MDCT-Based Perceptual Hashing for Compressed Audio Content Identification

Yuhua Jiao*, Bian Yang*, Mingyu Li† and Xiamu Niu*
*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
Email: {yuhua.jiao, bian.yang}@ict.hit.edu.cn, xiamu.niu@hit.edu.cn
†Dept. of Automatic Test and Control, Harbin Institute of Technology, Harbin, China
Email: mingyu.li@ict.hit.edu.cn

***Abstract*— In this paper, a perceptual audio hashing method in compressed domain is proposed for content identification, in which MDCT coefficients as the intermediate decoding result are selected for perceptual feature extraction and hash generation. The perceptual feature extraction is based on psychoacoustic model and exhibits good discrimination ability for different audio contents but robustness against common audio signal processing operations. Via feature extraction in the compressed domain, the MDCT-based compressed audios, such as MP3, AAC, etc., could be efficiently identified without complete decoding which facilitates those practical applications with strict requirements of memory and computational complexity, such as online audio retrieval, indexing of massive compressed audio data, audio identification by mobile phone, etc. The algorithm is highly robust against MDCT compression which is widely used in audio coding. Experiments demonstrate the effectiveness of the proposed scheme.**

I. Introduction

Perceptual audio hashing algorithm generates from the original audio data a robust and compact digital hash to discriminates different contents from those signal processed copies of the original audio. The generated hash exhibits high perceptual similarity (or even equality) to the hash generated after the original audio undergoes some content-preserving manipulations, such as lossy compression, resampling, filtering etc. The discrimination ability and the robustness are the most important requirements for perceptual hashing algorithm, which has been used for multimedia content identification and content authentication [1], [2].

Perceptual audio hashing exploits audio perceptual features and features extracted from raw PCM audio samples which have been well studied [3], [4]. However, in practical applications, there are still some problems to extract perceptual features from PCM audios. Digital audio in practical applications is usually encoded in compressed formats such as MPEG Layer 3(MP3), MPEG Advanced Audio Coding(AAC), and therefore needs feature extraction directly from compressed domain considering computational efficiency for practical applications. Furthermore, features extracted directly from compressed coefficients are independent of the information which is removed in lossy compression. That is, the features survive lossy compression. Thus, the compressed domain algorithm has high robustness against lossy compression.

In this paper, we propose a MDCT-based perceptual audio hashing algorithm, because Modified Discrete Cosine Transform (MDCT) is widely used in audio coding [5]. The perceptual hash value is calculated from the MDCT coefficients decoded from the compressed audio taking into account of the psychoacoustic model. The proposed perceptual hashing algorithm is inherently robust to MDCT-based audio coding and easy to implement. And it is also robust to some content-preserving operations. The background of the MDCT and MPEG audio coding scheme is described briefly in Section II.The details of the proposed method are presented in section III. Experimental results are discussed in section IV. Section V gives conclusions and discussions.

II. MDCT and MPEG Audio

A. A Short Introduction to MDCT Properties

There is a relationship between MDCT and DFT established via Shifted Discrete Fourier Transforms (SDFTs) [6]. Because of this, MDCT coefficients could be used to calculate compressed domain features, and to achieve the perceptual hash value of compressed audio data.

However, the non-orthogonal property of MDCT, which causes a problem called MDCT-DFT mismatch phenomenon [6], affects the robustness of the perceptual hashing algorithm. So, some modification is performed on the coefficients to optimize the proposed algorithm in this paper.

B. A Short Introduction to MPEG Audio

According to MPEG standards, audio data are encoded frame-by-frame. A MP3 frame consists of 2 granules, where each granule contains 576 samples per channel. And MPEG AAC files always have 1024 samples per frame [7], [8]. MDCT coefficients are also processed by frame in this paper.

MDCT coefficients is achieved after MDCT in encoding. It does not bring to any extra computational overhead for time-frequency transformation of the perceptual hashing algorithm. While in decoding, MDCT coefficients is obtained prior to IMDCT and synthesis filtering which take up almost 90% of decoding time. Therefore, the perceptual hash value of a compressed audio file could be calculated without complete decoding. It saves lots of computational and memory resources. It is desirable for the real-time applications such as online audio retrieval, indexing of massive compressed audio data, audio identification by mobile phone, etc.

