

# Performance Analysis of Robust Audio Hashing

Félix Balado\*, *Member, IEEE*, Neil J. Hurley, Elizabeth P. McCarthy and Guénolé C.M.  
Silvestre, *Member, IEEE*

## Abstract

We present a novel theoretical analysis of the Philips audio fingerprinting method proposed by Haitsma, Kalker and Oostveen [1]. Although this robust hashing algorithm exhibits very good performance, the method has only been partially analyzed in the literature. Hence, there is a clear need for a more complete analysis which allows both performance prediction and systematic optimization. We examine here the theoretical performance of the method for Gaussian inputs by means of a statistical model. Our analysis relies on formulating the unquantized fingerprint as a quadratic form, which affords a systematic way to compute the model parameters. We provide closed-form analytical upper bounds for the probability of bit error of the hash for two relevant scenarios: noise addition and desynchronization. We show that these results are useful when applied to real audio signals.

**EDICS Category: APP-IDEN**

\*Corresponding author

This work was supported by Enterprise Ireland Advanced Technologies Research Program, research grant ATRP2002/230 and by the European Commission through the IST Programme under Contract IST-2002-507609 SIMILAR. Parts of this manuscript were presented at the SPIE conference Multimedia Content Analysis, Management, and Retrieval, January 2006.

F. Balado: E-mail: felix@ihl.ucd.ie, Phone: +353 1 716 2454; N.J. Hurley: E-mail: neil.hurley@ucd.ie, Phone: +353 1 716 2903; E.P McCarthy: E-mail: liz@ihl.ucd.ie, Phone: +353 1 716 2454; G.C.M. Silvestre: E-mail: guenole.silvestre@ihl.ucd.ie, Phone: +353 1 716 2852. The authors are with the UCD School of Computer Science and Informatics, University College Dublin (National University of Ireland). Common address and fax number: UCD School of Computer Science, Belfield Campus, Dublin 4, Ireland, Fax: +353 1 269 7262.

# Performance Analysis of Robust Audio Hashing

## I. INTRODUCTION

A multimedia fingerprint is a compact representation of a multimedia signal which is linked to its perceptual content<sup>1</sup>. This representation is parallel to a cryptographic hash in the sense that it almost uniquely represents the signal, but it is different from it in the sense that different instances of the same signal which are perceptually equivalent must approximately lead to the same hash value. Due to this, fingerprinting is also known as *robust* hashing, and these two terms will be used interchangeably throughout this paper. Fingerprinting helps to identify multimedia signals in noisy environments. Due to this property, it finds application in areas of forensic interest such as content tracking in peer-to-peer networks [2] and multimedia authentication [3]. In the first scenario, fingerprinting allows to pin down the true identity of files which have been possibly transcoded and/or renamed before making them available on a public network. With respect to the second one, robust hashing plays a role in the so-called quantize-and-embed authentication schemes [4], in which a distortion-resistant low-dimensional descriptor of a signal is embedded within the same signal by means of data hiding. In order to assess these scenarios, it is fundamental to investigate the performance of the fingerprinting methods they rely on.

There are inevitable trade-offs between the size and the properties of the fingerprint, or robust hash. Ideally, it should be as small as possible without loss of discriminatory power, and robust to imperceptible distortions of the original signal. In many cases synchronization—that is, guaranteeing that different instances of the same signal at the input of the fingerprinting algorithm are properly aligned—becomes an important concern for fingerprinting. In other cases an added difficulty is the desire to identify the hashed signal from small portions of the original. In fact, it is these particularities that distinguish fingerprinting from classical rate-distortion problems. One particular audio fingerprinting scheme which has proved to be remarkably robust is the so-called Philips method (also referred to as Streaming Audio Fingerprinting) proposed by Haitsma et al. [1]. A measure of its success is that this algorithm has been implemented in different commercial products already. The method is based on quantizing differences of energy measures from overlapped short-term power spectra. This staggered and overlapped arrangement

<sup>1</sup>Notice that the term fingerprinting is also used with different meaning in data hiding applications.

allows for excellent robustness and synchronization properties, apart from allowing identification from subfingerprints computed from short segments of the original signal.

In this paper we examine the theoretical performance of the Philips method through a statistical model. This approach allows in some cases the influence of the system parameters to be studied, and optimization strategies aimed at minimizing the probability of bit error of the hash to be tackled. The good performance of Philips method is achieved at the cost of a rather large fingerprint size; in [1] the method parameters were chosen heuristically, which leaves open the possibility of tuning them to decrease this hash size while maintaining performance, or to improve performance for a fixed hash size. To our knowledge, the only prior work aimed at improving the performance of the Philips method is the one by Park et al. [5]. This work however is empirical and not founded on a theoretical basis, and it relies on the addition of an extra filtering stage.

There is nevertheless some previous work oriented to obtain statistical models and performance analyses of the Philips method. A performance analysis for false positives was already presented by Haitsma et al. in [6], under the assumption that the hash bits are independent and identically distributed (i.i.d.) and compensating for the shortcomings of this hypothesis with a correction factor. A more elaborate model of the algorithm was proposed by Doets and Lagendijk [7]–[10], for the case in which the signal to be hashed is uncorrelated Gaussian noise. This model hinges on an equivalent rearrangement of the Philips algorithm, and exploits its similarity with the Welch method of power spectrum estimation [11] to produce a model of conditioned probabilities. Of these works, only [10] tackles the issue of evaluating the performance of the fingerprinting method under distortion, but the results therein only apply to i.i.d. sources. Our paper tackles the case with correlation for stationary signals, from which that previous analysis follows as a particular case. Also, the important issue of performance analysis under desynchronization, which to our knowledge has not been previously tackled, constitutes the other main contribution of this paper.

Our analysis follows the work of Doets and Lagendijk in their rearrangement of the algorithm, and in their use of periodograms for estimating power spectra. We then take an approach inspired in the work of Johnson and Long [12], who model overlapped periodograms as quadratic forms. We show that this enables the parameters of the model to be obtained straightforwardly—even for correlated sources and arbitrary windows—which was one of the problems of the approach in [8]. Using the model proposed, we analyze the probability of bit error of the hash for the two previously mentioned scenarios.

The remainder of this paper is structured as follows. In Section II we develop the statistical model of the Philips method. In Section III we exploit this model to undertake the performance analysis of the fingerprinting method under noise distortion and desynchronization. The results are empirically verified

in Section IV and, finally, Section V draws the conclusions of our analysis.

*a) Notation:* Lower case bold face letters such as  $\mathbf{x}$  represent column vectors, while matrices are represented by upper case Roman letters such as  $X$ . The real eigenvalues of a symmetric  $L \times L$  matrix are denoted by  $\lambda_0 \cdots \lambda_{L-1}$ , sorted in an arbitrary order.  $\text{diag}(\mathbf{x})$  is a matrix with the elements of  $\mathbf{x}$  in the diagonal and zero elsewhere.  $\text{tr} X$  denotes the trace of  $X$ . The 2-norm of  $\mathbf{x}$  is denoted as  $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$ , where the superindex  $T$  denotes transposition, while the same notation for matrices refers to the Frobenius norm,  $\|X\| = \sqrt{\text{tr}[X^T X]}$ . The symbol  $\odot$  denotes the Hadamard (entry-wise) product of two matrices or vectors of the same size, and  $\otimes$  denotes the Kronecker (or direct) product. Null vectors and matrices are denoted by  $\mathbf{0}$  and  $O$ , respectively, while  $\mathbf{1}$  denotes an all-ones vector.

## II. STATISTICAL MODEL OF PHILIPS METHOD

We will first describe the operation of the Philips method. As aforementioned, we will use the equivalent rearrangement of the algorithm described in [9] and schematically illustrated in Figure 1. Let the input signal to the hash method, or hashed signal, be denoted by the vector of samples  $\mathbf{x} = (x[1], \dots, x[N])^T$ . This signal is divided into overlapped frames before processing it. If  $L$  is the number of samples in a single frame and  $\Delta$  the number of nonoverlapping samples between frames, then the  $L$ -length vector  $\mathbf{x}_n$  formed by the elements of  $\mathbf{x}$  used in the computations corresponding to the  $n^{\text{th}}$  frame is given by

$$\mathbf{x}_n \triangleq (x[n \cdot \Delta + 1], \dots, x[n \cdot \Delta + L])^T, \quad n = 0, 1, 2, \dots$$

If the samples in  $\mathbf{x}$  have been taken every  $1/f_s$  seconds, the duration of a frame is  $T_f \triangleq L/f_s$  seconds.

We will find convenient to define the degree of overlap as

$$\theta \triangleq 1 - \frac{\Delta}{L},$$

where  $\theta \in (0, 1)$ , and higher  $\theta$  corresponds to greater overlap. We will assume that the number of frames processed is  $N_f \triangleq (N - L + \Delta)/\Delta = (\frac{N}{L} - 1)\frac{1}{1-\theta} + 1$ , which we take to be integer for simplicity. Each framed signal  $\mathbf{x}_n$  is weighted next by a window of weights  $\mathbf{w} \triangleq (w[1], \dots, w[L])^T$  before taking its fast Fourier transform (FFT), as depicted schematically in Figure 1. Then, the vector at the input of the FFT for frame  $\mathbf{x}_n$  is just  $\mathbf{w} \odot \mathbf{x}_n = \text{diag}(\mathbf{w}) \mathbf{x}_n$ .

The spectrum is then divided into  $N_b + 1$  bands according to a logarithmic law. In the remainder of this paper we use the frequency band division given in [1], for which  $N_b = 32$ . Denoting by  $E_n(m)$  the energy of band  $m$  for input frame  $\mathbf{x}_n$ , the variables used to compute the corresponding hash are obtained as

$$D_n(m) \triangleq [E_n(m) - E_n(m + 1)] - [E_{n-1}(m) - E_{n-1}(m + 1)], \quad (1)$$

with  $m = 0, 1, \dots, N_b - 1$  and  $n = 0, 1, 2, \dots$ . Notice that these variables 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

$$F_n(m) \triangleq u(D_n(m)), \quad (2)$$

with  $u(\cdot)$  the unit step function. In the *acquisition stage*, that is, the first time a given signal is hashed, these values are stored for later comparison in the *identification stage*, in which a signal to be recognized is hashed using the fingerprinting method. We will assume in Section III that two scenarios may take place: a) the signal  $\mathbf{x}$  is distorted by noise addition; and b)  $\mathbf{x}$  is not distorted, but it is desynchronized at the input of the hash method, i.e. the frame boundaries do not match exactly those for which the corresponding stored fingerprints were computed.

*b) Model:* Our strategy will be based on modeling the continuous random variables  $D_n(m)$ . As (1) involves spectral energy measures, we need an estimator of the power spectrum for modeling purposes. As in [7], we use the periodogram estimator  $S_n(k)$  of the power spectrum of the windowed signal at the  $n^{\text{th}}$  frame, which is given by

$$S_n(k) = \frac{1}{\|\mathbf{w}\|^2} \left| \sum_{i=0}^{L-1} x_n[i+1]w[i+1] \exp\left(-j2\pi i \frac{k}{L}\right) \right|^2, \quad (3)$$

for  $k = 0, \dots, L-1$ , and where  $\|\mathbf{w}\|^2$  acts as an energy normalization factor. Now, following [12], (3) can be rewritten in matrix form as

$$S_n(k) = \mathbf{x}_n^T \mathbf{M}(k) \mathbf{x}_n, \quad (4)$$

where the  $L \times L$  matrix  $\mathbf{M}(k)$  is defined as

$$\mathbf{M}(k) \triangleq \frac{1}{\|\mathbf{w}\|^2} \Omega \mathbf{N}(k) \Omega.$$

The  $L \times L$  matrix  $\mathbf{N}(k)$  is defined such that its entry at position  $(i, j)$  is given by  $\cos(2\pi(i-j)k/L)$ , and  $\Omega \triangleq \text{diag}(\mathbf{w})$ . The imaginary parts of the exponentials in (3) are not included in the definition of  $\mathbf{N}(k)$ , as they cancel out in the positive semidefinite quadratic form (4). Notice that  $\mathbf{N}(k)$  is symmetric Toeplitz, as its entries only depend on  $|i-j|$ , and circulant, as  $\cos(2\pi ik/L) = \cos(2\pi(i-L)k/L)$ .

