

# STOCHASTIC MODEL OF A ROBUST AUDIO FINGERPRINTING SYSTEM

*P.J.O. Doets, R.L. Lagendijk*

Dept. of Mediamatics, Information and Communication Theory group,  
Delft University of Technology, P.O. Box 5031, 2600 GA Delft  
{p.j.doets, r.l.lagendijk}@ewi.tudelft.nl

## ABSTRACT

An audio fingerprint is a compact representation of the perceptually relevant parts of audio content. A suitable audio fingerprint can be used to identify audio files, even if they are severely degraded due to compression or other types of signal processing operations. When degraded, the fingerprint closely resembles the fingerprint of the original, but is not identical. We plan to use a fingerprint not only to identify the song but also to assess the perceptual quality of the compressed content. In order to develop such a fingerprinting scheme, a model is needed to assess the behavior of a fingerprint subject to compression. In this paper we present the initial outlines of a model for an existing robust fingerprinting system to develop a more theoretical foundation. The model describes the stochastic behavior of the system when the input signal is a stationary (stochastic) signal. In this paper the input is assumed to be white noise. Initial theoretical results are reported and validated with experimental data.

## 1. INTRODUCTION

Downloading music is very popular. Music identification is usually done by searching in the metadata describing the music content. Metadata like song title, artist, etc., however, is often incoherent or misleading, especially on Peer-to-Peer (P2P) file-sharing networks, calling for content-based identification. Identification, however, is often not enough. The perceptual quality of a song compressed using MP3 at 32 kbps is totally different from the perceptual quality of the CD-recording of the same song. Therefore, a content-based indication for the perceptual quality is needed. The Music2Share project proposes to use audio fingerprints for both identification and quality assessment of unknown content on a P2P network [1].

Audio fingerprints are compact representations of the perceptually relevant parts of audio content that can be used to identify music based on the content instead of the

metadata. A fingerprinting system consists of two parts: fingerprint extraction and a matching algorithm. The fingerprints of a large number of songs are usually stored in a database. A song is identified by comparing its fingerprint with the fingerprints in the database. Well-known applications of audio fingerprinting are broadcast monitoring, connected audio, and filtering for file sharing applications [2], [3]. In this contribution we focus only on the fingerprint extraction.

For a well-designed fingerprint it holds that fingerprints of two arbitrary selected pieces of music are very different, while fingerprints originating from the same audio file, encoded using different coding schemes or bit rates, are only slightly different. We aim to exploit these small fingerprint differences due to compression to assess the perceptual quality of the compressed audio file.

By modelling a successful existing audio fingerprinting scheme, we aim to gain more insight in the behavior of fingerprint differences when the audio is compressed. The fingerprint generation is modelled when the system is subject to an input with known spectrum. Once this is better understood, a differential model which gives an indication of the perceptual quality of the compressed version with respect to the original will be made.

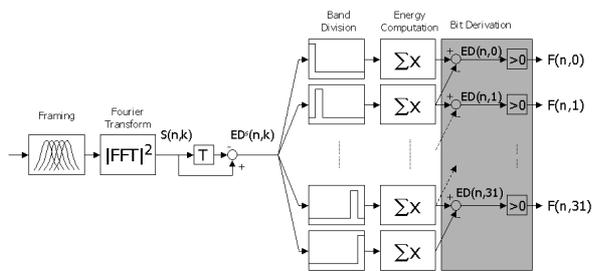
Several audio fingerprinting methods exist, using e.g. spectral flatness features [4] or Fourier coefficients [2], [5] and [6]. A good overview can be found in [3]. We choose to model the Philips audio fingerprinting system [2] because it is well documented, highly robust against compression and it can be modelled using stochastic models.

This paper consists of five sections: Section 2 presents details of the audio fingerprinting system to be modelled, Section 3 analyzes how to model the functional blocks of the fingerprinting system, Section 4 describes initial results for the fingerprint of white noise, while Section 5 draws conclusions and outlines future work.

