A Robust Perceptual Audio Hashing Using Balanced Multiwavelets

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 robust perceptual hashing of audio content using balanced multiwavelets (BMW). The framework for generating robust perceptual hash values (or fingerprints) is described. The generated hash values are used for identifying, searching, and retrieving audio content from large audio databases. Furthermore, we illustrate, through extensive computer simulation, the robustness of the proposed framework to efficiently represent audio content and withstand several signal processing attacks and manipulations.**

I. Introduction

This paper describes the details of a novel framework for robust perceptual hashing of audio content. Perceptual hashing, unlike conventional hashing used in cryptography, represents a unique binary string or code that uniquely identifies a segment of audio content (such as music) similar to personal fingerprints used to identify human beings. Several applications can be foreseen for audio hashing 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 for robust perceptual hashing consists of two sub-systems: The first system generates of extracts the hash values from the audio content, while the second sub-system applies an efficient search scheme to identify the extracted hash value from an existing multimedia database that represents the stored content by their extracted hash values. Therefore, instead of using the audio content in the search/identification/retrieval operations, we will base these operations on the extracted hash values which allow for efficient database queries. In the proposed solution, hash values are extracted using an algorithm that generates bit sequences of 32 bits length for every 11.72 ms of audio content. Then, the hash representations of 256 different audio sequences are combined to uniquely represent every 3 sec of audio content. Such a representation contains enough "*consistent*" information to allow for reliable content identification. It is worth noting that with such representations, the search space has a dimension of $2^{256\times 32}$. Such a space dimension dictates an efficient design for the hash search algorithm to enable efficient matching of hash blocks in the queried database. Obviously, this is not a trivial task, because hash blocks are expected to have different binary representations in the database. Many factors contribute to such differences. For instance, we may be searching for a content that has been represented in the database using either different file format or compression rates such as in the case of WAV versus MP3 formats or different audio quality (CD-Quality, FM-Quality, etc.). The foundation of the proposed system is the statistical modeling of balanced multiwavelet representations of audio signals. The results of the statistical model are used to build novel robust hash values for identifying manipulated audio content. In most of the tested cases, the proposed framework has been evaluated and informed by conducting several performance evaluation tests. New contributions of this work to the area of "*Computer Audition*" include: robust feature extraction from audio content, analysis of audio content using balanced multiwavelets [1], robust technique for audio content identification using perceptual hashing or fingerprinting. Fig. 1 shows a schematic diagram of the underlying principles of the proposed framework for robust perceptual audio hashing.

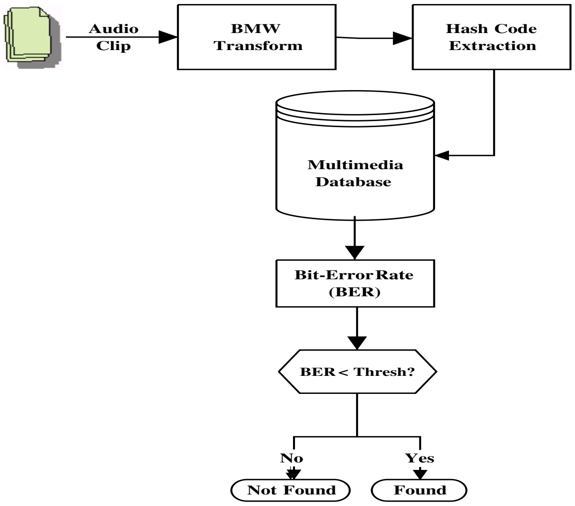

Fig. 1. Proposed framework for robust perceptual audio hashing.

II. A Robust Perceptual Audio Hashing Algorithm

In the remaining of this paper, we focus on the design and implementation of a robust perceptual audio hashing algorithm. For notational convenience, we will inherently assume that the hashing extraction procedure is secured through the

L. Ghouti and A. Bouridane are with the School of Computer Science, Queen's University of Belfast, Belfast BT7 1NN, UK. Email: {LGhouti01,A.Bouridane@qub.ac.uk}

[1] To the best of our knowledge, we are unaware of audio analysis using balanced multiwavelets reported in the literature.

use of secret keys, and therefore, we will drop the subscript K in the representation of the hash 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 achieve Eqs. 1-2for an extensive range of attacks.

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

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

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

A. Design Principles

Perceptual audio hashing aims at extracting the relevant perceptual features from an audio content. At the same time, implementation requirements impose that extracting and searching hash 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 [2] vs. non-semantic features).
- Hash representations.
- Hash granularity.

B. Hash Extraction Algorithm

Most hashing 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 (MFFC). 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-hash string that is not sufficient to identify a complete audio clip. However, the combination of several sub-hash strings will produce a global hash value that will sufficiently describe the overall audio clip. Fig. 2 gives an overview of the hash extraction procedure proposed for the perceptual audio hashing system.

