A HIGHLY ROBUST AUDIO HASHING SYSTEM USING AUDITORY-BASED FRONT-END PROCESSING

Abderraouf Ben Salem[1], *Sid-Ahmed Selouani*[1], *Habib Hamam*[2,3], *and Jean Caelen*[4]

[1]Université de Moncton, campus de Shippagan E8S 1P6 NB, Canada
bensalem@creatis.insa-lyon.fr, selouani@umcs.ca
[2]Université de Moncton, campus de Moncton E1A 3E9 NB, Canada
[3]Canadian University of Dubai, Dubai, UAE
habib.hamam@umoncton.ca
[4]IMAG (UJF & CNRS) BP 53, 38041, Grenoble Cedex 9, France
caelen@imag.fr

ABSTRACT

In this paper, a robust perceptual audio hashing system is presented. A model of the human auditory system is used to extract robust features from the outputs of a non-linear filter bank that mimics the human basilar membrane. Experiments on various audio excerpts show that this new ear-based front-end processing provides very effective hash values. The proposed audio hashing system performs very satisfactorily in identification and it turned out very resilient to a large variety of severe audio attacks.

Index Terms— audio hashing, ear model, filter bank, audio attacks, hash value, bit error rate.

1. INTRODUCTION

The increasing of audio material leads to the need of identifying a given audio clip throughout a huge database or metabase without a specific retrieval method. One method consists of using hash functions to extract a hash value (fingerprint) which ensures both content integrity and ease of identification. Such functions are used to obtain a final binary representation of the corresponding audio clip. However, many of the conventional hash functions are not efficient in such multimedia applications due to content-based audio signal representation. In this paper, we propose a new framework which summarizes a given long audio signal into a concise and robust signature sequence called $hash\ value$. For this purpose, a model of the human auditory system is used to extract robust features from the outputs of a non-linear filter bank that mimics the human basilar membrane. A perceptual audio hashing function is then obtained to reflect the perceptual component of the content. Such perceptual hash functions can be used for several applications such as searching an audio record related to a specific track (artist, title, etc.). Other possibilities include content identification, copyright-related applications to prove rightful ownership, monitoring the distribution, indexing multimedia libraries, detecting content attacks, etc. An ideal audio hashing system should fulfill several requirements. It should be as invariant as possible under severe signal degradations such as compression. The size of each hash value must be short enough in order to be stored in a database while providing a content sufficient to characterize and to identify an individual audio document. Hence, an audio hashing system must derive a set of significant perceptual features of a recording in a robust and concise form. The most important requirements according to [2] include the discrimination power over huge numbers of other hash values, the robustness, the efficacy of the representation, and the less computational complexity.

Haitsma et al. [4] extract 32-bit sub-hash values for every frame of a specific audio excerpt of 3 seconds. These frames are overlapped using a 31/32 Hanning window with the same overlapping factor. Through a Fast Fourier Transform the output is then shared out between 33 logarithmically spaced frequency bands. The resulting hash value is computed by comparing adjacent energy bands. Such results were combined to build a method providing a hash value of an audio-visual clip using only the audio-content signal part [1].

Three perceptual audio hashing algorithms are proposed in [6]. Two of them use periodicity-based hash functions that exploit the periodicity measured by either a least-squares estimation or by a correlation-based analysis. The third algorithm uses the time-frequency domain transform, namely the MFCC-based feature extraction method. An alternative solution, proposed in [3], consists of using a balanced multiwavelets transformation for each audio frame using 5 decomposition levels. The hash values are computed by comparing the mean of log variances of each audio frame for each subband. Several perceptual audio hashing algorithms found in the literature are summarized in [2].

This paper is further organized as follows. In section 2,

the main objectives and technical requirements of our system are outlined. Next, in section 3, we present the proposed auditory-based framework that permits the computation of the features providing robust hash values. Then, in section 4, the experimental results are presented and discussed. Finally, in section 5, we conclude our work.

