EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Low-order auditory Zernike moment: a novel approach for robust music identification in the compressed domain

Wei Li[1*], Chuan Xiao[1] and Yaduo Liu[2]

Abstract

Audio identification via fingerprint has been an active research field for years. However, most 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 music on personal computers and/or transmit it over the Internet. It will be interesting if a compressed unknown audio fragment could be directly recognized from the database without decompressing it into the wave format at first. So far, very few algorithms run directly on the compressed domain for music information retrieval, and most of them take advantage of the modified discrete cosine transform coefficients or derived cepstrum and energy type of features. As a first attempt, we propose in this paper utilizing compressed domain auditory Zernike moment adapted from image processing techniques as the key feature to devise a novel robust audio identification algorithm. Such fingerprint exhibits strong robustness, due to its statistically stable nature, against various audio signal distortions such as recompression, noise contamination, echo adding, equalization, band-pass filtering, pitch shifting, and slight time scale modification. Experimental results show that in a music database which is composed of 21,185 MP3 songs, a 10-s long music segment is able to identify its original near-duplicate recording, with average top-5 hit rate up to 90% or above even under severe audio signal distortions.

Keywords: Music identification; MPEG; Compressed domain; Zernike moment; Robustness; Auditory image

1 Introduction

As an emerging entertainment fashion, online music business such as listening, downloading, identification, and searching have become one of the hottest applications on the World Wide Web for several years. According to the statistical report in [1], online music ranks third in all network applications, and 75.2% Internet users have ever used the above services.

Among various online applications, music identification based on audio fingerprinting technique has attracted much attention from both the research community and the industry. By comparing the fingerprint of an unknown music segment, which is 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, related metadata such as lyrics and singer's name are returned. The fingerprint must characterize the nature of the music content to differentiate from each other, possess strong robustness to various kinds of severe audio signal degradations, and typically use only a several-second music fragment for identification in the database. To date, a number of algorithms have been published with rather high retrieval performance, most of them operate on the PCM wave format, and commercially deployed software systems have also appeared [2].

However, with the mature of CD-quality audio compression techniques at lower bit rate and the fast growing of the Internet, compressed audio content is increasingly ubiquitous and has become the dominant fashion of storing and transmitting in either music libraries or personal electronic equipment. It will be interesting and meaningful in practice if audio features are

* Correspondence: weili-fudan@fudan.edu.cn
[1]School of Computer Science and Technology, Fudan University, Shanghai 201203, China
Full list of author information is available at the end of the article

directly extracted from the compressed domain and used for music identification in the database.

So far, only a few algorithms that perform music information retrieval (MIR) directly on the compressed domain have been proposed. Liu and Tsai [3] calculated the compressed domain energy distribution from the output of the polyphase filters as a feature to index songs. For each song in the dataset, they use its refrain as the query example to retrieve all similar repeating phases, obtaining an average 78% recall and 32% precision. They claim that to their knowledge, this is the first compressed domain MIR algorithm. Lie and Su [4] directly used selected modified discrete cosine transform (MDCT) spectral coefficients and derived sub-band energy and 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 and Hung [5] described a query-by-example algorithm using 176 MP3 songs of the same singer as the database. They calculate spectrum energy from sub-band coefficients (SBC) to simulate the melody contour and use it to measure the 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 and Wang's paper [6], they used scale factors (SCF) and sub-band coefficients in an 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%, 100%), the query item can be obtained in top-5, 10, 15 results, respectively. Pye [7] designed a new parameterization referred to as an MP3 cepstrum based on a partial decompression of MPEG-1 Layer III audio to facilitate the management of a typical digital music library. It is approximately six times faster than conventional Mel frequency cepstrum coefficient (MFCC) for music retrieval while the average hit rate is only 59%. Jiao et al. [8] designed a robust compressed domain audio fingerprinting algorithm, taking the ratio between the sub-band energy and full-band energy of a segment as intra-segment feature and the difference between continuous intra-segment features as inter-segment feature. Experiments show that such fingerprints are robust against transcoding, down sampling, echo adding, and equalization. However, the authors do not show any results on the retrieval hit rate. Zhou and Zhu [9] designed an MP3 compressed domain audio fingerprinting algorithm. By exploiting long-term time variation information based on the modulation frequency analysis, it is reported to be especially robust against time scale modification (TSM) at the cost of higher computation complexity. However, in experiment, the defined detection rate and accuracy rate are different from the top-*n* style measures of other algorithms; thus, it is difficult to judge whether it really outperforms other methods as stated. In [10], Liu and Chang calculated four kinds of compressed domain features, i.e., MDCT, MFCC, MPEG-7, and chroma vectors from the compressed MP3 bit stream. PCA is applied to reduce the high dimensional feature space, and QUC-tree and inverted lists of MP3 signatures are constructed to perform more efficient search. However, the experiments are only performed on MP3 fragments, which is not enough to reflect the song-level performance in real application environment.

The above methods achieve certain retrieval achievements, while they do not consider or obtain convincing results to the most central problem in audio fingerprinting, i.e., robustness. In practical application scenarios, for example, transmitting an unknown music clip through cell phone and wireless telecom network, the audio might often be contaminated by various audio distortions like lossy compression, environmental noise, echo adding, time stretching, and pitch shifting. 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 audio fingerprinting? It is well known that Zernike moment has been widely used in many image-related research fields such as image recognition [11], image watermarking [12], human face recognition [13], and image analysis [14] due to its prominent property of strong robustness and rotation, scale, and translation (RST) invariance. So far, various compressed domain audio features including scale factors [15,16], MP3 window-switching pattern [17,18], 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 [19-24] 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 image-related research fields for years, to the authors' knowledge, Zernike moment has not yet been applied to music information retrieval. This motivated our initial idea of developing compressed domain Zernike moments for audio fingerprinting technique. Two important properties of Zernike moment, i.e., strong robustness and translation invariance, are utilized to respectively resolve the problems of noise interference and desynchronization to some extent. Note that in the one-dimensional (1D) audio circumstance, properties of rotation and scale invariance are of no use.

