakage of individual $i$ can be evaluated as

$$I(S(i);J(1),\cdots,J(M_i)) = I(S(i);J(i))$$

From the generation of codeword $J(i)$ and secret data $S(i)$, only $J(i)$ can depend on $S(i)$ for all $i \in I$.

**Theorem 1** The capacity region for the BIS is given by

$$\mathcal{R} = \{ (R_1, R_S, R_J) : 0 \leq R_i \leq I(Z; U), \quad R_J \geq I(Y; U), \quad 0 \leq R_S \leq H(Y|U) \text{ with some } U \text{ s.t. } Z - X - Y - U \}$$  \hspace{0.5cm} (11)

where $U$ is an auxiliary RV taking a value in some finite alphabet $\mathcal{U}$ with $|\mathcal{U}| \leq |Y| + 1$ and satisfying the Markov chain $Z - X - Y - U$ [5] so that the joint distribution satisfies

$$P_{X/Y/Z/U}(x,y,z,u) = P_X(x)P_{Y/X}(y|x)P_{Z/X}(z|x)P_{U/Y}(u|y)$$

and show that the capacity region is contained in the right-hand-side of Eq. (5). Because in our scenario we consider the maximum IEP, Eq. (5) for all $i \in I$ is a more rigorous condition so its capacity region is smaller than the region determined by assuming that $W$ is uniformly distributed and the average IEP is used.

We assume that rate tuple $(R_1, R_S, R_J)$ is achievable so that there exists a pair of an encoder and a decoder $(f,g)$ such that all conditions in Definition 1 with Eq. (5) replaced by Eq. (14) are satisfied for any $\delta > 0$ and large enough $n$. In the proof, we employ Fano’s inequality [5], which states that $H(W|\hat{W}) \leq 1 + \Pr[\hat{W} \neq W] \log M_I$. We start by expanding the identification rate of individuals as

$$\log M_I = H(W)$$

and show that the capacity region is contained in the right-hand-side of Eq. (5). Because in our scenario we consider the maximum IEP, Eq. (5) for all $i \in I$ is a more rigorous condition so its capacity region is smaller than the region determined by assuming that $W$ is uniformly distributed and the average IEP is used.

We assume that rate tuple $(R_1, R_S, R_J)$ is achievable so that there exists a pair of an encoder and a decoder $(f,g)$ such that all conditions in Definition 1 with Eq. (5) replaced by Eq. (14) are satisfied for any $\delta > 0$ and large enough $n$. In the proof, we employ Fano’s inequality [5], which states that $H(W|\hat{W}) \leq 1 + \Pr[\hat{W} \neq W] \log M_I$. We start by expanding the identification rate of individuals as

$$\log M_I = H(W)$$

and show that the capacity region is contained in the right-hand-side of Eq. (5). Because in our scenario we consider the maximum IEP, Eq. (5) for all $i \in I$ is a more rigorous condition so its capacity region is smaller than the region determined by assuming that $W$ is uniformly distributed and the average IEP is used.

We assume that rate tuple $(R_1, R_S, R_J)$ is achievable so that there exists a pair of an encoder and a decoder $(f,g)$ such that all conditions in Definition 1 with Eq. (5) replaced by Eq. (14) are satisfied for any $\delta > 0$ and large enough $n$. In the proof, we employ Fano’s inequality [5], which states that $H(W|\hat{W}) \leq 1 + \Pr[\hat{W} \neq W] \log M_I$. We start by expanding the identification rate of individuals as

$$\log M_I = H(W)$$

and show that the capacity region is contained in the right-hand-side of Eq. (5). Because in our scenario we consider the maximum IEP, Eq. (5) for all $i \in I$ is a more rigorous condition so its capacity region is smaller than the region determined by assuming that $W$ is uniformly distributed and the average IEP is used.
From Eqs. (14) and (15), we obtain

$$(1-\delta) \log M_I - 1 \leq I(W; J(1), \ldots, J(M_I), Z^n)$$

(c) follows because $W$ is independent of $J(1), \ldots, J(M_I)$, and (d) holds since only $Z_t$ is dependent on $J(1), \ldots, J(M_I), W$.