2. GOALS

As mentioned previously, an audio hashing system must obey several requirements. As two different signal-based content audio clips may include the same information according to the human auditory sense, it is interesting to build an efficient audio hashing system that mimics the human auditory system. Herein, we propose a new framework using the ear model. For a given audio clip A, let $\hat{A}$ denotes a modified recording of this clip which is perceptually the same as A. Let B a perceptually different audio clip. $H_K(.)$ represents a hash function secured by the subscript secret key K and takes as an input the excerpt audio signal. Therefore, an attacker cannot forge the audio signature. Our aim is to achieve the following probabilities:

$$Pr[H_K(A) = H_K(\hat{A})] \approx 1, \tag{1}$$

and

$$Pr[H_K(A) = H_K(B)] \approx 0. \tag{2}$$

Throughout this paper, we use the Hamming distance, normalized by the size of the hash, as a metric to prove the robustness of the proposed audio hashing system submitted to severe audio attacks. Let $D(.,.)$ denotes this metric. Two thresholds, T_l and T_h, have been defined to decide if two audio clips are perceptually the same or not [5][3].

$$D(H_K(A), H_K(\hat{A})) \leq T_l, \tag{3}$$

and

$$D(H_K(A), H_K(B)) \geq T_h. \tag{4}$$

where $T_h > T_l$. The lower threshold T_l is used as a criterion to evaluate the robustness under audio attacks. The upper threshold T_h allows evaluating the discrimination power of a given audio hashing system. The thresholds used in our experiments are given in section 4.

3. PROPOSED ROBUST AUDIO HASHING SYSTEM

The human auditory system (HAS) consists of three parts which simulate the behavior of the ear. As depicted in Fig. 1, the external and middle ear are modeled using a bandpass filter that can be adjusted to signal energy to take into account the various adaptive motions of ossicles. The next part of the model simulates the behavior of the basilar membrane (BM), the most important part of the inner ear, that acts substantially as a non-linear filter bank. Each location along the BM has a specific frequency, at which it vibrates maximally for a given input sound. In our experiments we have considered 32 filters. This number depends on the sampling rate of the signals (16 kHz) and on other parameters of the model such as the overlapping factor of the bands of the filters, or the quality factor of the resonant part of the filters. The final part of the model deals with the electro-mechanical transduction of hair-cells and afferent fibers and the encoding at the level of the synaptic endings. In order to not over-emphasize the problem of electro-mechanical transduction in hair cells and fibers, only the coupling effects are taken into account in the model. Thus, the main feature of the model retained for hair cells and fibers is supplied by coupling parameters and used by the sample-by-sample algorithm described in [7]. $y'_i(k)$ provided by the algorithm can be regarded as the resulting stimulus after the passage through the mid-external ear, the basilar membrane with the effect of hair cells, and afferent fibers. This stimulus is used to calculate the energy at each output of the cochlear filters providing 32 features that will be used to compute the hash value.

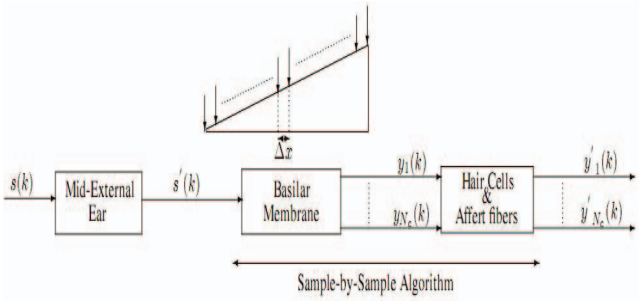

Fig. 1. Ear-Based model presentation and the basilar membrane modeled as a triangle divided into 32 cochlear filters.

3.1. Front-End Processing

The energy of the stimulus propagated through the nerve fibers along each portion Δx of the cochlea is calculated and lightly smoothed in order to be exploited for extracting pertinent information. The absolute energy of each channel, m, is given by:

$$E'(n,m) = 20\, log \sum_{k=1}^{L} |y'_{n,m}(k)|. \tag{5}$$

In equation (5), n refers to the frame index and L to the frame size. Between the current and the previous frame, a smoothing function is applied to smooth the energy fluctuations. The smoothing equation is:

$$E(n,m) = c_0\, E(n-1,m) + c_1\, E'(n,m), \tag{6}$$

where $E(n,m)$ is the smoothed energy, and c_0 and c_1 are coefficients for averaging the terms $E(n-1,m)$ and $E'(n,m)$ such that the sum of the two coefficients is unity.

The block diagram of our proposed method is given in Fig. 2. First, an audio signal input is segmented into 3 seconds audio excerpts. The process will be done only on a unique excerpt. The 3-seconds excerpt is then downsampled to obtain a sampling rate of 16000 Hz. This sampling frequency has been chosen because the human audition is more sensitive to the frequencies under 8000 Hz. A framing division is performed in order to extract the ear-based features. In order to avoid signal discontinuities, and to achieve a basic-time invariance with respect to content similarity (as perceived by the HAS) an overlap factor of 31/32 is used, as proposed in [4]. All overlapping frames are weighted by a Hanning window having the same overlap factor to avoid borders effect between frames.

