PERFORMANCE OF PHILIPS AUDIO FINGERPRINTING UNDER ADDITIVE NOISE

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

University College Dublin
Belfield Campus, Dublin 4, Ireland

ABSTRACT

We present a theoretical analysis of the Philips audio fingerprinting method under Gaussian white noise distortion for correlated stationary Gaussian sources. Prior analyses were for white Gaussian sources, which do not model realistically real audio signals. Our approach 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, and we apply these expressions to real audio signals.

Index Terms— Audio fingerprinting, error analysis

1. INTRODUCTION

An audio fingerprint is a compact representation (hash) of an audio signal linked to its perceptual content. Perceptually equivalent instances of the same audio signal must approximately lead to the same hash value. Fingerprinting (also known as *robust* hashing) helps to identify audio signals in noisy environments, and then it finds application in content tracking in peer-to-peer networks, authentication, and efficient indexing of multimedia databases [1]. An audio fingerprinting scheme that has proved to be remarkably robust is the so-called Philips method proposed by Haitsma et al. [2], based on quantizing differences of energy measures from overlapped short-term power spectra. Here we examine its theoretical performance through a statistical model. A performance analysis for false positives was presented in [3]. A more elaborate model was proposed by Doets and Lagendijk [4, 5, 6], for the uncorrelated Gaussian input signals. Of these works, only [6] tackles the issue of evaluating the performance of the fingerprinting method under random additive distortion. Here we address the more general case with stationary correlated Gaussian input signals, from which that previous analysis follows as a particular case. We also apply this analysis to real audio signals.

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. For a symmetric $L \times L$ matrix X its eigenvalues are denoted by $\lambda_1(\mathrm{X}) \leq \cdots \leq \lambda_L(\mathrm{X})$. $\mathrm{diag}(\mathbf{x})$ is a matrix with the elements of $\mathbf{x}$ in the diagonal and zero elsewhere. $\mathrm{tridiag}(a, b, c)$ is a Toeplitz tridiagonal matrix with constant diagonal elements b, and constant superdiagonal and subdiagonal elements c and a, respectively. $\mathrm{tr}\,\mathrm{X}$ 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. The symbol $\otimes$ denotes the Kronecker (or direct) product.

This work was kindly supported by Enterprise Ireland, research grant ATRP2002/230 and by the European Commission through IST-2002-507609 SIMILAR.

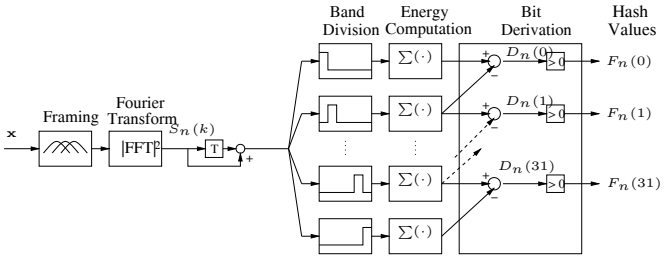

Fig. 1. Philips hashing algorithm, rearranged as in [5].

2. STATISTICAL MODEL OF PHILIPS METHOD

We describe firstly the operation of the Philips method using its equivalent rearrangement given in [5] (see Fig. 1). The input signal $\mathbf{x} = (x[1], \ldots, x[N])^T$ is divided into overlapped frames before hashing it. If L is the number of samples in a single frame and Δ the number of non-overlapping samples between two frames, then the L-length vector $\mathbf{x}_n$ formed by the elements of $\mathbf{x}$ used in the computations corresponding to the n^{th} frame is given by $\mathbf{x}_n \triangleq (x[n \cdot \Delta + 1], \cdots, x[n \cdot \Delta + L])^T$ for $n = 0, 1, 2, \ldots$ Each framed signal $\mathbf{x}_n$ is weighted next by a window of weights $\mathbf{w} \triangleq (w[1], \cdots, w[L])^T$ before taking its fast Fourier transform (FFT). Then, the vector at the input of the FFT for frame $\mathbf{x}_n$ is just $\mathrm{diag}(\mathbf{w})\ \mathbf{x}_n$. The spectrum is then divided into $N_\mathrm{b} + 1$ bands; we will assume the logarithmic frequency band division given in [2], for which $N_\mathrm{b} = 32$. Denoting by $E_n(m)$ the energy of band m for input frame $\mathbf{x}_n$, an unquantized hash value is given by