Now we introduce an RV $U_t$ ($t = 1, 2, \ldots, n$) as $U_t = (Z_t^{i-1}, J(W))$. Using the chain rule of entropies in Eq. (16), we obtain

$$(1-\delta) \log M_I - 1 \leq \sum_{i=1}^{n} \{H(Z_t) - H(Z_t|Z_1^{i-1}, J(W))\}$$

By denoting $U = (U_t, Q), Z = Z_t, X = X_t, Y = Y_t$. Eq. (18) becomes

$$\sum_{i=1}^{n} I(Z_t; U_t) = n I(Z_t; U_t, Q)$$

Substituting Eqs. (6) and (19) into Eq. (17), we obtain

$$R_I \leq \frac{1}{1-\delta} I(Z; U) + \left\{ \delta + \frac{1}{n(1-\delta)} \right\}$$

Next, we analyze the compression rate. To avoid confusion, we introduce the notation $Y'_t(i) = (Y_{t1}, Y_{t2}, \ldots, Y_{ti})$ as the partial sequence of the sequence $Y'_t$ from the first symbol to the $i$th symbol. It follows from Eq. (7) that

$$n(R_I + \delta) \geq \log M_I \geq H(J(W))$$

$$\geq I(J(W); Y'_t(W))$$

$$= H(Y'_t(W)) - H(Y'_t(W)|J(W))$$

$$= \sum_{i=1}^{n} [H(Y_{ti(W))} - H(Y_{ti(W)}|J(W), Y_{t-1}(W))]$$

From the fact that $Y_{ti(W)} - (J(W), Y_{t-1}(W)) - Z_t^{i-1}$ forms a
Markov chain (cf. [3]), we can further bound \( n(R_J + \delta) \) as

\[
\begin{aligned}
n(R_J + \delta) & \geq \sum_{i=1}^{n} [H(Y_{W_i}) - H(Y_{W_i}|J(W), Y_{W_i}^{-1}(W), Z_{W_i}^{-1})] \\
& \geq \sum_{i=1}^{n} [H(Y_{W_i}) - H(Y_{W_i}|J(W), U_i)] \\
& = \sum_{i=1}^{n} [H(Y_{W_i}) - H(Y_{W_i}|U_i)] \\
& = nI(Y; U) 
\end{aligned}
\]

(22)

where (f) follows because conditioning reduces entropy [5]. Moreover, by applying a similar argument to that in Eqs. (18) and (19), Eq. (22) becomes

\[
n(R_J + \delta) \geq H(J(W)) \geq nI(Y; U) \tag{23}
\]

Finally, we focus on the secrecy rate that the system achieves. For a deterministic function (encoder) \( f \), \( H(L) \geq H(f(L)) \) holds , where \( L \) is an RV. Applying this inequality to our system model, we have

\[
nH(Y) \geq H(e(Y^n_w)) = H(S(w), J(w)) \\
= H(J(w)) + H(S(w)) - I(S(w); J(w)) 
\]

(24)

Furthermore, from Eqs. (8) and (9) in Definition 1,

\[
nH(Y) \geq H(J(w)) + H(S(w)) - n\delta \\
\geq H(J(w)) + nR_S - 2\delta - n\delta \\
= H(J(w)) + nR_S - 3n\delta 
\]

(25)

Since \( w \) can be any value in \( I \), we obtain

\[
nH(Y) \geq \sum_{j=1}^{M_J} \frac{\delta n}{M_J} H(J(w)) + nR_S - 3n\delta \\
= H(J(W)) + nR_S - 3n\delta 
\]

(26)

Substituting Eq. (23) into Eq. (26) and dividing both sides by \( n \), we obtain

\[
R_S \leq H(Y) - I(Y; U) + 3\delta \leq H(Y|U) + 3\delta 
\]

(27)

By letting \( n \to \infty \) and \( \delta \downarrow 0 \) in Eqs. (20), (23) and (27), the converse part of Theorem 1 is proved.

**Achievability (Direct) Part:** For details of the proof of this part, readers should refer to [6]. We omit the details and give the scheme only. We make use of the argument developed by Tuncel [3] for codeword generation and the argument developed by Ignatenko and Willems [4] for secret data generation.

First, we fix \( \delta > 0, \epsilon > 0 \) arbitrarily and block length \( n \). We choose a test channel \( P_{UX} \) satisfying \( Z = X - Y - U \) to determine the joint distributions \( P_{X,Y,U,Z} \). We set the values of each rate as \( R_J = I(Z; U) - 2\delta, R_S = H(Y|U) - \delta, \) and \( R_J = I(Y; U) + 2\delta \). We also set \( M_U = 2^{R_U}, M_S = 2^{R_S}, \) and \( M_J = 2^{R_J} \).

**Random Code Generation:**

Codewords \( \{U^n(j)\}_{j=1}^{M_J} \) are generated i.i.d. from \( P_U \). We randomly place any sequence in the set \( A^{(n)}_s(Y) \) into \( M_S \) bins. For each \( j \), the corresponding secret label \( s(j) \) is assigned randomly and uniformly from \( S = \{1, \cdots, M_S\} \).

