Robust Audio Identification for MP3 Popular Music

Wei Li, Yaduo Liu, Xiangyang Xue
School of Computer Science and Technology, Fudan University
825 Zhangheng Road, Shanghai 201203, P.R.China
weili-fudan@fudan.edu.cn duoyal@gmail.com xyxue@fudan.edu.cn

ABSTRACT

Audio identification via fingerprint has been an active research field with wide applications for years. Many technical papers were published and commercial software systems were also employed. However, most of these previously reported methods work on the raw audio format in spite of the fact that nowadays compressed format audio, especially MP3 music, has grown into the dominant way to store on personal computers and transmit on the Internet. It would be interesting if a compressed unknown audio fragment is able to be directly recognized from the database without the fussy and time-consuming decompression-identification-recompression procedure. So far, very few algorithms run directly in the compressed domain for music information retrieval, and most of them take advantage of MDCT coefficients or derived energy type of features. As a first attempt, we propose in this paper utilizing compressed-domain spectral entropy as the audio feature to implement a novel audio fingerprinting algorithm. The compressed songs stored in a music database and the possibly distorted compressed query excerpts are first partially decompressed to obtain the MDCT coefficients as the intermediate result. Then by grouping granules into longer blocks, remapping the MDCT coefficients into 192 new frequency lines to unify the frequency distribution of long and short windows, and defining 9 new subbands which cover the main frequency bandwidth of popular songs in accordance with the scale-factor bands of short windows, we calculate the spectral entropy of all consecutive blocks and come to the final fingerprint sequence by means of magnitude relationship modeling. Experiments show that such fingerprints exhibit strong robustness against various audio signal distortions like recompression, noise interference, echo addition, equalization, band-pass filtering, pitch shifting, and slight time-scale modification etc. For 5s-long query examples which might be severely degraded, an average top-five retrieval precision rate of more than 90% can be obtained in our test data set composed of 1822 popular songs.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications – *signal processing*

General Terms

Algorithms, Experimentation

Keywords

Audio identification, compressed domain, MDCT spectral entropy robustness, fragment retrieval

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. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
Copyright 2010 ACM 978-1-60558-896-4/10/07...$10.00.

1. INTRODUCTION

Music identification is the most typical application of audio fingerprinting technique. By comparing the fingerprint of an unknown music query fragment, usually transmitted from mobile phones on the wireless telecom network or from personal computers on the Internet, with those previously calculated and stored in a fingerprint database, matching results and related metadata are returned. The fingerprint must characterize the nature of music content to differentiate from each other, and possess strong robustness against various kinds of audio signal degradations. Also, the query music fragment is usually required to be only a few seconds long, namely fit the demand of fragment retrieval. To date, a number of algorithms such as [1, 2, 3, 4, 5, 6] have been published with rather high retrieval precision. Most of them run on the raw wave format, and commercially deployed software systems also appeared [7].

However, with the mature of CD quality audio compression techniques at low bit rate and the fast growing of the Internet, compressed audio signals are increasingly ubiquitous and have become the dominant fashion of storing in personal electronic equipments and transmitting on the Internet. It would be interesting and meaningful in practice if audio features are directly extracted from the compressed domain and used for music identification in the database.

So far, only a few algorithms are designed to perform music information retrieval (MIR) directly in the compressed domain, research in this field is still in its infancy. Liu *et al*. [8] calculate the compressed-domain energy distribution from the output of the poly-phase filters as features to index songs. For each song in the data set, they use its refrain as the query example to retrieve all similar repeating phases, obtaining an average 78% recall and 32% precision rate. They claim that, to their knowledge, this is the first compressed-domain MIR algorithm. Lie *et al*. [9] directly use selected modified discrete cosine transform (MDCT) spectral coefficients and derived sub-band energy as well as its variation to represent the tonic characteristic of a short-term sound and to match between two audio segments. The retrieving probability achieves up to 76% among the top 5 matched. Tsai *et al*. [10] describe a query-by-example algorithm using 176 MP3 songs of a same singer as the database. They calculate spectrum energy from sub-band coefficients (SBC) to simulate the melody contour and use it to measure similarity between the query example and those database items. By summing up the sub-band coefficients in every 12 frames (about one tone duration) to obtain tone energy lines, the melody contour is represented by a string sequence with two letters (U, D), where 'D' means the current tone energy is smaller than its preceding one and 'U' means greater. With 40 frames assembled as a query example, the accuracy achieves 74% within top 4 and 90% within top 5. In Tsai's another paper [11], they use scale factors (SCF) and sub-band coefficients in a MP3 bit stream frame as features to characterize and index the object. All SCF and SBC values are divided into 26 bins using a tree-structured quantizer, values in the same bin are accumulated to form a histogram as the final indexing patterns. Due to its statistical nature, this approach can tolerate certain length variance between the query example and database items. When length variance is between [0%, 10%), [10%, 15%), [15%,), the query item can be obtained in top 5, 10, 15 results, respectively. Pye *et al*. [12] design a new compressed-domain audio feature referred to as MP3 cepstrum (MP3CEP) based on partial decompression of MPEG layer-3 audio signals to facilitate the management of a typical digital music library. It is approximately six times faster than conventional MFCC coefficient for music retrieval while the average precision is only 59%. Jiao *et al*. [13] design a robust compressed-domain audio fingerprinting algorithm, taking the ratio between the sub-band energy and the full-band energy of a segment as intra-segment feature and difference between continuous intra-segment features as inter-segment feature. Experiments show that such fingerprints are robust against transcoding, down sampling, echo addition and equalization. However, the authors didn't show any results on the retrieval precision rate.