In this paper, we first group 90 granules, the basic processing unit in decoding the MP3 bit stream, into a relatively big block for the statistical purpose, then calculate low-order Zernike moments from extracted MDCT coefficients located in selected low to middle sub-bands, and finally obtain the fingerprint sequence by modeling the relative relationship of Zernike moments between consecutive blocks. Experimental results show that this low-order Zernike moment-based audio feature achieves high robustness against common audio signal degradations like recompression, noise contamination, echo adding, equalization, band-pass filtering, pitch shifting, and slight TSM. A 10-s music fragment, which is possibly distorted, is able to retrieve its original recording with an average top-5 hit rate of 90% or beyond in our test dataset composed of 21,185 popular songs.

The remainder of this paper is organized as follows. Section 2 introduces the basic principles of MPEG-1 Layer III, bit stream data format, the concept of Zernike moment, and its effectiveness as a robust audio compressed domain feature. Section 3 details the steps of deriving MDCT low-order Zernike moment-based audio fingerprint and the searching strategy. Experimental results on retrieval hit rate under various audio signal distortions are given in Section 4. Finally, Section 5 concludes this paper and points out some possible ways for future work.

2 Compressed domain auditory Zernike moment

2.1 Principles of MP3 compression and decoding

An illustration of the MPEG-1 Layer III encoder is shown in Figure 1. In the first step, a sequence of 1,152 PCM audio samples are filtered through a polyphase filter bank into 32 Bark scale-like sub-bands, which simulate the critical bands in the human auditory system (HAS), and then decimated by a factor 32. Each sub-band will thereby contain 36 sub-band samples that are still in the time domain [25,26].

Next, the sub-bands are further subdivided to provide better spectral resolution by MDCT transform using long or short window depending on the dynamics within each sub-band which is controlled by the psychoacoustic model. If the time-domain samples within a given sub-band show a stationary behavior, a long window (e.g., 25 ms) is chosen in order to enhance the spectral resolution in the following MDCT. If the sub-band samples contain transients, three consecutive short windows (e.g., each is 4 ms) are applied in order to enhance the time resolution in the following MDCT. Moreover, start window and stop window are also defined in order to obtain better adaption when window transients appear. Figure 2 shows an example of a sequence of windows applied to a sub-band.

MDCT transform on a sub-band will produce 18 frequency lines if using a long window and three groups of frequency lines each having six frequency lines at different time intervals if using three consecutive short windows. Fifty percent overlap between adjacent windows is adopted in both cases. Therefore, MDCT transform on a granule will always produce 576 frequency lines, which are organized in different ways in the cases of long windowing and short windowing.

Combined with other adjuvant techniques including psychoacoustic model, scale-factor, *Huffman* coding, and quantization, the final compressed bit stream is generated. Figure 3 displays the frame format of MP3 bit stream [27]. Each frame has two granules to exploit further redundancies.

In MP3 decoder, the basic processing unit of the input bit stream is a frame of 1,152 samples, approximately 26.1 ms at the sampling rate of 44.1 kHz (note that each granule can be dealt with independently) [28]. One granule of compressed data is first unpacked and dequantized into 576 MDCT coefficients then mapped to the polyphase filter coefficients in 32 sub-bands by inverse modified discrete cosine transform. Finally, these sub-band polyphase filter coefficients are inversely transformed and synthesized into PCM audio, as shown in Figure 4 [26]. Therefore, access of transformation coefficients in Layer III can be either at the MDCT or the

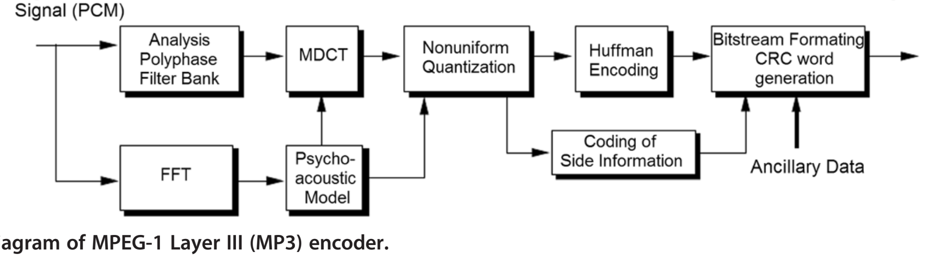

Figure 1 Block diagram of MPEG-1 Layer III (MP3) encoder.

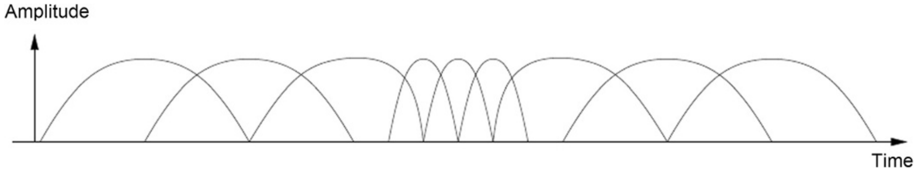

Figure 2 Illustration of a typical sequence of windows to be applied to a sub-band.

filter bank level, the latter is obviously more time-consuming.

2.2 A brief introduction of the Zernike moment

Zernike moment was originally designed as a powerful tool for image processing applications due to its robustness and RST invariant property. It has been demonstrated to outperform other image moments such as geometric moments, Legendre moments, and complex moments in terms of sensitivity to image noise, information redundancy, and capability for image representation [29].

