

Statistical Modeling and Analysis of Content Identification

Pierre Moulin

University of Illinois

Beckman Inst., Coord. Sci. Lab., & ECE Dept.
405 N. Mathews Ave., Urbana, IL 61801, USA

Email: moulin@jfp.uiuc.edu

Abstract—A number of hash-based algorithms for audio and video identification (ID) have been studied in recent literature, and some have been deployed as mobile phone applications and on file sharing sites. A fundamental question is what is the relationship between database size, hash length, and robustness, that any reliable content ID system should satisfy. This paper presents some answers under a simple statistical model for the signals of interest.

Index Terms—Content identification, audio, video, hashing, fingerprinting, decoding, error exponents, capacity, strong converse

I. INTRODUCTION

The problem of *content identification* (ID) has received considerable attention in the literature starting circa 2000 and was originally motivated by applications such as broadcast monitoring and connected audio, as described in [1] and references therein. The problem has also gained interest as a potential filtering technology for file sharing [1], [2]. Such technology is currently deployed on sites such as YouTube and Dailymotion and aims at identifying (automatically and in real time) copyrighted uploaded content (audio and video). Hash-based algorithms allow for real-time operation. Instead of matching the content itself, one matches short fingerprints extracted from it, using robust hashing methods.

Content ID is closely related to the problem of *content retrieval*, in which a user submits a query (probe) to a database and retrieves similar content. Again the use of hashing methods is desirable when the database is large and real-time operation is required.

While considerable efforts have been invested constructing signal processing primitives for robust hashes as well as efficient string matching algorithms, the fundamental limits of content ID remain largely unexplored. A recent exception is Altug *et al.* [3] which studies information-theoretic limits for decoding the output of a discrete memoryless channel using asymmetric iid (independent and identically distributed) codebooks. The full received signal is used for decoding.

The current paper derives a fundamental relationship between database size, hash length, and robustness that holds for any reliable, fingerprint-based, content ID system. A general framework is introduced for analyzing practical systems. The content ID problem may be formulated as a decoding problem, albeit a nonstandard one. Connections with the problems of

erasure and list decoding [4] and their extension for unknown channels [5] are pointed out.

II. STATEMENT OF THE CONTENT ID PROBLEM

For the purposes of this paper, a *content database* is defined as a collection of M elements (content items)

$$\mathbf{x}(m) \in \mathcal{X}^N, \quad m = 1, 2, \dots, M,$$

each of which is a sequence of N frames $\{x_1(m), x_2(m), \dots, x_N(m)\}$. Here \mathcal{X} is the set of possible values of a frame. A frame could be a short video segment (say a few seconds), a short sequence of image blocks (say 32×32 pixels) or a short audio segment. Frames may be overlapping spatially, temporally, or both. For instance, the audio fingerprinting paper [1] use overlapping time windows that are 370 ms long and start every 11.6 ms (overlap = 31/32). Thus a 3-minute second song is represented by $N = 15,500$ frames. In [2] random sampling times are used. It is desired that the audio be identifiable from a short segment, say 3 sec long, corresponding to $L = 258$ frames. This is called the granularity of the audio ID system [1].

The problem is to determine whether a given *probe* consisting of L frames,

$$\mathbf{y} \in \mathcal{X}^L,$$

is related to some element of the database, and if so, identify which one. To this end, an algorithm $\psi(\cdot)$ must be designed, returning the decision

$$\psi(\mathbf{y}) \in \{0, 1, 2, \dots, M\}$$

where $\psi(\mathbf{y}) = 0$ indicates that \mathbf{y} is unrelated to any of the database elements. The probe could be very short, i.e., $L \ll N$, which clearly reduces the reliability of the decision.

Algorithm performance is evaluated using several metrics:

- Execution time;
- Probability of false positives (incorrectly decide that \mathbf{y} is related to an element of the database);
- Probability of false negatives (incorrectly decide that \mathbf{y} is unrelated to any element of the database);
- Robustness;
- Granularity (L);
- Database size (linear in M);
- Storage requirements (linear in MN).