With (4) in hand, we will see next that it is straightforward to express  $D_n(m)$  in (1) as a quadratic form as well. Defining first  $B(m)$  as the set of integers indexing the periodogram samples in frequency band  $m$ , we can write an estimate of the energy in this band as

$$E_n(m) = \sum_{k \in B(m)} S_n(k) = \mathbf{x}_n^T \left[ \sum_{k \in B(m)} \mathbf{M}(k) \right] \mathbf{x}_n. \quad (5)$$

We define next the  $L \times L$  matrices

$$\mathbf{R}(m) \triangleq \frac{1}{\|\mathbf{w}\|^2} \left[ \sum_{k \in B(m)} \mathbf{N}(k) - \sum_{k \in B(m+1)} \mathbf{N}(k) \right], \quad (6)$$

and

$$\mathbf{P}(m) \triangleq \Omega \mathbf{R}(m) \Omega. \quad (7)$$

As  $\mathbf{R}(m)$  is the summation of symmetric circulant Toeplitz matrices it is also symmetric circulant Toeplitz, but  $\mathbf{P}(m)$  is only symmetric in general. Now, plugging (5) in (1) and using (7) we obtain

$$D_n(m) = \mathbf{x}_n^T \mathbf{P}(m) \mathbf{x}_n - \mathbf{x}_{n-1}^T \mathbf{P}(m) \mathbf{x}_{n-1}. \quad (8)$$

In order to write (8) as a single quadratic form, we define next the extended vector

$$\tilde{\mathbf{x}}_n \triangleq (x[(n-1) \cdot \Delta + 1], \dots, x[n \cdot \Delta + L])^T, \quad n = 0, 1, 2, \dots, \quad (9)$$

which includes all the components of the overlapping vectors  $\mathbf{x}_n$  and  $\mathbf{x}_{n-1}$  and which is of length  $M \triangleq L + \Delta$ . Notice that we assume the convention of padding with zeros (9) for indices less than or equal to zero. We also define now the  $M \times M$  matrix

$$\mathbf{Q}(m) \triangleq \begin{bmatrix} -\mathbf{P}(m) & \vdots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \mathbf{P}(m) \end{bmatrix}, \quad (10)$$

which is formed by adding  $-\mathbf{P}(m)$  at the position  $(1, 1)$  of an empty  $M \times M$  matrix with  $\mathbf{P}(m)$  at the position  $(\Delta + 1, \Delta + 1)$ . The matrix (10) is formally obtained as explained in Appendix A. Notice that the matrix structure in (10) is depicted with almost no overlap between  $-\mathbf{P}(m)$  and  $\mathbf{P}(m)$  for visualization purposes; the actual overlap is usually a lot higher. Also,  $\mathbf{Q}(m)$  is symmetric because  $\mathbf{P}(m)$  is so. Using (10) and (9) we can finally write (8) as

$$D_n(m) = \tilde{\mathbf{x}}_n^T \mathbf{Q}(m) \tilde{\mathbf{x}}_n. \quad (11)$$

For  $\tilde{\mathbf{x}}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{Z})$ , (11) is consequently a quadratic form in a Gaussian vector. The distribution of this r.v. may be expressed exactly as a weighted sum of  $\chi^2$  distributions [13]. Nevertheless, it can be verified empirically that, for large  $M$ , a Gaussian distribution suffices to approximate the probability density function (pdf) of (11). This approximation is supported by the Central Limit Theorem (CLT) for i.i.d. signals; for locally correlated signals a broader version of the CLT can also be invoked. Note in any case that, like the Gaussian distribution, the pdf of (11) is even due to the symmetry of the setting. Last but not least, the Gaussian approximation is a lot easier to handle analytically than the weighted sum of  $\chi^2$ . The adequacy of this assumption will be confirmed in Section III by means of the results of the performance analysis undertaken.

The relevance of the quadratic form (11) is that it enables us to easily compute the parameters of the Gaussian model. The expectation and variance of the variables (11) for a Gaussian input are simply computed as [14]

$$E[D_n(m)] = \text{tr} [Z Q(m)], \quad (12)$$

$$\text{Var}[D_n(m)] = 2 \text{tr} [(Z Q(m))^2]. \quad (13)$$

These expressions hold for zero-mean  $\tilde{\mathbf{x}}_n$ , but they are extensible to nonzero mean variables. Also, (12) holds regardless the actual distribution of  $\tilde{\mathbf{x}}_n$ . Notice that the model's parameters can be obtained using (12) and (13) for any arbitrary window  $\mathbf{w}$ .

*c) Dependencies:* A rigorous theoretical analysis will have to tackle the dependencies between the variables  $D_n(m)$ . Notice that the vectors  $\tilde{\mathbf{x}}_n$  overlap for several consecutive frame indices  $n$ , and this overlap is higher the smaller  $\Delta$  is. A small value of  $\Delta$  is required for resilience to desynchronization. Then, for a fixed band  $m$  the consecutive r.v.'s  $D_n(m)$  are intuitively strongly dependent. In fact, developing (8) recursively we arrive at

$$D_n(m) = \mathbf{x}_n^T \mathbf{P}(m) \mathbf{x}_n - \sum_{i=0}^{n-1} D_i(m). \quad (14)$$

From (14) it is clear that the random variables  $\{D_n(m)\}$ , with  $m$  fixed, do not conform to a Markov chain. Nevertheless, the dependence evident in (14) is weak in the long term, that is, between random variables roughly spaced  $L$  samples apart.

Similarly, given a fixed frame  $n$ , the dependencies among the r.v.'s  $\{D_n(m)\}$ , with  $m = 0, 1, \dots, N_b - 1$ , needs consideration. As can be observed in (11), the random vector  $\tilde{\mathbf{x}}_n$  which generates all these random variables is exactly the same. Although it is well known that the FFT is good at decorrelating the signal over the spectral components, the overlap of different bands and frames requires some attention. For  $m' \neq m$ , the two corresponding quadratic forms in the same Gaussian vector in (11) are independently distributed iff  $Q(m) Z Q(m') = \mathbf{O}$  [14]. In practice, as we have argued about the fact that (11) is well approximated by a Gaussian distribution, it is enough to verify that the random variables are uncorrelated. For zero-mean Gaussian  $\tilde{\mathbf{x}}_n$  the correlation (covariance) between the quadratic forms  $D_n(m)$  and  $D_n(m')$  is given by [14]

$$\text{Cov}[D_n(m), D_n(m')] = 2 \text{tr} [Z Q(m) Z Q(m')]. \quad (15)$$

Notice by substituting the independence condition into (15) that two independent quadratic forms on the same Gaussian variables are uncorrelated, although the reverse is not true in general. Applying (15) it is possible to see that the dependencies are not negligible between adjacent indices.

As we will see in the next section, we will overcome the difficulties posed to our performance analysis by these dependencies by resorting to upper bounding strategies.

### III. PERFORMANCE ANALYSIS

As discussed in Section I, a signal presented to the algorithm in the identification stage may differ from the corresponding original indexed in the database during the acquisition stage, and it is the objective of robust hashing to overcome the problems associated with identifying the original in this situation. In this section we use the model proposed in Section II to examine the probability of bit error of the hash ( $P_e$ ) in such a situation. We identify two scenarios which result in a nonzero probability of error: a) the signal is distorted by noise addition; b) the signal is not distorted, but it is desynchronized at the input of the hashing algorithm, i.e. the frame boundaries do not match exactly those for which the corresponding stored fingerprints were computed.

#### A. Noise Addition

In this section we examine the probability of error following the addition of zero mean white Gaussian noise  $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{I})$  to a Gaussian hashed signal assumed zero-mean and stationary. Then,  $\tilde{\mathbf{x}}_n \sim \mathcal{N}(\mathbf{0}, Z)$ , where the  $M \times M$  covariance matrix  $Z = \mathbb{E}[\tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T]$  is Toeplitz with diagonal elements  $\sigma_x^2$ . The nonstationary case will be discussed later using our analysis of the stationary case. In order to establish a point of operation we define the signal-to-noise ratio (SNR)  $\xi \triangleq \frac{\sigma_x^2}{\sigma_g^2}$ . If the signal presented to the system in the identification stage at frame  $n$  is  $\tilde{\mathbf{x}}_n + \mathbf{g}$  instead of  $\tilde{\mathbf{x}}_n$ , the hash value (11) becomes

$$\begin{aligned} D'_n(m) &= (\tilde{\mathbf{x}}_n + \mathbf{g})^T \mathbf{Q}(m) (\tilde{\mathbf{x}}_n + \mathbf{g}) \\ &= D_n(m) + 2 \tilde{\mathbf{x}}_n^T \mathbf{Q}(m) \mathbf{g} + \mathbf{g}^T \mathbf{Q}(m) \mathbf{g}. \end{aligned}$$

Before proceeding, let us define for notational simplicity the variables  $S \triangleq D_n(m)$  and  $T \triangleq 2 \tilde{\mathbf{x}}_n^T \mathbf{Q}(m) \mathbf{g} + \mathbf{g}^T \mathbf{Q}(m) \mathbf{g}$ . The expectation of  $S$  is computed using (12), which, using (41) in Appendix A to expand  $\mathbf{Q}(m)$  and applying the circular property of the trace, is

$$\begin{aligned} \mathbb{E}[S] &= \text{tr}[Z \mathbf{Q}(m)] \\ &= -\text{tr}[\mathbf{U}Z\mathbf{U}^T \mathbf{P}(m)] + \text{tr}[\mathbf{V}Z\mathbf{V}^T \mathbf{P}(m)] = 0, \end{aligned}$$

where the second expression is zero due to the fact that the covariance matrix  $Z$  is Toeplitz and then  $\mathbf{U}Z\mathbf{U}^T = \mathbf{V}Z\mathbf{V}^T$ . For the same reason  $\mathbb{E}[D'_n(m)] = 0$ , and therefore  $\mathbb{E}[T] = 0$ . It is straightforward to show that  $S$  and  $T$  are uncorrelated, that is,  $\mathbb{E}[S \cdot T] = 0$ , because  $\mathbf{g}$  is zero-mean and white. We

have argued that we can model  $D_n(m)$  (and then  $D'_n(m)$ ) by means of the Gaussian distribution. In this case  $T$  must be modeled as a Gaussian too, and then, from uncorrelation, it follows that  $S$  and  $T$  are independent.