In this section, we give a brief introduction of the basic concept of Zernike moment. Zernike moments are constructed by a set of complex polynomials which form a complete orthogonal basis set defined on the unit disk $x^2 + y^2 \leq 1$. These polynomials have the form

$$P_{nm}(x,y) = V_{n,m}(\rho,\theta) = R_{nm}(\rho)\exp(jm\theta), \tag{1}$$

where n is a non-negative integer, m is a non-zero integer subject to the constraints that $(n - |m|)$ is non-negative and even, ρ is the length of vector from the origin to the pixel (x, y), and θ is the angle between the vector and x-axis in counter-clockwise direction. $R_{nm}(\rho)$ is the Zernike radial polynomials in (ρ, θ) polar coordinates defined as

$$R_{nm}(\rho) = \sum_{s=0}^{n-\frac{|m|}{2}} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s}. \tag{2}$$

Note that $R_{n,m}(\rho) = R_{n,-m}(\rho)$, so $V_{n,-m}(\rho,\theta) = V^*_{n,m}(\rho,\theta)$.

Zernike moments are the projection of a function onto these orthogonal basis functions. The Zernike moment of order n with repetition m for a continuous two-dimensional (2D) function $f(x, y)$ that vanishes outside the unit disk is defined as

$$A_{nm} = \frac{n+1}{\pi} \iint_{x^2+y^2\leq 1} f(x,y)V^*_{n,m}(x,y)dxdy. \tag{3}$$

For 2D signal-like digital image, the integrals are replaced by summations to

$$A_{nm} = \frac{n+1}{\pi}\sum_x \sum_y f(x,y)V^*_{n,m}(x,y), x^2 + y^2 \ll 1. \tag{4}$$

2.3 Compressed domain auditory Zernike moment

The inconvenience of directly applying Zernike moment to the audio case lies in that audio is inherently a time-variant 1D data, while the Zernike moments are only applicable for 2D data. Therefore, we must map the audio signals to 2D form before making them suitable for calculating the moment. In this paper, we construct a series of consecutive granule-MDCT 2D images to directly calculate the Zernike moment sequence in the MP3 compressed domain. In light of the frame format of the MP3 bit stream, one granule corresponds to about 13 ms, which means that it is indeed an alternative representation of time. On the other hand, MDCT coefficients can be roughly mapped to actual frequencies [30]. Therefore,

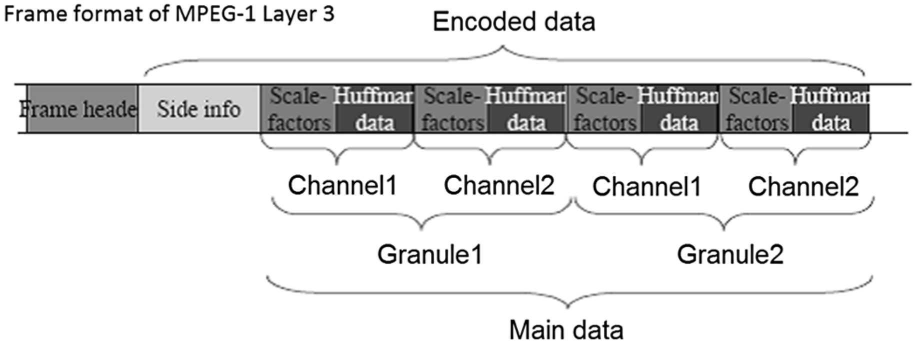

Figure 3 Frame format of MPEG-1 Layer III bit stream.

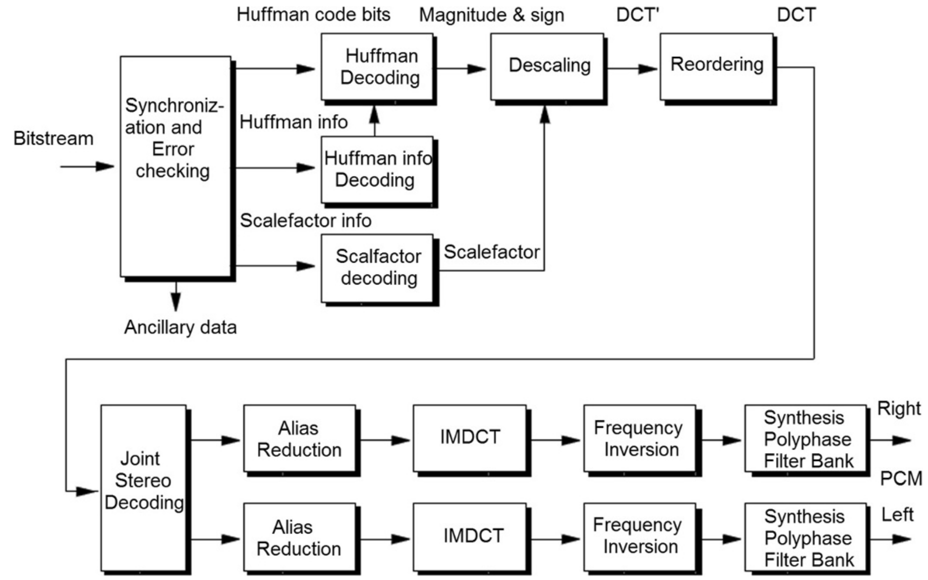

Figure 4 Block diagram of MPEG-1 Layer III (MP3) decoder.

the way we construct granule-MDCT images is virtually done on the time-frequency plane. Human audition can be viewed in parallel with human vision if the sound is converted from a one-dimensional wave to a two-dimensional pattern distributed over time along a frequency axis, and the two-dimensional pattern (frequency vs. time) constitutes a 2-D auditory image [31]. This way, we may seek to explore alternative approaches to audio identification by making recourse to mature technical means applied in computer vision. Although the link between computer vision and music identification has been made in several published algorithms, which all take the short-time Fourier transform of time-domain audio slices to create the spectrograms for the time-frequency representation using only the magnitude components [32-35], methods based on visualization of compressed domain time-MDCT images have not yet been demonstrated for music identification. We argue that mature techniques in computer vision such as Zernike moment may in fact be useful for computational audition; the detailed calculation procedures of the proposed method will be described in the next section.

