MODEL-BASED DECODING METRICS FOR CONTENT IDENTIFICATION

Rohit Naini and Pierre Moulin

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

ABSTRACT

In this paper, decoding metrics are designed for statistical fingerprint-based content identification. A fairly general class of structured codes is considered, and a statistical model for the resulting fingerprints and their degraded versions (following miscellaneous content distortions) is proposed and validated. The Maximum-Likelihood fingerprint decoder derived from this model is shown to considerably improve upon previous decoders based on the Hamming metric. A GLRT test is also proposed and evaluated to deal with unknown distortion channels.

***Index Terms*—** Content identification, audio, video, hashing, fingerprinting, maximum likelihood decoding

1. INTRODUCTION

Hash-based content identification (ID) is an emerging research area. Applications include broadcast monitoring, connected audio, content tracking, asset management, contextual advertisement, and filtering for user-generated content websites [1, 2]. Content ID technologies are currently deployed on sites such as YouTube and Dailymotion and aim 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.

An impressive variety of algorithms have already been developed for constructing signal processing primitives for robust hashes as well as efficient string matching algorithms. Recently there have been attempts to formulate a scientific framework for content ID, aiming at discovering the fundamental limits of content ID and ways to achieve them. For instance, the paper [3] derives an information-theoretic relationship between database size, hash length, and robustness that holds for any reliable, fingerprint-based, content ID system, under some structural assumptions on the fingerprinting code and a statistical assumption on the signals of interest. Decoding of correlated binary fingerprints is studied in [4] and the related problem of physical object identification is studied in [5]. On the algorithmic side, the papers [6, 7] have reported excellent ID performance using structured content ID codes for audio and video ID, respectively, and a new hash design algorithm called *Symmetric Pairwise Boosting* (SPB).

In this paper, we formulate a novel statistical model for original and distorted fingerprints and validate this model on the state-of-the-art audio fingerprinting algorithm of [6]. The model is then used to construct a new decoder based on the maximum-likelihood principle. This decoder is shown to vastly improve upon [6], which uses the Hamming decoding metric, and over [8], which learns a weighted L^2 decoding metric.

2. STATEMENT OF THE CONTENT ID PROBLEM

A *content database* is defined as a collection of M elements (content items) $\mathbf{x}(m) \in \mathcal{X}^N, \quad m = 1, 2, \cdots, M$, each of which is a sequence of N *frames* $\{x_1(m), x_2(m), \cdots, x_N(m)\}$. Here $\mathcal{X}$ is the set of possible values of a frame. A frame could be a short video segment, a short sequence of image blocks, or a short audio segment. Frames may be overlapping spatially, temporally, or both. For instance, the audio fingerprinting paper [6] uses overlapping time windows that are 2 sec long and start every 185 ms; the temporal overlap is 15/16. A 3-minute second song is represented by $N = 1000$ frames. It is desired that the audio be identifiable from a short segment, say 5 sec long, corresponding to $L = 16$ frames. This is called the granularity of the audio ID system [6]. Typically $L \ll N$.

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

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

where $\psi(\mathbf{y}) = 0$ indicates that $\mathbf{y}$ is unrelated to any of the database elements.

Algorithm performance is evaluated using several metrics [1], including execution time, probability of false positives, probability of false negatives, robustness, granularity (L), database size (linear in M), and storage requirements (linear in MN).

3. STRUCTURED CONTENT ID CODES

In this paper, we restrict our attention to the following fairly general class of content ID codes. The codes of [1, 6, 7], among others, fall in this category.

Definition 3.1 *A* (M, N, L) **content ID product code** *for a size-*M *database populated with* $\mathcal{X}^N$*-valued content items, and granularity* L*, is a pair consisting of a mapping* $\phi : \mathcal{X} \to \mathcal{F}$ *and a decoding function* $\psi : \mathcal{F}^L \to \{0, 1, \cdots, M\}$*, such that (i) a content item* $\mathbf{x}$ *is encoded into a fingerprint* $\mathbf{f} \in \mathcal{F}^N$ *with components* $f_i = \phi(x_i)$*,* $1 \le i \le N$*; (ii) a probe* $\mathbf{y}$ *is encoded into a probe fingerprint* $\mathbf{g} \in \mathcal{F}^L$ *with components* $g_i = \phi(y_i)$*,* $1 \le i \le L$*; (iii) the decoder returns* $\hat{m} = \psi(\mathbf{g})$*.*