1-4244-1274-9/07/$25.00 ©2007 IEEE

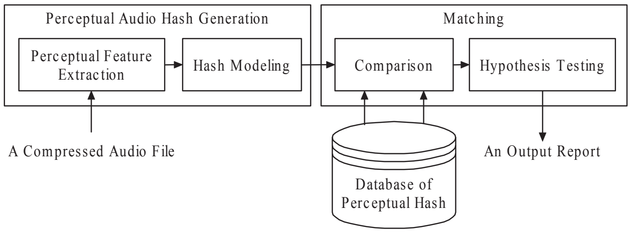
Perceptual Audio Hash Generation
Perceptual Feature Extraction
Hash Modeling
Matching
Comparison
Hypothesis Testing
A Compressed Audio File
An Output Report
Database of Perceptual Hash

Fig. 1. Content-based Audio Identification System

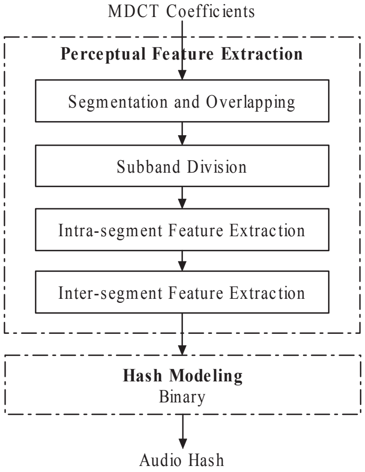
MDCT Coefficients
Perceptual Feature Extraction
Segmentation and Overlapping
Subband Division
Intra-segment Feature Extraction
Inter-segment Feature Extraction
Hash Modeling
Binary
Audio Hash

Fig. 2. Algorithm Flowchart

III. Proposed Method

Fig. 1 is the block diagram of a content-based audio identification system . There are two processes: the perceptual audio hash generation and the matching process. The flowchart of the proposed algorithm is shown in Fig. 2.

A. Perceptual Feature Extraction

The chosen features are the perceptual essential parts of audio content perceived by Human Auditory System. The basilar membrane in the hearing mechanism performs subband filtering for incoming sound [9] and human auditory system is sensitive to intensity [10]. Therefore, Subband energy in MDCT domain is used to calculate the perceptual audio features.

1) Segmentation And Overlapping: To eliminate the effect of the MDCT-DFT mismatch phenomenon, which has been introduced in section II, MDCT coefficients of the whole audio file are segmented before feature extraction. The segmentation has an average effect on the fluctuation of the MDCT domain energy. Coefficients in one MP3 granule or one AAC frame must be in the same segment. And overlapping is applied to improve the robustness against shifting.

2) Subband Division: Scale factor bands might have been selected as the subbands. However, the number of scale factors bands depends on sample rate and block size [7], [8]. Furthermore, in order to reduce the bit rate of hash value, some experiments are performed to optimize the subband division.

3) Intra-segment Feature: For each segment, the ratio of the subband energy to the energy of segment is defined as intra-segment feature in this paper. It represents the static feature of the audio signal in each segment. The energy of the n_{th} subband of the m_{th} frame of the s_{th} segment is denoted by $E(s,m,n)$. And the intra-segment feature is denoted by $F_A(s,n)$ and given by (1).

$$F_A(s,n) = \frac{\sum_{m=1}^{M_{Gr}} E(s,m,n)}{\sum_{m=1}^{M_{Gr}} \sum_{n=1}^{N_{SB}} E(s,m,n)} \tag{1}$$

Where, M_{Gr} is the segment size, which denotes the number of frames or granules in one segment. N_{SB} is the number of subbands.

4) Inter-segment Feature: Difference between the intra-segment features of contiguous segments is used as the inter-segment feature and denoted by $F_E(s,n)$. It represents the dynamic characteristic of audio signal and is given by (2).

$$F_E(s,n) = F(s,n) - F(s-1,n) \tag{2}$$

B. Hash Modeling

The feature value is quantized to binary code in this paper as (3).

$$B(s,n) = \begin{cases} 1, & if F_E\left(s,n\right) \geq T \\ 0, & if F_E\left(s,n\right) < T \end{cases} \tag{3}$$

Where, $B(s,n)$ denotes the binary code of the perceptual hash value corresponding to the s_{th} segment n_{th} subband. T is the binary threshold. In this paper, it is set to be 0. Some energy-adaptive methods could be used to optimize it.

The hash value of one segment could not be used for identification and a hash value of a identifiable audio is consist of several hash values of contiguous overlapped segments.

C. Matching

After hash value generation, a database of hash values is created. Those in database are called templates, while the hash value extracted from input compressed audio in matching process is common named sample.

The similarity of two audios is measured by Hamming distance (i.e. the number of different bits of two hash values). Bit error rate(BER), which is the ratio of the number of different bits to the total number of bits, is used in our experiments. The sample is recognized as a copy of the template, when BER is below the threshold. Threshold varies from one system to another according to requirements. In [4], the threshold is set by modeling the relationship of threshold and false positive rate (FPR).

