PERFORMANCE OF PHILIPS AUDIO FINGERPRINTING UNDER DESYNCHRONISATION

Neil J. Hurley, Félix Balado, Elizabeth P. McCarthy, Guénolé C.M. Silvestre
School of Computer Science and Informatics
University College Dublin, Ireland

ABSTRACT

An audio fingerprint is a compact representation (robust hash) of an audio signal which is linked to its perceptual content. Perceptually equivalent instances of the signal must lead to the same hash value. Fingerprinting finds application in efficient indexing of music databases. We present a theoretical analysis of the Philips audio fingerprinting method under desynchronisation for correlated stationary Gaussian sources.

1 INTRODUCTION

There are inevitable trade-offs between the size and the properties of a robust hash or fingerprint. An audio fingerprinting scheme which has proved to be remarkably robust is the Philips method, proposed by Haitsma *et al* [1] based on quantizing differences of energy measures from overlapped short-term power spectra. In this paper we examine the theoretical performance of the Philips method under desychonisation through a statistical model. This approach allows the influence of the system parameters to be studied and optimization strategies to minimize the probability of bit error of the hash to be tackled. Some previous work has tackled performance analysis of the Philips method through a statistical model. For example a model was proposed by Doets and Lagendijk [2], for the case in which the signal to be hashed is uncorrelated Gaussian noise. This was used to evaluate the performance of the fingerprinting method under distortion, but the results only apply to i.i.d. sources. The important issue of performance analysis under desynchronization, which to our knowledge has not been previously tackled, constitutes the main contribution of this paper.

2 DESYNCHRONIZATION ERROR ANALYSIS

In the Philips method, the length N input signal $\mathbf{x}$ is divided into overlapped frames before hashing. Let L be the number of samples in a single frame, Δ be the number of non-overlapping samples between two frames and $\mathbf{x}_n$ be the input signal corresponding to the n^{th} frame. We define the degree of overlap as $\theta \triangleq 1 - \Delta/L$, where $\theta \in (0,1)$, and higher θ corresponds to greater overlap.

A window $\mathbf{w}$ is applied to $\mathbf{x}_n$ before computing the power spectrum. The spectrum is divided into 32 frequency bands on a logarithmic scale. Denoting by $E_n(m)$ the energy of frequency band m for input frame $\mathbf{x}_n$, an unquantised hash value is given by $D_n(m) \triangleq [E_n(m) - E_n(m+1)] - [E_{n-1}(m) - E_{n-1}(m+1)]$, with $m = 0, 1, \cdots, 31$ and frames $n = 0, 1, 2, \cdots$. The variables $D_n(m)$ completely determine the system, as the binary hash value $F_n(m) \in \{0,1\}$ corresponding to frame n and band m is computed as $F_n(m) \triangleq u\left(D_n(m)\right)$, with $u(\cdot)$ the unit step function. This method can be expressed as a quadratic form on the extended vector $\tilde{\mathbf{x}}_n \triangleq (x[(n-1)\cdot \Delta + 1], \cdots, x[n\cdot\Delta + L])^T$ for $n = 0, 1, 2, \ldots$, which includes all the components of the overlapping vectors $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$ and which is of length $M \triangleq L + \Delta$, such that $D_n(m) = \tilde{\mathbf{x}}_n^T\ \mathrm{Q}(m)\ \tilde{\mathbf{x}}_n$, for a band and window dependent matrix $\mathrm{Q}(m)$, whose derivation is described in [3].

Desynchronization is the potential lack of alignment between the original framing used in the acquisition stage and the framing that takes place in the identification stage. The Philips algorithm has a high degree of overlapping in order to counteract desynchronization. Nevertheless, this strategy has a cost of generating a long hash sequence, which may be costly to store and compare. Consider a situation in which the signal fed to the system is desynchronized by k samples, with $k \in \{-\Delta/2+1, \cdots, \Delta/2\}$ and assuming $\Delta/2$ integer for simplicity. It is sufficient to consider this range since a desynchronization of Δ just shifts all the fingerprint bits one position. A desynchronization by k samples results in a distorted hash value $D'_n(m)$ and then a certain probability of bit error. It is convenient to write $D_n(m)$ and $D'_n(m)$ as quadratic forms in the same extended Gaussian vector

$$\underline{\mathrm{x}}_n \triangleq (x[(n-1)\Delta - \Delta/2 - 1], \cdots, x[n\Delta + L + \Delta/2])^T, \tag{1}$$

of length $M + \Delta - 1$, which we assume is distributed as $\underline{\mathrm{x}}_n \sim \mathcal{N}(\mathbf{0}, \underline{\mathrm{Z}})$. We write these quadratic forms as $D_n(m) = \underline{\mathrm{x}}_n^T\ \underline{\mathrm{Q}}_0(m)\ \underline{\mathrm{x}}_n$ and $D'_n(m) = \underline{\mathrm{x}}_n^T\ \underline{\mathrm{Q}}_k(m)\ \underline{\mathrm{x}}_n$.

