BIOMETRIC SEARCH CODES

Frans M.J. Willems

Eindhoven University of Technology

IEEE EURASIP Spain Seminar on Signal Processing, Communication and Information Theory,
Universidad Carlos III de Madrid,
December 11, 2014

European Association for Signal Processing

Procedure

A biometric identification system identifies individuals based on physical features. Let M individuals be indexed $w \in \{1, 2, \cdots, M\}$. There are three operational phases:

1. **Generation phase:** A biometric sequence $z^N(w)$ is generated for each individual w, hence

$$\Pr\{Z^N(w) = z^N\} = \prod_{n=1,N} Q(z_n), \text{ for all } z^N \in \mathcal{Z}^N.$$

2. **Enrollment phase:** Each individual is observed via an enrollment channel. The resulting enrollment-sequence $x^N(w)$, is added to a **database**. Now

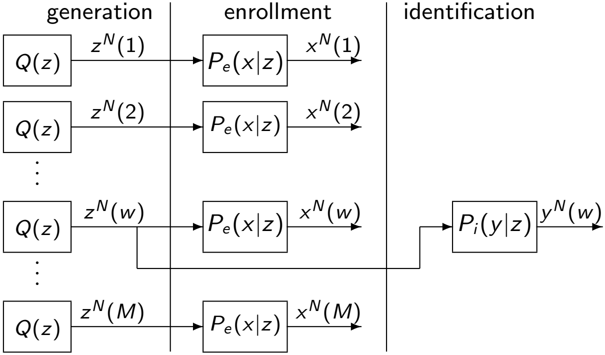
$$\Pr\{X^N(w) = x^N | Z^N(w) = z^N(w)\} = \prod_{n=1,N} P_e(x_n|z_n(w)) \text{ for all } x^N \in \mathcal{X}^N.$$

3. **Identification phase:** An unknown individual is observed via an identification channel. For individual w identification sequence y^N occurs with probability

$$\Pr\{Y^N = y^N | Z^N(w) = z^N(w)\} = \prod_{l=1,N} P_i(y_n|z_n(w)) \text{ for all } y^N \in \mathcal{Y}^N.$$

The observed identification-sequence y^N is now "compared" to all sequences x^N in the database and an **estimate** $\widehat{w}$ of the unknown individual is given.

Model


generation
enrollment
identification
$Q(z)$
$z^N(1)$
$P_e(x|z)$
$x^N(1)$
$Q(z)$
$z^N(2)$
$P_e(x|z)$
$x^N(2)$
$Q(z)$
$z^N(w)$
$P_e(x|z)$
$x^N(w)$
$P_i(y|z)$
$y^N(w)$
$Q(z)$
$z^N(M)$
$P_e(x|z)$
$x^N(M)$

Probabilistic structure

- Note that for all $w \in \{1, 2, \cdots, M\}$, and $x^N \in \mathcal{X}^N$,

$$\begin{aligned}\Pr\{X^N(w) = x^N\} &= \prod_{n=1,N} Q_b(x_n)\\ \text{with } Q_b(x) &= \sum_{z \in \mathcal{Z}} Q(z)P_e(x|z) \text{ for all } x \in \mathcal{X},\end{aligned}$$

  hence all enrollment sequences are IID with $Q_b(x)$.

- For all $w \in \{1, 2, \cdots, M\}$, $x^N \in \mathcal{X}^N$ and $y^N \in \mathcal{Y}^N$,

$$\begin{aligned}\Pr\{Y^N = y^N|X^N(w) = x^N\} &= \prod_{n=1,N} Q_c(y_n|x_n)\\ \text{with } Q_c(y|x) &= \frac{\sum_{z \in \mathcal{Z}} Q(z)P_e(x|z)P_i(y|z)}{\sum_{z \in \mathcal{Z}} Q(z)P_e(x|z)},\\ &\quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y},\end{aligned}$$

  hence the channel $Q_c(y|x)$ between enrollment sequence and observation sequence is a DMC.