In order to complete the statistical characterization we just need the variances of the random variables involved. As  $D_n(m)$  and  $D'_n(m)$  are just quadratic forms on zero mean Gaussian variables we may use (13) to obtain their variances. Doing so we obtain  $\sigma_S^2 \triangleq \text{Var}[S] = 2 \text{tr} [(Z Q(m))^2]$ . Similarly, as  $\text{Var}[D'_n(m)] = 2 \text{tr} [(Z + \sigma_g^2 \mathbf{I}) Q(m)]^2$  it follows from independence between  $S$  and  $T$  that

$$\sigma_T^2 \triangleq \text{Var}[T] = 2\sigma_g^2 \text{tr} [(2Z + \sigma_g^2 \mathbf{I}) Q^2(m)].$$

Denoting the error event at frame  $n$  and band  $m$  as  $\epsilon_n(m) \triangleq \{F'_n(m) \neq F_n(m)\}$ , we are ready now to compute the probability of error at an individual frame and band, which, recalling (2), may be written as

$$\Pr[\epsilon_n(m)] = \frac{1}{2} (\Pr[S + T > 0 | S \leq 0] + \Pr[S + T \leq 0 | S > 0]), \quad (16)$$

as  $\Pr[S \leq 0] = \Pr[S > 0] = 1/2$ . Using the Gaussian  $\mathcal{Q}$ -function  $\mathcal{Q}(x) \triangleq 1/\sqrt{2\pi} \int_x^\infty \exp(-v^2/2) dv$ , we can compute (16) as

$$\Pr[\epsilon_n(m)] = \int_{-\infty}^0 \mathcal{Q}\left(\frac{-s}{\sigma_T}\right) f_S(s) ds + \int_0^\infty \mathcal{Q}\left(\frac{s}{\sigma_T}\right) f_S(s) ds,$$

with  $f_S(s)$  a Gaussian distribution with zero mean and variance  $\sigma_S^2$ . Through a change of variable we have finally that

$$\begin{aligned} \Pr[\epsilon_n(m)] &= 2 \int_0^\infty \mathcal{Q}\left(\frac{s}{\sigma_T}\right) f_S(s) ds = \frac{1}{\pi} \arctan\left(\frac{\sigma_T}{\sigma_S}\right) \\ &= \frac{1}{\pi} \arctan\left(\sqrt{\frac{\text{tr}[(2\bar{Z} + \mathbf{I}) Q^2(m)]}{\text{tr}[(\bar{Z} Q(m))^2]}}\right), \end{aligned} \quad (17)$$

with  $\bar{Z} \triangleq \frac{1}{\sigma_g^2} Z$ . It is shown in Appendix B that (17) is bounded by

$$\Pr[\epsilon_n(m)] \leq \frac{1}{\pi} \arctan\left(\sqrt{\left(2 + \frac{\gamma_u}{\xi}\right) \frac{\gamma_u}{\xi}}\right), \quad (18)$$

where  $\gamma_u$  only depends on  $\sigma_x^2$  and on the minimum eigenvalue of  $Z$ . Note that (17) and (18) apply to the particular stationary situation in which the hashed signal is i.i.d. This case can also be expressed exactly in terms of the SNR using  $\bar{Z} = \xi \mathbf{I}$  in (17). In this case we have that

$$\Pr[\epsilon_n(m)] = \frac{1}{\pi} \arctan\left(\sqrt{\left(2 + \frac{1}{\xi}\right) \frac{1}{\xi}}\right). \quad (19)$$

The equation (19) was also previously obtained by Doets and Lagendijk in [10] through a frequency analysis. Notice however that our analysis highlights that both the bound (18) and the exact expression (19) are independent of  $Q(m)$ , and then both of the type of window used and of the band  $m$ . This means that the same expressions would be obtained placing any arbitrary full-rank real matrix in (11), which sets a bound on the performance of any hashing system whose unquantized hash values can be put as a quadratic form on the input signal.

It is also important to realize that, as the variables  $\{D_n(m)\}$  —and therefore also  $\{F_n(m)\}$ — exhibit dependencies (see the end of Section II), the probabilities (17) or (19) cannot just be averaged to obtain the overall probability of bit error. Nevertheless, we can resort to a union-bound argument that allows the average probability of bit error to be upper bounded as

$$P_e = \frac{1}{N_f} \frac{1}{N_b} \Pr[\cup_{n,m} \epsilon_n(m)] \leq \frac{1}{N_f} \frac{1}{N_b} \sum_{n=0}^{N_f-1} \sum_{m=0}^{N_b-1} \Pr[\epsilon_n(m)]. \quad (20)$$

As the error events are uncorrelated in the long term then the bound must be reasonably sharp for large  $N_f$ . Also, using the results above, the union bound is bounded in turn by the upper bound in (18) for generic stationary Gaussian signals, and given exactly by (19) for i.i.d. Gaussian signals.

### B. Desynchronization

Even in the absence of additive distortions, it is unlikely that the frame alignment of the input signal will correspond exactly to the one used for computing the original hash stored in a reference database. We mean by desynchronization this lack of alignment between the original framing used in the acquisition stage and the framing that takes place in the identification stage. Counteracting desynchronization is the reason why the Philips robust hashing algorithm has such a high degree of overlapping. Nevertheless, this strategy pays the price of generating a long hash sequence, which may be costly to store and compare. In this section we analyze the system performance under desynchronization. We investigate how we can tune the system parameters in order to minimize the probability of bit error caused by desynchronization for a given overlap, or, equivalently, to minimize the hash length for a given target probability of error.

Consider a situation in which the signal fed to the system to compute (11) is desynchronized by  $k$  samples, with  $k \in \{-\Delta/2 + 1, \dots, \Delta/2\}$  and assuming  $\Delta/2$  integer for simplicity. Notice that there is no loss of generality in choosing this restricted range, as stronger desynchronization is dealt with at the application level (i.e., database search) [6]. For instance, a desynchronization of  $\Delta$  just shifts all the fingerprint bits one position (cf. (9)). Assuming desynchronization by  $k$  samples, the vector used in the computation of the hash value corresponding to  $D_n(m)$  uses the  $L$  indices of  $\mathbf{x}$  between  $(n-1)\Delta + 1 + k$

and  $n\Delta + L + k$  instead of the correct original ones as in (9). The result is a distorted hash value  $D'_n(m)$  and then a certain probability of bit error.

Following a similar procedure as in Section II to obtain (11), we may write  $D_n(m)$  and  $D'_n(m)$  as quadratic forms in the same Gaussian vector by defining the extended vector

$$\underline{\mathbf{x}}_n \triangleq (x[(n-1)\Delta + 2 - \Delta/2], \dots, x[n\Delta + L + \Delta/2])^T \quad (21)$$

of length  $M + \Delta - 1$ , that includes all possible desynchronization indices for  $\tilde{\mathbf{x}}_n$  when  $k \in \{-\Delta/2 + 1, \dots, \Delta/2\}$ , and that is distributed as  $\underline{\mathbf{x}}_n \sim \mathcal{N}(\mathbf{0}, \underline{\mathbf{Z}})$ . We just need now two  $(M + \Delta - 1) \times (M + \Delta - 1)$  matrices

$$\underline{\mathbf{Q}}_0(m) \triangleq \begin{bmatrix} \boxed{\mathbf{Q}(m)} \end{bmatrix}, \quad \underline{\mathbf{Q}}_k(m) \triangleq \begin{bmatrix} \boxed{\mathbf{Q}(m)} \end{bmatrix}, \quad (22)$$

obtained by placing  $\mathbf{Q}(m)$  at the positions  $(\Delta/2, \Delta/2)$  and  $(\Delta/2 + k, \Delta/2 + k)$ , respectively, of an empty  $(M + \Delta - 1) \times (M + \Delta - 1)$  matrix. The way to obtain analytically the matrices (22) can be found in Appendix A. Using (21) and (22) we can achieve our objective of writing

$$D_n(m) = \underline{\mathbf{x}}_n^T \underline{\mathbf{Q}}_0(m) \underline{\mathbf{x}}_n, \quad D'_n(m) = \underline{\mathbf{x}}_n^T \underline{\mathbf{Q}}_k(m) \underline{\mathbf{x}}_n. \quad (23)$$

Letting again  $S \triangleq D_n(m)$  and  $V \triangleq D'_n(m)$ , our model is based now on the fact that  $S$  and  $V$  are clearly correlated. Assuming again that  $S$  and  $V$  can be modeled by zero-mean Gaussian distributions, then they can also be jointly modeled as a bivariate normal distribution centered at the origin, that is, with pdf

$$f_{S,V}(s, v) = \frac{1}{2\pi\sigma_S\sigma_V\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{s^2}{\sigma_S^2} - \frac{2\rho sv}{\sigma_S\sigma_V} + \frac{v^2}{\sigma_V^2}\right)\right). \quad (24)$$

In order to compute the parameters of this model we just need to observe (23). The variances required are just  $\sigma_S^2 = \sigma_V^2 = 2 \operatorname{tr}[(\underline{\mathbf{Z}}\mathbf{Q}(m))^2]$ . In order to compute the correlation coefficient  $\rho$  we need the covariance between  $S$  and  $V$ , which is just (cf. (15))

$$\operatorname{Cov}[S, V] = 2 \operatorname{tr}[\underline{\mathbf{Z}}\underline{\mathbf{Q}}_0(m)\underline{\mathbf{Z}}\underline{\mathbf{Q}}_k(m)].$$

As  $\rho = \operatorname{Cov}[S, V]/(\sigma_S \cdot \sigma_V)$ , then

$$\rho_k(m) = \frac{\operatorname{tr}[\underline{\mathbf{Z}}\underline{\mathbf{Q}}_0(m)\underline{\mathbf{Z}}\underline{\mathbf{Q}}_k(m)]}{\operatorname{tr}[(\underline{\mathbf{Z}}\mathbf{Q}(m))^2]}, \quad (25)$$

where we have made explicit the dependence of  $\rho$  on  $k$  and  $m$ . We can now write the probability of bit error of the hash resulting from a desynchronization of  $k$  samples. Defining  $\epsilon_n^k(m) \triangleq \{F'_n(m) \neq F_n(m) \mid k\}$  we have that

$$\Pr[\epsilon_n^k(m)] = \frac{1}{2} (\Pr[S > 0 | V \leq 0] + \Pr[S \leq 0 | V > 0]). \quad (26)$$

Using (24) to compute (26), this probability is

$$\begin{aligned} \Pr[\epsilon_n^k(m)] &= \int_{-\infty}^0 \left( \int_0^{\infty} f_{S,V}(s,v) ds \right) dv + \int_0^{\infty} \left( \int_{-\infty}^0 f_{S,V}(s,v) ds \right) dv \\ &= 2 \int_0^{\infty} \int_{-\infty}^0 f_{S,V}(s,v) ds dv. \end{aligned} \quad (27)$$

Finally, evaluating the integral (27) we obtain

$$\Pr[\epsilon_n^k(m)] = \frac{1}{\pi} \arccos(\rho_k(m)). \quad (28)$$

However, (28) is still conditioned to a particular desynchronization value  $k$ . In order to average (28) over  $k$ , we will assume that  $k$  is uniformly distributed, that is,  $\Pr[k] = 1/\Delta$ , for  $k = -\Delta/2+1, \dots, \Delta/2$ . As this expectation is difficult to handle analytically due to the nonlinearity of  $\arccos(\cdot)$ , we will resort to an upper bound based on assuming  $\rho_k(m) \geq 0$ . This is not a very demanding assumption, because as long as the system does not work catastrophically (that is, as long as  $\Pr[\epsilon_n^k(m)] \leq 1/2$ ) then (28) implies that the correlation coefficient has to be nonnegative. Within the indicated range  $\arccos(\cdot)$  is a concave function, and then we may apply Jensen's inequality [15] to upper bound the probability of bit error at frame  $n$  and band  $m$  as

$$\Pr[\epsilon_n(m)] = \mathbb{E} \left[ \frac{1}{\pi} \arccos(\rho(m)) \right] \leq \frac{1}{\pi} \arccos(\mathbb{E}[\rho(m)]). \quad (29)$$

The expectation required in (29) can be readily computed from (25) and the uniform characterization of the random variable  $k$ , as  $\mathbb{E}[\rho(m)] = \frac{1}{\Delta} \sum_k \rho_k(m)$ . The expression (29) can be used in turn to upperbound the probability of bit error of the hash under desynchronization by means of the union bound, just as in (20).

It is possible to lower bound (25) for any  $Q(m)$  using the lower bound to a Rayleigh quotient, proceeding similarly to Appendix B for upper bounding (44). In this way we obtain an upper bound to (28), as  $\arccos(\cdot)$  is strictly decreasing. Unfortunately, this bound on (25) is not guaranteed to be positive, and then it cannot be readily applied to (29).