The above methods acquired certain retrieval achievements, whereas they have two critical limitations. The first is that none of them works in the way of fragment retrieval, which is a basic requirement of audio fingerprinting systems; the second is that robustness is not considered in the above methods except reference [13] which shows some basic robustness results without taking account of the two most challenging distortions i.e. tempo-reserved pitching shifting and pitch-reserved time-scale modification (TSM). Moreover, previously used features principally follow the line of MDCT coefficient and its derived spectral energy, then can we develop a new type of compressed-domain feature to achieve high robustness in music identification? It's well known that entropy plays an important role in information theory as a measure of information, choice and uncertainty, it has been widely used in many fields like automatic speech recognition (ASR) [14], uncompressed-domain robust audio identification [15] and speech/music classification [16]. So far, various compressed-domain audio features including scale factors [17, 18], MP3 window-switching pattern (WSP) [19, 20], basic MDCT coefficients and derived spectral energy, energy variation, duration of energy peaks, amplitude envelope, spectrum centroid, spectrum spread, spectrum flux, roll-off, RMS, rhythmic content like beat histogram etc [21, 22, 23, 24, 25, 26] have been used in different applications such as retrieval, segmentation, genre classification, speech/music discrimination, summarization, singer identification, watermarking and beat tracing/tempo induction. However, in spite of the extensive use in various research fields for years, to the authors' knowledge, we are not aware of the usage of entropy for music identification in the compressed domain. This motivated our initial idea of developing MDCT spectral entropy as a novel audio feature for robust music identification.

In this paper, we first group 22 MP3 bit stream granules, the basic processing unit in decoding, into a longer window which is called 'block' for the statistical purpose. Then we remap the 576 MDCT coefficients extracted from a granule of partially decoded MP3 songs into 192 new frequency lines to unify the frequency distribution of long and short windows. After dividing these new MDCT values into 9 new subbands in terms of the scale-factor bands of short windows, we calculate the spectral entropy of every block and acquire the final fingerprint sequence via magnitude relationship modeling. Experimental results show that this MDCT spectral entropy based fingerprint achieves high robustness against common audio signal distortions like recompression, noise contamination, echo addition, equalization, band-pass filtering, pitch shifting and slight time-scale modification etc. A query example of 5s-long music, even if severely degraded by various time-frequency distortions, is sufficient to retrieve its original recording with an average top-five precision rate of more than 90% in our test data set composed of 1822 popular songs.

The remaining paper is organized as follows. Section 2 introduces the basic principles of MPEG Layer-3, bit stream data format, and the concept of compressed-domain spectral entropy. Section 3 details the steps of deriving MDCT entropy based audio fingerprint and the searching strategy. Experimental results on retrieval precision under audio signal distortions are given in section 4. Finally, section 5 concludes this paper and points out possible ways for future research.

2. COMPRESSED-DOMAIN SPECTRAL ENTROPY

2.1 Principles of MP3 Compression and Decoding

A simplified illustration of MPEG Layer-3 encoder is shown in Figure 1. The input PCM signal is mapped into 32 equal-bandwidth subbands through a polyphase filterbank, which simulate the critical bands in the human auditory system (HAS). The sub-band outputs are further subdivided by MDCT transform using long or short window to provide better spectral resolution. The long window allows greater frequency resolution for audio signals with stationary characteristics, while the short window provides better time resolution for transients. Combined with other adjuvant techniques including psychoacoustic model, scale-factor, *Huffman* coding, quantization etc, the final compressed bit stream is generated [27]. Figure 2 displays the frame format of MP3 bit stream, and each frame has two granules to exploit further redundancies.

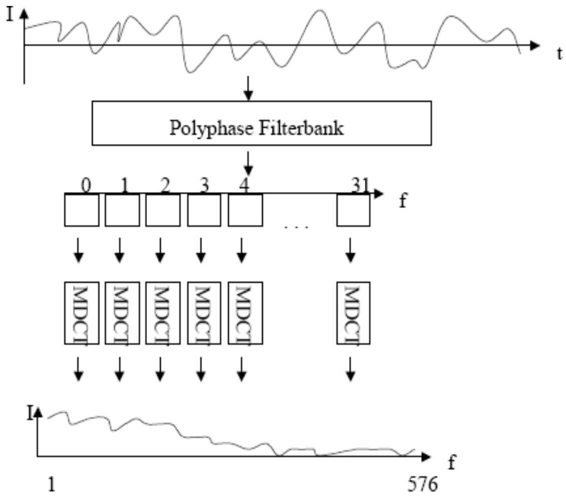

Figure 1. Simplified frequency transformations of MPEG Layer-3 encoder

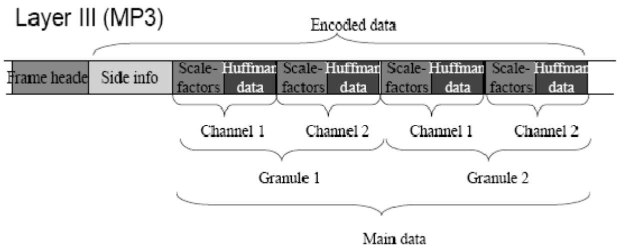

Figure 2. Frame format of MPEG Layer-3 bit stream

In MP3 decoder, the basic unit of input bit stream is a granule of 576 samples, approximately 13 ms at the sampling rate of 44.1 kHz. One granule of compressed data is first unpacked and dequantized into 576 MDCT coefficients, then mapped to the polyphase filter coefficients in 32 subbands by IMDCT. Finally, these sub-band polyphase filter coefficients are inverse transformed and synthesized into PCM audio. Therefore, access of transformation coefficients in Layer 3 can be either at the MDCT or the filterbank level, the latter is obviously more time consuming.

