



PERFORMANCE OF PHILIPS AUDIO FINGERPRINTING UNDER DESYNCHRONISATION

Neil J. Hurley, Félix Balado, Elizabeth P. McCarthy and Guéno   C.M. Silvestre

UCD School of Computer Science and Informatics
University College Dublin, Ireland

neil.hurley@ucd.ie

Abstract

We present a theoretical analysis of the Philips audio fingerprinting method under desynchronisation for correlated stationary Gaussian sources. We provide closed-form analytical upper bounds for the probability of bit error of the hash and verify these bounds on real audio signals.

1. Introduction

- **Fingerprint (a.k.a., robust hash)** of an audio signal:
 - compact representation of the signal, linked to its perceptual content
 - perceptually equivalent instances of same audio signal must *approximately* lead to same robust hash
 - **applications**: content tracking in p2p networks; authentication with distortion constraints; indexing of multimedia databases
- Robust audio fingerprinting scheme: **Philips method** [1]
- **Contributions of this work**:
 1. we study the bit error rate of the hash of Philips method under desynchronisation, using stationary correlated Gaussian input signals and show that this analysis applies to real audio signals

2. Statistical Model of Philips Audio Fingerprinting

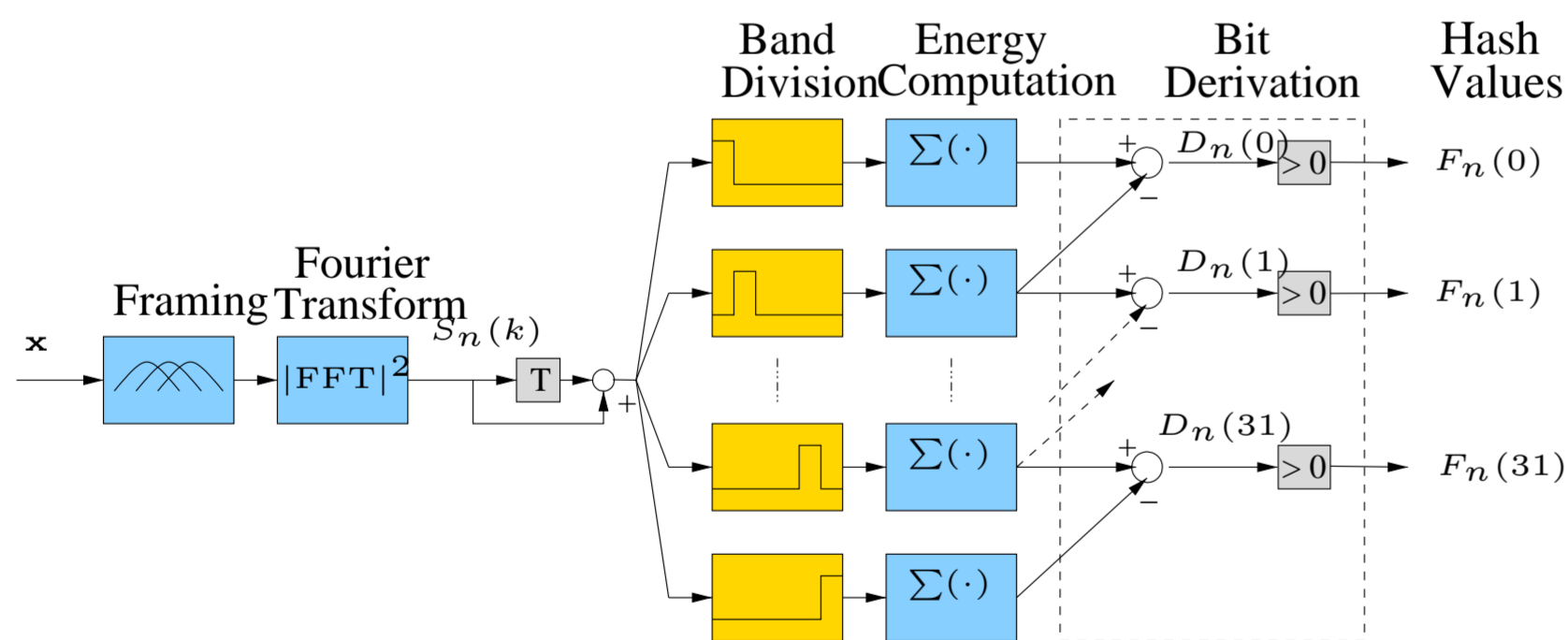


Figure 1: Philips audio fingerprinting

- Input signal x divided into overlapped frames x_n and multiplied by a window vector w . The degree of overlap is given by θ , such that $\theta = 1$ implies full overlap. In practise $\theta \approx 0.97$
- Binary hash value $F_n(m)$ (frame n , frequency band m) \approx sign of double differential of energy measure $D_n(m)$ with respect to time and frequency
- Operation flow: 1) *acquisition stage*: signals of interest hashed and $F_n(m)$ stored in database; 2) *identification stage*: input signal hashed and compared to stored values.
- **Our approach**: modeling $D_n(m)$ by expressing it as a **quadratic form**

$$D_n(m) = \mathbf{x}_n^T \mathbf{Q}(m) \mathbf{x}_n \quad (1)$$

- For $\mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{Z})$: pdf of $D_n(m)$ is a weighted sum of χ^2 distributions
- Approximation: $D_n(m)$ Gaussian-distributed with parameters

$$E[D_n(m)] = \text{tr}[\mathbf{Z}\mathbf{Q}(m)], \quad \text{Var}[D_n(m)] = 2 \text{tr}[(\mathbf{Z}\mathbf{Q}(m))^2]$$

3. Theoretical Performance Analysis