## 2. DETAILS OF THE EXISTING SYSTEM

Figure 1 shows an overview of the fingerprint extraction, which is functionally equivalent to the Philips fingerprinting system [2]; some blocks, however, have been reshuffled to facilitate the modelling process described in the following sections. The audio signal is first segmented into frames of 0.37 seconds with an overlap factor of 31/32,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.  
© 2004 Universitat Pompeu Fabra.



**Figure 1.** Functional equivalent of the Philips audio fingerprinting system [2].

weighted by a Hanning window. The compact representation of a single audio frame is called a sub-fingerprint. In this way, it extracts 32-bit sub-fingerprints for every interval of 11.6 ms. Due to the large overlap, subsequent sub-fingerprints have a large similarity and slowly vary in time. A song consists of a sequence of sub-fingerprints, which are stored in a database. The system is capable of identifying a segment of about 3 seconds of music - generating 256 sub-fingerprints - in a large database, even if the segment is degraded due to signal processing operations.

To extract a 32-bit sub-fingerprint for every frame, 33 non-overlapping frequency bands are selected from the estimated Power Spectral Density (PSD). These bands range from 300 Hz to 2000 Hz and are logarithmically spaced, matching the properties of the Human Auditory System (HAS). Haitsma and Kalker report that experiments have shown that the sign of energy differences is a property that is very robust to many kinds of processing [2].

Using the notation of [2], we denote the energy of frequency band  $m$  of frame  $n$  by  $E(n, m)$ . Differences in these energies are computed in time and frequency:

$$ED(n, m) = E(n, m) - E(n, m+1) - (E(n-1, m) - E(n-1, m+1)). \quad (1)$$

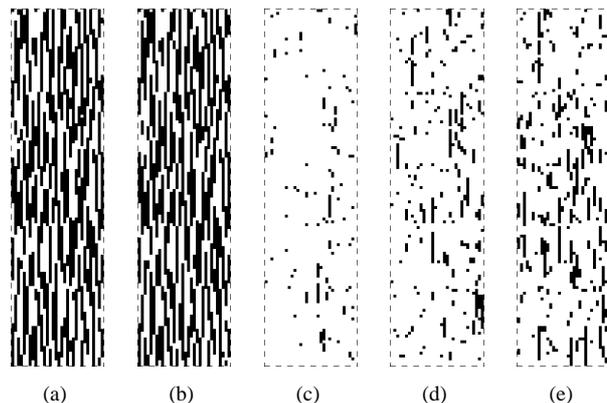
The bits of the sub-fingerprint are derived by

$$F(n, m) = \begin{cases} 1 & ED(n, m) > 0 \\ 0 & ED(n, m) \leq 0 \end{cases}, \quad (2)$$

where  $F(n, m)$  denotes the  $m^{\text{th}}$  bit of the sub-fingerprint of frame  $n$ .

Figure 2(a) shows an example of a fingerprint. White parts indicate positive energy differences, black parts indicate negative energy differences. The small side of the fingerprint block is the frequency direction, consisting of the 32 bits corresponding to the differences between the 33 frequency bands. The long side of the block corresponds to the temporal dimension.

The bit error rate between the extracted fingerprint and the fingerprint in the database is used as the similarity measure. When the song is subject to signal degradations such as compression, the fingerprint changes slightly. To indicate the effect of MP3 compression on the fingerprint extraction, Figure 2 shows the difference pattern of the fingerprint of a recording at different bit-rates relative to



**Figure 2.** Fingerprints for an excerpt of 'Anarchy in the U.K.' by the Sex Pistols. Fingerprints of (a) the original and (b) of an MP3 compressed version encoded at 128 kbps; (c-e) Differences between the fingerprints of the original and an MP3 compressed version encoded at (c) 128 kbps (d) 80 kbps and (e) 32 kbps. The black positions mark the differences.

the fingerprint of the CD-quality recording of the same song. The black sections mark the fingerprint differences due to compression, white positions indicate similarity between the fingerprints. Generally speaking, compression at a lower bit rate increases the number of fingerprint bit-errors. The bit-error pattern also shows a significant number of bursts of errors.