Regarding storage requirements and execution time, the cost of the algorithm could be enormous for large databases, especially for Internet-scale video databases. This has motivated the development of fingerprinting methods [1], [2], where a moderately short binary fingerprint $\mathbf{g} = \phi(\mathbf{y})$ is computed for the probe and compared with the M fingerprints $\mathbf{f}(m) = \phi(\mathbf{x}(m))$ associated with the database elements. Then only the *fingerprint database* $\{\mathbf{f}(m), 1 \leq m \leq M\}$ needs to be stored, and the algorithm seeks the fingerprint that matches the probe fingerprint \mathbf{g} . This matching is particularly fast if \mathbf{g} perfectly matches the corresponding $\mathbf{f}(m)$, or differs from it in a few positions [1].

To meaningfully define the error probabilities, one needs to define a statistical model for this problem. Then one would like to express these probabilities in terms of the problem parameters M, N, L . To simplify the notation, we find it convenient to assume the fingerprints are length- N sequences over a finite alphabet \mathcal{F} . The storage cost of the fingerprint is thus at most $N \log |\mathcal{F}|$ bits. As discussed above, the alphabet for \mathcal{X} is generally huge, so representing a signal \mathbf{x} by its fingerprint $\mathbf{f} = \phi(\mathbf{x})$ entails tremendous compression.

We propose the following fairly general definition of a fingerprint-based content ID code.

Definition 2.1: A (M, N, L) content ID code for a size- M database populated with \mathcal{X}^N -valued content items, and granularity L , is a pair consisting of an encoding function $\phi : \mathcal{X}^N \rightarrow \mathcal{F}^N$ returning a fingerprint $\mathbf{f} = \phi(\mathbf{x})$, and a constrained decoding function $\psi : \mathcal{X}^L \rightarrow \{0, 1, \dots, M\}$ returning $\hat{m} = \psi(\mathbf{y})$, where the dependency on input \mathbf{y} is via the fingerprint $\phi(\mathbf{y})$.

Definition 2.2: The rate of the code is

$$R \triangleq \frac{1}{L} \log(M(N - L)). \quad (1)$$

The code rate will turn out to be the fundamental scaling parameter for the content ID problem. The operational significance of R will be detailed in Secs. VI and VII. We shall consider sequences of codes where M, N, L tend to infinity in such a way that $\frac{1}{L} \log(M(N - L))$ tends to a limit R and that $L \leq N$ (but not necessarily $L \ll N$). Note in particular that M and N could be of the same order of magnitude, or one could dominate the other.

III. STATISTICAL MODEL FOR CONTENT ID

Our statistical model consists of the following ingredients:

- A statistical model for the content database;
- A statistical model for the relationship between the probe $\mathbf{y} \in \mathcal{X}^L$ and a corresponding content $\mathbf{x}(m) \in \mathcal{X}^N$;
- An assumption about the encoding function ϕ .

Database. We assume the M database elements $\mathbf{x}(m) \in \mathcal{X}^N$ are drawn independently from a common stationary distribution. As a special case, the N frames associated with each database element $\mathbf{x}(m) \in \mathcal{X}^N$ could be drawn *independently* from a common probability distribution $P(x)$, $x \in \mathcal{X}$. This would be a reasonable approximation provided the clips are not too short (e.g., this model would fail for a video clip

consisting of just a few consecutive images) and the frames have weak or no overlap.

Probe. In the event the probe \mathbf{y} is related to some element $\mathbf{x}(m)$ of the database, we assume this relationship takes the following form. Let N_0 be an integer in $\{0, 1, 2, \dots, N - L - 1\}$ and $Q(y|x)$ a conditional probability mass function (pmf) for $x, y \in \mathcal{X}$. Then the conditional probability of \mathbf{y} given $\mathbf{x}(m)$ and N_0 is the product conditional pmf

$$Q^L(\mathbf{y}|\mathbf{x}(m), N_0) \triangleq \prod_{n=1}^L Q(y_n|x_{n+N_0}(m)). \quad (2)$$

Hence \mathbf{y} depends only on samples $x_{N_0+1}, \dots, x_{N_0+L}$ of $\mathbf{x}(m)$, and the process $(\mathbf{X}(m), \mathbf{Y})$ is jointly stationary.

In the event the probe \mathbf{y} is unrelated to any database element, we assume that \mathbf{y} is drawn from $P_{\mathbf{Y}}$, the marginal distribution of $P_{\mathbf{X}}Q^L$.