- **Dysynchronisation**: the potential lack of alignment between original framing in the acquisition stage and that in the identification stage.
- Signal is desynchronized by k samples, with $k \in \{-\Delta/2 + 1, \dots, \Delta/2\}$, assuming $\Delta/2 \in \mathbb{Z}$ and $D_n(m)$ computed using the indices of x between $(n-1)\Delta + 1 + k$ and $n\Delta + L + k$ instead of the correct original ones.
- The result is a distorted hash value $D'_n(m)$ and then a certain probability of bit error.
- We find matrices $\mathbf{Q}_0(m)$ and $\mathbf{Q}_k(m)$ such that

$$D_n(m) = \mathbf{x}_n^T \mathbf{Q}_0(m) \mathbf{x}_n, \quad D'_n(m) = \mathbf{x}_n^T \mathbf{Q}_k(m) \mathbf{x}_n. \quad (2)$$

- Model joint distribution of $D_n(m)$ and $D'_n(m)$ as **bivariate normal distribution**.
- Error event $\epsilon_n(m) \triangleq \{F'_n(m) \neq F_n(m)\}$; using the model above

$$\Pr[\epsilon_n(m)] = \frac{1}{\pi} \arccos(\rho_k(m)) \quad (3)$$

with $\rho_k(m)$ the correlation coefficient, given by

$$\rho_k(m) = \frac{\text{tr}[\mathbf{Z}\mathbf{Q}_0(m)\mathbf{Z}\mathbf{Q}_k(m)]}{\text{tr}[(\mathbf{Z}\mathbf{Q}(m))^2]}, \quad (4)$$

3.1 Optimal Window

- The window vector w is contained in (3) in the matrices $\mathbf{Q}_0(m)$ and $\mathbf{Q}_k(m)$
- Minimizing (3) wrt w leads to a non-linear system of equations in terms of matrices \mathbf{B}_j that can be solved through the following iterative system using an eigenvalue solver:

$$\frac{1}{\Delta} \sum_k \left[2\mathbf{B}_k^{(i-1)} - \mathbf{B}_{\Delta+k}^{(i-1)} - \mathbf{B}_{\Delta-k}^{(i-1)} \right] \mathbf{w}^{(i)} = 2\rho^{(i)} \left[\mathbf{B}_0^{(i-1)} - \mathbf{B}_{\Delta}^{(i-1)} \right] \mathbf{w}^{(i)} \quad (5)$$

3.2 Asymptotic Performance

- Solving 5 exactly for the case $L \rightarrow \infty$ and the overlap $\theta \rightarrow 1$ leads to an asymptotic upper bound for the synchronisation error:

$$P_e \leq \frac{1}{\pi} \arccos\left(\frac{\sin((1-\theta)\pi)}{(1-\theta)\pi}\right). \quad (6)$$

4. Experimental Results

- Our analysis is applied to 5-second excerpts of three real audio signals used : ‘‘O Fortuna’’ by Carl Orff, ‘‘Say what you want’’ by Texas, and ‘‘Whole lotta Rosie’’ by AC/DC (16 bits, 44.1 kHz).
- 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.
- We can use (6) to predict accurately performance for any signal, especially when frames sizes have realistic (large) values.