*1) Optimal Window for Band  $m$ :* Notice that (25) implies that, even for i.i.d. input, the bound (29) on the probability of bit error is now dependent on the window  $\mathbf{w}$  and on the band  $m$ , differently to what happened in Section III-A (cf. (19)). We will center next our attention on minimizing (29) with respect to the window  $\mathbf{w}$  in the i.i.d. case. This study may not seem at first sight relevant to scenarios with correlation. Nevertheless, we will show in Section III-B3 and Appendix D that this optimization can also be exploited to obtain a closed-form bound on  $P_e$  solely dependent on the overlap level  $\theta$ . It is necessary to stress that this bound will apply to any Gaussian signal regardless of its autocovariance,

as the i.i.d. scenario is the worst case for the synchronization problem. Indeed, correlated signals will naturally lead to less hashing errors in the presence of desynchronization. For the same reason one may think that desynchronization must lead to more errors with realizations of i.i.d. Gaussian signals than with real audio signals. Before proceeding, we must remark that in the original proposal of the Philips algorithm a von Hann window was employed without further considerations about its optimality for the hashing problem. Other different windows were studied in [9], but mainly for tractability reasons.

In the case we are considering in this subsection  $\underline{Z} = \sigma_x^2 \mathbf{I}$ , and (25) becomes

$$\rho_k(m) = \frac{\text{tr} [\underline{Q}_0(m) \underline{Q}_k(m)]}{\|\underline{Q}(m)\|^2}, \quad (30)$$

where  $\|\cdot\|$  is the Frobenius norm. As  $\arccos(\cdot)$  is strictly decreasing, maximizing  $E[\rho(m)]$  as a function of  $\mathbf{w}$  minimizes the bound (29). In order to make this dependence explicit, we will rewrite the average correlation coefficient in terms of the matrix  $\mathbf{R}(m)$  defined in (6). For notational simplicity, we will drop the index  $m$  in this section whenever the dependence is clear. Using the lower shift matrix  $\mathbf{C}_k$  defined in Appendix C it is easy to see that  $\mathbf{C}_k \mathbf{w}$  is just the window  $\mathbf{w}$  with its elements shifted  $k$  places and padded with zeros. Then,

$$\mathbf{v}_k \triangleq \mathbf{w} \odot \mathbf{C}_k \mathbf{w} \quad (31)$$

is the Hadamard product of  $\mathbf{w}$  with itself shifted  $k$  places. Relying on this definition it is shown in Appendix C that the average autocorrelation correlation coefficient can be written in terms of the quadratic forms

$$A_k \triangleq \mathbf{v}_k^T (\mathbf{R} \odot \mathbf{R}) \mathbf{v}_k \quad (32)$$

as

$$E[\rho] = \frac{1}{\Delta} \sum_k \frac{2A_k - A_{\Delta+k} - A_{\Delta-k}}{2(A_0 - A_\Delta)}. \quad (33)$$

Note that  $A_k$  is quadratic in  $\mathbf{v}_k$  but fourth order in  $\mathbf{w}$ . To maximize, we require the gradient of  $A_k$  with respect to  $\mathbf{w}$ . Using the chain rule, this can be written as

$$\begin{aligned} \nabla_{\mathbf{w}} A_k &= \nabla_{\mathbf{v}_k} A_k \cdot \mathbf{J}_{\mathbf{w}} \mathbf{v}_k \\ &= 2\mathbf{v}_k^T (\mathbf{R} \odot \mathbf{R}) \cdot \mathbf{J}_{\mathbf{w}} \mathbf{v}_k \end{aligned}$$

where  $\mathbf{J}_{\mathbf{w}} \mathbf{v}_k$  is the Jacobian matrix of the transformation  $\mathbf{v}_k(\mathbf{w})$  in (31), which, using matrix differential analysis, can be shown to be

$$\mathbf{J}_{\mathbf{w}} \mathbf{v}_k = \text{diag}(\mathbf{C}_k \mathbf{w}) + \text{diag}(\mathbf{w}) \mathbf{C}_k.$$

Hence, defining  $\mathbf{b}_k \triangleq 2(\mathbf{R} \odot \mathbf{R})\mathbf{v}_k$  the gradient can be put as

$$\begin{aligned}\nabla_{\mathbf{w}} A_k &= \mathbf{b}_k^T (\text{diag}(\mathbf{C}_k \mathbf{w}) + \text{diag}(\mathbf{w})\mathbf{C}_k) \\ &= \mathbf{w}^T (\mathbf{C}_k^T \text{diag}(\mathbf{b}_k) + \text{diag}(\mathbf{b}_k)\mathbf{C}_k).\end{aligned}\quad (34)$$

Define next the matrix  $B_k(\mathbf{w}) \triangleq \mathbf{C}_k^T \text{diag}(\mathbf{b}_k) + \text{diag}(\mathbf{b}_k)\mathbf{C}_k$ , where the dependence of  $B_k$  on  $\mathbf{w}$  is explicitly shown. Using (34) to obtain the gradient of (33) and equating to zero, we have that any extremum  $\rho^*$  (that is,  $\rho^*(m)$ ) of (33) must satisfy the third-order system of equations

$$\frac{1}{\Delta} \sum_k [2B_k(\mathbf{w}) - B_{\Delta+k}(\mathbf{w}) - B_{\Delta-k}(\mathbf{w})] \mathbf{w} = 2\rho^* [B_0(\mathbf{w}) - B_{\Delta}(\mathbf{w})] \mathbf{w}. \quad (35)$$

This nonlinear system can be solved iteratively as follows. Define  $\mathbf{w}^{(i)}$  to be the window at the  $i^{\text{th}}$  iteration and  $B_k^{(i)} \triangleq B_k(\mathbf{w}^{(i)})$ . At the  $i^{\text{th}}$  step, solve the generalized eigenvalue problem:

$$\frac{1}{\Delta} \sum_k [2B_k^{(i-1)} - B_{\Delta+k}^{(i-1)} - B_{\Delta-k}^{(i-1)}] \mathbf{w}^{(i)} = 2\rho^{(i)} [B_0^{(i-1)} - B_{\Delta}^{(i-1)}] \mathbf{w}^{(i)} \quad (36)$$

for the maximum generalized eigenvalue  $\rho^{(i)}$  and associated generalized eigenvector  $\mathbf{w}^{(i)}$ . This iterative method converges experimentally to an optimal window  $\mathbf{w}^*(m)$ , using, for example, a von Hann window as the initial window.

Since  $\mathbf{R} \odot \mathbf{R}$  is circulant, the coefficients  $\mathbf{b}_k$  can be computed in  $O(L \log L)$  operations using only the first row of  $\mathbf{R} \odot \mathbf{R}$  and the matrix multiplication algorithm for circulant matrices based on the FFT. Furthermore, for any  $k$  the matrix  $B_k$  is nonzero only in the  $k^{\text{th}}$  subdiagonal and superdiagonal. Hence, an eigenvalue solver for large sparse systems such as ARPACK [16] can be used to compute the largest eigenvalue/eigenvector pair of (36).

Note that the optimal window according to this strategy depends on  $\mathbf{R}(m)$  and hence on the frequency band  $m$ . In Figure 2, optimal windows obtained for a number of different frequency bands  $m$  are shown. Examining the figure we note the similarity of the optimal windows on different frequency bands. The reason for this similarity is due to the near-independence of the optimal window with respect to the band for large  $L$ , which is discussed in depth in Appendix D. The windows are generated for  $L = 13,300$  samples, corresponding to  $T_f \approx 0.3$  seconds for the sampling frequency  $f_s = 44.1$  kHz, which we assume by default.

2) *A Band-Independent Window*: The actual use of a different window for each frequency band — implied by the optimum given in the previous section— would require  $N_b$  times more computation than the original Philips algorithm. It is therefore not a tractable approach in practice. On the other hand, note that the average correlation coefficient over all bands and desynchronization levels  $\mu \triangleq \frac{1}{N_b} \sum_m \mathbb{E}[\rho(m)]$

determines a bound of the overall probability of bit error that can be used for optimization purposes. Indeed,

$$P_e \leq \frac{1}{N_f} \frac{1}{N_b} \sum_n \sum_m \Pr[\epsilon_n(m)] \leq \frac{1}{\pi} \arccos(\mu), \quad (37)$$

where the first inequality is just the union bound (cf. (20)) and the second one is due to the aforementioned concavity of  $\arccos(\cdot)$  for positive arguments. Unfortunately, the problem of finding a single window  $\mathbf{w}^*$  that yields a maximum average correlation coefficient  $\mu^*$ , that is, that minimizes (37), turns out to be both analytically and computationally difficult.

In order to overcome the difficulties entailed by this optimization, we resort to a suboptimal approach. This approach consists of employing the average window given by

$$\bar{\mathbf{w}}^* = \frac{1}{N_b} \sum_m \mathbf{w}^*(m), \quad (38)$$

with  $\mathbf{w}^*(m)$  the optimal normalized window for band  $m$ , whose corresponding average correlation coefficient is  $\rho^*(m)$ . This window works well empirically over all frequency bands. Notice that, due to the higher degree of freedom, it always holds that  $\mu^* \leq \frac{1}{N_b} \sum_m \rho^*(m)$ , that is, the result of maximizing the average correlation coefficient using a single window for all bands can never be better than when we use a different window per band. Then, we can use this inequality to gauge how good the approach (38) is. Letting  $\bar{\mu}^*$  denote the average correlation coefficient corresponding to the window given by (38), if  $\frac{1}{N_b} \sum_m \rho^*(m) \approx \bar{\mu}^*$ , then a bound with the solution (38) must be close to the globally optimum bound.

3) *Asymptotic Performance of Optimal Window:* As in Section III-A, where we made the probability of bit error explicitly dependent on  $\xi$  (SNR), it is desirable in this case to make  $P_e$  explicitly dependent on the degree of overlap  $\theta$ . In Appendix D we obtain such a relationship through an expression for  $E[\rho]$  in terms of  $\theta$  which holds for the optimal window as  $L \rightarrow \infty$ . According to the derivation in that appendix, which proceeds by approximating the shifted and zero-padded window  $C_k \mathbf{w}$  by a circularly shifted window  $\hat{C}_k \mathbf{w}$ , and by using the optimization equation (35), it is possible to write

$$P_e \leq \frac{1}{\pi} \arccos \left( \frac{\sin((1-\theta)\pi)}{(1-\theta)\pi} \right), \quad (39)$$

which represents an asymptotic upper bound on the bit error due to desynchronization in terms of the overlap, that holds as  $L \rightarrow \infty$  and  $\theta \rightarrow 1$ . This expression has been derived for i.i.d. Gaussian signals, but can be used as a guide for setting the system parameters on real signals, whose performance should be better in practice due to short-term correlations. It also suggests that provided  $L$  and  $\theta$  are sufficiently large, the system performance under desynchronization is independent of the frequency band.

Finally, it is also possible to apply the derivation of Appendix D to the von Hann window. As indicated therein, it is possible to verify that, provided the band-dependent term  $\frac{\lambda_0}{8(\lambda_0 + \lambda_1)}$  is small, the von Hann window is asymptotically close to optimal. Consequently, its performance will be in this case close to (39).

#### IV. EXPERIMENTAL RESULTS

##### A. Gaussian Signals

In this section we verify empirically the theoretical analyses undertaken for zero-mean stationary Gaussian signals. We present first the results in Section III-A obtained for zero-mean additive i.i.d. Gaussian noise independent of the hashed signal. Figure 3 shows the union upper bounds using (18) and (19) compared to corresponding empirical data. The band division used in the simulations is the original one in Philips method, but frames of  $T_f = 0.01$  seconds are used. The two correlated cases correspond to the  $M \times M$  tridiagonal covariance matrices  $Z^\pm = \sigma_x^2 \left( \mathbf{I} \pm \frac{1}{4}(\tilde{\mathbf{C}}_1 + \tilde{\mathbf{C}}_{-1}) \right)$ , which have the same minimum eigenvalue and, consequently, the same upper bound. This illustrates the fact that the bound (18) may be tighter or looser depending on the particular autocovariance matrix. We must also point out that the upper bound may not be useful if the minimum eigenvalue of  $Z$  is too small.