As stated in the introduction, the goal of calculating MDCT-based Zernike moment is to use it as an audio fingerprint after necessary modeling, for direct compressed domain music identification. As an effective audio feature, will it be steady enough under various audio signal distortions? We did some experiments to check it. Figure 5 shows an example of MDCT 2-order Zernike moment sequence calculated from a 5-s clip of an MP3 song. The calculation includes several steps like granule grouping, sub-bands selection, and auditory image construction. It is a complex procedure and will be depicted in detail in the next section. It can be clearly seen that the Zernike moment curve is rather stable, keeping its basic shape at the same time positions under common audio signal distortions like MP3 recompression at 32 kbps, echo adding, band-pass filtering, noise contamination, volume modulation, equalization, and pitch shifting up to ±10%. When the sample excerpt is slightly time scale-modified, the curve only translates a small distance along the time axis with little change to the basic shape. These observed phenomena confirm our initial motivation. Herein, low-order Zernike moment of time-MDCT auditory image displays great potential to become a powerful audio fingerprint.

3 Algorithm description

As described above, the main difficulty of applying Zernike moment to audio is the dimension mismatching. So, we first depict how to create a 2D auditory image from 1D compressed domain MP3 bit stream. The detailed procedure of this proposed algorithm is described as follows.

3.1 MDCT-granule auditory image construction

3.1.1 Y-axis construction: frequency alignment

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 576 values represent 576 frequency lines and are assigned into 32 Bark scale-like sub-bands, that is, each sub-band includes 18 frequency lines. If a granule is compressed via a short window, these values

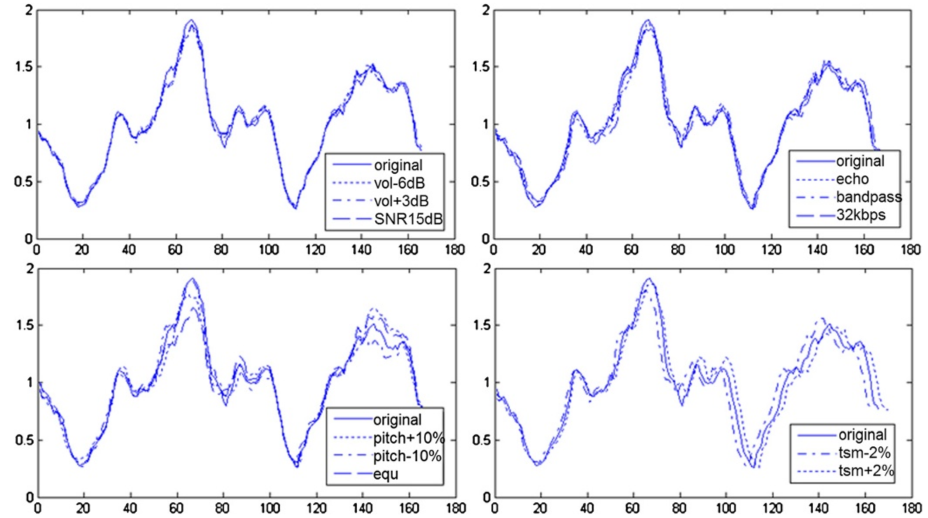

Figure 5 An example MDCT Zernike moments curve under various audio signal degradations. Order = 2, block = 50 granules, hop size = 2 granules.

only stand for 192 frequency lines, and each line includes three values that belong to three consecutive windows respectively, see Figure 6.

In order to construct the *Y*-axis of the auditory images to calculate the Zernike moment, we must unify the frequency distribution of both long- and short-window cases by adapting the original MDCT-granule relationship to achieve approximately the same frequency resolution. For long-window cases, we group every three consecutive MDCT coefficients of one granule into a new sub-band value, which is equal to the mean of the absolute value of the original three MDCT coefficients, considering that MDCT coefficients could be positive or negative, see Equation (5). 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 Equation (6). In this way, all MDCT values in a granule are uniformly divided into 192 new sub-bands for both long- and short-window

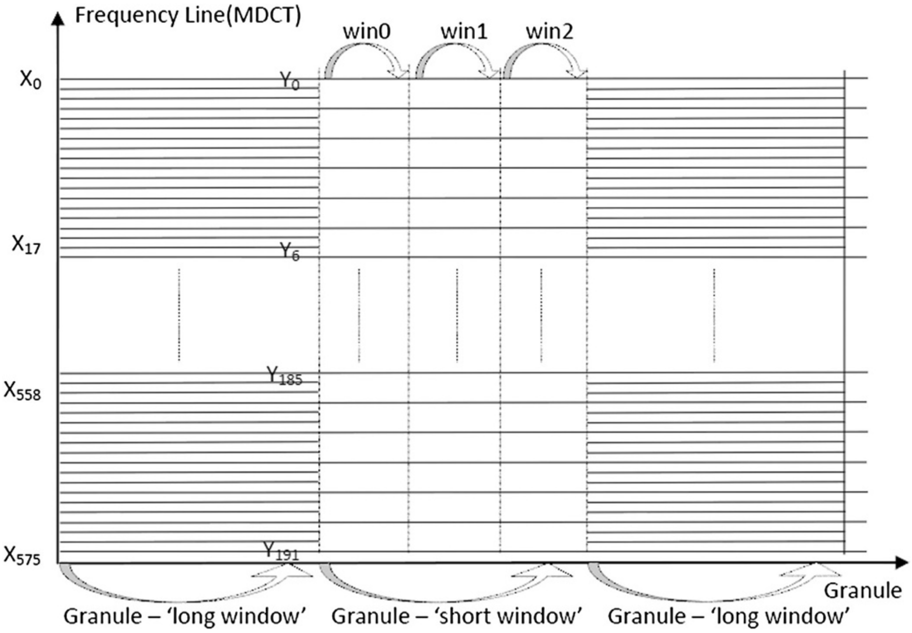

Figure 6 Distribution of MDCT coefficients in 'long window' and 'short window' types of granule.