### 3. PROPOSED MODEL

We analyze the behavior of the system when it is subject to the input of a well-understood, stationary signal such as white noise or a  $p^{\text{th}}$  order auto-regressive (AR- $p$ ) process. From this analysis we will derive a fingerprint model.

Although for these kinds of signals the PSD is well-known, we need to keep two things in mind. First, the PSD is estimated using the periodogram, which has certain statistical properties. Second, we have to take into account the strong overlap between the subsequent frames, which causes the slow time-variation of the fingerprint. Therefore, even the fingerprint of a white noise process is non-white. In fact, white noise fingerprints look similar to the fingerprint shown in Figure 2(a).

The outcome of our proposed model is the probability that a fingerprint bit is equal to one, given the previous fingerprint bit corresponding to the same frequency position  $m$ , the frame length  $L$ , frame overlap length  $L - \Delta L$ , the number of frequency bands  $N$  and the autocorrelation function of the input signal  $R(l)$ :

$$P(F(n, m) = 1 | F(n-1, m); L, \Delta L, N, R(l)) \quad (3)$$

Other parameters, such as frequency range and sampling frequency are kept similar to the original system [2] (in this system  $L = 0.37$  seconds and  $\Delta L = 11.6$  ms).

The modelling procedure leading to equation (3) follows the steps show in Figure 1. Ignoring the (Hanning) window function, the periodogram estimator of the PSD of frame  $n$  is given by:

$$S(n, k) = \frac{1}{L} \left| \sum_{i=0}^{L-1} x(i + n\Delta L) e^{-j2\pi \frac{k}{L} i} \right|^2, \quad (4)$$

$$k = 0, \dots, L-1$$

The energy difference between two successive (overlapping) frames per spectral sample is given by:

$$ED^s(n, k) = S(n, k) - S(n-1, k) \quad (5)$$

The differential energy within a single frequency band  $m$  is then obtained by summing the samples belonging to that frequency band:

$$E(n, m) - E(n-1, m) = \sum_{k \in B_m} ED^s(n, k), \quad (6)$$

where  $B_m$  indicates the set of samples belonging to frequency band  $m$ . From  $ED^s(n, k)$  the energy differences between frequency bands can be computed by:

$$ED(n, m) = \sum_{k \in B_m} ED^s(n, k) - \sum_{k \in B_{m+1}} ED^s(n, k) \quad (7)$$

The fingerprint bits are derived according to Equation (2).

The statistical properties of four steps of the fingerprint generation expressed by equations (4)-(7), can be described in terms of Probability Density Functions (PDFs), starting with the PDF describing the statistical behavior of individual samples within the periodogram estimate of the PSD:

$$f_{S(n,k)}(y). \quad (8)$$

We may use this to derive a PDF for samples in  $ED^s(n, k)$ :

$$f_{ED^s(n,k)}(y). \quad (9)$$

From this we can extend the analysis to the statistical behavior of samples from *successive* energy difference spectra. This analysis leads to a PDF of the form

$$f_{ED^s(n,k), ED^s(n+1,k)}(y, z). \quad (10)$$

Introducing frequency bands leads to the following PDF:

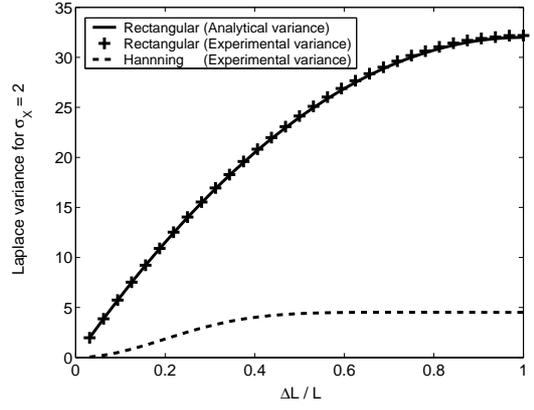
$$f_{ED(n,m), ED(n+1,m)}(y, z). \quad (11)$$

We can easily convert this PDF into probabilities for the fingerprint bits in two successive sub-fingerprints, e.g.:

$$P[F(n, m) = 1, F(n+1, m) = 1] \quad (12)$$

$$= \int_0^\infty \int_0^\infty f_{ED(n,m), ED(n+1,m)}(y, z) dy dz.$$

Equation (12) yields the model outlined in Equation (3).