2.2 Compressed-domain Spectral Entropy

Entropy is a fundamental concept in information theory to measure the uncertainty associated with a random variable [28]. If a random variable X is discrete with possible values $\{x_i: i=1, \dots, n\}$, the *Shannon* information entropy of X can be explicitly defined as,

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i) \qquad (1)$$

where $p(x)$ denotes the probability mass function (PMF) of X, b is the base of the logarithm used and typically adopted as 2. With respect to a random signal, its entropy is a measure of how unpredictable it is. The entropy will be minimum when the signal is constant since it is most predictable and on the contrary maximum if the signal has a uniform distribution so that sample values are most unpredictable. For example, voiced sounds with clear formants have lower entropy; oppositely, flatter spectrum corresponding to non-speech or noisy regions has higher entropy [14]. Entropy is therefore expected to be different from those usual features derived from spectral energies and to capture the nature of audio content with a distinctive capability.

Follow this line of thought, we propose to calculate the MDCT spectral entropy as follows. Without loss of generality, we only take a set of MDCT coefficients $X=\{x_1, \dots, x_n\}$ as an illustration, detailed formulation compliant with MP3 bit stream format will be described in the next section. In light of formula (1), the biggest problem to compute the entropy of a spectrum distribution lies in that spectrum itself does not possess the necessary property of a PMF, namely all the spectrum elements don't sum up to 1. In order to convert the original MDCT spectral coefficients into a PMF-like function, we divide every individual MDCT coefficient by the sum of all components, i.e. by sum normalization, to approximate its probability as shown in formula (2). Then formula (1) can be used to compute the MDCT spectral entropy.

$$p(x_i) = \frac{x_i}{\sum_{i=1}^{n} x_i} \qquad \{x_i : i = 1, \dots, n\} \qquad (2)$$

As stated in the introduction, the goal we calculate MDCT spectral entropy is to use it as an audio feature for compressed-domain music identification, then is it steady enough under various audio signal distortions? We did some experiments to check it. Figure 3 shows the MDCT entropy sequence calculated as above from a 6s clip of compressed song after granule grouping and sub-band division as detailed in the next section. It can be clearly seen that the entropy curve is rather stable under volume modulation, echo addition, band-pass filtering, noise interference and MP3 recompression at 32kbps. For pitch shifting up to ± 10% and 10-band equalization, though relatively bigger entropy variation is introduced compared with the above, the basic profiles are maintained. When the example excerpt is slightly time-scale modified, ±2% in experiment, the entropy curve only translates a small distance along the granule/time axis with little change to the spectral entropy contour. These observed phenomena confirm our initial motivation, and herein MDCT spectral entropy displays great potential to become a powerful audio fingerprint.

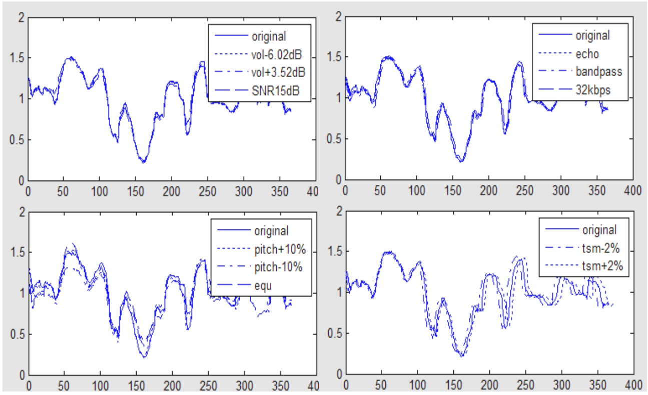

Figure 3. MDCT spectral entropy curve under various audio signal degradations. From the first to the second row, the curve of a clip of 6s original audio and distorted versions under volume modulation, noise addition, band-pass filtering, lossy recompression, echo addition, pitch shifting, equalization and time-scale modification are drawn for comparison

3. PROPOSED METHOD

The whole algorithm includes five steps stated as follows.

3.1 Granule Grouping

Compared with conventional content-based music information retrieval, fragment input and robustness are known to be two crucial constraints on audio fingerprinting schemes to retrieve in a music database. If modeling with audio signal operations, the input fragmented query example can be obtained from the original music via random cropping plus other types of audio processing. Random cropping causes serious desynchronization problem between the input fingerprint sequence and its stored original version, posing a serious threat on the retrieval accuracy. In general, there are two effective mechanisms to resist time-domain misalignment: one is invariant feature, the other is implicit synchronization which might be more powerful than the former [29]. However, in the MPEG compressed domain, due to the inherent data nature of compressed bit stream and the fixed frame structure, it is almost impossible to extract meaningful salient points serving as anchors as in the uncompressed domain [30]. Therefore, designing statistically stable audio features and adopting heavy overlap between adjacent time-domain processing units become the only two ways to counter desynchronization and fulfill the task of fragment retrieval in audio fingerprinting.

To achieve this end, we first need to calculate the audio feature in longer time duration to alleviate its fluctuation over time. In the proposed method, we group 22 granules, the basic processing unit in the MP3 compressed bit stream, into a so-called ‘block’ as the basic time unit to compute audio features. Next, a hop size of 1 granule between two consecutive blocks is adopted, i.e. approximately 95% overlapped. Heavy overlapping is expected to alleviate the problem of time-domain misalignment. In the above case, the displacement between the query fragment and those original items will be no more than half the hop size, i.e. 1/2 granule will be the worst case. Figure 4 shows an illustration of the mechanism: for an original compressed bit stream, *H*(*i*) represents the start section of the *i*th block, *H*(*i+1*) represents that of the (*i*+*1*)th block (the whole block length is unable to be depicted due to space limitation). The hop size is 1 granule (95% overlap), e.g. *A*+*B* (*A*=*B* here). By reason of the time-domain desynchronization caused by random cropping etc, the input query example is scarcely possible to be exactly aligned to *H*(*i*) or *H*(*i*+*1*). When the query fragment lies on the left of the dashed line, for example *QH*(*j*), it resembles *H*(*i*) more; while when it lies on the right, for example *QH*(*j'*), it looks more like *H*(*i*+*1*). Only when the query fragment happens to lies in the middle, i.e. half of the hop size, it comes to the worst synchronization situation. We hope that a suitably designed audio fingerprint with statistical characteristics will only be slightly or even not changed under this scope of misalignment.