We present next empirical results for the desynchronization analysis undertaken in Section III-B. First, we verify the accuracy of the analysis for one band and a fixed desynchronization level. In Figure 4 (28) is shown for the normalized values of the desynchronization level  $k$ , i.e.,  $\alpha \triangleq k/\Delta$ , and for degrees of overlap  $\theta$  between 0 and 1. We see that the analytic expression (28) provides a good fit for the whole desynchronization range. The results are obtained using a von Hann window and averaging frames of length  $L = 441$  samples ( $T_f \approx 0.01$  seconds) with i.i.d. Gaussian hashed signal. Figure 4 also illustrates the relationship between the probability of bit error and the degree of overlap, i.e., when the number of nonoverlapping samples  $\Delta$  is large, the hash size is small and less robust to desynchronization, resulting in a higher probability of error.

We address next in Figure 5 the results for the overall union bound on  $P_e$  with desynchronization. The empirical plots illustrate the results for both the von Hann window and the optimized window (38), obtained by averaging the band-optimal windows. The performance achieved with these windows is compared with the bound on the union bound given by (37) using  $\bar{\mu}^*$ . The expression given by  $\frac{1}{\pi} \arccos\left(\frac{1}{N_b} \sum_m \rho^*(m)\right)$  (labeled as ideal union bound) is also illustrated; the closeness of these two theoretical plots confirms the accuracy of the suboptimal approach, as discussed in Section III-B2. Note that the bound on the union bound is relatively sharp, although it looks somewhat loose in Figure 5 due

to the scale. It can be seen that the suboptimal window based on the analytical bound (i.e., the average of band-optimal windows) results in a reduction of  $P_e$  with respect to the von Hann window for any given degree of overlap  $\theta$ . Alternatively, for a fixed  $P_e$  the optimized window allows the size of the hash to be reduced. For instance, looking at Figure 5 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, as reasoned at the end of Section III-B3.

### B. Real Audio Signals

Although the analysis in this paper has been centered on Gaussian signals, the robust hashing method by Haitsma et al. is a practical tool. Then, it is interesting to determine the validity of our analysis for real audio signals. In order to undertake this study, observe firstly that our performance analyses have been obtained as *averages* over the ensemble of signals with a given distribution. Nonetheless, when dealing with real audio signals we will have particular realizations, and so these averages (and the concept of stationarity) will have to be interpreted in an ergodic sense. Also, in general, real audio signals neither present Gaussian statistics nor are they stationary. Assuming that  $D_n(m)$  can still be modelled as Gaussian, the first issue above will cause inevitable inaccuracies with respect to the computation of (13), but (12) will still be exact for arbitrary distributions. With respect to the second issue, we will assume that it is possible to approximate a high percentage of the length of real audio signals by locally stationary stretches, so that our stationary analysis can be approximately applied. In this case, we will have for each frame  $n$  a possibly different autocovariance matrix  $Z_n$ , and  $\Pr[\epsilon_n(m)]$  will depend now on  $n$  unlike in the stationary case that we have analyzed.

A short preliminary discussion will help us to determine what behavior we can expect, showing at the same time the insights for real signals afforded by our Gaussian analysis. We have seen that the performance of the Philips hashing method is governed both for noise addition and desynchronization by ratios of matrix traces (cf. (44) and (25)). In order to shed light on the effect of these ratios on real signals, it is useful to assume for a moment that  $Q(m) = I$ . In this particular case, it is trivial to see that the noise addition performance depends on  $(2 \operatorname{tr} \bar{Z} + M)/\|\bar{Z}\|^2 = (2\xi + 1)/(\|\bar{Z}\|^2/M)$  whereas the desynchronization performance depends on  $L/(L + \Delta) = 1/(2 - \theta)$  (actually, a lower bound on the average correlation coefficient, and hence an upper bound to performance). That is, for fixed SNR ( $\xi$ ), the noise addition case will depend on the normalized Frobenius norm of the autocovariance matrix, whereas the upper bound in the desynchronization case will only depend on the degree of overlap for

any autocovariance matrix. Assuming that this discussion applies to the case with the full-matrices  $Q(m)$ , it suggests highly variable performance results for different real audio signals with noise addition, as the locally stationary autocovariance matrices can be quite different for different audio tracks. On the other hand, it also suggests that in the desynchronization case we should expect less inter-signal performance variability, owing to a weaker performance dependence on the autocovariance matrices.

Following the discussion above, in order to predict performance for a particular signal —especially, in the noise addition case— it will be necessary to estimate the autocovariance matrices corresponding to each locally stationary stretch. Note that it is necessary that these matrix estimates be positive definite. If the estimated autocovariances were not positive definite we could certainly obtain spurious predictions, as we might have negative or zero eigenvalues in (45) and the traces in (44) might take negative values, which bears no real-world meaning. Finally, there is an inherent tradeoff in the estimation of these autocovariance matrices. Notice that the estimation interval length must be short enough to guarantee local stationarity within intervals of that duration (whenever possible), but at the same time long enough in order to guarantee a good quality estimate.

We present next the results of predicting the hashing performance by applying our analysis to 5-second excerpts of three real audio signals already used in [1]: “O Fortuna” by Carl Orff, “Say what you want” by Texas, and “Whole lotta Rosie” by AC/DC (16 bits, 44.1 kHz). We may observe in Figure 6 the disparity of results for these three signals, that we have discussed is due to noticeably different autocovariance matrices. The theoretical results in the figure have been obtained using (17) and (20), as unfortunately the bound (18) turned out to be too loose with the real data used. The estimation interval lengths ( $T_a$ ) used are given in the plot; the best estimation interval size is of course dependent on the local stationarity features of the actual signal considered. The use of the full expression (17) limits the practical applicability of the theoretical predictions to small frame sizes, due to the matrix multiplication involved; nevertheless, the good fit of the theoretical results encourages the future development of practical approximations to that equation. The theoretical result for i.i.d. Gaussian is also plotted in order to stress the fact that it is not possible to predict performance for real signals using exclusively the SNR. We have only shown the adequacy of our analysis for three signals with noticeably different performance results; nevertheless, we have verified similar goodness of fit for a number of other signals. These results suggest that the true value of the variance of  $D_n(m)$  is never far from (13) with real signals, and that our assumption of Gaussianity of  $D_n(m)$  and local stationarity are also sufficient with real signals.

In the desynchronization case depicted in Figure 7 we observe the opposite effect discussed at the beginning of this section. It is remarkable that, for the same three real audio signals as before, the empirical

results are now very similar to each other and very similar to the i.i.d. Gaussian case. As discussed in Section III-B3, the performance of the latter case acts as a natural upper bound for desynchronization. This bound is tight due to the weak dependence of the results with respect to the autocovariance matrix that we have discussed. Therefore, we can use (39) to predict accurately performance for any signal, especially when frame 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. Other real signals tested exhibited empirical results very close to those in Figure 7, confirming the weak effect of the unstationarity of real audio signals in the desynchronization case suggested by the upper bound discussed.

## V. CONCLUSIONS

We have presented a new model of the Philips robust audio fingerprinting scheme [1]. Using this model on Gaussian hashed signals, we have undertaken an analysis of the probability of bit error of the hash under both i.i.d. Gaussian noise addition and desynchronization of the hashed signal. In the case of noise addition, we have derived an upper bound on the average probability of bit error for stationary Gaussian signals, which is tight for i.i.d. input. We have shown that it is possible to use these expressions to predict with reasonable accuracy performance with real audio signals, although frame sizes need to be small for undertaking the computations. In the desynchronization case, we have obtained an upper bound for correlated Gaussian signals. For the i.i.d. case, which is a worst case for desynchronization, we have studied a strategy to obtain a window that minimizes the previous upper bound. As an approximation to the solution of this problem, we have obtained a suboptimal window that performs slightly better than the Hann window used in the original method, which we have shown to be close to optimal. More importantly, this optimization procedure has allowed a closed-form theoretical expression to be obtained that can be used to predict the performance under desynchronization with generic input signals, in particular with real audio signals.

Lastly, although the model and the subsequent analysis have been developed with the method of Haitsma et al. [1] in mind, it may find application in other similar methods. Particularly, a similar analysis is applicable whenever the generation of the hash involves linear combinations of energy measures obtained by overlapping linear transforms. A linear transform leads to a quadratic form in terms of energy measurement, and the overlap can be taken into account following the steps undertaken in our proposal. This is the case, for example, in the audio fingerprinting methods proposed by Mihçak and Venkatesan [17] and Seo et al. [18].

## APPENDIX A

## COMPUTATION OF SHIFTED AND OVERLAPPED MATRICES

In this appendix we explain how to obtain analytically some of the matrices used throughout the paper. The way to achieve this is by means of auxiliary matrices defined by blocks. We define first the following  $L \times M$  auxiliary matrices ( $M = L + \Delta$ ):

$$\mathbf{U} \triangleq [\mathbf{I}_{L \times L} \mid \mathbf{O}_{L \times \Delta}], \quad \mathbf{V} \triangleq [\mathbf{O}_{L \times \Delta} \mid \mathbf{I}_{L \times L}], \quad (40)$$

with  $\mathbf{I}$  the identity matrix and  $\mathbf{O}$  the null matrix of size given by the subindices. Using (40) the matrix (10) can be written as

$$\mathbf{Q}(m) = -\mathbf{U}^T \mathbf{P}(m) \mathbf{U} + \mathbf{V}^T \mathbf{P}(m) \mathbf{V}. \quad (41)$$

In order to place a matrix  $M \times M$  at the position  $(\Delta/2 + j, \Delta/2 + j)$  of an empty  $(M + \Delta - 1) \times (M + \Delta - 1)$  matrix we just define the  $M \times (M + \Delta - 1)$  matrix

$$\underline{\mathbf{U}}_j \triangleq [\mathbf{O}_{M \times (\Delta/2 + j - 1)} \mid \mathbf{I}_{M \times M} \mid \mathbf{O}_{M \times (\Delta/2 - j)}]. \quad (42)$$

Then, the matrices (22) are computed using (42) as

$$\underline{\mathbf{Q}}_j(m) = \underline{\mathbf{U}}_j^T \mathbf{Q}(m) \underline{\mathbf{U}}_j, \quad (43)$$

using  $j = 0$  and  $j = k$ , respectively.

## APPENDIX B

## BOUND ON THE PROBABILITY OF ERROR FOR STATIONARY GAUSSIAN SIGNALS WITH NOISE ADDITION

We pursue here an upper bound to the probability (17) for a fixed SNR and over all covariance matrices. As  $\arctan(\cdot)$  is strictly increasing, bounds can be simply obtained by bounding the argument inside the square root in (17). This argument is

$$\psi \triangleq \frac{\text{tr} [(2\bar{\mathbf{Z}} + \mathbf{I}) \mathbf{Q}^2(m)]}{\text{tr} [(\bar{\mathbf{Z}} \mathbf{Q}(m))^2]} = \frac{(\text{vec } \mathbf{Q}(m))^T ((2\bar{\mathbf{Z}} + \mathbf{I}) \otimes \mathbf{I}) \text{vec } \mathbf{Q}(m)}{(\text{vec } \mathbf{Q}(m))^T (\bar{\mathbf{Z}} \otimes \bar{\mathbf{Z}}) \text{vec } \mathbf{Q}(m)}, \quad (44)$$

where we have applied  $\text{tr } \mathbf{ABCD} = (\text{vec } \mathbf{D})^T \mathbf{A} \otimes \mathbf{C}^T \text{vec } \mathbf{B}^T$ . The operator  $\text{vec}(\cdot)$  just stacks the columns of an  $M \times M$  matrix to form an  $M^2 \times 1$  column vector. Notice that, according to (44), the ratio of traces on the left-hand side is equivalent to a ratio of quadratic forms on  $M^2 \times M^2$  matrices. Therefore, our strategy will use bounds to a ratio of quadratic forms. Firstly, the covariance matrix  $\bar{\mathbf{Z}}$  is trivially symmetric positive definite, and we assume it is full rank. Then all its eigenvalues are positive, and it can

also be decomposed as  $\bar{Z} = W^T W$  for some square matrix  $W$ . Applying elementary properties of the Kronecker product we have that  $\bar{Z} \otimes \bar{Z} = (W \otimes W)^T (W \otimes W)$ , and defining next  $\mathbf{v} \triangleq (W \otimes W) \text{vec } Q(m)$  we can rewrite (44) as