**Encoding (Enrollment):**

When encoder \( f \) observes the biometric sequence \( y^n \in Y^n \), it determines codeword \( j(i) \) and secret data \( s(i) \). Codeword \( j(i) \) is defined as the smallest \( j \) that satisfies \( (y^n, U^n(j)) \in A^{(n)}_s(Y) \). We set \( s(i) = s(j) \). The codeword \( j(i) \) is stored at position \( i \) in the database, and the corresponding secret data is handed back to individual \( i \). In the case that no such \( j(i) \) exists, we let \( j(i) = 1 \) and \( s(i) = s(1) \).

**Decoding (Identification):**

When the decoder observes \( \tilde{z}^n \), the noisy version of the identified individual sequence \( x^n_w \), it compares the sequence \( x^n_w \) with code \( \{j(1), \cdots, j(M_J)\} \) in the database. We define \( g(\tilde{z}^n, J(1), \cdots, J(M_J)) \) as the smallest \( w \) such that \( (\tilde{z}^n, U^n(j(\tilde{w}))) \in A^{(n)}_s(Z) \). If no such \( w \) exists, we let \( g(\tilde{z}^n, J(1), \cdots, J(M_J)) = 1 \). We set \( \hat{w} = g(\tilde{z}^n, J(1), \cdots, J(M_J)) \).

\( \Box \)

**4. Discussion and Conclusion**

In this work, we analyzed the fundamental tradeoff among the identification, secrecy, and compression rates provided that the prior distribution of an identified individual is unknown. We showed that when the secrecy rate is zero, the capacity region is characterized as the same form as the one given by Tuncel [3]. The capacity region in our scenario is larger than the region derived in [4]. This is because, unlike the model proposed in [4], we assumed that only the index of an individual is estimated by the decoder in the Identification Phase, and not the secret data. This means that our problem is more relaxed than the one in [4] in terms of the IEP, which leads to the larger capacity region.

**Acknowledgements**

This study was supported in part by JSPS KAKENHI Grant Number 16K06340 and Grant Number 17K00020 and by JSPS and NUS under the Japan-Singapore Research Cooperative Program.

**References**

[1] J. A. O’Sullivan and N. A. Schmid: Large deviations performance analysis for biometrics recognition, Proc. 40th Annual Allerton Conf. on Communication, Control, and Computing, IL, USA, 2002.
[2] F. Willems, T. Kalker, S. Baggen, and J. P. Linnartz: On the capacity of a biometric identification system. Proc. IEEE Int. Symp. Inform. Theory, Yokohama, Japan, p. 82, 2003.

[3] E. Tuncel: Capacity/storage tradeoff in high-dimensional identification systems, IEEE Trans. Inf. Theory, Vol. 55, No. 5, pp. 2097–2016, 2009.

[4] T. Ignatenko and F. Willems: Fundamental limits for biometric identification with a database containing protected templates, Proc. 2010 IEEE Int. Symp. on Inform. Theory and Its Appl. (ISITA2010), Taichung, Taiwan, pp. 54–59, 2010.

[5] T. M. Cover and J. A. Thomas: Elements of Information Theory, 2nd ed., John Wiley & Sons, 2006.

[6] V. Yachongka and H. Yagi: The capacity region of identification, secrecy and compression rates in biometric identification systems, IEICE Tech Rep., Vol. 117, No. 394, IT2017-60, pp. 31–36, 2018.

Vamoua Yachongka received his B.E. degree and M.E. degree in communication engineering and informatics from the University of Electro-Communications, Japan, in 2015 and 2017, respectively. He is currently a Ph.D. candidate at the University of Electro-Communications under the supervision of Assoc. Prof. Hideki Yagi. His research interests include information theory and information-theoretic security.

Hideki Yagi received his B.E. degree, M.E. degree, and Ph.D. degree in industrial and management systems engineering from Waseda University, Tokyo, Japan, in 2001, 2003, and 2005, respectively. He is currently an Associate Professor at the Department of Computer and Network Engineering, University of Electro-Communications, Tokyo, Japan. He was with Media Network Center, Waseda University as a Research Associate from 2005 to 2007 and an Assistant Professor from 2007 to 2008. In the winter of 2008 and from July 2010 to January 2011, he was a Visiting Fellow at Princeton University. His research interests include information and coding theory and information-theoretic security. Dr. Yagi is a member of Research Institute of Signal Processing, IEICE, and IEEE.

(Received July 2, 2018; revised September 8, 2018)