Letting $S \triangleq D_n(m)$ and $V \triangleq D'_n(m)$, we note that stationarity implies that Z is Toeplitz and hence the mean of S and V is zero. Assume that S and V can be jointly modeled as a bivariate normal distribution centered at the origin with correlation coeficient ρ, which can be written as

$$\rho_k(m) = \frac{\mathrm{tr}\left[\underline{\mathrm{Z}}\,\underline{\mathrm{Q}}_0(m)\,\underline{\mathrm{Z}}\,\underline{\mathrm{Q}}_k(m)\right]}{\mathrm{tr}\left[(\mathrm{Z}\,\mathrm{Q}(m))^2\right]}. \tag{2}$$

Defining $\epsilon_n^k(m) \triangleq \{F'_n(m) \neq F_n(m) \mid k\}$ we have that

$$\begin{aligned}\Pr[\epsilon_n^k(m)] &= \frac{1}{2}(\Pr[S > 0|V \leq 0] + \Pr[S \leq 0|V > 0]) \\ &= \frac{1}{\pi}\ \arccos(\rho_k(m))\end{aligned}$$

In order to average over k, we assume that k is uniformly distributed. An upper bound is based on assuming that $\rho_k(m) \geq 0$ which holds as long as $\Pr\left[\epsilon_n^k(m)\right] \leq 1/2$. Hence, $\arccos(\cdot)$ is a concave function and we may apply Jensen's inequality [4] to upper bound the probability of bit error at frame n and band m as

$$\Pr[\epsilon_n(m)] = \mathrm{E}\left[\frac{1}{\pi}\ \arccos(\rho(m))\right] \leq \frac{1}{\pi}\ \arccos(\mathrm{E}[\rho(m)]). \tag{3}$$

2.1 Optimal Window and Asymptotic Performance

Notice that (2) implies that, even for i.i.d. input, the bound (3) on the probability of bit error is dependent on the window $\mathbf{w}$ (through the matrices $\underline{\mathrm{Q}}_i(m)$) and on the band m. It is possible to minimize (3) with respect to the window $\mathbf{w}$ in the i.i.d. case. In [3], we show that the optimal window satisfies a non-linear system of equations that can be solved numerically using a generalised eigenvalue solver. Furthermore, this optimisation can be exploited to obtain a closed-form bound on P_e solely dependent on the overlap level θ, that holds as $L \to \infty$ and $\theta \to 1$, namely,

$$P_e \leq \frac{1}{\pi} \arccos\left(\frac{\sin((1-\theta)\pi)}{(1-\theta)\pi}\right). \tag{4}$$

3 EXPERIMENTAL RESULTS

Firstly, we obtain results on zero-mean Gaussian i.i.d. signals. For a range of values of overlap θ, P_e is averaged over all desynchronization levels k in the range $-\Delta/2 + 1, \ldots, \Delta/2$ and over all bands. In Figure 1, this is illustrated both for the von Hann window and for a window obtained by averaging the band-dependent optimal windows over all bands. The empirical values are obtained by averaging over 2×10^5 frames. We see that with $\theta = 0.945$ and the optimized window we can get the same P_e as with $\theta = 0.955$ and the von Hann window. This overlap decrease accounts for a reduction of approximately 20% in the hash size. In any case, these results show that the von Hann window is very close to optimal. In Figure 2, we also apply our analysis to 5-second excerpts of three real audio signals used in [1]. We observe that the empirical results are very similar to each other and very similar to the i.i.d. Gaussian case. The performance of the i.i.d. case acts as a natural upper bound for desynchronization. This bound is tight due to the weak dependence of the results on the autocovariance matrix. Therefore, we can use (4) to predict accurately performance for any signal, especially when frames sizes have realistic (large) values. Notice that this expression has been obtained for the best possible window in the i.i.d. case, and for this reason the plot lies below the i.i.d. Gaussian empirical values which correspond to the von Hann window.

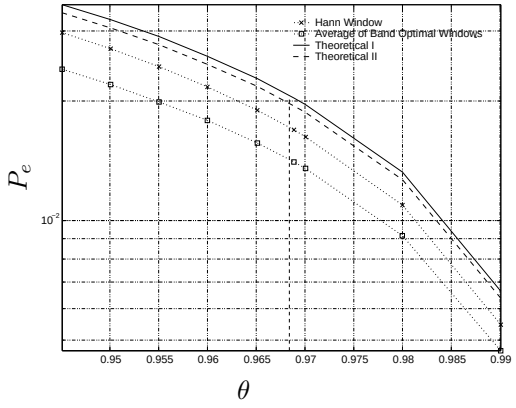

Figure 1. Probability of bit error under uniform desynchronization versus overlap level, using an i.i.d. Gaussian hashed signal. Empirical results correspond to the von Hann window and to the averaged band-optimal windows, respectively. Frame duration, $T_f \approx 0.3$ seconds. Theoretical I and II refer to two theoretical upper bounds applicable to this situation (see [3]).

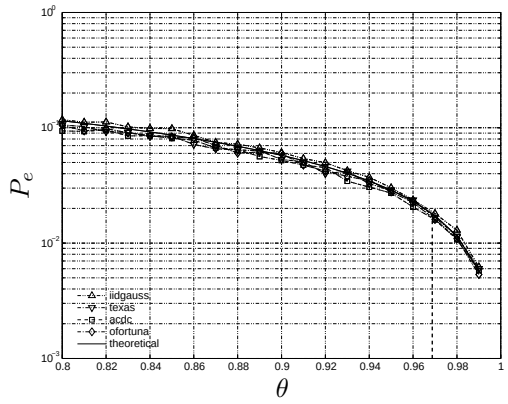

Figure 2. Probability of bit error under uniform desynchronization versus overlap level, using 5-second excerpts of three real audio signals and i.i.d. Gaussian signal. Frame duration $T_f = 0.3$ seconds, von Hann window. The theoretical result is the asymptotic performance for an optimal window.

4 REFERENCES

[1] J. Haitsma, T. Kalker, and J. Oostven, "Robust audio hashing for content identification," in *Procs. of the International Workshop on Content-Based Multimedia Indexing*, (Brescia, Italy), October 2001.

[2] P. J. O. Doets, "Modelling a robust audio fingerprinting system," in *Technical Report, Delft University of Technology*, June 2004.

[3] F. Balado, N. Hurley, E. McCarthy, and G. Silvestre, "Performance analysis of robust audio hashing," *IEEE Trans. on Information Forensics and Security*, vol. 2, pp. 1556–6013, June 2007.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.