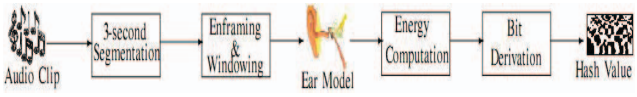

Fig. 2. Simplified proposed framework front-end.

3.2. Bit derivation process

The 32 sub-channels outputs are used in the bit derivation process. As it was experimentally verified in [4], the sign of energy differences is very robust to many kinds of processing. Therefore, our sub-hash extraction scheme is based on thresholding the energy differences between frequency cochlear bands. If we denote the energy of band m of frame n by $E(n,m)$, then m and n will respectively refer to the m-th bit of the n-th frame of the hash value $H(n,m)$. The bits of the hash string are extracted using the following formula:

$$H(n,m) = \begin{cases} 1 & \text{if } \Delta E_n - \Delta E_{n-1} > 0 \\ 0 & \text{if } \Delta E_n - \Delta E_{n-1} \leq 0, \end{cases} \quad (7)$$

where $\Delta E_n = E(n,m) - E(n,m+1)$,
and $\Delta E_{n-1} = E(n-1,m) - E(n-1,m+1)$.

4. RESULTS AND DISCUSSIONS

In order to demonstrate the robustness of our method, a set of simulation experiments were performed. As mentioned in section 2, the robustness degree is evaluated through a threshold comparison of the bit error rate. The bit error rate (BER) is computed by the following formula :

$$\mathrm{B}ER = \frac{number\ of\ bit\ errors}{number\ of\ bits\ in\ a\ hash\ value} \quad (8)$$

A threshold of $T_l = 0.25$ (see Eq. 3) is usually used [5][3]. In addition, the Hamming distance must be under this threshold for audio clips that are considered as perceptually identical. To test the robustness, hash values were extracted from three 16 bits coded stereo pieces of music and digitized at a 16000 Hz sampling rate. Each one belongs to a specific music type: a soft Arabic music track "*Lawala Albi*" by Fadel Shaker, the famous 20th century's metal song "*One*" by the rock band Metallica, and "*The Barber Of Seville*" a classical piece by Mozart. Then, different severe audio attacks are applied.In our settings each frame has 256 samples. Hence, a hash value is representing by a 256x32 bits string. A comparison between the original audio clip and the modified one is made by calculating the corresponding Hamming distance. The attacks we considered are given below.

- **MP3 compression**: 128 kbps and 32 kbps compression rates.
- **Amplitude compression**: compression ratios are: 8.94:1 for |A|≤ -28.6 dB; 1.73:1 for -46.4 dB <|A|<-28.6 dB; 1:1.61 for |A|≥ -46.4 dB.
- **Amplitude fading**: at ±3 dB.
- **Band-pass filtering**: 100 Hz and 6000 Hz cut-off frequencies using a second order Butterworth filter.
- **Dynamic delay**: 5 ms right delayed with 70% for original signal.
- **Echo addition**: a decay of 41% and a delay of 98 ms with an initial volume of 100%.
- **Equalization**: 10-band equalizer :

Freq.(Hz)	31	62	125	250	500	1k	2k	4k	8k	16k
Gain (dB)	-3	3	-3	3	-3	3	-3	3	-3	3

- **Denoising**: 15 dB noise reduction.
- **Noise addition**: pending an additive white noise with various SNR ratios in different positions through the original excerpt.
- **Pitching-stretching**: 90%-110% ratio with an overlap of 33%.
- **Silence addition**: pending 2-millisecond silence duration within the signal in many positions.

Table 1 summarizes the obtained Hamming distance after a comparison between the original excerpt and the corresponding altered version. It is clear that the computed BERs are under the threshold. Pitching and stretching process yield to BERs that are close to the fixed threshold. In fact, these two audio degradations affect the song's quality, especially the sound of the singer. This could alter the perception by an human ear, and consequently our ear-based model. For the same attacks, the BER value of Mozart's piece remains under the threshold. The only audio signal degradation that leads to

