A Fingerprinting System for Musical Content
Lahouari Ghouti and Ahmed Bouridane

Abstract—Digital multimedia content (especially audio) is becoming a major part of the average computer user experience. Large digital audio collections of music, audio and sound effects are also used by the entertainment, music, movie and animation industries. Therefore, the need for identification and management of audio content grows proportionally to the increasing widespread availability of such media virtually "any time and any where" over the Internet. In this paper, we propose a novel framework for musical content fingerprinting using balanced multiwavelets (BMW). The framework for generating robust perceptual fingerprinting (or hash) values is described. The generated fingerprints are used for identifying, searching, and retrieving audio content from large digital music databases. Furthermore, we illustrate, through extensive computer simulation, the robustness of the proposed framework to efficiently represent musical content and withstand several signal processing attacks and manipulations.

I. INTRODUCTION

This paper describes the details of a novel framework for robust perceptual fingerprinting of musical content. Perceptual fingerprinting (or hashing), unlike conventional hashing used in cryptography, represents a unique binary string or code that uniquely identifies a segment of musical content similar to personal fingerprints used to identify human beings. Several applications can be foreseen for perceptual fingerprinting such as: 1) content identification; 2) broadcast monitoring; 3) connected audio; 4) filtering technology for peer-to-peer (P2P) networks; 5) automatic music library organization. The proposed framework consists of two sub-systems: The first system generates of extracts the fingerprint values from the audio content, while the second sub-system applies an efficient design for the fingerprint search algorithm to enable efficient matching of fingerprint blocks in the queried database. In the remaining of this paper, we focus on the design and implementation of a robust perceptual fingerprinting algorithm. For notational convenience, we will inherently assume that the fingerprint extraction procedure is secured through the use of secret keys, and therefore, we will drop the subscript $K$ in the representation of the fingerprint function; it will be denoted by $H_X$ for an input content $X$.

II. A ROBUST MUSICAL FINGERPRINTING ALGORITHM

In the remaining of this paper, we focus on the design and implementation of a robust perceptual fingerprinting algorithm. For notational convenience, we will inherently assume that the fingerprint extraction procedure is secured through the use of secret keys, and therefore, we will drop the subscript $K$ in the representation of the fingerprint function; it will be denoted by $H_X$ for an input content $X$. We present the
corresponding design algorithm and some simulation results. We experimentally show that the proposed algorithm achieves Eqs. 1-2 for an extensive range of attacks.

\[
Pr \left[ D(H_K(X), H_K(\hat{X})) = 0 \right] \approx 1, \quad (1)
\]

\[
Pr \left[ D(H_K(X), H_K(Y)) > 0 \right] \approx 1, \quad (2)
\]

where \( H_K(X) \) and \( H_K(\hat{X}) \) represent the fingerprint representations for the input \( X \) and a similar input \( \hat{X} \), respectively. \( D(\ldots) \) represents a distance metric such as Hamming distance [1].

A. Design Principles

Perceptual audio fingerprinting aims at extracting the relevant perceptual features from an audio content. At the same time, implementation requirements impose that extracting and searching fingerprint values should be fast and easy, preferably with a small granularity to allow system scalability in highly demanding applications (e.g. mobile-enabled music identification). However, many issues should be addressed before proceeding with the design and implementation of such systems. The most prominent ones are [1]:

- Features’ selection (semantic vs. non-semantic features).
- Fingerprint representations.
- Fingerprint granularity.

B. Fingerprint Extraction Algorithm

Most fingerprinting extraction algorithms process audio content in a similar way to the techniques in audio coding and processing [2]. In order to reduce the computational load, we reduce the sampling rate of the processed audio clip by a factor of 8. It is assumed that the input audio is sampled at a CD-Quality rate, i.e., 44100Hz. Then, the audio signal is segmented into audio frames. For every frame, a set of features is computed as explained below. It should be kept in mind that the extracted features are chosen to achieve perceptual invariance to content degradations. Unlike [1] where the content features are extracted from Fourier coefficients, we propose the use of subband coefficients of balanced multiwavelets. Features derived from Fourier coefficients are also proposed in [1], [3]. In [4], Logan proposes the use of non-semantic features based on the Mel Frequency Cepstral Coefficients (MFCC). Allamanche et al. [5] propose the use of spectral flatness, sharpness, and coefficients of the linear predictive coding (LPC). Mathematically-derived quantities are also used to represent the content features. These derived quantities such as the means and variances of the features are mapped into more compact representations using either hidden Markov models (HMM) [6] or quantization [1]. We extract from each audio frame an sub-fingerprint string that is not sufficient to identify a complete audio clip. However, the combination of several sub-fingerprint strings will produce a global fingerprint value that will sufficiently describe the overall audio clip. Fig. 2 gives an overview of the fingerprint extraction procedure proposed for the perceptual audio fingerprinting system.

Each audio frame is transformed using a one-dimensional balanced multiwavelet transform. Then, for each audio frame (time interval of 11.8 milliseconds), a sub-fingerprint string of 32 bits is extracted. For an audio clip of 3 sec, we will have 256 different audio frames. Therefore, the system guarantees a granularity of 3 seconds. It is worth noting that in order to avoid signal discontinuities, we propose to use overlapping frames with an overlap factor of 31/32 [1]. All the audio frames are weighted with a Hanning window having the same overlap factor of 31/32. Based on this approach, we obtain a sub-fingerprint binary for every 11.6 milliseconds. Furthermore, in order to capture the most important perceptual features, we propose to use a multisolution decomposition based on the balanced multiwavelet transform. Such a decomposition has the merits to provide an excellent model of the human auditory system (HAS). On the other side, frame boundaries have a negative effect on the content phase. However, because the HAS system is relatively insensitive to phase, we propose to use the coefficients’ magnitude represented by the estimation quantization (EQ) scheme [7]. For the extraction of the sub-fingerprint binary strings, we divide the decomposition subbands into 32 different frequency bands. In Fig. 3 shows the approach used in the selection of the frequency blocks for the lowpass subband at decomposition level 5. Similar approach is used for the remaining subbands with a varying number of blocks. Unlike the system proposed in [1], the algorithm described here allows for a frequency decomposition that is similar to that performed by the HAS system which operates on the Bark scale [8]. In [1], to obtain a decomposition similar to the HAS-based one, a logarithmic spectrum division is carried out to obtain 32 different blocks.