Hence the mapping ϕ is applied independently to each frame. It might be convenient to impose additional structure on the code. For instance, the mapping $\phi : \mathcal{X} \to \mathcal{F}$ in [6,7] is obtained by applying a set of J optimized filters to each frame and quantizing each of the J real-valued filter outputs to four levels, as illustrated in Fig. 1. Hence $\mathcal{F}$ takes the form $\mathcal{F} = \mathcal{A}^J$ with $\mathcal{A} = \{a, b, c, d\}$. In this case we view the fingerprint as an array $\mathbf{F} = \{F_{ij}, 1 \le i \le N,\ 1 \le j \le J\}$ and the probe fingerprint as an array $\mathbf{G} = \{G_{ij}, 1 \le i \le L,\ 1 \le j \le J\}$. We also write ϕ in vector form as $\phi = \{\phi_j,\ 1 \le j \le J\}$.

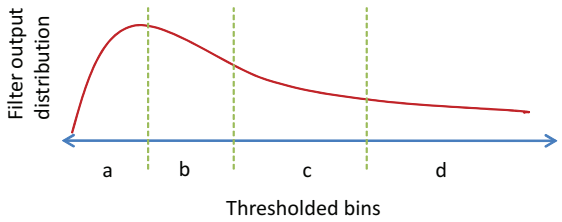


Fig. 1. 4-level quantization of each filter output in fingerprinting algorithm of [6]

Frame overlap causes strong dependencies between successive fingerprint components. The challenge is to develop a realistic and tractable model to account for such dependencies. See [3] for a general stationary model, and [4] for a binary fingerprint model (i.e., $|\mathcal{F}| = 2$).

4. STATISTICAL MODEL FOR ORIGINAL FINGERPRINT F

Since the frames are strongly temporally correlated, so are the fingerprint components. Our statistical model is as follows.

- $\mathbf{X}$ is a stationary process, hence so is the fingerprint process $\mathbf{F}$.
- For each time $1 \le i \le N$, the fingerprint components $F_{ij} = \phi_j(X_i), 1 \le j \le J$ are iid with marginal pdf P over the alphabet $\mathcal{A}$. (For the example of Fig. 1 applied to our audio database, we obtain $P = [0.281, 0.259, 0.248, 0.212]$.)
- The fingerprint components $\phi_j(X_i)$ and $\phi_{j'}(X_{i'})$ are mutually independent for each pair (i, i') and for each $j \ne j'$.
- $F_{ij} = \phi_j(X_i), 1 \le i \le N$ is an homogeneous Markov process for each j. The probability associated with fingerprint $\mathbf{f}$ is then

$$p(\mathbf{f}) = \prod_{j=1}^{J} \left[P(f_{1j}) \prod_{i=1}^{L-1} Q(f_{i+1,j} | f_{ij}) \right] \qquad (1)$$

where Q denotes the $|\mathcal{A}| \times |\mathcal{A}|$ transition probability matrix for $F_{i+1,j} = \phi_j(X_{i+1})$ given $F_{ij} = \phi_j(X_i)$. In our example,

$$Q = \begin{pmatrix} 0.7816 & 0.1650 & 0.0432 & 0.0101 \\ 0.1770 & 0.6029 & 0.1888 & 0.0313 \\ 0.0517 & 0.2108 & 0.5918 & 0.1458 \\ 0.0162 & 0.0460 & 0.1915 & 0.7463 \end{pmatrix}.$$

5. STATISTICAL MODELS FOR DEGRADED FINGERPRINT G

In the event the probe is related to some element of the database, we assume this relationship takes the following form. Let N_0 be an integer in $\{0, 1, 2, \cdots, N - L - 1\}$ representing a time offset. We assume the degradation channel from $\mathbf{X}$ to $\mathbf{Y}$ is a stationary stochastic mapping. Hence so is the channel from $\mathbf{F}$ to $\mathbf{G}$. We consider two models.