Hash function. Let $\mathbf{F} = \phi(\mathbf{X}) \in \mathcal{F}^N$ and $\mathbf{G} = \phi(\mathbf{Y})$ where \mathbf{X} is drawn from the stationary distribution $P_{\mathbf{X}}$ and \mathbf{Y} is related to \mathbf{X} by the model (2). While many possible constructions for ϕ are possible, we will assume that

- The samples F_i , $1 \leq i \leq N$ are iid with pmf p_F .
- Analogously to (2), the conditional probability of \mathbf{g} given $\mathbf{f}(m)$ and N_0 is the product conditional distribution

$$p_{G|F}^L(\mathbf{g}|\mathbf{f}(m), N_0) \triangleq \prod_{n=1}^L p_{G|F}(g_n|f_{n+N_0}(m)). \quad (3)$$

Hence \mathbf{g} depends only on samples $f_{N_0+1}, \dots, f_{N_0+L}$ of $\mathbf{f}(m)$, and the pairs (F_i, G_i) , $1 \leq i \leq N$ are iid with pmf p_{FG} .

If \mathbf{X} and \mathbf{Y} are drawn independently from the stationary distributions $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$, respectively, then the pairs (F_i, G_i) , $1 \leq i \leq N$ are iid with product pmf $p_F p_G$.

These statistical assumptions about the fingerprints are motivated by the practical designs in [1], [2] where $\mathbf{F} = \phi(\mathbf{X})$ is obtained by constructing a robust sparse feature sequence from \mathbf{X} and using hashing methods based on random permutations.

Hypothesis Testing. The ID problem is viewed as a statistical test with $M + 1$ hypotheses H_0, H_1, \dots, H_M . The probability of false positives is

$$P_{FP} \triangleq Pr[\psi(\mathbf{Y}) > 0 | H_0],$$

and the probability of false negatives is

$$P_{FN} \triangleq \frac{1}{M} \sum_{m=1}^M Pr[\psi(\mathbf{Y}) \neq m | H_m].$$

IV. LIST DECODER

We now propose a list decoder that is slightly easier to analyze than the single-output decoders of Def. 2.1.

Define a “decoding metric” $d : \mathcal{F}^2 \rightarrow \mathbb{R}$, extended additively to pairs of subsequences $\{f_{i+N_0}, g_i, 1 \leq i \leq L\}$ as follows:

$$d(\mathbf{f}, \mathbf{g}|N_0) \triangleq \sum_{i=1}^L d(f_{i+N_0}, g_i).$$

Also define a decision threshold τ . The decoder outputs the list \mathcal{L} of all m such that the minimum distance (over all N_0) between the corresponding fingerprint subsequence starting at time N_0 and the extracted fingerprint is below the threshold:

$$\mathcal{L}(\mathbf{g}) \triangleq \{m \in \{1, \dots, M\} : \min_{0 \leq N_0 < N-L} d(\mathbf{f}(m), \mathbf{g}|N_0) < k\tau\}. \quad (4)$$

V. ERROR ANALYSIS

Since the sequences $\mathbf{X}(m)$ are drawn independently from the same distribution, the error probabilities are independent of m . Without loss of generality, assume that the true $m = 1$. There are two error events of interest for the list decoder:

- *Miss*: The correct m does not appear on the decoder's list:

$$\forall N_0 \in \{0, \dots, N-L-1\} : d(\mathbf{f}(1), \mathbf{g}|N_0) > k\tau.$$

- *Incorrect Decoding*: One or more incorrect m appear on the decoder's list:

$$\exists m > 1, N_0 \in \{0, \dots, N-L-1\} : d(\mathbf{f}(m), \mathbf{g}|N_0) < k\tau.$$

The number of incorrect messages on the list is

$$N_i \triangleq \sum_{m>1} \mathbb{1}\{\min_{0 \leq N_0 < N-L} d(\mathbf{f}(m), \mathbf{g}|N_0) < k\tau\}.$$

Corresponding to these two events are the probability of miss

$$P_{miss} = Pr \left[\min_{0 \leq N_0 < N-L} d(\mathbf{f}(1), \mathbf{g}|N_0) > k\tau \right] \quad (5)$$

and the expected number of incorrect items on the list:

$$\mathbb{E}[N_i] = M Pr \left[\min_{0 \leq N_0 < N-L} d(\mathbf{f}(2), \mathbf{g}|N_0) < k\tau \right]. \quad (6)$$