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

with $m = 0, 1, \cdots, N_\mathrm{b} - 1$ and $n = 0, 1, 2, \cdots$. The variables (1) completely determine the system, as the binary hash value $F_n(m) \in \{0, 1\}$ corresponding to frame n and band m is just

$$F_n(m) \triangleq u\left(D_n(m)\right), \tag{2}$$

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. Our approach is based on modeling the continuous random variables $D_n(m)$. As in [7], we rely on the periodogram estimator of the power spectrum of the windowed signal at the n-th frame, which is given by

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

for $k = 0, \cdots, L-1$, and where $\|\mathbf{w}\|^2$ is a normalization factor. Now, following [8], $S_n(k)$ can be rewritten as

$$S_n(k) = \mathbf{x}_n^T\ \mathrm{M}(k)\ \mathbf{x}_n, \tag{3}$$

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

$$\mathrm{M}(k) \triangleq \frac{1}{\|\mathbf{w}\|^2}\ \Omega\ \mathrm{N}(k)\ \Omega. \tag{4}$$

The $L \times L$ matrix $\mathrm{N}(k)$ is defined such that its entry at position (i, j) is given by $\cos(2\pi(i-j)k/L)$, and $\Omega \triangleq \mathrm{diag}(\mathbf{w})$. Using the quadratic form (3) it is straightforward to express $D_n(m)$ 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

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

Considering (3) and (4), we define next the $L \times L$ matrix $\mathrm{R}(m) \triangleq \frac{1}{\|\mathbf{w}\|^2}[\sum_{k\in B(m)}\mathrm{N}(k) - \sum_{k\in B(m+1)}\mathrm{N}(k)]$. Now, plugging (5) in (1) and defining $\mathrm{P}(m) \triangleq \Omega\ \mathrm{R}(m)\ \Omega$ we obtain

$$D_n(m) = \mathbf{x}_n^T\ \mathrm{P}(m)\ \mathbf{x}_n - \mathbf{x}_{n-1}^T\ \mathrm{P}(m)\ \mathbf{x}_{n-1}. \tag{6}$$

In order to write (6) as a single quadratic form, we define next the extended vector $\tilde{\mathbf{x}}_n \triangleq (x[(n-1)\cdot\Delta+1], \cdots, x[n\cdot\Delta+L])^T$, for $n = 0, 1, 2, \cdots$, which includes all the components of the overlapping vectors $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$ and which is of length $M \triangleq L+\Delta$. We assume the convention of padding with zeros for indices out of range. Defining next the $L \times M$ auxiliary matrices

$$\mathrm{U} \triangleq [\ \mathrm{I}_{L\times L}\ |\ \mathrm{O}_{L\times\Delta}\ ],\quad \mathrm{V} \triangleq [\ \mathrm{O}_{L\times\Delta}\ |\ \mathrm{I}_{L\times L}\ ],$$

with I the identity matrix and O the null matrix of size given by the subindices, we can build the matrix

$$\mathrm{Q}(m) \triangleq \mathrm{U}^T\mathrm{P}(m)\mathrm{U} + \mathrm{V}^T\mathrm{P}(m)\mathrm{V}. \tag{7}$$

which is formed by adding $-\mathrm{P}(m)$ at the position $(1, 1)$ of an empty $M \times M$ matrix with $\mathrm{P}(m)$ at the position $(\Delta+1, \Delta+1)$. $\mathrm{Q}(m)$ is symmetric because $\mathrm{P}(m)$ is so. Using (7) and $\tilde{\mathbf{x}}_n$ we can finally write (6) as

$$D_n(m) = \tilde{\mathbf{x}}_n^T\ \mathrm{Q}(m)\ \tilde{\mathbf{x}}_n. \tag{8}$$