Order-0 Model. The simplest model for the channel is of the form

$$p_0(\mathbf{g}|\mathbf{f}, N_0) = \prod_{j=1}^{J} \prod_{i=1}^{L} W(g_{ij} | f_{i+N_0,j}) \qquad (2)$$

where W is the conditional distribution of G_{ij} given F_{ij}. This model implies that the errors on the fingerprint symbols are conditionally iid given $\mathbf{F}$. We refer to (2) as the order-0 (or memoryless) degradation model. To simplify the notation, we assume that $N_0 = 0$ below. For illustration, the transition probability matrix W for the 20% echo distortion attack considered later in this paper is

$$W = \begin{pmatrix} 0.8333 & 0.1505 & 0.0149 & 0.0012 \\ 0.1338 & 0.6991 & 0.1585 & 0.0085 \\ 0.0156 & 0.1900 & 0.6881 & 0.1063 \\ 0.0023 & 0.0184 & 0.1807 & 0.7986 \end{pmatrix}.$$

This memoryless model is relatively simple but not accurate. Indeed the errors on the fingerprint symbols are not only correlated over time (for each j), but also correlated conditioned on $\mathbf{F}$.

Order-1 Model. Our proposed model for capturing the correlation on the errors on fingerprint symbols is as follows:

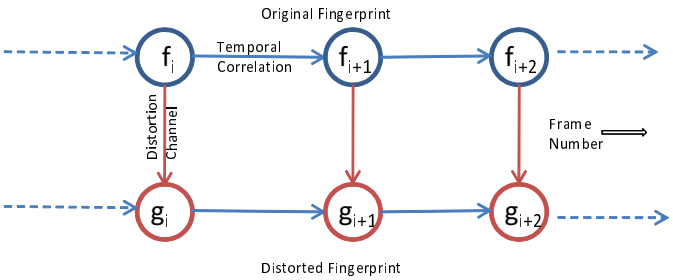


Fig. 2. Graphical model for order-1 statistical model on matched fingerprints.

- The J processes $\{(F_{ij}, G_{ij}),\ 1 \le i \le L\}$ are mutually independent and have the same distribution for all $j \in \mathcal{J}$.
- $(\mathbf{F}_{1:i}, G_{1:i-1,j}) \to (F_{i+1,j}, G_{ij}) \to G_{i+1,j}$ forms a Markov chain for each $j \in \mathcal{J}$.

We use the symbol V to denote the conditional pmf of $G_{i+1,j}$ given $(F_{i+1,j}, G_j)$. Hence V is an $|\mathcal{A}| \times |\mathcal{A}|^2$ stochastic matrix.

Based on the above model, the conditional distribution of $\mathbf{g}$ given $\mathbf{f}$ factors as

$$p_1(\mathbf{g}|\mathbf{f}) = \prod_{j=1}^{J} \left[W(g_{1j}|f_{1j}) \prod_{i=1}^{L-1} V(g_{i+1,j}|f_{i+1,j}, g_{ij}) \right] \quad (3)$$

where the subscript 1 on p denotes the order-1 model.

What we have done above is to define a graphical model (1) (3) for the joint process $(\mathbf{F}, \mathbf{G})$ which is loop-free and therefore lends itself to optimal inference. It is easy to estimate the conditional probability matrices W, V, etc. when the total number of unknown parameters ($O(|\mathcal{A}|^3)$) is low relative to the number of training data. Note that V depends on the type of distortion used, e.g., audio compression or equalization.

Hybrid Model. Different models present different advantages, and it can be beneficial to combine them. Consider for instance the geometrically weighted combination of p_0 and p_1 in (2) and (3),

$$p_\lambda(\mathbf{g}|\mathbf{f}) = \frac{1}{C(\lambda)} p_0^{1-\lambda}(\mathbf{g}|\mathbf{f})\, p_1^{\lambda}(\mathbf{g}|\mathbf{f}) \quad (4)$$

where $\lambda \in [0,1]$ is a tradeoff parameter and $C(\lambda) = \int p_0^{1-\lambda}\, p_1^{\lambda}$ is a normalization constant.