cases; this forms the basis for further construction of auditory image.

$$sn(i,j) = \begin{cases} sn^l(i,j) = \dfrac{1}{3}\displaystyle\sum_{n=3j}^{3j+2} |s(i,n)| j = 0,1,2\ldots,191 & (5) \\ sn^s(i,j) = \dfrac{1}{3}\displaystyle\sum_{m=0}^{2} |s^m(i,j)| j = 0,1,2\ldots,191, & (6) \end{cases}$$

where $s(i, n)$ is the original MDCT coefficient in the ith granule, nth frequency line for the long-window case; $s^m(i, j)$ is the original MDCT coefficient in the ith granule, jth frequency line, mth window for the short-window case; and $sn^l(i, j)$ and $sn^s(i, j)$ are the new MDCT values in the ith granule, jth frequency line for the long- and short-window cases, respectively.

3.1.2 X-axis construction: granule grouping

After the above Y-direction operations, we have to go on setting up the X-axis to form the final auditory images. In the proposed method, N continuous granules ($N = 90$ in experiment) are partitioned into a block and act as the X-axis of one auditory image. The first block slides forward with M granules ($M = 1$ in experiment) as the hop size to form the X-axis of the following images.

Besides being a necessary step to construct the X-axis of auditory images, overlapping between two consecutive blocks is also a radical means to alleviate the time-domain desynchronization between the original and the extracted fingerprint calculated later. Figure 7 shows an illustration of the mechanism: $H(i)$ represents the start section of the ith block, $H(i + 1)$ represents that of the $(i + 1)$th block (the whole block length is unable to be depicted due to the space limitation). The hop size is for example 2 granules, i.e., $A + B$ ($A = B$ here). By reason of the time-domain desynchronization caused by random cropping etc., query bit stream is scarcely possible to be exactly aligned to $H(i)$ or $H(i + 1)$. When the query clip 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 lie in the middle, it comes to the worst synchronization performance, i.e., half hop size. We hope that the designed audio fingerprint benefits from the transformation invariance property of the Zernike moment and will only be slightly or even not changed under this scope of misalignment.

3.1.3 Auditory image construction

With the above definition of the X- and Y-axes, we are now to construct the auditory images for calculating Zernike moments. Figure 8 is an image for illustration, where its pixels constitute an $M \times N$ matrix. N pixels along the Y-axis represent N new MDCT coefficients calculated in terms of Equations (5) and (6), and M pixels at the X-axis mean the M time-domain granules, i.e., a block. It is known that sounds located in the low-middle frequency area cover the main content most vital to the HAS and are usually much more robust against various audio distortions than high frequency components. Therefore, we pick out the second to the fifty-first new sub-band MDCT values in this method to act as the Y-axis, which roughly correspond to 300 to 5,840 Hz of real frequency according to Table 1 [17]. M is set to 90 granules to form the X-axis and mitigate the problem of desynchronization.

Consequently, the (x, y) coordinates of a pixel in the kth constructed auditory image are shown in Equation (7):

$$\boldsymbol{f^k(x,y) = sn(i,j)} \qquad \begin{array}{c} \boldsymbol{k} = 0,1,2\ldots,\boldsymbol{N}_{\text{block}} \\ \boldsymbol{i} = 0,1,2\ldots,89 \\ \boldsymbol{x} = \boldsymbol{k} \times \text{hop size} + \boldsymbol{i} \\ \boldsymbol{y} = \boldsymbol{j} = 2,3,\ldots,51 \end{array}, \qquad (7)$$

where k means the kth auditory image, and N_{block} is the total number of blocks of the query clip or the

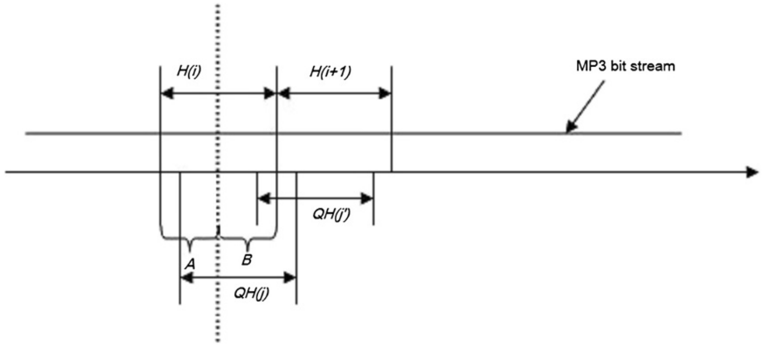

Figure 7 Desynchronization alleviation by overlapping between contiguous blocks.

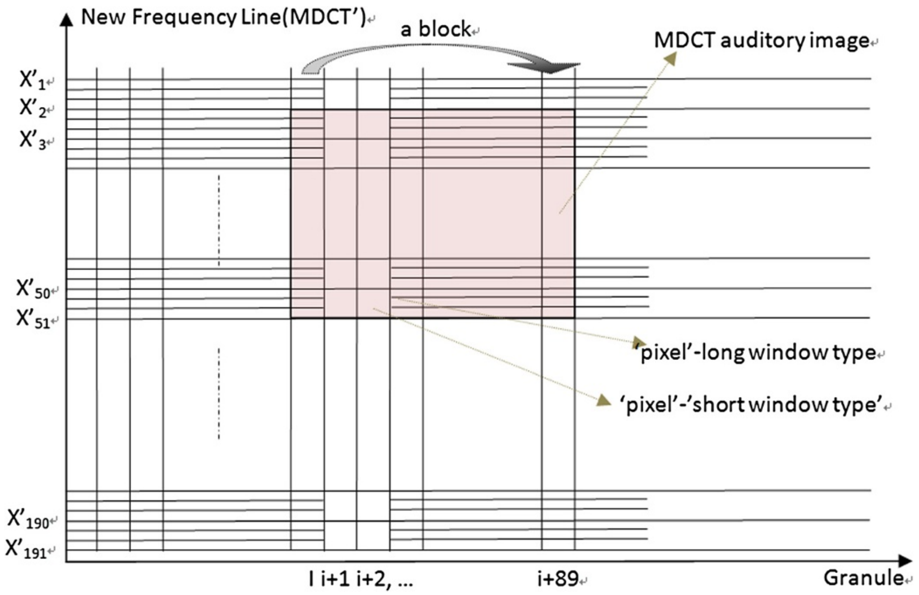

Figure 8 **An illustration of the constructed auditory image.**

original music piece, which is variable and determined by the audio length.