$$\begin{aligned} \psi &= \frac{\mathbf{v}^T (W \otimes W)^{-T} ((2\bar{Z} + \mathbf{I}) \otimes \mathbf{I}) (W \otimes W)^{-1} \mathbf{v}}{\|\mathbf{v}\|^2} \\ &= \frac{\mathbf{v}^T ((2\mathbf{I} + \bar{Z}^{-1}) \otimes \bar{Z}^{-1}) \mathbf{v}}{\|\mathbf{v}\|^2} \end{aligned}$$

with  $W^{-T} = (W^{-1})^T = (W^T)^{-1}$ . Now, this is a Rayleigh quotient [19] which is upper and lower bounded for any  $\mathbf{v} \neq \mathbf{0}$  by the maximum and minimum eigenvalues  $\lambda_{\max}(\cdot)$  and  $\lambda_{\min}(\cdot)$  of the matrix of the quadratic form in the numerator. The upper bound is

$$\begin{aligned} \psi &\leq \lambda_{\max}((2\mathbf{I} + \bar{Z}^{-1}) \otimes \bar{Z}^{-1}) = \lambda_{\max}(2\mathbf{I} + \bar{Z}^{-1}) \lambda_{\max}(\bar{Z}^{-1}) \\ &\leq (2 + \lambda_{\max}(\bar{Z}^{-1})) \lambda_{\max}(\bar{Z}^{-1}) = (2 + \lambda_{\min}^{-1}(\bar{Z})) \lambda_{\min}^{-1}(\bar{Z}), \end{aligned} \quad (45)$$

where we have applied the following facts: a) the eigenvalues of a Kronecker product are the products of the eigenvalues of the matrices in the product, which only have positive eigenvalues; b) for any two symmetric matrices  $A$  and  $B$  it holds  $\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$  [19]; and c) the (positive) eigenvalues of  $\bar{Z}^{-1}$  are the inverses of the eigenvalues of  $\bar{Z}$ . If we define next  $\gamma_u \triangleq \lambda_{\min}^{-1}(Z) \sigma_x^2$  then we can write

$$\psi \leq \left(2 + \frac{\gamma_u}{\xi}\right) \frac{\gamma_u}{\xi}, \quad (46)$$

with  $\xi$  the SNR. A lower bound can be similarly obtained using the maximum eigenvalue of  $Z$ , but only (46) is relevant for bounding the union bound in (20).

## APPENDIX C

### EXPRESSION OF THE CORRELATION COEFFICIENT IN TERMS OF QUADRATIC FORMS

We show next how (30)—and consequently  $E[\rho(m)]$ —may be rewritten in the i.i.d. case in terms of the quadratic form (32). As in Section III-B1, we will drop the band index  $m$  for notational simplicity. We will expand firstly the Frobenius norm of  $Q$  in the denominator of (30). Using (41),  $\|Q\|^2 = \text{tr}[Q^T Q] = \text{tr}[Q^2]$  is just

$$\text{tr}[Q^2] = \text{tr}[(U^T P U)^2] + \text{tr}[(V^T P V)^2] - 2 \text{tr}[U^T P U V^T P V],$$

where we have applied the cyclic permutation of the trace to obtain the last summand. Using  $UU^T = VV^T = I$  and  $\text{tr}[U^T A U] = \text{tr}[V^T A V] = \text{tr} A$ , and applying again the cyclic permutation of the trace to the last summand, we have

$$\text{tr}[Q^2] = 2 \text{tr}[P^2] - 2 \text{tr} [P C_\Delta P C_\Delta^T], \quad (47)$$

where the  $L \times L$  matrix  $C_\Delta \triangleq UV^T$  is just a subdiagonal of ones at distance  $\Delta$  from the diagonal (lower shift matrix), and hence  $C_\Delta^T = C_{-\Delta}$  is the corresponding upper shift matrix. Then, the  $L \times L$  matrix  $C_\Delta P C_\Delta^T$  is just  $P$  with rows and columns shifted  $\Delta$  and clipped. Next we obtain an expression parallel to (47) for the numerator of (30), which, using (43), can be written as

$$\text{tr} [\underline{Q}_0 \underline{Q}_k] = \text{tr} [Q \tilde{C}_k Q \tilde{C}_k^T] \quad (48)$$

with the  $M \times M$  lower shift matrix  $\tilde{C}_k \triangleq \underline{U}_0 \underline{U}_k^T$  a subdiagonal of ones at distance  $k$  from the diagonal. We can now develop (48) using (41) as

$$\begin{aligned} \text{tr} [Q \tilde{C}_k Q \tilde{C}_k^T] &= \text{tr} [P U \tilde{C}_k U^T P U \tilde{C}_k^T U^T] - \text{tr} [P U \tilde{C}_k V^T P V \tilde{C}_k^T U^T] \\ &\quad - \text{tr} [P U \tilde{C}_k^T V^T P V \tilde{C}_k U^T] + \text{tr} [P V \tilde{C}_k V^T P V \tilde{C}_k^T V^T]. \end{aligned} \quad (49)$$

In this result we have used the linearity of the trace and its cyclic property. As  $C_k \triangleq U \tilde{C}_k U^T = V \tilde{C}_k V^T$  is an  $L \times L$  subdiagonal of ones at distance  $k$  from the diagonal, the first and last summands are equal. It is also straightforward to see that  $C_{\Delta+k} \triangleq U \tilde{C}_k V^T$  and  $C_{\Delta-k} \triangleq U \tilde{C}_k^T V^T$  are  $L \times L$  subdiagonals of ones at distances  $\Delta+k$  and  $\Delta-k$  from the diagonal, respectively. Then, (49) can be written as (cf. (47))

$$\text{tr} [\underline{Q}_0 \underline{Q}_k] = 2 \text{tr} [P C_k P C_k^T] - \text{tr} [P C_{\Delta+k} P C_{\Delta+k}^T] - \text{tr} [P C_{\Delta-k} P C_{\Delta-k}^T]. \quad (50)$$

Substituting next (7) in any of the matrix shifts in (50) we note that

$$\begin{aligned} C_k P C_k^T &= C_k \Omega R \Omega C_k^T \\ &= C_k \Omega R_{-k} \Omega C_k^T = \Omega_k R \Omega_k^T, \end{aligned} \quad (51)$$

where the second equality is due to the fact that  $R$  is Toeplitz and that the two outer matrices shift the rows and columns of the inner matrix  $k$  positions. Applying (51) to (50) we have that

$$\text{tr} [\underline{Q}_0 \underline{Q}_k] = 2 \text{tr} [(\Omega \Omega_k R)^2] - \text{tr} [(\Omega \Omega_{\Delta+k} R)^2] - \text{tr} [(\Omega \Omega_{\Delta-k} R)^2] \quad (52)$$

using  $\Omega_k \Omega = \Omega \Omega_k$  because diagonal matrices commute. Similarly, we can write (47) as

$$\text{tr}[Q^2] = 2 \text{tr} [(\Omega^2 R)^2] - 2 \text{tr} [(\Omega \Omega_\Delta R)^2]. \quad (53)$$

We note next that for any diagonal matrix  $\mathbf{D}$  [19]

$$\text{tr}[(\mathbf{D}\mathbf{R})^2] = \mathbf{d}^T(\mathbf{R} \odot \mathbf{R})\mathbf{d}, \quad (54)$$

with  $\odot$  the Hadamard product and  $\mathbf{d} \triangleq \mathbf{D}\mathbf{1}$ , that is,  $\mathbf{D} = \text{diag}(\mathbf{d})$ . Notice that  $\mathbf{R} \odot \mathbf{R}$  is trivially real and symmetric, because so is  $\mathbf{R}$ . Using (54) we may write  $\text{tr}[(\Omega\Omega_k\mathbf{R})^2] = (\Omega\Omega_k\mathbf{1})^T(\mathbf{R} \odot \mathbf{R})(\Omega\Omega_k\mathbf{1})$ . Applying this expression to the right hand sides of (52) and (53) and using  $\Omega\Omega_k\mathbf{1} = \mathbf{w} \odot \mathbf{C}_k\mathbf{w} = \mathbf{v}_k$  it is now straightforward to obtain (33) departing from the expectation of (30).

#### APPENDIX D

##### ASYMPTOTIC EXPRESSION FOR THE OPTIMAL CORRELATION COEFFICIENT IN TERMS OF $\theta$

In this appendix we will obtain an asymptotic expression for the correlation coefficient employing the optimal window obtained in Section III-B1. Firstly, in order to achieve our objective we approximate the shift operation, defined in terms of the matrix  $\mathbf{C}_k$ , by a *circular* shift. Specifically, consider the  $L \times L$  circular shift matrix  $\hat{\mathbf{C}}_k$  defined by

$$\hat{\mathbf{C}}_k \triangleq \mathbf{C}_k + \mathbf{C}_{L-k}^T. \quad (55)$$

As  $L \rightarrow \infty$ , we have that  $k \ll L$  and hence  $\mathbf{C}_{L-k}^T \rightarrow \mathbf{O}$  and  $\hat{\mathbf{C}}_k \rightarrow \mathbf{C}_k$ , understanding this convergence in the sense of the Frobenius norm of the difference normalized by  $L$ . Proceeding in the same manner as (31) through to (35), we define the correlation coefficient  $\hat{\rho}_k$  of the circularly shifted system by replacing  $\mathbf{C}_k$  with  $\hat{\mathbf{C}}_k$  in (31) and arrive at the set of equations that must be satisfied by any window that maximizes  $\hat{\rho}_k$ :

$$\frac{1}{\Delta} \sum_k \left[ 2\hat{\mathbf{B}}_k(\mathbf{w}) - \hat{\mathbf{B}}_{\Delta+k}(\mathbf{w}) - \hat{\mathbf{B}}_{\Delta-k}(\mathbf{w}) \right] \mathbf{w} = 2\hat{\rho}^* \left[ \hat{\mathbf{B}}_0(\mathbf{w}) - \hat{\mathbf{B}}_{\Delta}(\mathbf{w}) \right] \mathbf{w}, \quad (56)$$

where  $\hat{\mathbf{B}}_k(\mathbf{w}) \triangleq \hat{\mathbf{C}}_k^T \text{diag}(\hat{\mathbf{b}}_k) + \text{diag}(\hat{\mathbf{b}}_k)\hat{\mathbf{C}}_k$ ,  $\hat{\mathbf{b}}_k \triangleq 2(\mathbf{R} \odot \mathbf{R})\hat{\mathbf{v}}_k$  and  $\hat{\mathbf{v}}_k \triangleq \mathbf{w} \odot \hat{\mathbf{C}}_k\mathbf{w}$ . Consider next the family of windows

$$\mathbf{w}_\epsilon \triangleq \mathbf{1} - \epsilon\boldsymbol{\omega}_p, \quad (57)$$

where  $\boldsymbol{\omega}_p \triangleq (\omega_p(0) \dots \omega_p(L-1))^T$  and  $\omega_p(i) \triangleq \cos(2\pi p i/L)$ , with  $p = 1, \dots, L-1$ . Both  $\mathbf{1}$  and  $\boldsymbol{\omega}_p$  are eigenvectors of any real-valued circulant matrix [20], in particular of  $\mathbf{R} \odot \mathbf{R}$ . We will demonstrate that  $\mathbf{w}_\epsilon$  satisfies (56) as  $\epsilon \rightarrow 0$ . To see this, note that  $\hat{\mathbf{v}}_k(\mathbf{w}_\epsilon) = \mathbf{1} - \epsilon(\boldsymbol{\omega}_p + \hat{\mathbf{C}}_k\boldsymbol{\omega}_p) + O(\epsilon^2)$  and hence

$$\hat{\mathbf{b}}_k(\mathbf{w}_\epsilon) = \lambda_0\mathbf{1} - \epsilon\lambda_p(\boldsymbol{\omega}_p + \hat{\mathbf{C}}_k\boldsymbol{\omega}_p) + O(\epsilon^2), \quad (58)$$

where  $\lambda_0, \lambda_p$  are the eigenvalues of  $\mathbf{R} \odot \mathbf{R}$  corresponding to eigenvectors  $\mathbf{1}, \boldsymbol{\omega}_p$  respectively. It follows that

$$\begin{aligned} \text{diag}(\hat{\mathbf{b}}_k)\hat{\mathbf{C}}_k\mathbf{w}_\epsilon &= \text{diag}(\hat{\mathbf{b}}_k)\hat{\mathbf{C}}_k(\mathbf{1} - \epsilon\boldsymbol{\omega}_p) = \text{diag}(\hat{\mathbf{b}}_k)(\mathbf{1} - \epsilon\hat{\mathbf{C}}_k\boldsymbol{\omega}_p) \\ &= \lambda_0\mathbf{1} - \epsilon\lambda_p\boldsymbol{\omega}_p - \epsilon(\lambda_0 + \lambda_p)\hat{\mathbf{C}}_k\boldsymbol{\omega}_p + O(\epsilon^2), \end{aligned} \quad (59)$$

from which we get

$$\hat{\mathbf{B}}_k(\mathbf{w}_\epsilon)\mathbf{w}_\epsilon = \lambda_0\mathbf{1} - \epsilon\lambda_p\boldsymbol{\omega}_p - \epsilon(\lambda_0 + \lambda_p)(\hat{\mathbf{C}}_k + \hat{\mathbf{C}}_k^T)\boldsymbol{\omega}_p + O(\epsilon^2). \quad (60)$$