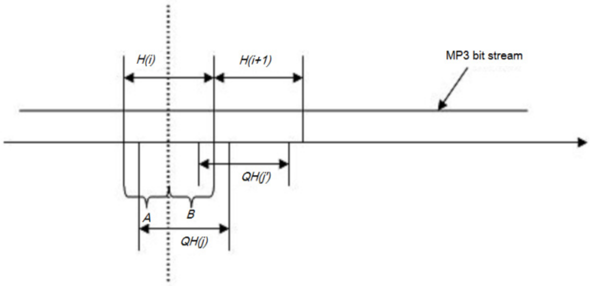

Figure 4. Desynchronization alleviation by overlapping between contiguous blocks

3.2 Frequency Alignment between Long and Short Windows

MP3 encoded bit stream is divided into many frames, which are the basic processing unit in decoding. Each frame is further subdivided into two independent granules, each with 576 values. If a granule is encoded using a long window, these values represent 576 frequency lines which are equally assigned into 32 subbands. That is, each subband includes 18 frequency lines. If a granule is compressed via a short window, these values only stand for 192 frequency lines and each line includes 3 values that belong to three consecutive windows respectively, see Figure 5.

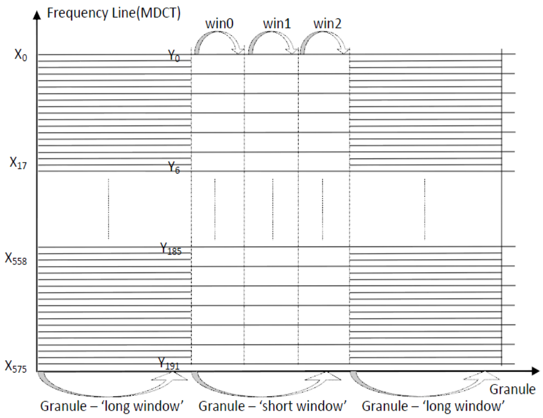

Figure 5. Distribution of MDCT coefficients in long- and short-window type of granules

Granule grouping makes any block a mixture of long-window type of and short-window type of granules, the former are usually in the majority and the latter in the minority. In order to calculate the MDCT spectral entropy, the original frequency distribution of long and short windows must be rearranged to achieve approximately the same frequency resolution. For long-window cases, we group every three consecutive MDCT coefficients of one granule into a new value, which is equal to the mean of the absolute value of the original three MDCT coefficients considering that MDCT coefficients may be positive or negative, see formula (3). For short-window cases, we substitute the original three MDCT values belonging to different windows at the same frequency line with the mean of their absolute value, see formula (4). In this way, all MDCT values in a granule are uniformly divided into 192 new frequency lines for both long- and short-window cases, this forms the basis for further processing.

$$sn(i,j)=\begin{cases} sn^{l}(i,j)=\dfrac{1}{3}\sum_{n=3j}^{3j+2}|s(i,n)| \qquad j=0,1,2,...191 \qquad (3)\\ sn^{s}(i,j)=\dfrac{1}{3}\sum_{m=0}^{2}|s^{m}(i,j)| \qquad j=0,1,2,...191 \qquad (4)\end{cases}$$

where *s*(*i*, *n*) is the original MDCT coefficient in the *i*th granule, *n*th frequency line for the long-window cases; *s*^{m}(*i*, *j*) is the original MDCT coefficient in the *i*th granule, *j*th frequency line, *m*th window for the short-window cases; *sn*^{l}(*i*, *j*) and *sn*^{s}(*i*, *j*) are the new MDCT value in the *i*th granule, *j*th frequency line for the long- and short-window cases, respectively.

3.3 New Subband Division

After time-domain granule grouping and the definition of new frequency lines, MDCT coefficients between the process of reordering and alias reduction of decoding are extracted for further processing. In MPEG Layer-3 compressed bit stream, most of the 576 MDCT coefficients in a granule are fed into a set of scale-factor bands, each band covers several coefficients and approximates certain critical band. The number of scale-factor bands depends on the sampling rate and the window type used in encoding. Granules using short window usually correspond to music edges such as transients or percussive instruments that are most crucial to auditory perception, therefore we divide the above newly defined 192 frequency lines into 9 subbands which cover the main frequency spectrum of popular music in terms of the scale-factor bands defined for short windows, as shown in Table 1.

Table 1. Division of new subbands

New subband	Index of new MDCT coefficients	Long window	Short window
		Index of MDCT coefficients	Index of frequency lines
0	0-3	0-11	0-3
1	4-7	12-23	4-7
2	8-11	24-35	8-11
3	12-15	36-47	12-15
4	16-21	48-65	16-21
5	22-29	66-89	22-29
6	30-39	90-119	30-39
7	40-51	120-155	40-51
8	52-65	156-197	52-65

3.4 MDCT Spectral Entropy Calculation and Fingerprint Modeling

With the above preparation, we first calculate the sub-band energy of the i^{th} block j^{th} new subband, each block includes 22 granules in this research. Then the PMF like ratio between the j^{th} subband energy and the sum of all subbands in the same block is approximately computed as shown in formula (5) and (6),

$$SBE(i,j)=\sum_{m=2i-1}^{2i+20}\sum_{n=MDCTB_j}^{MDCTT_j}\left\|sn^2(m,n)\right\|$$

$$i=1,2,...,N_{block}\,,\ j=0,1,...,8 \qquad (5)$$

$$P(i,j)=\frac{SBE(i,j)}{\sum_{j=0}^{8}SBE(i,j)}$$

$$i=1,2,...,N_{block}\,,\ j=0,1,...,8 \qquad (6)$$

where *sn(m, n)* denotes the n^{th} new MDCT coefficient in the m^{th} granule, $MDCTT_j$ and $MDCTB_j$ represent the top and bottom index of MDCT coefficients which belong to the j_{th} subband respectively, and N_{block} is used to indicate the maximum granule number of the input query example or original recordings stored in the database.