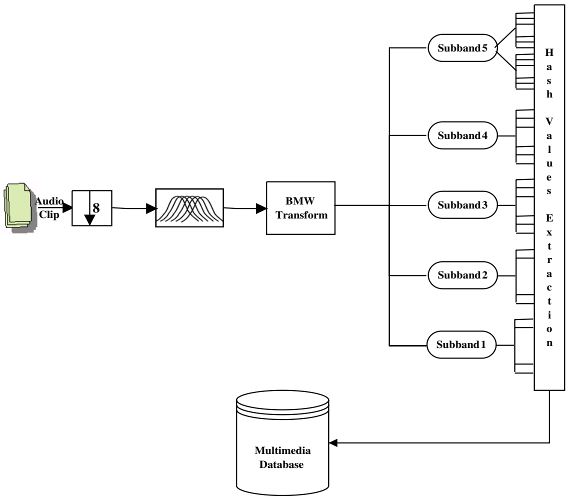

Fig. 2. Overview of audio hashing extraction.

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-hash string of 32 bits is extracted. For an audio clip of 3 seconds, 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-hash binary for every 11.6 milliseconds. Furthermore, in order to capture the most important perceptual features, we propose to use a multiresolution 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-hash 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*. In [1], to obtain a decomposition similar to the HAS-based one, a logarithmic spectrum division is carried out to

[2] We propose to use the class of non-semantic features for their mathematical tractability and ease of computation.

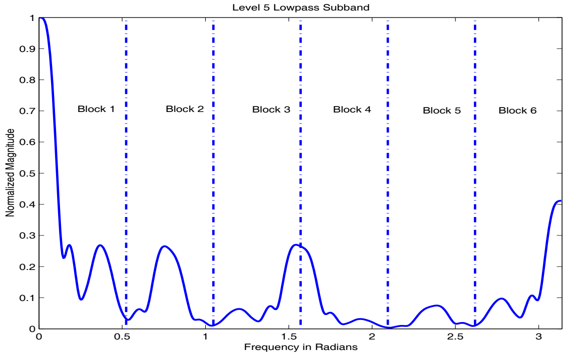

Fig. 3. Frequency block selection for lowpass subband at decomposition level 5.

obtain 32 different blocks.

Details of the extraction algorithm are given below:

1) Downsample the input audio content of 3 sec to obtain a sampling rate of 5512 Hz.
2) Apply the framing division on the downsampled 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, μ_{EQ}^{n}, 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-hash bit:

$$H(n,m) = \begin{cases} 1 & \text{if} \quad \sigma_{EQ}^{(m,n)} > \mu_{EQ}^{n} \\ 0 & \text{if} \quad \sigma_{EQ}^{(m,n)} < \mu_{EQ}^{n} \end{cases}$$
$$1 \le n \le 256 \quad \text{and} \quad 1 \le m \le 32 \qquad (3)$$

where $H(n,m)$ is the n^{th} bit in the sub-hash 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 hash 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 hash 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]:

$$P\left\{\mathbf{y} = k\right\} = \left(\begin{array}{c} n \\ k \end{array}\right) p^{k} \left(p-1\right)^{n+k} k = 0, 1, 2, \ldots, n \qquad (4)$$

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]:

$$\left(\begin{array}{c} n \\ k \end{array}\right) p^{k} \left(p-1\right)^{n+k} \approx \frac{1}{\sqrt{2\pi np(p-1)}} e^{-(k-np)^2/2np(p-1)} \qquad (5)$$

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

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

where α denotes the BER rate and 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]:

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

B. System Robustness

In order to assess the robustness of the proposed hashing 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], Haitsma 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 Breadth*" 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

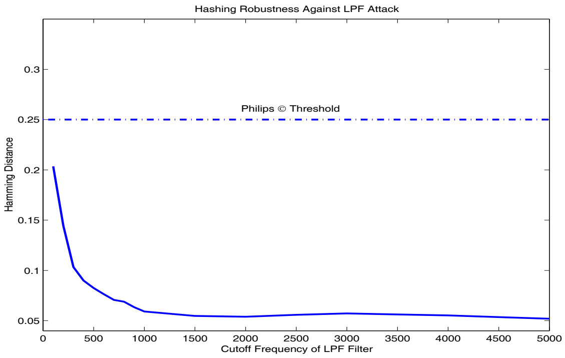

Fig. 4. System performance in the presence of lowpass filtering attack.