3.2 Compressed domain audio features: MDCT Zernike moments

Fragment input and robustness are known to be two crucial constraints on audio fingerprinting schemes. If modeling with audio signal operations, this is equal to imposing random cropping plus other types of audio signal processing on the input query example. Random cropping causes serious desynchronization between the input fingerprint sequence and those stored ones, bringing a great threat to the retrieval hit rate. Usually, there are two effective mechanisms to resist time-domain misalignment, one is invariant feature, and the other is implicit synchronization which might be more powerful than the former [36]. However, in the MPEG compressed domain, due to its compressed bit stream data nature and fixed frame structure, it is almost impossible to extract meaningful salient points serving as anchors as in the uncompressed domain [37]. Therefore, designing a statistically stable audio feature becomes the main method to fulfill the task of fragment retrieval and resisting time-domain desynchronization in audio fingerprinting.

With the preparations above, we substitute $f(x, y)$ in Equation (4) with $f^k(x, y)$ in Equation (7) and calculate the Zernike moment of the kth auditory image as below

$$A_{nm}^{k} = \frac{n+1}{\pi} \sum_{x} \sum_{y} f^{k}(x, y) V_{n,m}^{*}(x, y), \tag{8}$$

where n is the moment order, and m must be subject to the condition that $(n - |m|)$ is non-negative and even.

3.2.1 Effect of moment orders

The order n plays a crucial role in the above moment calculation. A carefully selected order will directly determine the robustness of this feature and the running speed. Generally

Table 1 Map between MDCT coefficients and actual frequencies for long and short windows sampled at 44.1 kHz

Long window		Short window	
Index of MDCT coefficient	Frequency (Hz)	Index of MDCT coefficient	Frequency (Hz)
0 to 11	0 to 459	0 to 3	0 to 459
12 to 23	460 to 918	4 to 7	460 to 918
24 to 35	919 to 1,337	8 to 11	919 to 1,337
36 to 89	1,338 to 3,404	12 to 29	1,338 to 3,404
90 to 195	3,405 to 7,462	30 to 65	3,405 to 7,462
196 to 575	7,463 to 22,050	66 to 191	7,463 to 22,050

speaking, low-order moments characterize the basic shape of an audio or image signal, while higher-order ones depict the high-frequency details [14]. Thereby, we naturally conjecture that low-order Zernike moments will perform better than high-order moments in our application. In order to verify this assumption and help obtain the most suitable order for strong robustness, we did some comparative experiments shown below. In Figure 9, the Zernike moment of orders 2, 6, 10, and 16 are first calculated and then compared with those values under two typical distortions, i.e., equalization and noise addition. It can be clearly seen that with the order increasing, the moment envelope fluctuates more and more dramatically. The Zernike moment curve of the order 2 is the most stable one in the experiment and is chosen as the final value in this algorithm. An affiliated benefit brought by this order is that the computation speed of its corresponding Zernike moment is much faster than any other higher-order situations.

3.3 Fingerprint modeling

On the basis of those Zernike moments calculated from a series of auditory images sliding along the granule axis, we sum up all Zernike moments with order $n \leq 2$ as the final feature to further increase the invariance as shown in Equation (9). The final audio fingerprint sequence is derived according to Equation (10). This method is straightforward yet effective by omitting the exact moment values and only retaining their relative magnitude relationship. Similar methods have been used in query-by-humming systems to model the progressive tendency of the melody line.

$$\boldsymbol{Z}_{\boldsymbol{mn}}^{\boldsymbol{k}} = \sum_{\substack{0 \leq n \leq 2 \\ (n-|m| \geq 0 \\ (\boldsymbol{n}-|\boldsymbol{m}|)\%2 = 0}} \boldsymbol{A}_{\boldsymbol{mn}}^{\boldsymbol{k}} \tag{9}$$

$$\boldsymbol{S}(\boldsymbol{k}) = \begin{cases} 0 \text{ if } \boldsymbol{Z}_{\boldsymbol{mn}}^{\boldsymbol{k}} < \boldsymbol{Z}_{\boldsymbol{mn}}^{\boldsymbol{k}+1} \\ 1 \text{ if } \boldsymbol{Z}_{\boldsymbol{mn}}^{\boldsymbol{k}} \geq \boldsymbol{Z}_{\boldsymbol{mn}}^{\boldsymbol{k}+1} \end{cases} \quad \boldsymbol{k} = 0, 1, 2 \ldots, \boldsymbol{N}_{\text{slot}} - 1. \tag{10}$$

3.4 Fingerprint matching

The emphasis of this paper is taking compressed domain audio Zernike moment as the key feature for audio fingerprinting. As stated in Section 2, such kind of feature is rather stable under common audio signal distortions and slight time-domain misalignment like time scale modification. By further modeling with the fault-tolerant magnitude relationship between moments of successive auditory images, the steadiness of the 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 10. More specifically, let $\{x_1, x_2, \ldots, x_n\}$ be the input query fingerprint sequence which belongs to the kth song, $\{x_1^i, x_2^i, \ldots, x_N^i\}$ be the stored fingerprint sequence of the ith song $(\boldsymbol{n} \ll \boldsymbol{N})$, N_{song} be the number of songs stored in the