VI. ERROR EXPONENTS

For any sequence of (M, N, L) content ID codes such that $\lim_{\frac{1}{L} \log(M(N-L))} = R$, define the miss exponent

$$E_{miss}(P_F, P_{G|F}, \tau) = \liminf_{L \rightarrow \infty} -\frac{1}{L} \ln P_{miss} \quad (7)$$

and the incorrect-item exponent

$$E_i(P_F, P_{G|F}, R, \tau) = \liminf_{L \rightarrow \infty} -\frac{1}{L} \ln \mathbb{E}[N_i]. \quad (8)$$

Also define the following set of joint pmf's over \mathcal{F}^2 :

$$\Gamma(\tau) \triangleq \left\{ Q : \sum_{f, g \in \mathcal{F}} Q(f, g) d(f, g) < \tau \right\}, \quad (9)$$

and denote its closure by $\overset{\circ}{\Gamma}(\tau)$.

Proposition 6.1: The error exponents (7) and (8) are given by

$$E_{miss}(P_F, P_{G|F}, \tau) = \min_{P'_{FG}} \left[D(P'_{FG} \| P_F P_{G|F}) + \min_{Q \in \overset{\circ}{\Gamma}(\tau)} D(P'_{FG} \| Q) \right] \quad (10)$$

$$E_i(P_F, P_{G|F}, R, \tau) = \min_{P'_{FG}} \left[D(P'_{FG} \| P_F P_G) + \min_{Q \in \overset{\circ}{\Gamma}(\tau)} D(P'_{FG} \| Q) - R \right] \quad (11)$$

which are respectively nondecreasing and nonincreasing functions of τ .

Remarks:

- $E_{miss}(P_F, P_{G|F}, \tau)$ does not depend on R .
- $E_i(P_F, P_{G|F}, R, \tau)$ can be negative if R is too large, in which case the expected number of incorrect items on the decoder's output list grows exponentially with L .
- $E_i(P_F, P_{G|F}, R, \tau)$ depends on P_F and $P_{G|F}$ only via the product of the marginals P_F and p_G .

Sketch of the Proof. Monotonicity follows from the definition of the constrained set $\Gamma(\tau)$ in (9). From (5), the probability of miss is given by

$$\begin{aligned} P_{miss} &= Pr \left[\min_{0 \leq N'_0 < N-L} d(\mathbf{F}(1), \mathbf{G}|N'_0) > k\tau \right] \\ &\leq Pr[d(\mathbf{F}(1), \mathbf{G}|N_0) > k\tau] \\ &\stackrel{(a)}{=} P_{FG}^L \left[\sum_{i=1}^L d(F_i, G_i) > k\tau \right] \\ &\stackrel{(b)}{\leq} 2^{-LE_{miss}(\tau)} \end{aligned} \quad (12)$$

where (a) holds because the pairs $(F_i(1), G_i)$, $1 \leq i \leq L$ are iid with joint pmf p_{FG} ; and (b) follows from Sanov's theorem.

From (6), the expected number of incorrect items on the list is given by

$$\begin{aligned} \mathbb{E}[N_i] &= M Pr \left[\min_{0 \leq N_0 < N-L} d(\mathbf{F}(2), \mathbf{G}|N_0) < k\tau \right] \\ &\stackrel{(a)}{\leq} M(N-L) \max_{0 \leq N_0 < N-L} Pr[d(\mathbf{F}(2), \mathbf{G}|N_0) < k\tau] \\ &\stackrel{(b)}{=} M(N-L) Pr[d(\mathbf{F}(2), \mathbf{G}|N_0 = 0) < k\tau] \\ &\stackrel{(c)}{=} M(N-L) P_F^L P_G^L \left[\sum_{i=1}^L d(F_i, G_i) < k\tau \right] \\ &\stackrel{(d)}{\leq} 2^{-LE_i(\tau)} \end{aligned} \quad (13)$$

where (a) follows from the union bound; (b) from the fact that the probability in the right side of (a) is independent of N_0 ; (c) because the sequences $F_i(2)$, $1 \leq i \leq L$ and G_i , $1 \leq i \leq L$ are independent with respective distributions P_F^L and P_G^L ; and (d) follows from Sanov's theorem and the definition of R in (1).

VII. CONTENT ID CAPACITY

We now ask what is the maximum achievable rate for the content ID system analyzed in the previous section, and thus whether the decoder can be improved. The hash function ϕ is fixed in this analysis.