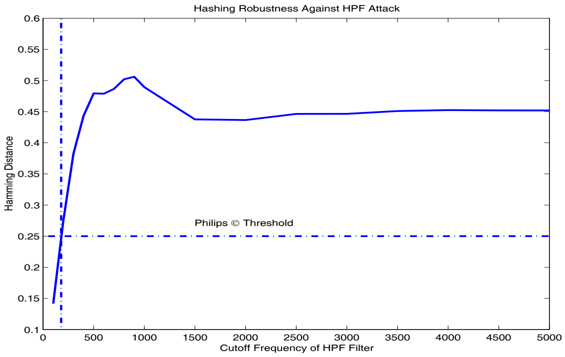

Fig. 5. System performance in the presence of highpass filtering attack.

For illustration purposes, we report the results related to Manchild clip only. Fig. 4 shows the robustness of the proposed system against lowpass filtering attacks. The performance curve indicates that the system is performing adequately as long as the main content energy is not removed by the filtering attack. Fig. 5 shows the robustness of the proposed system against highpass filtering attack. An interesting interpretation of the obtained performance indicates that the content frequency from 100 Hz to 4000 Hz is crucial for a reliable perceptual representation of the content. Fig. 6 shows the robustness of the proposed system against MP3 compression attacks. By analyzing the performance reported, we can conclude that under MP3 compression attack, the proposed system yields an excellent performance for compression rates higher or equal to 24 KBits/sec. However for compression rates lower than this threshold value, the system performance degrades gradually. The system robustness to various audio attacks is illustrated in Fig. 7. It is clear from this figure that the system performs robustly in the presence of various audio attacks with the exception of the double stretching attack.

IV. Conclusions

In this paper, we have proposed a novel framework for robustly identifying audio content by using short robust hashing codes. We have proven, through extensive computer simulation, that these hashing codes are robust against various audio alterations and attacks. Furthermore, the proposed hash extraction mechanism yields totally different hash sequences for different audio content.

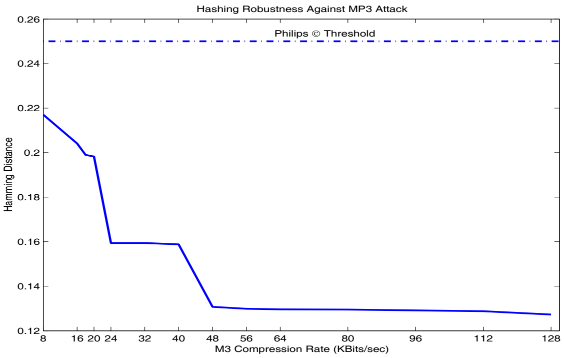

Fig. 6. System performance in the presence of MP3 compression attack.

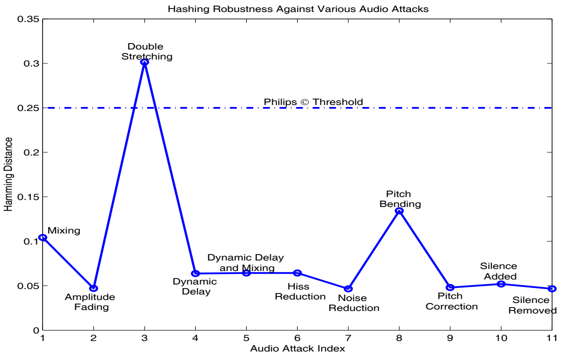

Fig. 7. System performance in the presence of various audio attacks.

V. Acknowledgment

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

References

[1] J. A. Haitsma, “Audio Fingerprinting: A New Technology to Identify Music,” *Unclassified Report 2002/824*, Philips Electronics N.V., Eindhoven, Netherlands, Aug. 2002.

[2] B. Gold and N. Morgan, “ Speech and Audio Signal Processing: Processing and Perception of Speech and Music,” *John Wiley and Sons*, 2000.

[3] D. Fragoulis, G. Rousopoulos, T. Panagopoulos, C. Alexiou and C. Papaodysseus, “On the Automated Recognition of Seriously Distorted Musical Recordings,” *EEE Trans. on Signal Processing*, vol. 49, no. 4, pp. 898-908, Apr. 2001.

[4] B. Logan, “ Mel Frequency Cepstral Coefficients for Music Modeling,” *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, Oct. 2000.

[5] E. Allamanche, J. Herre , O. Hellmuth, B. Bernhard Fr´obach and M. Cremer, “AudoID: Towards Content-Based Identification of Audio Material,” *Proceedings of 100th AES Convention*, May 2001.

[6] H. Neuschmied, H. Mayer and E. Battle, “Identification of Audio Titles on the Internet,” *Proceedings of International Conference on Web Delivering of Music*, Nov. 2001.

[7] L. Ghouti, A. Bouridane, M. K. Ibrahim, and S. Boussakta, “Digital image watermarking using balanced multiwavelets,” *IEEE Trans. on Signal Processing*, vol. 54, no. 3, pp. XXX–XXX, Mar. 2006.

[8] A. Papoulis and S. U. Pillai, “Probability, Random Variables and Stochastic Processes,” *McGraw Hill*, 2002.