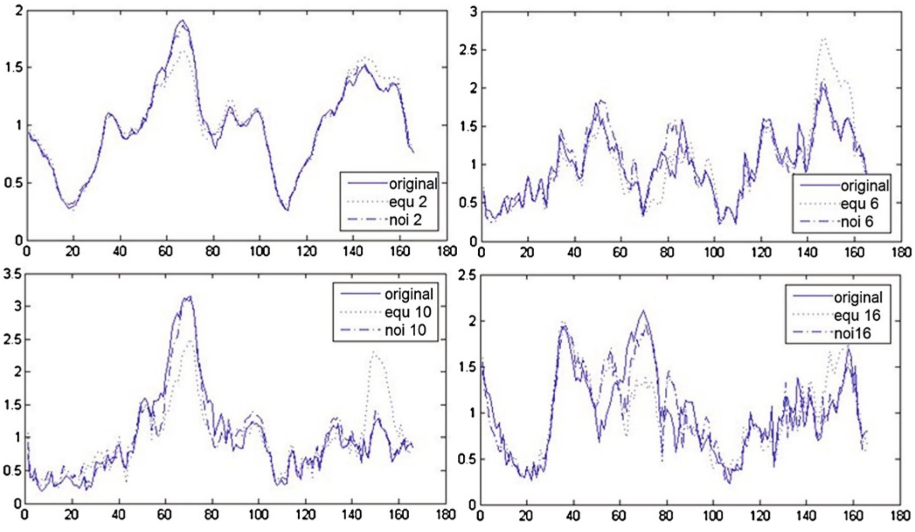

Figure 9 Stability of MDCT Zernike moments at different orders (2, 6, 10, and 16).

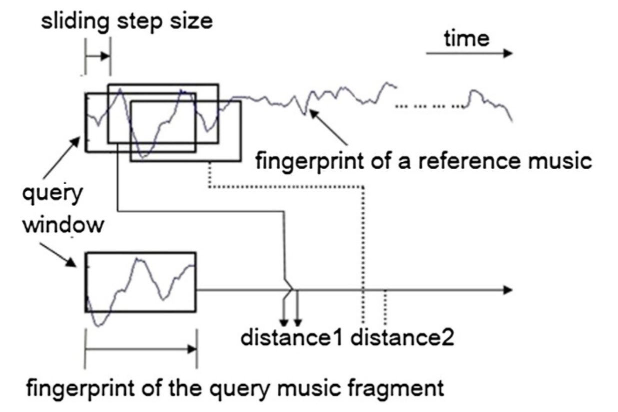
sliding step size
time
fingerprint of a reference music
query
window
distance1 distance2
fingerprint of the query music fragment

Figure 10 An illustration of the fingerprint matching procedure.

database, and Equation (11) be the minimum bit error rate (BER) of matching within a song.

$$\mathrm{BER}(\boldsymbol{i}) = \frac{1}{\boldsymbol{n}}\min\left((\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{\boldsymbol{n}})\otimes\left(\boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{j}}, \boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{j}+1}, \ldots, \boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{j}+\boldsymbol{n}-1}\right)\right) \quad \boldsymbol{i} = 1, \ldots, \boldsymbol{N}_{\mathrm{song}} \quad \boldsymbol{j} = 1, \ldots, \boldsymbol{N}-\boldsymbol{n}+1. \tag{11}$$

The total number of comparison within the database is $(N - n + 1) \times N_{\mathrm{song}}$.

Given a reasonable false positive rate (FPR), the threshold of the bit error rate T can be acquired from both theoretical and practical ways to indicate under what condition a match can be called a hit. 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}), then the final retrieval results are summarized in Equation (12), and more details are shown in the flow diagram of Algorithm 1 below.

$$\text{result} = \begin{cases} \text{top1} & \text{if } k = 1' \\ \text{top5} & \text{else if } k \in \{2', 3', 4', 5'\} \\ \text{top10} & \text{else if } k \in \{6', 7', 8', 9', 10'\} \\ \text{failed} & \text{else} \end{cases} \tag{12}$$

Algorithm 1. Fragment Retrieval

```
Input: fingerprint of unknown query music fragment, fingerprint database
Initialize NTop1 = 0, NTop5 = 0, NTop10 = 0
for $k$ = 1 to $N_{song}$ do
for $j$ = 1 to $N - n+1$ do
$BER(k) = min((x^0{}_1,x^0{}_2,...,x^0{}_n) \oplus (x^k{}_j,x^k{}_{j+1},...,x^k{}_{j+n-1}))/n$
$POS(k)$ = j
end for
resort $BER(k)$ incrementally to $BER'(k')$
$k'$ = index($k$)
end for
$k'$ = 1; $k = index^{-1}(k')$
if $k$==$k^0$ and $BER'(k') \leq T$
  NTop1 = NTop1 + 1, return ($k$, $POS$(k))
end if
for $k'$ = 2 to 10
$k$ = index$^{-1}$($k'$)
 if $k$==$k^0$ and $2 \leq k' \leq 5$
   NTop5 = NTop5 + 1, return ($k$, $POS$($k$))
 else if $k$==$k^0$ and $6 \leq k' \leq 10$
NTop10 = NTop10 + 1, return ($k$, $POS$($k$))
end if

end for
hit rate-top1 = NTop1/number of queries
hit rate-top5 = NTop5/number of queries
hit rate-top10 = NTop10/number of queries
```

4 Experiments

The experiments include a training stage and a testing stage. In the training step, three parameters (i.e., hop size, block size, and BER threshold) that affect the algorithm's performance are experimentally tuned to get the best retrieval results. To achieve this end, a small training music database composed of 100 distinct MP3 songs is set up. In the testing stage, the algorithm with the obtained parameters from training is tested on a large dataset composed of 21,185 different MP3 songs to thoroughly evaluate the retrieval performance and robustness. All songs in the two databases are mono, 30 s long, originally sampled at 44.1 kHz, and compressed to 64 kbps, with a fingerprint sequence of 672 bits. In both stages, audio queries are prepared as follows. For each song in the training (testing) database, a 10-s long query segment is randomly cut and distorted by 13 kinds of common audio signal manipulations to model the real-world environment, and hence, 1,400 (296,590) query segments (including the original segments) are obtained, respectively.