Capacity

Error probability is defined as

$$P_e \triangleq \sum_{w=1,M} \frac{1}{M} \Pr\{\widehat{W} \neq w | W = w\}$$

We say that the capacity of a biometric system is C if for any $\delta > 0$ there exist, for all large enough N, decoders that achieve

$$\begin{aligned} \frac{1}{N} \log_2 M &\geq C - \delta, \\ P_e &\leq \delta. \end{aligned}$$

Theorem

O'Sullivan & Schmid [Allerton 2002], W., Kalker, Goseling & Linnartz [ISIT 2003]: The capacity of a biometric identification system is given by

$$C = I(X;Y),$$

where $P(x,y) = Q_b(x)Q_c(y|x) = \sum_{z \in \mathcal{Z}} Q(z)P_e(x|z)P_i(y|z)$ *for all* $x \in \mathcal{X}, y \in \mathcal{Y}$.

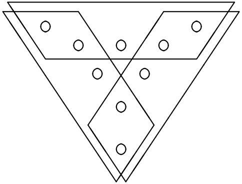
Proof

- Observe that the enrollment sequences $x^N(1), x^N(2), \cdots, x^N(M)$ form a **random code**.
- Ones of these codewords is observed via a DMC. The decoder looks for the unique index w such that $(x^N(\widehat{w}), y^N) \in \mathcal{A}_\varepsilon^N(XY)$.
- Standard arguments apply directly here and result in achievability.
- Converse is standard.

Can we somehow decrease the search complexity?

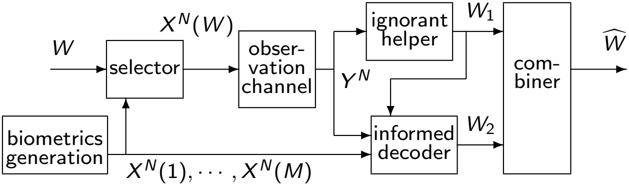
Note that to do the identification, the decoder has to check all enrollment sequences $\{x^N(w), w = 1, 2, \cdots, M\}$ to find out whether $(x^N(w), y^N) \in \mathcal{A}_\varepsilon^N(XY)$.

- **QUESTION:** Can we speed up this process?
- **IDEA:** First we determine the "cluster" to which the unknown individual belongs, then we find out which individual "within the cluster" is the unknown individual we are looking for. (cluster-check must be as elementary as a refinement-check).
- **EXAMPLE:** 9 individuals in 3 clusters, 3 cluster-checks and 5 refinement-checks needed, 8 checks in total (6 would be better):



- **QUESTION:** What is the **fundamental trade-off** between # of cluster-checks and # of refinement-checks here?

How do we model this search system?



- First all enrollment-sequences $x^N(1), x^N(2), \cdots, x^N(M)$ are generated, and this "code" is made available to informed decoder.
- An individual W is chosen uniformly. Its enrollment sequence $X^N(W)$ is transmitted via the observation channel, output is Y^N.
- The ignorant helper determines from Y^N the cluster index W_1, sends it to informed decoder and combiner.
- The informed decoder determines from Y^N and W_1 the refinement-index W_2 and sends it to the combiner.
- The combiner determines index $\widehat{W}$.

Why should the helper be ignorant?

If the helper would know the “code” it could do $\sqrt{M}$ cluster-checks, that each involve a typicality check for all $\sqrt{M}$ individuals in that cluster. In that case a cluster-check is not elementary anymore.

Fundamental trade-off

There are three rates. Rate R corresponds to the number of individuals, cluster-rate R_1 to the number of clusters, and refinement-rate R_2 to the number of individuals in a cluster.