Finally, the entropy of the i_{th} block is calculated in terms of formula (7) and the fingerprint sequence is obtained by comparing the magnitude of entropy between two adjacent blocks as shown in formula (8). This method is equivalent to a two-level quantization, maintaining the relative relationship regardless of the detailed magnitude. In this way, the final fingerprint sequence of ones and zeros will be not only more robust but also more convenient for fingerprint matching.

$$H(i)=-\sum_{j=0}^{8}P(i,j)\log_2 P(i,j) \qquad i=1,2,...,N_{block} \qquad (7)$$

$$S(i)=\begin{cases}0 & H(i)<H(i+1)\\ 1 & H(i)\geq H(i+1)\end{cases} \qquad i=1,2,...,N_{block}-1 \qquad (8)$$

In section 2, we have demonstrated the invariance of this proposed compressed-domain spectral entropy under various time-frequency audio distortions through some experiments. Herein we will derive from a simplified theoretical way that *H*(i), the MDCT spectral entropy of *P*(i, j) in a block, has approximately the same stability with the variance of *P*(i, j). Based on the intuitive experience of human auditory perception, there are always more and stronger notes distributed in the mid-frequency range than in the high and low frequency regions. Therefore, energy of MDCT coefficients $sn(i,j)$ and derived $P(i,j)$ in each block should reflect this phenomenon and can be approximately modeled with Gaussian distribution. Let the mean and variance of $P(i\ ,\ j)$ be $m(i)=\sum_{j=0}^{8}P(i,j)/9$ and $\sigma(i)=\sum_{j=0}^{8}(P(i,j)-m(i))^2/9$, after rearranging j into [-4, +4] and extending to ±∞ (P(i, j)=0, $j\notin[0,\ 8]$) , we have

$$H(i)\approx-\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma(i)^2}}e^{-\frac{(u-m(i))^2}{2\sigma(i)^2}}\ln[\frac{1}{\sqrt{2\pi\sigma(i)^2}}e^{-\frac{(u-m(i))^2}{2\sigma(i)^2}}]du$$

$$=E_u\{\ln[\frac{1}{\sqrt{2\pi\sigma(i)^2}}e^{-\frac{(u-m(i))^2}{2\sigma(i)^2}}]^{-1}\}$$

$$=E_u[\frac{1}{2}\ln(2\pi\sigma(i)^2)+\frac{(u-m(i))^2}{2\sigma(i)^2}]$$

$$=\frac{1}{2}\ln(2\pi\sigma(i)^2)+\frac{E_u(u-m(i))^2}{2\sigma(i)^2}$$

$$=\frac{1}{2}\ln(2\pi\sigma(i)^2)+\frac{\sigma(i)^2}{2\sigma(i)^2}$$

$$=\frac{1}{2}\ln(2\pi e\sigma(i)^2)$$

$$=\ln(\sqrt{2\pi e})+\ln(\sigma(i)))$$

$$<\ln(\sqrt{2\pi e})+\sigma(i)$$

Obviously, for *P*(i, j) of the *i*-th block, its entropy and variance are in the same magnitude of order. It is known that variance is a typical statistic reflecting the overall fluctuation of random data, it will keep almost invariant unless undergoes strong distortions. From this point of view, the stability of the entropy based compressed-domain fingerprint is also demonstrated.

3.5 Fingerprint Matching

The emphasis of this paper is taking compressed-domain MDCT spectral entropy as the key feature for audio fingerprint deriving. As demonstrated above and in section 2, such kind of feature is rather stable under common audio signal distortions and slight time-domain misalignment like ±2% time-scale modifications. By further modeling with the fault tolerant big and small relationship between entropy of successive blocks, the steadiness of derived fingerprints is further reinforced. Therefore, by right of the power of the stable fingerprint, we can adopt a relatively straightforward yet effective measure, i.e. *Hamming* distance, to perform exhaustive matching between the query example and those stored recordings. An illustration of the matching procedure is shown in Figure 6. To explain more clearly, let $\{x_1, x_2, \dots, x_n\}$ be the input query fingerprint sequence which belongs to the *k*-th song, $\{x_1^i, x_2^i, ..., x_N^i\}$ be the stored fingerprint sequence of the *i-th* song ($n \ll N$), N_{song} be the number of songs stored in the database, then the minimum bit error rate (BER) of matching within a song is denoted as formula (9). The total number of comparison within the database is $(N-n+1)\times N_{song}$.

$$BER(i)=\min((x_1,x_2,...,x_n)\oplus(x^i_j,x^i_{j+1},...,x^i_{j+n-1}))/n$$
$$i=1,2,...,N_{song}\quad j=1,2,...,N-n+1 \qquad (9)$$

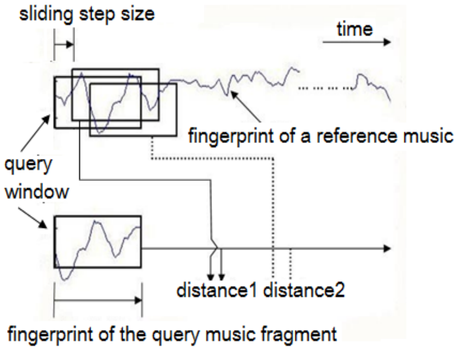
sliding step size
time
fingerprint of a reference music
query
window
distance1 distance2
fingerprint of the query music fragment