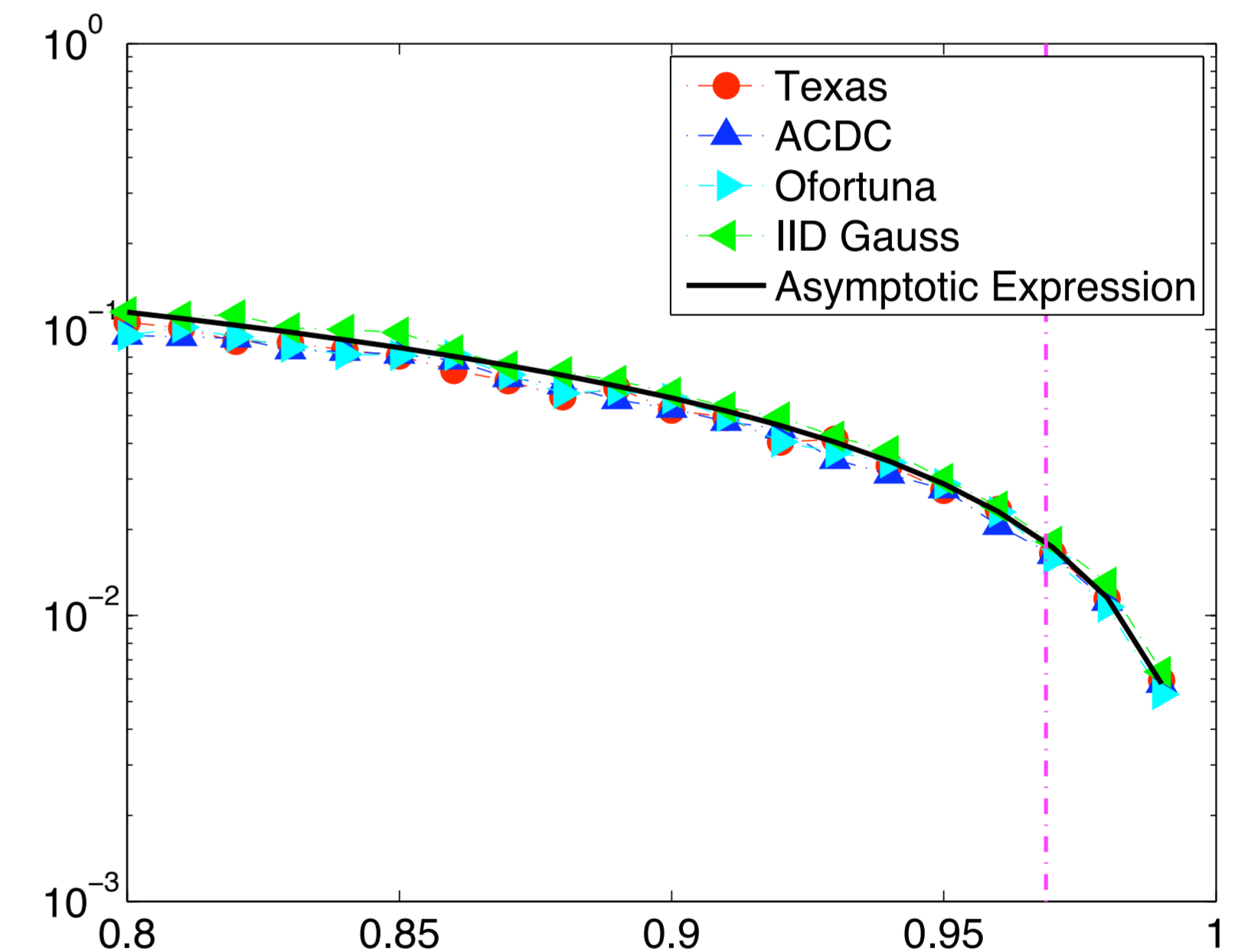


Figure 2: Probability of bit error under uniform desynchronization versus overlap level, using 5-second excerpts of three real audio signals and i.i.d. Gaussian signal. $T_f = 0.3$ seconds, von Hann window. The theoretical result is the asymptotic performance for an optimal window. The vertical dashed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

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.