For $\tilde{\mathbf{x}}_n \sim \mathcal{N}(\mathbf{0}, \mathrm{Z})$, (8) is consequently a quadratic form in a Gaussian vector. The distribution of this r.v. may be expressed exactly as a weighted sum of χ^2 distributions [9], which is unwieldy for analysis. Nevertheless, for large M, a Gaussian distribution suffices to approximate the probability density function (pdf) of (8). This approximation is supported by the Central Limit Theorem (CLT) for independent identically distributed (i.i.d.) signals; for locally correlated signals a broader version of the CLT can also be invoked. The adequacy of this assumption will be confirmed empirically in Section 4. The quadratic form (8) allows to easily compute the parameters of the Gaussian model; the expectation[1] and variance of the variables (8) for Gaussian $\tilde{\mathbf{x}}$ are [10]

$$\mathrm{E}[D_n(m)] = \mathrm{tr}\left[\mathrm{Z}\,\mathrm{Q}(m)\right], \tag{9}$$

$$\mathrm{Var}[D_n(m)] = 2\ \mathrm{tr}\left[(\mathrm{Z}\,\mathrm{Q}(m))^2\right]. \tag{10}$$

3. PERFORMANCE ANALYSIS

The signal presented to the algorithm in the identification stage may differ from the corresponding original indexed in the database during acquisition. Next, we use our model to examine the probability of bit error (P_e) of the hash when the distortion on $\mathbf{x}$ is assumed to be zero-mean additive white Gaussian noise $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\ \mathrm{I})$. We assume that $\tilde{\mathbf{x}}_n \sim \mathcal{N}(\mathbf{0}, \mathrm{Z})$, where the $M \times M$ covariance matrix $\mathrm{Z} = \mathrm{E}[\tilde{\mathbf{x}}_n\tilde{\mathbf{x}}_n^T]$ is Toeplitz (i.e., the signal is stationary) with diagonal elements σ_x^2. When the signal presented to the system is $\tilde{\mathbf{x}}_n + \mathbf{g}$ instead of $\tilde{\mathbf{x}}_n$, the hash value (8) becomes

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

Before proceeding, let us define for notational simplicity the variables $S \triangleq D_n(m)$ and $T \triangleq 2\ \tilde{\mathbf{x}}_n^T\mathrm{Q}(m)\mathbf{g} + \mathbf{g}^T\mathrm{Q}(m)\mathbf{g}$. The expectation of S is computed using (9), which, using (7) and applying the circular property of the trace, is

$$\mathrm{E}[S] = -\mathrm{tr}[\mathrm{U}\mathrm{Z}\mathrm{U}^T\mathrm{P}(m)] + \mathrm{tr}[\mathrm{V}\mathrm{Z}\mathrm{V}^T\mathrm{P}(m)] = 0,$$

which is zero because Z is Toeplitz and then $\mathrm{U}\mathrm{Z}\mathrm{U}^T = \mathrm{V}\mathrm{Z}\mathrm{V}^T$. For the same reason $\mathrm{E}[D'_n(m)] = 0$, and therefore $\mathrm{E}[T] = 0$. It is straightforward to show that S and T are uncorrelated, that is, $\mathrm{E}[S \cdot T] = 0$, because $\mathbf{g}$ is zero-mean and white. We have argued that we will 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

[1] The expectation is actually valid for an arbitrary zero-mean input pdf.