Details of the extraction algorithm are given below:

1) Downsampling the input audio content of 3 sec to obtain

\[
2\text{We propose to use the class of non-semantic features for their mathematical tractability and ease of computation.}
\]
a sampling rate of 5512 Hz.
2) Apply the framing division on the downscaled content using Hanning window with an overlap factor of 31/32.
3) Compute the forward balanced multiwavelet (BMW) transform for each audio frame using 5 decomposition levels.
4) Divide the subbands’ coefficients into 32 different blocks as illustrated in Fig. 3 for the case of the finer lowpass subband.
5) Apply the estimation quantization (EQ) scheme using a neighboring window of 5 audio samples.
6) Compute the log variances of the magnitudes of the subbands’ coefficients.
7) Compute the mean value, \( \mu^{n}_{EQ} \), of all the log variances for each audio frame where \( n = 1, 2, \ldots, 256 \).
8) For each of the 32 subband blocks, apply the following formula to extract the sub-fingerprint bit:

\[
H(n, m) = \begin{cases} 
1 & \text{if } \sigma^{(m, n)}_{EQ} > \mu^{n}_{EQ} \\
0 & \text{if } \sigma^{(m, n)}_{EQ} < \mu^{n}_{EQ} 
\end{cases} \\
1 \leq n \leq 256 \text{ and } 1 \leq m \leq 32
\] (3)

where \( H(n, m) \) is the \( n^{th} \) bit in the sub-fingerprint string representing the \( m^{th} \) audio frame.

III. PERFORMANCE ANALYSIS

A. Statistical Analysis

Based on the similarity metric, used in the proposed system, we declare two audio contents perceptually similar if the normalized Hamming distance is below a specific threshold, \( T_h \). The false positive rate, \( P_{fp} \), can be directly determined from the threshold \( T_h \) [8]. For an adequate choice of \( T_h \), we assume that the extracted fingerprint strings are independent and identically distributed (i.i.d) random variables. Therefore, the number of bit errors will follow a binomial distribution \((n, p)\) [8], where \( n \) represents the length of the extracted fingerprint strings and \( p \) represents the probability that the extracted bit is ‘0’ or ‘1’. For a random variable with \((n, p)\) binomial distribution, the probability is given by [8]:

\[
P\{ y = k \} = \binom{n}{k} p^k (1-p)^{n-k} k = 0, 1, 2, \ldots, n
\] (4)

In this case, the corresponding distribution is a staircase function [8]. An interesting asymptotic approximation for the binomial distribution is given by the normal distribution for sufficiently large values of \( n \) [8]. For a fixed value of \( p \) (we have \( p = 0.5 \)), and a large value of \( n = 8192 \), DeMoivre-Laplace theorem gives the following approximation [8]:

\[
\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}
\] (5)

where the mean is given by \( \mu = np \) and the standard deviation is given by \( \sigma = \frac{1}{\sqrt{np(1-p)}} \). Using Eq. 5, for a given fingerprint string \( H_1 \), the probability that a randomly selected fingerprint string \( H_2 \) has less than \( T = \alpha n \) bit errors with respect to \( H_1 \) is given by [1], [8]:

\[
P_{fp}(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{(1-2\alpha)\sqrt{n}}^{\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{2} \text{erfc} \left( \frac{(1-2\alpha)}{\sqrt{2}} \sqrt{n} \right)
\] (6)

where \( \alpha \) denotes the BER rate and \( \text{erfc} \) represents the error function [8]. Finally, to take into account the larger standard deviation of the BER distribution, Eq. 6 is modified as follows [1]:

\[
P_{fp}(\alpha) = \frac{1}{2} \text{erfc} \left( \frac{(1-2\alpha)}{3\sqrt{2}} \sqrt{n} \right)
\] (7)

B. System Robustness

In order to assess the robustness of the proposed fingerprinting system, we need to address the issue of the performance measure to be used during the evaluation experiments. It is legitimate to assume a given value for the threshold \( T_h \) for the evaluation purposes. In [1], Haitjema et al. assume a threshold of \( T_h = 0.35 \). To illustrate the robustness of the proposed algorithm, we will lower the latter threshold further to \( T_h = 0.25 \). This new threshold will be considered as the upper performance bound for the system described herein. Throughout the simulation carried out to assess the system robustness, we use two different audio clips, namely "Manchild" clip by Neneh Cherry and "Last Breath" clip by Abu Khater. During the evaluation experiments, both original audio clips will be subjected to the following signal degradations/attacks:

- MP3 Encoding/Decoding
- Lowpass and Highpass Filtering
- Noise Addition
- Silence Reduction and Addition
- Amplitude Alteration
- Echo Attacks
- Stretching and Pitch Bending
- Dynamic Delay
- Content Mixing
- Denoising and Hiss Reduction
posed fingerprint extraction mechanism yields totally different fingerprint sequences for different musical content.

V. ACKNOWLEDGMENT

The authors would like to thank J. Haitsma (Philips Company, Holland) and B. Coskun (Boğaziçi University, Turkey) for their valuable collaboration.

REFERENCES