6. MAXIMUM-LIKELIHOOD DECODER

We now consider the maximum-likelihood decoder matched to the proposed model. This would be the optimal decoder if the statistical model was exact. As in [3], it will be mathematically more convenient to consider a *list decoder* that returns all indices $m \ge 1$ for which the negative loglikelihood score

$$d(\mathbf{f}(m), \mathbf{g}) = -\ln p(\mathbf{g}|\mathbf{f}(m))$$

falls below a predetermined threshold $L\tau$. If no such m can be found, the decoder outputs a no-match decision, i.e., $\psi(\mathbf{g}) = 0$.

The ML decoding metric for the three statistical models (2), (3), and (4) admits a relatively simple form owing to the factorized forms. We respectively obtain

$$\begin{aligned} d_0(\mathbf{f}, \mathbf{g}) &= \sum_{j=1}^{J} \sum_{i=1}^{N} -\ln W(g_{ij}|f_{ij}) \\ d_1(\mathbf{f}, \mathbf{g}) &= \sum_{j=1}^{J} \Big[-\ln W(g_{1j}|f_{1j}) \\ &\qquad - \sum_{i=1}^{L-1} \ln V(g_{i+1,j}|f_{i+1,j}, g_{ij}) \Big] \\ d_\lambda(\mathbf{f}, \mathbf{g}) &= (1-\lambda) d_0(\mathbf{f}, \mathbf{g}) + \lambda d_1(\mathbf{f}, \mathbf{g}). \end{aligned}$$

The Hamming metric

$$d_H(\mathbf{f}, \mathbf{g}) = \sum_{j=1}^{J} \sum_{i=1}^{N} \mathbb{1}\{g_{ij} \ne f_{ij}\}$$

is obtained as a special case of the order-0 model when W is the so-called $|\mathcal{A}|$-ary symmetric channel.

7. GLRT

The stochastic matrices W and V are estimated for a given type of distortion (e.g., audio compression). In reality the decoder may not know the type of distortion channel used, in which case the decoding metric may not depend on the distortion channel. We consider two approaches.

- Average the matrices W and V over the different types of distortion used, i.e., produce "one size fits all" matrices W and V.
- Use a variant of the Generalized Likelihood Ratio Test (GLRT) where the maximum likelihood over all channels is evaluated: the list decoder outputs all m such that
$$-\max_{\theta} \ln p(\mathbf{g}|\mathbf{f}(m), \theta) < L\tau$$
where θ is the index of the distortion channel. This test uses the collection of matrices $\{W_\theta\}$ and $\{V_\theta\}$.

8. NUMERICAL RESULTS

In our simulations, we used the SPB fingerprints of [6] with $J = 8$ and $L = 4$. Three distortions are considered here: 64 kbps audio compression using WMA encoding, insertion of a 20% echo, and bandpass filtering in the 0.4 – 4 kHz range. Five other distortions were considered but only the stronger attacks are reported here.

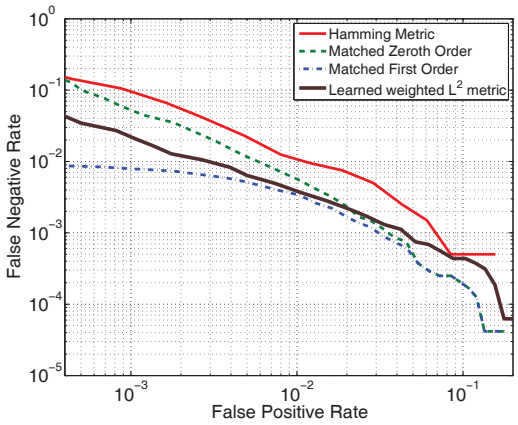
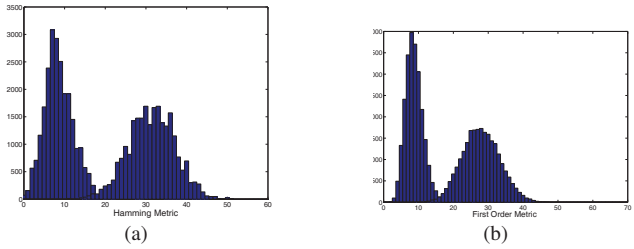


Fig. 3. ROC curves for 20% echo distortion for Hamming decoding metric, order-0 matched metric, hybrid order-1 matched metric, and learned weighted L^2 metric of [8].