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 (10) to obtain their variances. Doing so we obtain $\sigma_S^2 \triangleq \mathrm{Var}[S] = 2\,\mathrm{tr}\left[(\mathrm{Z}\,\mathrm{Q}(m))^2\right]$. Similarly, as $\mathrm{Var}[D'_n(m)] = 2\,\mathrm{tr}\left[\left((\mathrm{Z}+\sigma_g^2\,\mathrm{I})\,\mathrm{Q}(m)\right)^2\right]$ it follows from independence between S and T that

$$\sigma_T^2 \triangleq \mathrm{Var}[T] = 2\sigma_g^2\,\mathrm{tr}\left[(2\mathrm{Z}+\sigma_g^2\,\mathrm{I})\,\mathrm{Q}^2(m)\right].$$

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

$$\Pr[\epsilon_n(m)] = 2\int_0^\infty \mathcal{Q}\left(\frac{s}{\sigma_T}\right) f_S(s)\,ds = \frac{1}{\pi}\,\arctan\left(\frac{\sigma_T}{\sigma_S}\right).$$

Plugging the corresponding moments in, we have finally that

$$\Pr[\epsilon_n(m)] = \frac{1}{\pi}\,\arctan\left(\sqrt{\frac{\mathrm{tr}\left[(2\bar{\mathrm{Z}}+\mathrm{I})\,\mathrm{Q}^2(m)\right]}{\mathrm{tr}\left[(\bar{\mathrm{Z}}\,\mathrm{Q}(m))^2\right]}}\right), \quad (11)$$

with $\bar{\mathrm{Z}} \triangleq \frac{1}{\sigma_g^2}\mathrm{Z}$. As shown in Appendix A, (11) is bounded by

$$\Pr[\epsilon_n(m)] \leq \frac{1}{\pi}\,\arctan\left(\sqrt{\left(2+\frac{\gamma_u}{\xi}\right)\frac{\gamma_u}{\xi}}\right), \quad (12)$$

where γ_u only depends on σ_x^2 and on the minimum eigenvalue of Z and $\xi \triangleq \frac{\sigma_x^2}{\sigma_g^2}$ is the signal-to-noise ratio (SNR). Note that (11) and (12) apply to the particular situation in which the hashed signal is i.i.d. This case can be expressed exactly in terms of ξ substituting $\bar{\mathrm{Z}} = \xi\,\mathrm{I}$ in (11):

$$\Pr[\epsilon_n(m)] = \frac{1}{\pi}\,\arctan\left(\sqrt{\left(2+\frac{1}{\xi}\right)\frac{1}{\xi}}\right). \quad (13)$$

The expression (13) was also previously obtained by Doets and Lagendijk in [6]. Lastly, the probabilities (11) or (13) cannot just be averaged to obtain the overall probability of bit error due to the dependencies between the variables $\{D_n(m)\}$ —and therefore also $\{F_n(m)\}$— caused by the overlapping of frames. Nevertheless, we can resort to an upper bounding argument. The average probability of bit error will be

$$P_e = \frac{1}{N_\mathrm{f}}\frac{1}{N_\mathrm{b}}\Pr\left[\cup_{n,m}\,\epsilon_n(m)\right] \leq \frac{1}{N_\mathrm{f}}\frac{1}{N_\mathrm{b}}\sum_{n=0}^{N_\mathrm{f}-1}\sum_{m=0}^{N_\mathrm{b}-1}\Pr[\epsilon_n(m)], \quad (14)$$

which is just the union bound to the average probability of bit error. As the error events are uncorrelated in the long term then the bound must be reasonably tight for large N_f. It is remarkable that both the bound (12) and the exact expression (13) are independent of $\mathrm{Q}(m)$, and then both of the type of window used and of the band m. Then, the union bound is upperbounded by (12) for generic stationary Gaussian signals, and exactly given by (13) for i.i.d. Gaussian signals.