A. Achievable Rates

Define the set of conditional distributions

$$\mathcal{P}'_{G|F} \triangleq \{P'_{G|F} : P'_G = P_G, \mathbb{E}_{P_F P'_{G|F}} d(F, G) = \mathbb{E}_{P_{FG}} d(F, G)\} \quad (14)$$

and the *generalized mutual information*

$$I_{\text{GMI}}(P_F, P_{G|F}, d) \triangleq \min_{P'_{G|F} \in \mathcal{P}'_{G|F}} D(P_F P'_{G|F} \| P_F P_G) \quad (15)$$

which appears in information-theoretic analyses of channel capacity with mismatched decoders [7]–[9]. The range of values of M, N, L for which reliable ID is possible is given by the following proposition.

Proposition 7.1: The supremum of the values of R for which the error exponents (10) and (11) are positive is

$$R = I_{\text{GMI}}(P_F, P_{G|F}, d)$$

and is achieved when $\tau = \mathbb{E}_{P_{FG}} d(F, G)$.

Remark: by application of the weak law of large numbers to (12)(a), P_{miss} tends to 1 for all $\tau > \mathbb{E}_{P_{FG}} d(F, G)$ and to 0 for all $\tau < \mathbb{E}_{P_{FG}} d(F, G)$.

Sketch of the proof. From (10), the exponent vanishes when both divergences are zero, which occurs if $P'_{G|F} = P_{G|F}$ (first divergence is zero) and $P_F P'_{G|F} \in \Gamma^c(\tau)$, i.e., $\mathbb{E}_{P_{FG}} d(F, G) \leq \tau$. Now define the set of conditional distributions

$$\mathcal{P}'_{G|F}(\tau) \triangleq \{P'_{G|F} : P'_G = P_G, \mathbb{E}_{P_F P'_{G|F}} d(F, G) \leq \tau\}$$

which is disjoint with $\mathcal{P}'_{G|F}$ of (14) for all $\tau < \mathbb{E}_{P_{FG}} d(F, G)$, and contains $\mathcal{P}'_{G|F}$ for all other τ . Moreover, observe that $P'_{G|F} \in \mathcal{P}'_{G|F}(\tau)$ implies that $P_F P'_{G|F} \in \Gamma(\tau)$. An upper bound on the incorrect-item exponent (11) is obtained by constraining $P'_{G|F} \in \mathcal{P}'_{G|F}$ and Q to have the product form $P_F P''_{G|F}$ where $P''_{G|F} \in \mathcal{P}'_{G|F}(\tau)$ and $P_F P''_{G|F} \in \Gamma(\tau)$. Thus

$$\begin{aligned} 0 &\leq E_i(P_F, P_{G|F}, R, \tau) \\ &\leq \min_{P'_{G|F} \in \mathcal{P}'_{G|F}} \left[D(P_F P'_{G|F} \| P_F P_G) \right. \\ &\quad \left. + \min_{P''_{G|F} \in \mathcal{P}'_{G|F}(\tau)} D(P_F P'_{G|F} \| P_F P''_{G|F}) - R \right]. \end{aligned}$$

The right side is nonincreasing in τ , and $\mathbb{E}_{P_{FG}} d(F, G)$ is the supremum of all τ for which the miss exponent is positive. For this value of τ , the minimum over $P'_{G|F}$ is achieved by $P'_{G|F}$, hence the second divergence above is zero. Recalling (15), we obtain

$$0 \leq I_{\text{GMI}}(P_F, P_{G|F}, d) - R.$$

B. Matched Metric

If the degradation channel is known, the decoding metric that maximizes $I_{\text{GMI}}(P_F, P_{G|F}, d)$ is the negative loglikelihood

$$d(f, g) = -\log p_{G|F}(g|f)$$

and the generalized mutual information of (15) coincides with the ordinary mutual information $I(p_F, p_{G|F})$ between the random variables F and G [7], [8]. The corresponding threshold is

$$\tau = \mathbb{E}_{P_{FG}} d(F, G) = H(G|F) = H(G) - I(p_F, p_{G|F}).$$

If the degradation channel is known to belong to a convex class \mathcal{W} of conditional pmf's, let $p^*_{G|F}$ be the worst channel achieving

$$I(p_F, \mathcal{W}) \triangleq \min_{p_{G|F} \in \mathcal{W}} I(p_F, p_{G|F}). \quad (16)$$

Then the optimal decoding metric is the negative loglikelihood matched to $p^*_{G|F}$:

$$d(f, g) = -\log p^*_{G|F}(g|f)$$

and $I(p_F, \mathcal{W})$ is achievable.