(a) (b)

Fig. 4. Distributions of decoding metrics for 20% echo distortion for (a) Hamming decoding metric, and (b) Hybrid order-1 matched metric.

To characterize the performance of competing decoders, we show the Receiver Operating Characteristic (ROC) curve parameterized by the test threshold τ. We also show the distribution of the decoding metric for matching and non-matching pairs. The decoder is applied to pairs of audio sequences obtained from our audio database. The pairs are either matching (one audio segment is a distorted version of the other) or non-matching (the two audio segments are unrelated).

As shown in Fig. 3, the improvements of the order-0 decoding metric over the Hamming metric are strong. Additional improvements are obtained using the hybrid decoder which exploits order-1 memory. Improvements are also obtained relative to the decoder of [8]. It is seen in Fig. 4a,b that the distributions of the decoding metrics for matching and non-matching pairs are more concentrated and better separated when correlations are properly modeled, which explains the better ROC curves of Fig. 3.

Finally, the ROC performance of the decoders for an unknown degradation channel is shown in Fig. 5. The "averaged order-1 decoder" is obtained by using matrices W and V that are estimated from an ensemble of degradations (in this case the three degradations listed above). The best decoder is the GLRT decoder. Its performance compares well with that of the ML decoder that knows the degradation channel (called Genie-aided in the figure).

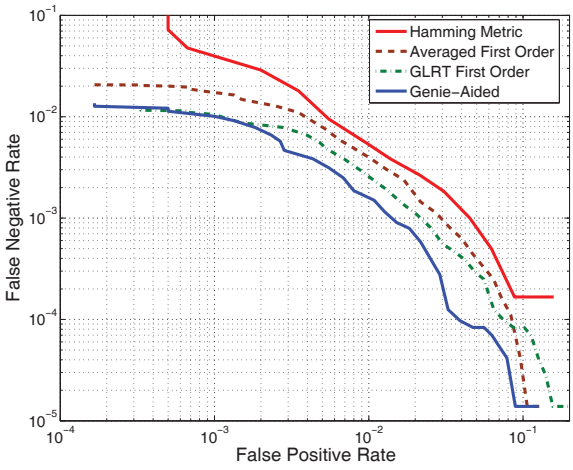


Fig. 5. ROC curve in case of an unknown degradation type.

9. CONCLUSION

Significant improvements in decoding performance have been obtained by selecting a decoding metric matched to the fingerprint statistics and the degradation process. This is already apparent from the order-0 (memoryless) degradation model, and further improvements are obtained using an order-1 model. The improvements are consistent with theory explaining the benefits of carefully choosing the decoding metric for content ID [3, 4].

10. 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] P. Moulin,"Statistical Modeling and Analysis of Content Identification," *Proc. IEEE Workshop on Information Theory and Applications (ITA)*, San Diego, CA, Jan-Feb. 2010.

[4] A. L. Varna and M. Wu, "Modeling and Analysis of Correlated Binary Fingerprints for Content Identification," *IEEE Trans. IFS*, Vol. 6, No. 3, pp. 1146—1159, Sep. 2011.

[5] T. Holotyak, S. Voloshynovskiy, O. Koval, and F. Beekhof, "Fast physical object identification based on unclonable features and soft fingerprinting," *Proc. ICASSP 2011*, Prague, Czech Republic, May 2011.

[6] D. Jang, C. D. Yoo, S. Lee, S. Kim, and T. Kalker, "Pairwise Boosted Audio Fingerprint," *IEEE Trans. Information Forensics and Security*, Vol. 4, No. 4, pp. 995—1004, Dec. 2009.

[7] S. Lee, C. D. Yoo, and T. Kalker, "Robust Video Fingerprinting Based on Symmetric Pairwise Boosting," *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 19, No. 9, pp. 1379—1388, Sep. 2009.

[8] D. Jang, C. D. Yoo, and T. Kalker, "Distance Metric Learning for Content Identification," *IEEE Trans. Information Forensics and Security*, Vol. 5, No. 4, pp. 932—944, Dec. 2010.