In this paper, we take use of two methods to determine it. The threshold is set according to equal error rate (EER) or false positive rate (FPR) in different circumstances. EER is the value of the crosspoint of false acceptance rate (FAR)

TABLE I

SUBBAND DIVISION

Subband Index	MDCT Coefficients (Long Window)	MDCT Coefficients (Short Window)
1	1-3	1
2	4-9	2-3
3	10-27	4-9
4	28-81	10-27
5	82-243	28-81

curve and false reject rate (FRR) curve. FAR is another name of FPR, while FRR is also named false negative rate.

IV. Experimental Results and Discussion

Experiments were performed to evaluate the discrimination ability and the robustness of the proposed perceptual hashing algorithm.

A. Experimental Implementation

The audio database in our experiment was composed of 1,150 MP3 audio excerpts with a sample rate of 44.1KHz and 128Kbps of about 3 seconds extracted from songs of different genres. These excerpts were subjected to some content-preserving manipulations such as transcoding, resampling, and bandpass filtering to generate modified versions of them. Thus, there were 15,000 hash values comparing with each other randomly. FAR-FRR curves were drawn based on the comparison results.

The segment is composed of 50 granules experientially and the shift is a granule because audio signal of a granule could be processed as steady-state signal. The first 243 coefficients for long window or 81 coefficients for short window are used and divided into 5 subbands as shown in Table I.

A identifiable audio excerpt is consist of 250 overlapped segments and 3.26 second long. Thus, the size of a hash template is $250 \times 5 = 1250$ bits and the bit rate of hash value is 0.384Kbps. It is one seventh of the bit rate of that in [4] which is 2.758Kbps.

B. Transcoding

The audio excerpts are transcoded. The bit rate of them is different in the range of 32 Kbps to 320 Kbps. They are 32Kbps, 64Kbps, 192Kbps, 256Kbps and 320Kbps. Experimental results are shown in Fig. 3.

There is no crosspoint of the FAR-FRR curves in Fig. 3, so the threshold is set by FPR with probability prediction. Fig. 4 illustrates the comparison of the distribution of BERs and the normal distribution. It shows that BERs has a normal distribution approximately. The mean value is 0.5000, and the standard deviation is 0.0293. And then, the false positive rate could be given in (4).

$$FPR = f(\alpha|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\alpha-\mu)^2}{2\sigma^2}} \tag{4}$$

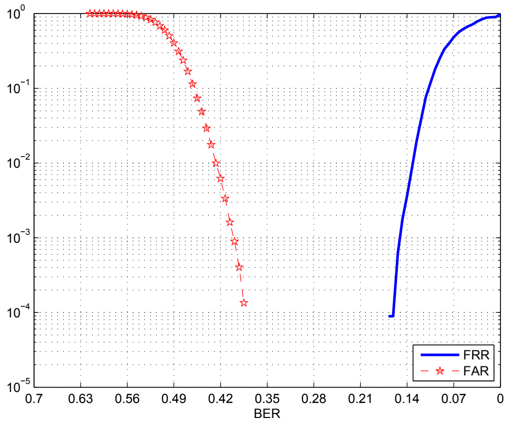

Fig. 3. FAR-FRR Curves

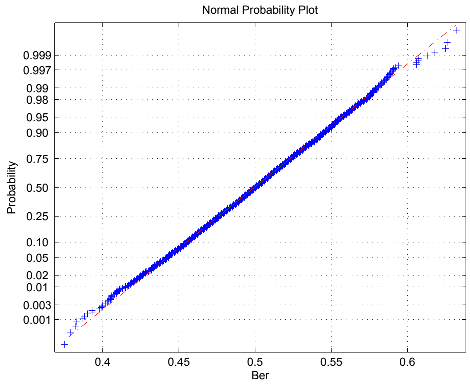

Fig. 4. Normal probability plot of BERs

According to Fig. 3, the threshold could be set in a range of 0.20 to 0.30. Table II shows the false positive rate varying with the different thresholds. FPR is similar to the results of the method proposed in [4],when the threshold is 0.20.

C. Robustness to Some Operations

The operations on audio data are as follows:

Downsampling to 22.05KHz and then upsampling back.

Change volume: -6.0206 dB.

Change volume: 3.5218 dB.

Band-pass filtering using a 10th order Butterworth filter with the band width from 60Hz to 8KHz.

TABLE II

FPR AND THRESHOLD

α	FPR
0.20	2.3048×10^{-22}
0.25	2.1146×10^{-15}
0.30	1.0385×10^{-10}

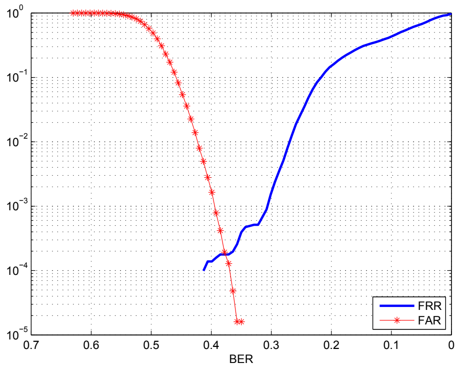

Fig. 5. FAR-FRR Curves

Equalization with a 7 bands equalizer, and the parameters is listed in Table III.