C. Converse

At first sight it seems trivial to prove that no rate exceeding $I(p_F, \mathcal{W})$ is achievable. However there is no guarantee that the decoder using the matched metric is optimal. Indeed such a decoder would just be a generalized maximum-likelihood decoder owing to the search over the unknown offset parameter $N_0 \in \{0, 1, \dots, N-L-1\}$, and the optimality of this strategy needs to be established.

Assume the offset parameter N_0 is drawn uniformly from $\{0, 1, \dots, N-L-1\}$. We consider codes that return a single $m \in \{0, \dots, M\}$ (as defined in Def. 2.1) and consider average error probability

$$\bar{P}_e \triangleq \frac{1}{M+1} \sum_{m=0}^M Pr[\psi(\mathbf{Y}) \neq m | H_m]$$

and maximum error probability

$$P_{e,\max} \triangleq \max_{0 \leq m \leq M} Pr[\psi(\mathbf{Y}) \neq m | H_m]$$

as performance metrics. The first metric is of questionable value because it gives a vanishing weight ($1/(M+1)$) to the null hypothesis. However it is used to establish the following proposition, which follows by application of Fano's inequality.

Proposition 7.2: For any sequence of (M, N, L) content ID codes such that

$$\lim_{L \rightarrow \infty} \frac{1}{L} \log M > I(p_F, \mathcal{W}),$$

\bar{P}_e does not vanish.

Note that the argument of the logarithm is M and not $M(N-L)$. When $M = N-L$, this implies there is a gap (by a factor of two) between the lower and upper bounds on

achievable R . The gap is even larger if $N-L > M$. The reason for this gap is that N_0 is a nuisance variable. There are only $M+1$ hypotheses to which Fano's inequality is applied, which yields the following necessary condition for P_e to vanish:

$$\begin{aligned} \log(M+1) &\leq I(M; \mathbf{G}) \\ &\stackrel{(a)}{=} I(M; \mathbf{G} | N_0) - I(N_0; M | \mathbf{G}) \\ &\leq I(M; \mathbf{G} | N_0) \\ &\stackrel{(b)}{\leq} I(\mathbf{F}_{N_0}^{N_0+L-1}(M); \mathbf{G} | N_0) \\ &= LI(p_F, p_{G|F}), \quad \forall p_{G|F} \in \mathcal{W}. \end{aligned}$$

where (a) holds because N_0 and M are independent random variables and (b) because $(M, N_0) \rightarrow \mathbf{F}_{N_0}^{N_0+L-1}(M) \rightarrow \mathbf{G}$ forms a Markov chain.

However, using a strong converse, we are able to eliminate this gap and use the more appropriate $P_{e,\max}$ as the performance metric.

Proposition 7.3: For any sequence of (M, N, L) content ID codes such that

$$\lim \frac{1}{L} \log M(N-L) = R > I(p_F, \mathcal{W}),$$

$P_{e,\max}$ tends to 1.

Sketch of the proof. Denote the decoding region for m by \mathcal{D}_m :

$$\mathbf{g} \in \mathcal{D}_m \Leftrightarrow \psi(\mathbf{g}) = m, \quad 0 \leq m \leq M.$$

The decoding regions form a partition of the fingerprint space \mathcal{F}^L . Fix an arbitrary $\epsilon > 0$ and define the set

$$T_m(\epsilon) = \left\{ \mathbf{g} : \begin{aligned} &\sum_{n_0=0}^{N-L-1} \frac{p_{G|F}^L(\mathbf{g} | \mathbf{f}(m), n_0)}{p_G^L(\mathbf{g})} \\ &< 2^{L[I(p_F, p_{G|F}) + \epsilon]} \end{aligned} \right\}, \quad 0 \leq m \leq M.$$