**Figure 3.** Laplace variance as function of the ratio  $\Delta L/L$ , both for rectangular windows and for Hanning windows.

#### 4. WHITE NOISE FINGERPRINT

The section describes the statistical behavior of the fingerprint generation of a white noise input signal in terms of the PDFs and probabilities of Equations (8)-(12). The fingerprint generation process of white noise is worth modelling for three reasons. First, the spectrum of an AR- $p$  process is the filtered version of corresponding white noise process. Second, differences in the audio spectrum due to compression might be modelled as (filtered) noise. Third, fingerprint generation process of white noise is the most simple example that still provides insight and shows the generally applicable fingerprint behavior.

This section is organized as follows: Section 4.1 gives an expression for the marginal PDF  $f_{ED^s(n,k)}(y)$ ; Section 4.2 investigates the joint-PDF for frequency band energies in successive frames,  $f_{ED(n,m), ED(n+1,m)}(y, z)$ . It shows that this distribution can be approximated well by a Gaussian PDF, resulting in the  $P[F(n, m), F(n+1, m)]$ . Rectangular windows are assumed unless stated otherwise.

##### 4.1. Marginal Distribution

Leon-Garcia shows that when the input signal  $x(i)$  is white noise, ( $x(i)$  thus is a zero-mean, Gaussian random variable (RV) with variance  $\sigma_X^2$ ), the samples of the periodogram estimate of the PSD using rectangular windows are uncorrelated and follow an exponential distribution:

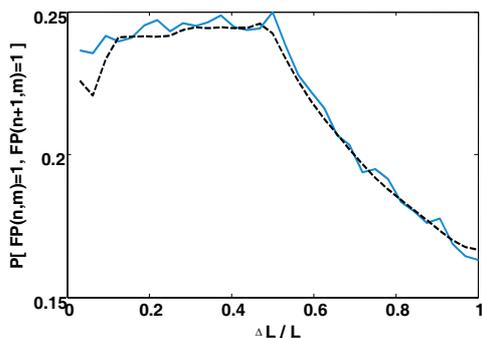
$$f_{S(n,k)}(y) = \frac{1}{\sigma_X^2} e^{-y/\sigma_X^2}, \quad y \geq 0. \quad (13)$$

Using this result and two properties of the Laplace distribution [8], it can be shown that the samples of  $ED^s(n, k)$  follow a Laplace PDF [9]. In terms of the parameters in Section 3:

$$f_{ED^s(n,k)}(y | L, \Delta L, R(l) = \sigma_X^2 \delta(l)) = \frac{1}{2s} e^{-|y|/s} \quad (14)$$

with the variance of the Laplace distribution,  $2s^2$ :

$$\text{VAR}[ED^s(n, k)] = 2s^2 = 2\sigma_X^4 \frac{(2L - \Delta L)\Delta L}{L^2} \quad (15)$$



**Figure 4.** Analytical (dashed) and experimental (solid) probability  $P[F(n, m) = 1, F(n + 1, m) = 1]$  for white noise input using windows.

Equations (14) and (15) have been verified experimentally. Figure 3 shows the variance as a function of the ratio  $\frac{\Delta L}{L}$ . The same graph also shows the behavior of the variance when Hanning windows are used. The behavior is slightly different, the derivation of (and the expression for) the variance much more complex [9].