Figure. 6 An illustration of the fingerprint matching procedure

Given a reasonable false positive rate (FPR), the threshold *T* of bit error rate can be derived from both theoretical and experimental ways to indicate under what condition a match can be viewed as true. Let *BER(i')* be the ascending reordered form of *BER(i)*, namely *BER(1')<BER(2')<BER(3')<BER(4')<BER(5')<…<BER(N$_{song}$')* in which $1'=\arg\min_i(BER(i)\le T)$, thus is deemed as the song that is most similar to the input query fragment, associated metadata like lyric is often returned to users in real software systems. Remember that *k* means the number of the original song from which the query is excerpted, the overall retrieval result is summarized in formula (10).

$$result=\begin{cases} top1 & if\quad k=1' \\ top5 & elseif\quad k\in\{2',3',4',5'\} \\ top10 & elseif\quad k\in\{6',7',8,9',10'\} \\ failed & else \end{cases} \qquad (10)$$

4. EXPERIMENTAL RESULTS

To test the proposed algorithm, we first set up a music database composed of 1822 distinct MP3 popular songs and a corresponding fingerprint database based on the procedures in section 3. Each song is mono, 30 seconds long, sampled at 44.1 kHz and compressed at 64 kbps, with a fingerprint sequence of 2280 bits long. To achieve a good tradeoff among fingerprint granularity, robustness and efficiency, we experimentally use 100 pieces of 5s-long audio signals as the query examples, which are composed of excerpts cut from selected database songs and their distorted copies. Each query example has a fingerprint of 366 bits.

4.1 BER Threshold Determination

Since we use BER as the metric to test fingerprint similarity (discrimination) and robustness, we have to first determine a reasonable threshold *T* based on the desired false positive rate (FPR) in real applications. It is insignificant to claim the robustness without taking FPR into consideration. For a query fingerprint and an equal-length part of a stored fingerprint, they are judged as similar in an auditory perceptual sense if the bit error rate is below the threshold *T*. Theoretical analysis on bit error rate has been studied in the literature, for example [1, 13]. However, this approach relies on the assumption that fingerprint bits are random *i.i.d.* (independent and identically distributed) and error bits can be further modeled by normal distribution. This is unfortunately not the case in reality due to the relevance incurred by large overlap between contiguous blocks. Therefore, we go through an experimental way to determine the BER threshold. For example, in experiment we exhaustively calculate the bit error rate of all possible pairs between 11 input distorted fingerprint sequence and those different fingerprints stored in database, obtaining 7951408 results as shown in Figure 7. Given a specific bit error rate as the threshold, we can count the number of falsely matched queries and then calculate the false positive rate. Some results are listed in Table 2, where we can see that FPRs corresponding to most thresholds are acceptable in practice, then which threshold is most appropriate?

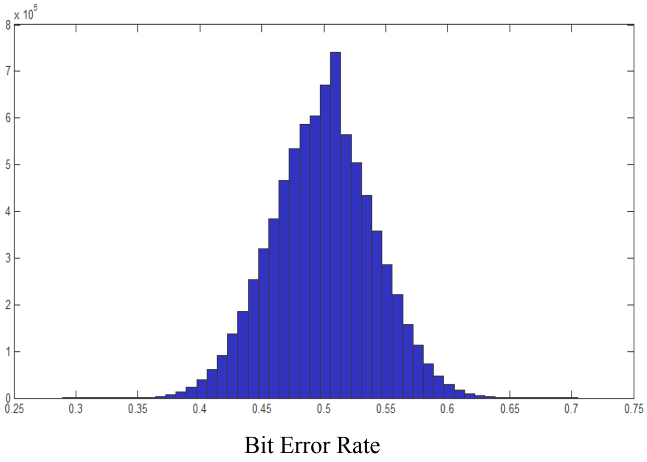

Figure 7. BER distribution of fingerprint pairs

Table 2. FPR vs. BER thresholds

BER threshold	False positive rate	BER threshold	False positive rate
0.3000	1.2576e-007	0.3400	3.5968e-005
0.3100	3.7729e-007	0.3500	1.0350e-004
0.3200	2.7668e-006	0.3600	2.5970e-004
0.3300	1.0187e-005	0.3700	6.8969e-004

To help this selection, we did some experiments from another point of view to investigate the relationship between top-1 identification precision and the BER threshold *T* as shown in Figure 8. It can be seen that the change of *T* (0.30~0.37) doesn't affect the identification performance under common audio signal distortions, while it shows obvious influence on the identification results under time-scale modifications. The precision lines under TSM (the lowest three lines) go upwards slowly and monotonously when *T* increases from 0.30 to 0.33 and hereafter keep horizontal. That is, thresholds bigger than 0.33 doesn't contribute to the identification precision any more.