Let  $\hat{\mathbf{T}}_k \triangleq \hat{\mathbf{C}}_k + \hat{\mathbf{C}}_k^T$ . Then, using (60), (56) evaluated at  $\mathbf{w}_\epsilon$  reduces to

$$\epsilon(\lambda_0 + \lambda_p)\frac{1}{\Delta}\sum_k(2\hat{\mathbf{T}}_k - \hat{\mathbf{T}}_{\Delta-k} - \hat{\mathbf{T}}_{\Delta+k})\boldsymbol{\omega}_p = 2\hat{\rho}^*\epsilon(\lambda_0 + \lambda_p)(2\hat{\mathbf{T}}_0 - \hat{\mathbf{T}}_\Delta)\boldsymbol{\omega}_p + O(\epsilon^2). \quad (61)$$

To  $O(\epsilon)$ , (61) is a generalized eigenvalue problem in real-valued circulant matrices and hence is approximately satisfied by  $\boldsymbol{\omega}_p$  for any  $p = 1, \dots, L-1$ . The maximum eigenvalue, which gives the maximum correlation in the generalized eigenvalue problem, occurs at  $p = 1$ , leading to the approximate solution

$$\mathbb{E}[\hat{\rho}^*] \approx \sum_k \frac{2\omega_1(k) - \omega_1(\Delta+k) - \omega_1(\Delta-k) + O(\epsilon^2)}{2\Delta(1 - \omega_1(\Delta)) + O(\epsilon^2)}, \quad (62)$$

obtained using  $\hat{\mathbf{v}}_k^T(\mathbf{R} \odot \mathbf{R})\hat{\mathbf{v}}_k$  and  $\hat{\mathbf{v}}_k = \mathbf{w}_\epsilon \odot \hat{\mathbf{C}}_k\mathbf{w}_\epsilon$  as in (33). As  $\epsilon \rightarrow 0$  we may neglect the terms on  $O(\epsilon^2)$  and this solution is independent of  $\mathbf{R}$ , that is, independent of the band. Notice also that in this case (61) becomes closer to an exact generalized eigenvalue problem, and that  $(\lambda_0 + \lambda_p)$  factors from the equation. Simplifying next (62) by using the identity

$$\sum_{k=1}^K \cos(kx) = \frac{1}{2} \left( -1 + \cos(Kx) + \frac{\sin(x)}{(1 - \cos(x))} \sin(Kx) \right), \quad (63)$$

and taking the limit as  $L \rightarrow \infty$ , holding the overlap  $\theta$  fixed, we get

$$\mathbb{E}[\hat{\rho}^*] \rightarrow \frac{\sin((1-\theta)\pi)}{(1-\theta)\pi}. \quad (64)$$

Finally, notice that the periodic von Hann window belongs to the family of windows defined by (57), just setting  $\epsilon = 1$  and  $p = 1$ . Using this window we may recompute  $\mathbb{E}[\hat{\rho}]$  to obtain an expression completely parallel to (62), but in which the terms on  $O(\epsilon^2)$  cannot be made now tend to zero (as  $\epsilon$  is fixed to 1). Nevertheless, it is possible to prove that, as long as  $\frac{\lambda_0}{8(\lambda_0 + \lambda_1)} \ll 1$ , these terms can also be neglected, and in this case (64) applies. Then, the von Hann window is close to optimal in these conditions.

#### ACKNOWLEDGMENT

The authors wish to thank the reviewers for their useful comments. Félix Balado thanks Noreen Barron for valuable remarks for obtaining the bound in Appendix B.

## REFERENCES

- [1] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification," in *Procs. of the International Workshop on Content-Based Multimedia Indexing*, Brescia, Italy, September 2001, pp. 117–125.
- [2] T. Kalker, D. H. J. Epema, P. H. Hartel, R. L. Lagendijk, and M. V. Steen, "Music2Share - copyright-compliant music sharing in P2P systems," *Proceedings of the IEEE*, vol. 92, no. 6, pp. 961–970, June 2004.
- [3] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *Journal of VLSI Signal Processing*, vol. 41, no. 3, pp. 271–284, November 2005.
- [4] E. Martinian, G. W. Wornell, and B. Chen, "Authentication with distortion criteria," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2523–2542, July 2005.
- [5] M. Park, H. R. Kim, Y. M. Ro, and M. Kim, "Frequency filtering for a highly robust audio fingerprinting scheme in a real-noise environment," *IEICE Transactions on Information and Systems*, vol. E89-D, no. 7, pp. 2324–2327, July 2006.
- [6] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *3rd International Conference on Music Information Retrieval (ISMIR)*, Paris, October 2002, pp. 107–115.
- [7] P. J. O. Doets and R. L. Lagendijk, "Stochastic model of a robust audio fingerprinting system," in *5th International Symposium on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.
- [8] —, "Theoretical modeling of a robust audio fingerprinting system," in *Procs. of the 4th IEEE Benelux Signal Processing Symposium*, Hilvarenbeek, The Netherlands, April 2004, pp. 101–104.
- [9] P. J. O. Doets, "Modelling a robust audio fingerprinting system," in *Technical Report, Delft University of Technology*, June 2004.
- [10] P. J. O. Doets and R. L. Lagendijk, "Extracting quality parameters for compressed audio from fingerprints," in *6th International Symposium on Music Information Retrieval (ISMIR)*, London, UK, September 2005, pp. 498–503.
- [11] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, June 1967.
- [12] P. E. Johnson and D. G. Long, "The probability density of spectral estimates based on modified periodogram averages," *IEEE Transactions on Signal Processing*, vol. 47, no. 5, pp. 1255–1261, May 1999.
- [13] J. Imhof, "Computing the distribution of quadratic forms in normal variables," *Biometrika*, vol. 48, no. 3/4, pp. 419–426, December 1961.
- [14] S. R. Searle, *Linear Models for Unbalanced Data*. John Wiley & Sons, 1987.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.
- [16] D. Sorensen, "Implicitly restarted Arnoldi/Lanczos methods for large scale eigenvalue calculations," Rice University, Tech. Rep., 1995.
- [17] M. K. Mihçak and R. Venkatesan, "A perceptual audio hashing algorithm: A tool for robust audio identification and information hiding," in *Procs. of the 4th Information Hiding Workshop*, ser. Lecture Notes in Computer Science, vol. 2137. Pittsburgh, USA: Springer, April 2001, pp. 51–65.
- [18] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo, "Audio fingerprinting based on normalized spectral subband moments," *IEEE Signal Processing Letters*, vol. 13, no. 4, pp. 209–212, April 2006.
- [19] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd ed. John Wiley & Sons, 1999.
- [20] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.



and digital communications.

**Félix Balado** graduated with an M.Sc. in Telecommunications Engineering from the University of Vigo (Spain) in 1996, and received a Ph.D. from the same institution in 2003, for his work in data hiding. He then joined the National University of Ireland (University College Dublin) as a post-doctoral fellow at the Information Hiding Laboratory. Previously he worked as a research and project engineer at the University of Vigo, in different research projects funded by the Galician and Spanish Governments, and by the European Union. His research interests lie in the areas of multimedia signal processing, data hiding,



**Neil J. Hurley** received the M.Sc. degree in mathematical science from University College Dublin (UCD), Dublin, Ireland, in 1988. In 1989, he joined Hitachi Dublin Laboratory, a computer science research laboratory at the University of Dublin, Trinity College, from which he received the Ph.D. degree in 1995, for his work in knowledge-based engineering and high-performance computing. He joined the National University of Ireland, University College Dublin, in 1999 where his present research activities lie in the areas of data-hiding, signal processing, secure and robust information retrieval and distributed computing.



**Elizabeth P. McCarthy** received the B.E. degree in electronic engineering from the National University of Ireland (University College Dublin), Dublin, Ireland, in 2002. She continues here as a member of the Information Hiding Laboratory, where she is pursuing the Ph.D. degree. Her research activities lie in the area of multimedia signal processing, particularly fingerprinting of audio data.



**Gu no  C.M. Silvestre** received the M.Sc. degree in electronic and electrical engineering in 1993. In 1996, he received the Ph.D. degree from the University of Dublin, Trinity College, Ireland, for his work in silicon-on-insulator materials. As a post-doctoral fellow with Trinity College, Dublin, he pursued research on digital image processing and watermarking. In 1997, he was appointed Research Scientist at the Philips Research Laboratories, Eindhoven, The Netherlands, and his research focus switched toward the study of polymer light-emitting materials. In 1999, he joined the National University of Ireland (University College Dublin), where his present research activities lie in the area of digital communications, data-hiding, and signal processing. Dr. Silvestre was the 1995 recipient of the Materials Research Society Graduate Student Award.

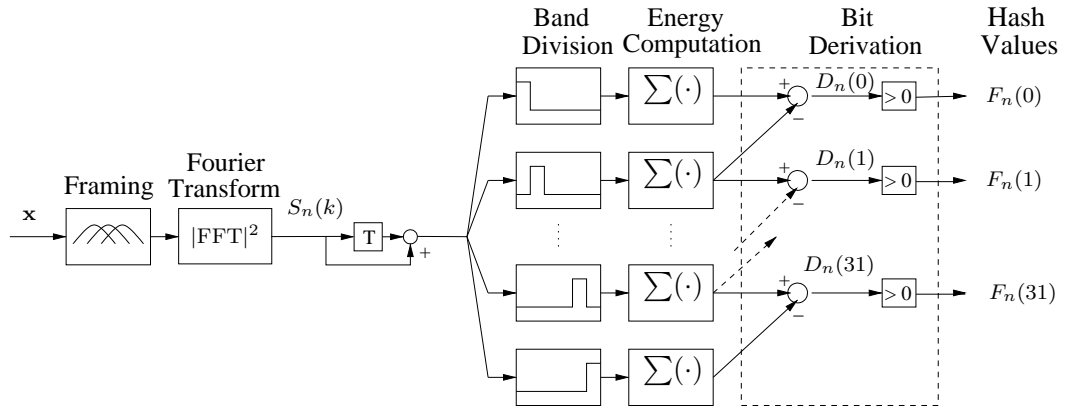


Fig. 1. Philips hashing algorithm, rearranged as in [9].