If the correct-decoding probability is at least $1 - \lambda$ (for any $\lambda > 0$) then we have, for each $m = 0, 1, \dots, M$,

$$\begin{aligned} 1 - \lambda &\leq \frac{1}{N-L} \sum_{\mathbf{g} \in \mathcal{D}_m} \sum_{N_0=0}^{N-L-1} p_{G|F}^L(\mathbf{g} | \mathbf{f}(m), N_0) \\ &= \frac{1}{N-L} \sum_{\mathbf{g} \in \mathcal{D}_m \cap T_m(\epsilon)} + \frac{1}{N-L} \sum_{\mathbf{g} \in \mathcal{D}_m \cap T_m^c(\epsilon)} \quad (17) \end{aligned}$$

Using the law of large numbers, the second term can be asymptotically upper-bounded by

$$\frac{1}{N-L} \sum_{\mathbf{g} \in \mathcal{D}_m \cap T_m^c(\epsilon)} \leq \frac{1}{N-L} \sum_{\mathbf{g} \notin T_m(\epsilon)} < \epsilon.$$

Substituting into (17) and using the definition of $T_m(\epsilon)$, we obtain, for any $0 < \epsilon < 1 - \lambda$,

$$\begin{aligned} 1 - \lambda - \epsilon &\leq \frac{1}{N-L-1} \sum_{\mathbf{g} \in \mathcal{D}_m \cap T_m(\epsilon)} \sum_{N_0=0}^{N-L} p_{G|F}^L(\mathbf{g} | \mathbf{f}(m), N_0) \\ &\leq \frac{1}{N-L} 2^{L[I(p_F, p_{G|F}) + \epsilon]} \sum_{\mathbf{g} \in \mathcal{D}_m \cap T_m(\epsilon)} p_G^L(\mathbf{g}). \end{aligned}$$

Summing over all m and using the fact that $\sum_m \sum_{\mathbf{g} \in \mathcal{D}_m} p_G^L(\mathbf{g}) = 1$, we obtain

$$(1 - \lambda - \epsilon)M(N-L) \leq 2^{L[I(p_F, p_{G|F}) + \epsilon]}$$

and thus

$$\frac{\log(1 - \lambda - \epsilon)}{L} \leq -\frac{\log(M(N-L))}{L} + I(p_F, p_{G|F}) + \epsilon.$$

Taking limits as $L \rightarrow \infty$, we obtain

$$0 \leq -R + I(p_F, p_{G|F}) + \epsilon, \quad \forall p_{G|F} \in \mathcal{W},$$

hence

$$0 \leq -R + I(p_F, \mathcal{W}) + \epsilon,$$

from which the claim follows by letting $\epsilon \downarrow 0$.

When combined with the achievability result of Sec. VII-B, Prop. 7.3 implies that $I(p_F, \mathcal{W})$ is the content ID capacity (under our statistical assumptions on the database, degradation channel, and hash function ϕ).

REFERENCES

- [1] J. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprinting System," *Proc. Int. Conf. Music Information Retrieval*, 2002.
- [2] S. Baluja and M. Covell, "Audio Fingerprinting: Combining Computer Vision & Data Stream Processing," *Proc. ICASSP*, Honolulu, HI, 2007.
- [3] Y. Altug, M. K. Mihcak, O. Ozyesil, and V. Monga: "Reliable Communication with Asymmetric Codebooks: An Information Theoretic Analysis of Robust Signal Hashing," arXiv:0809.1910v1 [cs.IT], Sep. 2008.
- [4] G. D. Forney, Jr., "Exponential Error Bounds for Erasure, List, and Decision Feedback Schemes," *IEEE Trans. Information Theory*, Vol. 14, No. 2, pp. 206–220, 1968.
- [5] P. Moulin, "A Neyman-Pearson Approach to Universal Erasure and List Decoding," *IEEE Trans. Information Theory*, Vol. 55, No. 10, pp. 4462–4478, Oct. 2009.
- [6] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd Ed., Wiley, 2000.
- [7] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "In Information Rates for Mismatched Decoders," *IEEE Trans. Information Theory*, Vol. 40, No. 6, pp. 1953–1967, 1994.
- [8] I. Csiszár and P. Narayan, "Channel capacity for a Given Decoding Metric," *IEEE Trans. Information Theory*, Vol. 41, No. 1, pp. 35–43, 1995.
- [9] E. Abbe, M. Médard, S. Meyn and L. Zheng, "Finding the Best Mismatched Detector for Channel Coding and Hypothesis Testing," *Proc. ITA*, San Diego, 2007.