#### 4.2. Bivariate Distribution: Gaussian approximation

Deriving an analytical expression for the bivariate PDF  $f_{ED^s(n,k), ED^s(n+1,k)}(y, z)$  is very difficult and not necessary. Experiments confirm the suitability of the Gaussian approximation of  $f_{ED(n,m), ED(n+1,m)}(y, z)$ . This approximate PDF is then fully defined by the marginal variances  $\sigma_{ED(n,m)}^2 = \sigma_{ED(n+1,m)}^2$  and the correlation coefficient  $\rho$ . Inspection of the system learns that the correlation between  $ED(n, m)$  and  $ED(n + 1, m)$  can be expressed in terms of variances:

$$\rho = \frac{1}{2} \frac{\text{VAR}[ED(n, m)|2\Delta L]}{\text{VAR}[ED(n, m)|\Delta L]} - 1 \quad (16)$$

So, expressing the variance of  $ED(n, m)$  in terms of  $\frac{\Delta L}{L}$  fully defines the Gaussian approximation. To compute this variance, however, we also have to take the correlation between the samples of  $ED^s(n, k)$  into account. Full derivations and expressions for this correlation function and the resulting variance can be found in [9].

Now we can compute the marginal variances and the correlation coefficient of the Gaussian approximation, the probabilities  $P[F(n, m), F(n + 1, m)]$  can be computed according to Equation (12). Figure 4 shows the computed probabilities  $P[F(n, m) = 1, F(n + 1, m) = 1]$  along with its experimental counterparts. This clearly shows the suitability of the Gaussian approximation for our purpose.

### 5. CONCLUSION AND FUTURE WORK

This paper presents initial outlines and results for a stochastic model of a state-of-the-art audio fingerprinting system. The longer-term aim of the modelling is to assess the perceptual quality of a compressed song with respect to the original high quality recording by using its fingerprint.

A model for the fingerprint generation of white noise input signals is presented. Its relevance lies in its possible extension to AR- $p$  signals, the possible modelling of compression effects by additive (filtered) noise and its relative simplicity. It provides analytical expressions and approximations for the statistics of the individual steps of the fingerprint generation process, resulting in fingerprint-bit probabilities assuming rectangular windows.

Future work concerning the white noise model will be the extension to Hanning windows and the incorporation of the correlation of fingerprint-bits in the spectral dimension. Future work concerning the overall model will be the extension to AR- $p$  input signals and the modelling of fingerprint differences when the input signal is compressed.

### 6. REFERENCES

- [1] Kalker, T., Epema, D.H.J., Hartel, P.H., Lagendijk, R.L. and van Steen, M. "Music2Share - Copyright-Compliant Music Sharing in P2P Systems", vol. 92, no. 6, pp. 961- 970, *Proceedings of the IEEE*, June 2004.
- [2] Haitsma, J. and Kalker, T. "A Highly Robust Audio Fingerprinting System", pp. 144-148, *Proc. of the 3rd Int. Symposium on Music Information Retrieval*, Oct. 2002.
- [3] Cano, P., Battle, E., Kalker, T. and Haitsma, J. "A Review of Algorithms for Audio Fingerprinting", pp. 169-173, *Proc. of the Int. Workshop on Multimedia Signal Processing*, Dec. 2002.
- [4] Allamanche, E., Herre, J., Hellmuth, O., Fröbach, B. and Cremer, M. "AudioID: Towards Content-Based Identification of Audio Material", *Proc. of the 100th AES Convention*, May 2001.
- [5] Cheng, Y. "Music Database Retrieval Based on Spectral Similarity", pp. 37-38, *Proc. of the 2nd Int. Symposium on Music Information Retrieval*, Oct. 2001.
- [6] Wang, A. "An Industrial Strength Audio Search Algorithm", *Proc. of the 4th Int. Symposium on Music Information Retrieval*, Oct. 2003.
- [7] Leon-Garcia, A. *Probability and Random Processes for Electrical Engineering*, 2nd Edition, ISBN 0-201-50037-X, Addison-Wesley Publishing Company, Inc., 1994.
- [8] Kotz, S., Kozubowski, T.J. and Podgórski, K. *The Laplace Distribution and Generalizations*. ISBN 3-7643-4166-1, Birkhäuser, 2001.
- [9] Doets, P.J.O., *Modelling a Robust Audio Fingerprinting System*. Technical Report, <http://ict.ewi.tudelft.nl>, 2004.