Reverberation to simulate a concert hall.

5 one second echoes addition.

TABLE III
THE PARAMETERS USED IN EQUALIZATION

Freq.(Hz)	**60**	**150**	**400**	**1000**	**2400**	**6000**	**15K**
Gain(dB)	12.0	7.0	-2.0	-2.0	-6.0	-2.0	8.0

The results are shown in Fig. 5.

The EER is 1.2758×10^{-4}, and the threshold should set to be about 0.37. According to those experimental results, the discrimination ability and robustness of the proposed algorithm satisfies the requirements of audio content identification.

Comparison of the proposed algorithm and the algorithm developed by Haitsma et al. [4] is done in this paper. 100 audio excerpts from songs of different musical genres are used and the average results are shown in Table IV. Robustness to lossy compression of the proposed algorithm is much better than that of the method in [4]. While, the proposed algorithm is more fragile to downsampling and equalization than the other. The MDCT-DFT mismatch phenomenon discussed in Section II results in the high Bers.

TABLE IV
COMPARISON OF DIFFERENT ALGORITHMS

Modification	Piano		Violin		Pop Song	
	I*	II†	I	II	I	II
MP3@192Kbps	0.026	0.102	0.033	0.104	0.040	0.116
MP3@64Kbps	0.073	0.135	0.074	0.143	0.094	0.154
Reverberation	0.149	0.138	0.153	0.139	0.155	0.139
Echo	0.233	0.231	0.232	0.266	0.216	0.200
Downsampling	0.050	0.002	0.055	0.004	0.061	0.004
Equalization	0.194	0.029	0.181	0.086	0.181	0.001

*The algorithm proposed in this paper †The algorithm developed by Haitsma

V. CONCLUSIONS AND FUTURE WORK

In this paper, we present a compression domain perceptual audio hashing algorithm. It has practicable applications on compressed audio identification, because:

1) The proposed algorithm utilized the MDCT coefficients for robust feature extraction without complete decoding of compressed audio files. Low bit rate of hash value and computational simplicity make it suitable to some real-time applications and usage of mobile devices.
2) The proposed algorithm has better robustness against audio transcoding than existing algorithm as shown in experimental results. It is also robust to some modifications which preserve the perceptual quality in some degree.

Because of the MDCT-DFT mismatch phenomenon, the robustness of the algorithm is not as good as that of some hash functions on PCM sources. A considerate adjusting to MDCT coefficients will be study to improve it.

The proposed method could also be used in content authentication systems.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Project Number: 60671064), the Science Foundation of Guangdong Province (Project Number: 05109511), the Foundation for the Author of National Excellent Doctoral Dissertation of China (Project Number: FANEDD-200238), the Foundation for the Excellent Youth of Heilongjiang Province, the Program for New Century Excellent Talents in University (Project Number:NCET-04-0330), and the Chinese national 863-Program (Project Number:2005AA733120).

REFERENCES

[1] F. Mapelli and R. Lancini, "Audio hashing technique for automatic song identification," *Proc. ITRE, International Conference on Information Technology*, pp. 84–88, 2003.

[2] B. B. Zhu, M. D. Swanson, and A. H. Tewfik, "When seeing isn't believing," *IEEE Signal Processing Magazine*, vol. 21, no. 2, pp. 40–49, 2004.

[3] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," *Journal of VLSI Signal Processing*, pp. 271–284, 2005.

[4] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," *Proc. ISMIR, International Conference on Music Information Retrieval*, pp. 107–115, Oct. 2002.

[5] H. C. Chiang and J. C. Liu, "Regressive implementations for the forward and inverse mdct in mpeg audio coding," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 116–118, April 1996.

[6] Y. Wang, L. Yaroslavsky, and M. Vilermo, "On the relationship between mdct, sdft and dft," *Proc. of the 5th International Conference on Signal Processing Beijing*, vol. 1, pp. 44–47, 2000.

[7] MPEG, *Coding of moving pictures and associated auido and digital storage media at up to about 1.5Mbit/s, Part3: Audio*. ISO/IEC11172-3, 1993.

[8] ——, *Information technology – Generic coding of moving pictures and associated audio, Part 7: Advanced Audio Coding*. ISO/IEC13818-7, 1997.

[9] R. A. GARCIA, "Digital watermarking of audio signals using a psychoacoustic auditory model and spread spectrum theory," in *Proceeding of the 107th Audio Engineering Society Convention*, 1999.

[10] J. H. Prout and G. R. Bienvenue, *Acoustics for You*. Kerieger Publishing Company, 1991.