Theorem

The region of achievable rate triples for our biometric identification system is given by

$$\begin{aligned}\{(R_1, R_2, R) \quad : \quad & R_1 \geq I(Y;U), \\ & R_2 \geq \max(0, R - I(X;U)), \\ & 0 \leq R \leq I(X;Y), \\ & \textit{for } P(x,y,u) = Q_b(x)Q_c(y|x)P(u|y), \\ & \textit{where } |\mathcal{U}| \leq |\mathcal{Y}| + 1\}.\end{aligned}$$

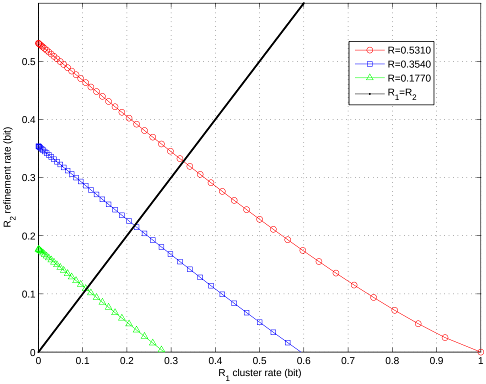
Proof outline

- Generate M_1 covering sequences $u^N(1), u^N(2), \cdots, u^N(M_1)$.
- The ignorant helper determines which covering sequence $u^N(w_1)$ is jointly typical with y^N, and outputs w_1. There is always such a sequence if $R_1 \geq I(U;Y)$.
- The informed decoder has a **list of individuals whose enrollment sequences are jointly typical with** $u^N(w_1)$. The log-size of this list is $N(R - I(U;X))$. It finds out which of these sequences is jointly typical with $(y^N, u^N(w_1))$, and outputs its index w_2 within the list.
- If $R \leq I(X;Y)$, the probability that the enrollment sequence of some other individual is jointly typical with $(y^N, u^N(w_1))$, is negligible.
- Converse.

Example: Binary uniform symmetric case

Consider a system with binary uniform biometric sequences and a binary symmetric observation channel with cross-over probability $q = 0.1$. Region of achievable triples:

$$\{(R_1, R_2, R) \quad : \quad R_1 \geq 1 - h(p), R_2 \geq \max(0, R - 1 + h(p * q)), \\ 0 \leq R \leq 1 - h(q), \text{ for } 0 \leq p \leq 1/2\}.$$



Excess Rate

Ideally $R_1 + R_2 = R$. However in general we can write for the excess rate Δ that

$$\begin{aligned}
\Delta = R_1 + R_2 - R &\geq I(U;Y) - I(U;X) \\
&= H(U|X) - H(U|Y,X) \\
&= I(U;Y|X) \\
&= H(Y|X) - H(Y|X,U).
\end{aligned}$$

For U such that $R \geq I(X;U)$ and for optimum cluster-refinement rate-pairs (R_1, R_2) we get

$$\Delta = H(Y|X) - H(Y|X,U) \leq H(Y|X).$$

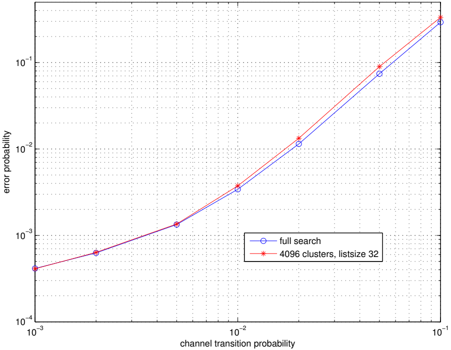
This maximum excess rate is achieved for $U = Y$, and this results in refinement rate $R_2 = 0$.
Note that the upper bound on the excess rate is larger for more noisy observation channels.
Noise-free observation channels allow for a zero-excess rate.

Concluding remarks

- Storage complexity (from the lists, which is $R_1 + R_2$) is not optimized here, and is Δ larger than R.
  Compressed data bases are considered by Westover & O'Sullivan [2008], and Tuncel [2009].
- Implementation: Helper should use structured vector quantizer. In that case checking all the clusters is not needed, and only the refinement rate is of interest.
- Three or more steps.

Search Complexity: Clustering Based on a Binary Golay Code

$N = 23$. Generate $M = 4096$ uniform binary biometric sequences ($R = 12/23$). Nr. of clusters $M_1 = 4096$. Refinement-list-size $M_2 = 32$. Error probability based on full search and on clustering. Complexity decrease $4096/32 = 128$.


error probability
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
channel transition probability
full search
4096 clusters, listsize 32