4. EXPERIMENTAL RESULTS

Fig. 2 shows the upper bounds using (12) and (13) compared to empirical data. The parameters used are the original ones in Philips method, but with shorter frames for computational reasons. The two correlated cases correspond to the $M \times M$ tridiagonal covariances $\mathrm{Z}^+ = \sigma_x^2\,\mathrm{tridiag}(\frac{1}{4}, 1, \frac{1}{4})$ and $\mathrm{Z}^- = \sigma_x^2\,\mathrm{tridiag}(-\frac{1}{4}, 1, -\frac{1}{4})$, which have the same minimum eigenvalue and, consequently, the same upper bound. This illustrates the fact that the sharpness of the union bound depends on the particular Z present. We will also see next that (12) may be loose if the minimum eigenvalue of Z is too small.

We examine now the validity of our Gaussian analysis for real audio signals. Our analysis is for *averages* over the ensemble of signals with a given Gaussian distribution, whereas audio signals are particular realizations. So, averages (and stationarity) will have to be interpreted in an ergodic sense. Also, neither real audio signals are Gaussian nor stationary. The first issue will be an inevitable source of inaccuracy. With respect to the second issue, it is possible to approximate audio signals by locally stationary stretches. For each frame n we will have a possibly different autocovariance matrix Z_n, and $\Pr[\epsilon_n(m)]$ will depend now on n unlike in the stationary case. Then, prediction for real signals will require the estimation of the (positive definite) autocovariance matrices corresponding to each locally stationary stretch. We present in Fig. 3 the results for 5-second excerpts of three real audio signals used in [2]: "O Fortuna" by Carl Orff, "Say what you want" by Texas, and "Whole lotta Rosie" by AC/DC (16 bits, 44.1 kHz). The theoretical plots in Fig. 3 have been obtained using (11) and (14), as (12) turned out to be too loose. The estimation interval lengths (T_a) are given in the plot. The use of (11) limits the applicability of the theoretical predictions to small frames, due to the matrix multiplication involved; nevertheless, the good fit of the theoretical results encourages the future development of more practical approximations to that expression. The results also show that the SNR alone does not suffice to predict performance for real signals.

A. APPENDIX

As $\arctan(\cdot)$ is strictly increasing, a bound on (11) can be obtained by bounding the argument ψ inside the square root:

$$\psi \triangleq \frac{(\mathrm{vec}\,\mathrm{Q}(m))^T\left((2\bar{\mathrm{Z}}+\mathrm{I})\otimes\mathrm{I}\right)\mathrm{vec}\,\mathrm{Q}(m)}{(\mathrm{vec}\,\mathrm{Q}(\mathrm{m}))^T(\bar{\mathrm{Z}}\otimes\bar{\mathrm{Z}})\,\mathrm{vec}\,\mathrm{Q}(m)}, \quad (15)$$

applying $\mathrm{tr}\,\mathrm{ABCD} = (\mathrm{vec}\,\mathrm{D})^T\mathrm{A}\otimes\mathrm{C}^T\,\mathrm{vec}\,\mathrm{B}^T$ [11]. The operator $\mathrm{vec}(\cdot)$ stacks the columns of an $M \times M$ matrix to

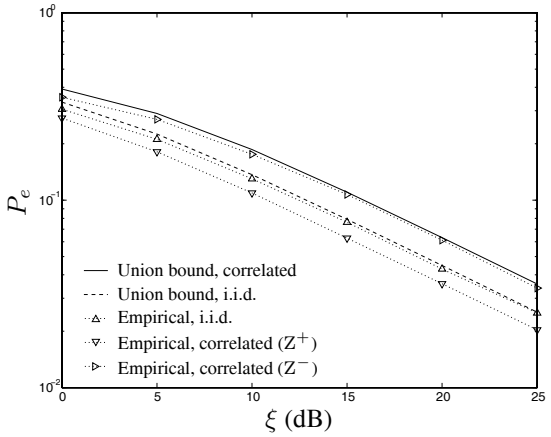

Fig. 2. Probability of bit error under additive independent i.i.d. Gaussian noise versus SNR, for stationary correlated and i.i.d. Gaussian hashed signals. Frame size 0.01 seconds.