Generally speaking, bigger BER threshold will give better robustness results and meanwhile make the performance of false positive rate worse. To balance these two aspects, we adopt 0.33 as the BER threshold in this paper. Its corresponding FPR equals 1.0187e-005 that is fairish for practical audio identification applications

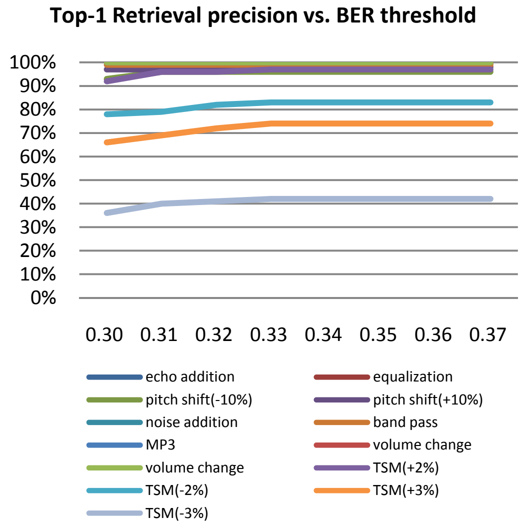

Figure 8. Identification results vs. BER thresholds

4.2 Retrieval Results under Distortions

To simulate the real world interference, we apply various audio signal operations on the compressed input query example using audio editing tools Cool edit and Gold wave. By their processing procedures to MP3 audio, the simulation is actually equivalent to decoding plus random cut plus audio signal processing plus recompression.

Given 100 randomly chosen query excerpts and 0.33 as the BER threshold, the top-1, 5 and 10 retrieval performance of this proposed algorithm under various audio distortions are averaged and shown in Figure 9. It can be seen that this MDCT entropy-based fingerprint shows very good identification precision, even under severe audio signal distortions like lossy compression of MP3@32kbps-64kbps, volume modulation, obvious echo addition (100ms, 50%), noise interference with the signal noise ratio (SNR) as low as 15dB, equalization with a 10-band equalizer, band-pass filtering from 200 to 6000 Hz, tempo-reserved pitch shifting up to ±10% and slight pitch-reserved time-scale modifications at ±2%. In the above cases, the averaged top-1 precision results are all over 80% and top-5 over 90%. Overall, the robustness under various audio signal distortions is pretty good. The only blemish is that the top-1 result under TSM@-3% is not very satisfying with only 42% averaged precision, inferior by comparison with those state-of-the-art uncompressed-domain algorithms such as [4] and [5] which can resist TSM up to ±15% and -21% - +26% respectively. The weakness of compressed-domain algorithms is essentially caused by the fixed data structure of the MP3 compressed bit stream. In this case, implicit synchronization methods based on salient local regions like in [30] can't be applied. The only way to resist serious time-domain desynchronization is to increase the overlap between contagious blocks and design more steady fingerprints, whereas the overlap has an upper limit of 100% (95% has been used in this paper) and discovering more powerful features is neither an easy work.

Compared with other related compressed-domain algorithms introduced in the introduction [8, 9, 10, 11] whose best top-5 precision rate is 90% [10] under a clean environment (no experimental results under any time-frequency distortions are reported in the above papers), our algorithm obviously outperforms those methods with the top-5 precision rates bigger than 90% even under severe audio signal distortions such as MP3@32kbps-64kbps, 10-band equalization, pitch shifting up to ±10% etc and the challenging time-scale modifications up to ±2% as shown in Figure 8. With respect to reference [13], to the authors' knowledge it is the only one published compressed-domain algorithm with some robustness reports including MP3@64Kbps, reverberation, echo, down sampling and equalization. However, it doesn't show any results against the two most challenging audio distortions, i.e. tempo-reserved pitch shifting and pitch-reserved time-scale modification, as we do in this paper. What is more important, the crucial parameter in audio identification, i.e. granularity, is not considered in any of the above algorithms. That is, they are not actually working in the way of fragment retrieval, this is an intrinsic deficiency compared with our method. Overall, this proposed method not only outperforms other existing compressed-domain audio retrieval systems, but also comes close to many state-of-the-art uncompressed domain algorithms for example [1, 4, 5, 6] in terms of robustness under common audio signal distortions except that results under serious time-scale modifications are worse due to the inability to use time-domain implicit synchronization method.

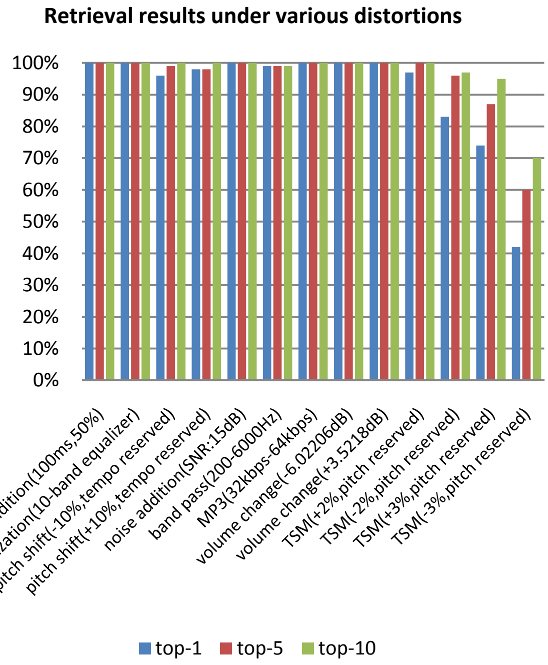

Figure 9. Retrieval results under various time-frequency distortions

5. CONCLUSION

In this paper, we propose a novel compressed-domain audio fingerprinting algorithm for MP3 popular music identification. In virtue of the short-time steady property of MDCT spectral entropy and large overlap, a 5s query excerpt is shown being able to achieve promising results on robustness and retrieval precision rates under various time-frequency audio signal distortions including the challenging pitch shifting and time-scale modification. For future work, designing more statistically stable compressed-domain features and combining with this MDCT entropy-based feature using information fusion will be our main approaches to improve the retrieval precision and robustness against large time-domain misalignment. Cover song identification performed right on the compressed-domain is our final aim to be accomplished.

6. ACKNOWLEDGMENTS

This work is jointly supported by NSFC(60873255), 973 Program(2010CB327906), Shanghai Leading Academic Discipline Project (B114).