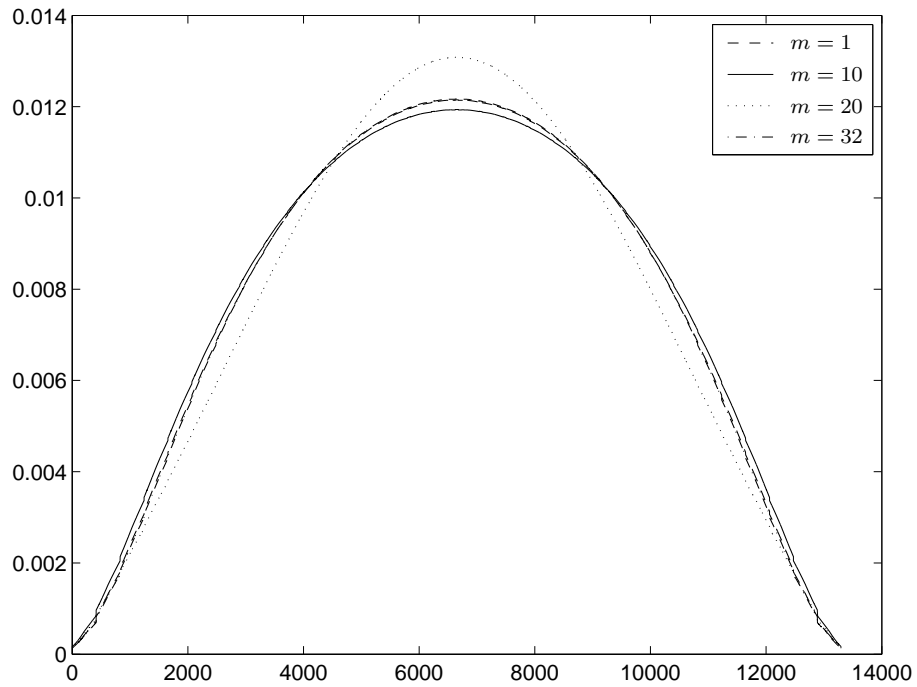


Fig. 2. Optimal windows obtained for a number of different frequency bands  $m$ .

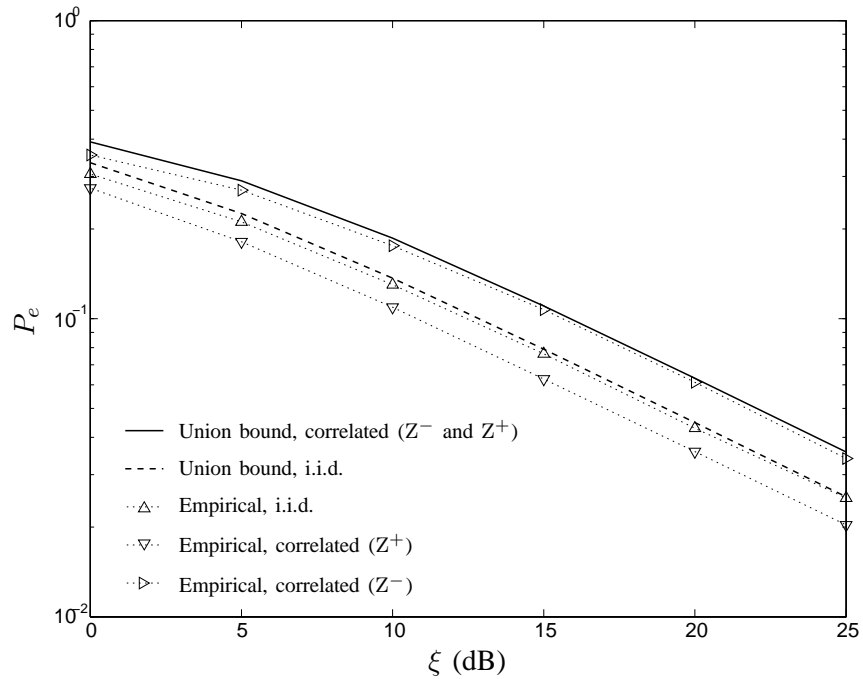


Fig. 3. Probability of bit error under additive independent i.i.d. Gaussian noise versus SNR, using both i.i.d and correlated stationary Gaussian hashed signal.

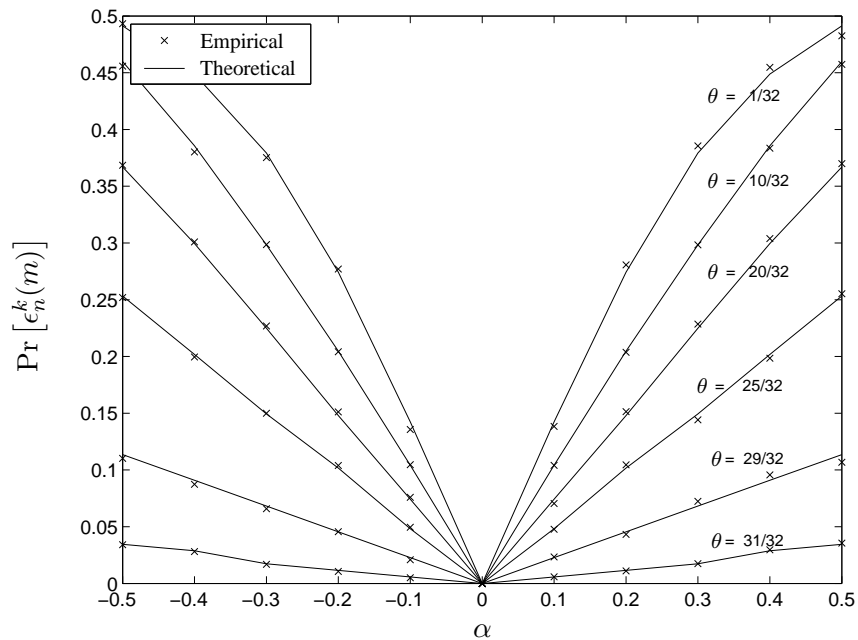


Fig. 4. Probability of bit error for one frame under desynchronization versus normalized desynchronization level  $\alpha = k/\Delta$ , for different levels of overlap  $\theta$  and using i.i.d. Gaussian hashed signal (band  $m = 5$ ).

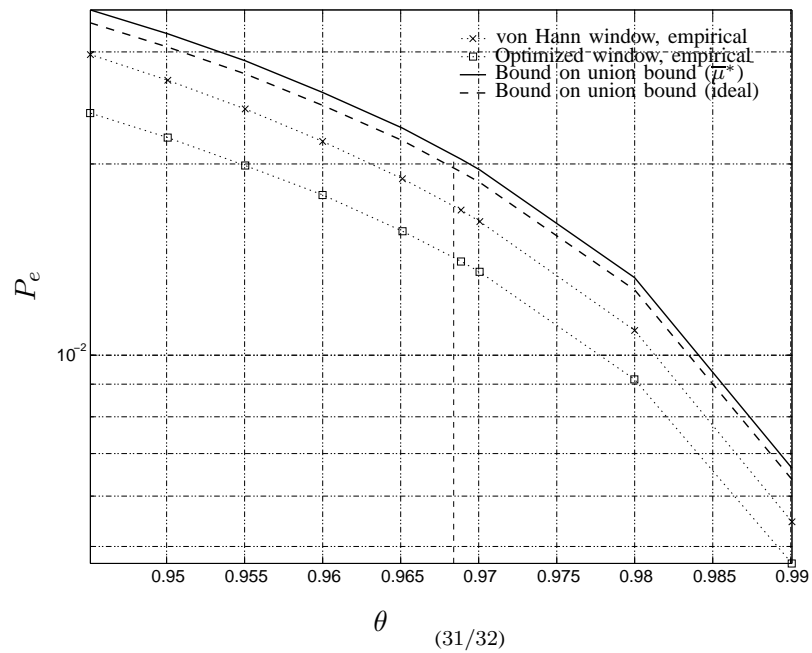


Fig. 5. Probability of bit error under uniform desynchronization versus overlap level, using i.i.d. Gaussian hashed signal. Empirical results correspond to the von Hann window and to the optimized window obtained as the average of band-optimal windows, respectively.  $T_f \approx 0.3$  seconds. The vertical dashed line shows the original overlap level proposed for the Philips algorithm.

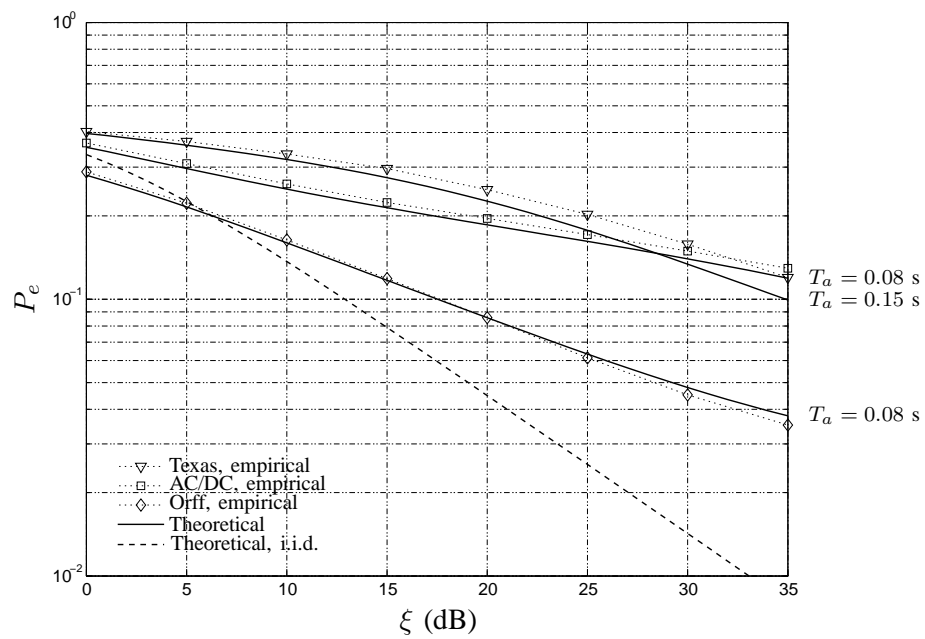


Fig. 6. Probability of bit error under additive independent i.i.d. Gaussian noise versus SNR, using 5-second excerpts of three real audio signals.  $T_f = 0.05$  seconds, von Hann window.

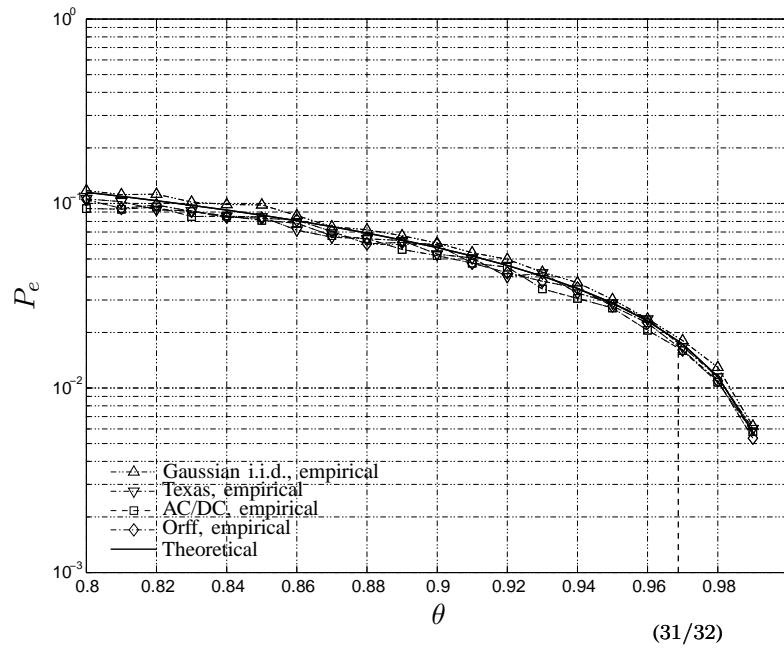


Fig. 7. Probability of bit error under uniform desynchronization versus overlap level, using 5-second excerpts of three real audio signals and i.i.d. Gaussian signal.  $T_f = 0.3$  seconds, von Hann window. The theoretical result is the asymptotic performance for an optimal window. The vertical dashed line shows the original overlap level proposed for the Philips algorithm.