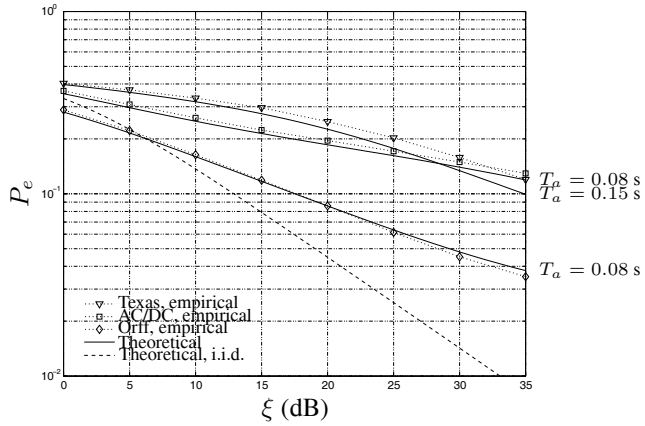

Fig. 3. Probability of bit error under additive independent i.i.d. Gaussian noise versus SNR, for 5-second excerpts of three real audio signals. Frame size 0.05 seconds.

form an $M^2 \times 1$ vector. As $\bar{\mathrm{Z}}$ is symmetric positive definite, assuming full rank all its eigenvalues are strictly positive, and it can be decomposed as $\bar{\mathrm{Z}} = \mathrm{W}^T\mathrm{W}$ for some square matrix W. Applying elementary properties of the Kronecker product we have that $\bar{\mathrm{Z}} \otimes \bar{\mathrm{Z}} = (\mathrm{W} \otimes \mathrm{W})^T(\mathrm{W} \otimes \mathrm{W})$, and defining next $\mathbf{v} \triangleq (\mathrm{W} \otimes \mathrm{W})\,\mathrm{vec}\,\mathrm{Q}(m)$ we can rewrite (15) as

$$\psi = \frac{\mathbf{v}^T\left((2\mathrm{I} + \bar{\mathrm{Z}}^{-1}) \otimes \bar{\mathrm{Z}}^{-1}\right)\mathbf{v}}{\|\mathbf{v}\|^2}. \qquad (16)$$

Now, this is a Rayleigh quotient [11] which is upperbounded for any $\mathbf{v} \neq \mathbf{0}$ by the maximum eigenvalue of the matrix in the numerator of (16). This is in turn upperbounded as

$$\psi \leq \left(2 + \lambda_1^{-1}(\bar{\mathrm{Z}})\right)\lambda_1^{-1}(\bar{\mathrm{Z}}),$$

applying: 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 $M \times M$ matrices A and B it holds $\lambda_M(\mathrm{A} + \mathrm{B}) \leq \lambda_M(\mathrm{A}) + \lambda_M(\mathrm{B})$ [11]; and c) the (positive) eigenvalues of Z^{-1} are the inverses of the eigenvalues of Z. If we define next $\gamma_u \triangleq \lambda_1^{-1}(\mathrm{Z})\sigma_x^2$ then we can write

$$\psi \leq \left(2 + \frac{\gamma_u}{\xi}\right)\frac{\gamma_u}{\xi}.$$

B. REFERENCES

[1] 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.

[2] 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.

[3] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *3rd International Conference on Music Information Retrieval (ISMIR)*, Brescia, Italy, October 2002.

[4] 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.

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

[6] 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)*, 2005, pp. 498–503.

[7] R. L. Lagendijk and P. J. O. Doets, "Fingerprinting of audio signals," in *Technical Report, Delft University of Technology*, April 2005.

[8] 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.

[9] J. Imhof, "Computing the distribution of quadratic forms in normal variables," *Biometrika*, vol. 48, no. 3/4, pp. 419–426, December 1961.

[10] S.R. Searle, *Linear Models for Unbalanced Data*, John Wiley & Sons, 1987.

[11] Jan R. Magnus and Heinz Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, 2nd edition, 1999.