7. REFERENCES

[1] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system, " proceeding of the international conference on music information retrieval, 2002, 107–115.

[2] S. Baluja and M. Covell, "Waveprint: efficient wavelet-based audio fingerprinting, " Pattern recognition, vol. 41, no. 11, 2008, 3467-3480.

[3] H. M. Yu, W. H. Tsai and H. M. Wang, "A query-by-singing system for retrieving karaoke music, " IEEE Transactions on multimedia, vol. 10, no. 8, 2008, 1626-1637.

[4] R. Bardeli and F. Kurth, "Robust identification of time-scaled audio, " proceeding of audio engineering society convention (AES 2004).

[5] F. Kurth, T. Gehrmann and M. Müller, "The cyclic beat spectrum: tempo related audio features for time-scale invariant audio identification, " proceeding of the international conference on music information retrieval (ISMIR 2006).

[6] J. Haitsma and T. Kalker, "Speed change resistant audio fingerprinting using autocorrelation, " proceeding of the international conference on acoustics, speech and signal processing (ICASSP 2003), 728-731.

[7] P. Cano, E. Batlle, T. Kalker and J. Haitsma, "A review of audio fingerprinting, " Journal of VLSI signal processing, vol. 41(3), 2005, 271-284.

[8] C. C. Liu and P. J. Tsai, "Content-based retrieval of MP3 music objects, " proceeding of the ACM international conference on information and knowledge management 2001, 506-511.

[9] W. N. Lie and C. K. Su, "Content-based retrieval of MP3 songs based on query by singing, " proceeding of the IEEE international conference on acoustics, speech, and signal processing (ICASSP 2004), 929-932.

[10] T. H. Tsai and J. H. Hung, "Content-based retrieval of MP3 songs for one singer using quantization tree indexing and melody-line tracking method, " proceeding of the IEEE international conference on acoustics, speech, and signal processing (ICASSP 2006), 505-508.

[11] T. H. Tsai and Y. T. Wang, "Content-based retrieval of audio example on MP3 compression domain, " proceeding of the IEEE workshop on multimedia signal processing (MSP 2004), 123- 126.

[12] D. Pye, "Content-based methods for the management of digital music," proceeding of the IEEE international conference on acoustics, speech and signal processing (ICASSP 2000), 24-27.

[13] Y. H. Jiao, B. Yang, M. Y. Li and X. M. Niu, "MDCT-based perceptual hashing for compressed audio content identification, " proceeding of the IEEE workshop on multimedia signal processing (MMSP 2007), 381-384.

[14] H. Misra, S. Ikbal, H. Bourlard and H. Hermansky, "Spectral entropy based feature for robust ASR, " proceeding of the IEEE international conference on acoustics, speech, and signal processing, 2004, 193-196.

[15] A. C. Ibarrola and E. Chavez, "A robust entropy-based audio fingerprint, " proceeding of the IEEE international conference on multimedia and expo,2006, 1729-1732.

[16] J. Pinquier and R. André-Obrecht, "Audio indexing: primary components retrieval and robust classification in audio documents, " Multimedia tools and applications, vol. 30, 2006, 313–330.

[17] R. Jarina, N. O'Connor, S. Marlow and N. Murphy, "Rhythm detection for speech-music discrimination in compressed domain," proceeding of the IEEE international conference on digital signal processing (DSP 2002), 129- 132.

[18] K. Takagi and S. Sakazawa, "Light weight MP3 watermarking method for mobile terminals," proceeding of the ACM international conference on multimedia (ACM Multimedia 2005), 443-446.

[19] Y. Wang and M. Vilermo, "A compressed domain beat detector using MP3 audio bit streams," proceeding of the ACM international conference on multimedia (ACM Multimedia 2001), 194-202.

[20] A. D'Aguanno and G. Vercellesim, "Tempo induction algorithm in MP3 compressed domain," proceeding of the ACM international conference on multimedia information retrieval (ACM MIR 2007), 153-158

[21] G. Tzanetakis and P. Cook, "Sound analysis using MPEG compressed audio, " proceeding of the IEEE international conference on acoustics, speech, and signal processing (ICASSP 2000), 761-764.

[22] C. C. Liu and C. S. Huang, "A singer identification technique for content-based classification of MP3 music objects, " proceeding of the ACM international conference on information and knowledge management 2002, 438 – 445.

[23] C. C. Liu and P. C. Yao, "Automatic summarization of MP3 music objects, " proceeding of the international conference on speech, acoustics, and signal Processing (ICASSP 2004), 921-924.

[24] X. Shao, C. S. Xu, Y. Wang and M. Kankanhalli, "Automatic music summarization in compressed-domain, " proceeding of the international conference on speech, acoustics, and signal Processing (ICASSP 2004), 261-264.

[25] R. Jarina, N. O'Connor, N. Murphy and S. Marlow, "An experiment in audio classification from compressed data, " proceeding of the international workshop on systems, signals and image processing (IWSSIP 2004).

[26] A. Rizzi, N. M. Buccino, M. Panella and A. Uncini, "Genre classification of compressed audio data," proceeding of the IEEE workshop on multimedia signal processing (MMSP 2008), 654-659.

[27] S. Pfeiffer and T. Vincent, "Formalisation of MPEG-1 compressed-domain audio features, " technical report number 01/196, CSIRO mathematical and information, sciences, Australia, 2001.

[28] http://en.wikipedia.org/wiki/Information_entropy.

[29] I. Cox, M. Miller, J. Bloom, J. Fridrich and T. Kalker, "Digital watermarking and steganography, " 2nd Edition, published by Morgan Kaufmann, 2007.

[30] C. W. Tang and H. M. Hang, "A feature-based robust digital image watermarking scheme, " IEEE Transactions on signal processing, vol. 51, no. 4, 2003, 950–959