Audio attack	Fadel	Metallica	Mozart
128 kbps MP3 comp.	0.097	0.114	0.072
32 kbps MP3 comp.	0.115	0.126	0.106
Ampli. comp.	0.087	0.105	0.067
-3 dB ampli. fading	0.010	0.012	0.009
+3 dB ampli. fading	0.016	0.011	0.011
Bandpass filtering	0.037	0.049	0.019
Dynamic delay	0.112	0.084	0.067
Distortion	0.032	0.025	0.029
Echo addition	0.063	0.088	0.095
Equalization	0.085	0.091	0.075
Denoising	0.052	0.097	0.068
Noise addition	0.070	0.109	0.544
Pitching	0.209	0.179	0.150
Silence addition	0.466	0.471	0.526
Stretching	0.200	0.212	0.174

Table 1. BERs for several audio signal attacks.

a Hamming distance above the threshold was the silence addition; this is due to the discontinuity between the processed frames. Further experiments have been carried out to prove the high level of robustness of the proposed audio hashing technique in noisy environments. Fig. 3 shows that the proposed scheme is robust in such environment. We have observed that classical music piece is more sensitive (than hard rock music) to the noise attack. This is due to the fact that the Mozart's track contains several low-magnitude regions that are affected by the noise more than the other parts of this signal. To measure the power of discrimination of our audio hashing method, comparative experiments based on Hamming distance are performed. These comparisons are made between two (perceptually) different audio clips. As given by Eq. 4, the Hamming distance for such excerpts should be above the threshold T_h. This threshold is fixed at 0.35. Table 2 shows that our proposed framework verifies this condition perfectly.

	Fadel	Metallica	Mozart
Fadel	0	0.464	0.455
Metallica	0.464	0	0.474
Mozart	0.455	0.474	0

Table 2. BER between different original audio clips.

5. CONCLUSION

A new audio hashing scheme, based on the human perception system, was presented. Experiments using various audio excerpts show that the proposed system is very resilient to a large variety of severe audio attacks. The discrimination power of the proposed audio hashing is also experimentally proven. This feature is very useful since it directly affects the efficiency of the method during a query search (by excerpt) for example. We are currently continuing the effort to incorporate this framework in Peer-to-Peer applications for content-based music retrieval. Besides this, tests that aims at showing if user opinions about perceptual similarity meet hash values will be carried out.

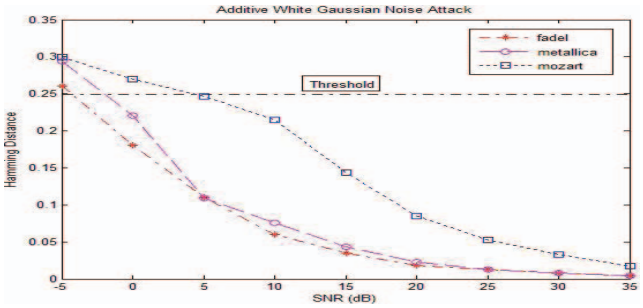

Fig. 3. Robusteness of the proposed audio hashing system in noisy environments: variations of Hamming distance with the respect of the SNR.

6. REFERENCES

[1] J.M. Bruck, S. Bres, and D. Pellerin, "Utilisation d'une signature audio pour l'indexation de documents audiovisuels", *CORESA Workshop*, 2004.

[2] P. Cano, E. Battle,T. Kalker, and J. Haitsma, "A Review of Algorithms for Audio Fingerprinting", *Proc. of the Int. Workshop on Multimedia Signal Processing*, pp. 169-173, 2002.

[3] L. Ghouti, and A. Bouridane, "A Robust Perceptual Audio Hashing Using Balanced Multiwavelets", *IEEE-ICASSP Conference*, pp. 209-212, 2006.

[4] J. Haitsma, T. Kalker, and J. Oostveen, "Robust Audio Hashing for Content Identification", *CBMI'01-International Workshop on Content Based Multimedia Indexing*, September 2001.

[5] M. Mihcak, and R. Venkatesan, "A perceptual Audio Hashing Algorithm: A Tool for Robust Audio Identification and Information Hiding", *In 4th Int. Information Hiding Workshop*, Pittsburg, PA, 2001.

[6] H. Ozer, B. Sankur, N. Memon, and E. Anarim, "Perceptual Audio Hashing Functions", *EURASIP Journ. on Applied Signal Proces.*, Vol. 62, pp. 1780-1793, Dec. 2005.

[7] S.A. Selouani, D. O'Shaughnessy, and J. Caelen, "Incorporating Phonetic Knowledge Into an Evolutionary Subspace Approach for Robust Speech Recognition", *International Journal of Computers and Applications*, Vol. 29, No. 2, pp. 143-147, 2007 .