4.1 Parameter tuning

First, we describe the parameter tuning procedure. Note that when the combination of the parameters varies, the associated fingerprint database is named in accordance

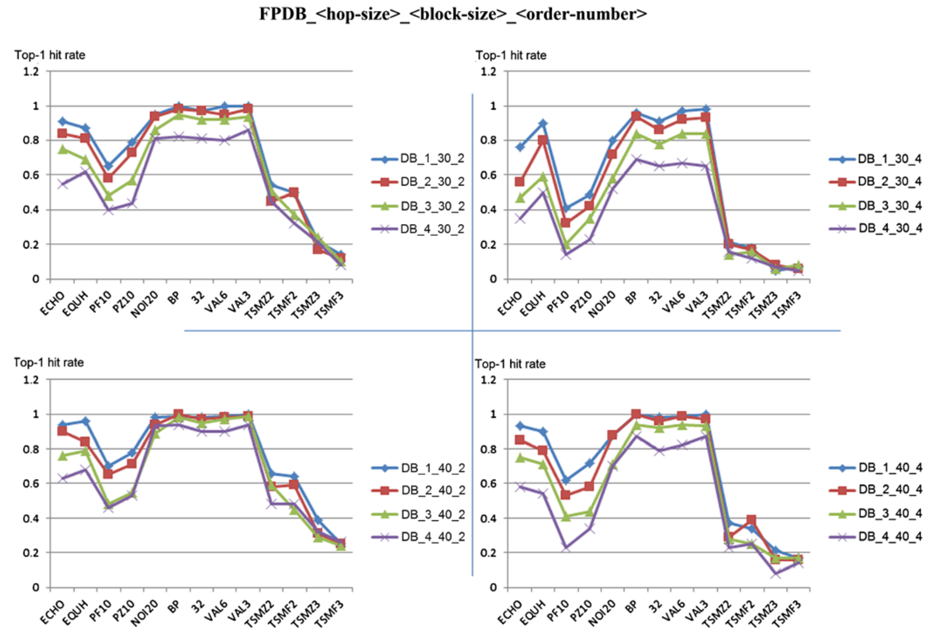

Figure 11 Influence of various hop sizes on top-1 hit rate.

with the following rule, i.e., FPDB_ < hop-size > _ < block-size > _ < order-number > .

4.1.1 Effect of hop size

Hop size is the interval between two adjacent blocks in the time axis. Smaller hop size is beneficial to alleviate the desynchronization between the query segment and its true counterpart in the original audio. Since each block is concatenated by granules, theoretically, one granule of hop size will lead to the minimal displacement. This conclusion is also experimentally demonstrated in Figure 11, where hop size varies from 1 to 4, the block size are fixed at 30 or 40, and the Zernike moment order fixed at 2 or 4. It can be clearly seen that when the hop size is 1, the corresponding blue curves are always above other curves. More precisely, when the

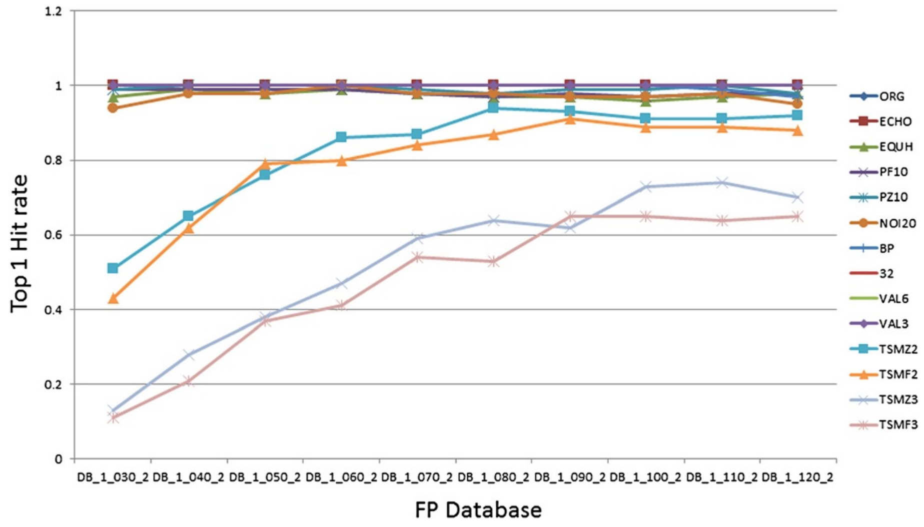

Figure 12 Influence of various block sizes on top-1 hit rate.

hop size becomes bigger, the top-1 hit rate curve moves downwards, namely the identification accuracy becomes worse.

4.1.2 Effect of block size

As stated in Section 3, a block is assembled by a set of granules in order to endure small variations in the time domain. Generally, longer block will generate steadier Zernike moment value at the cost of lowering local sensitivity and discriminability of fingerprints. To investigate the effect of block size on top-1 hit rate, we first fix the hop size at 1 and Zernike moment order at 2 and then vary the block size from 30 to 120 by increment of 10. From Figure 12, it can be seen that for the common audio signal distortions such as lossy compression, echo adding, and resampling, the top-1 hit rates are not obviously affected by the increase of the block size. However, for time scale modifications (±2% and ±3% in the experiment), the corresponding four curves (in the middle of the figure) go up monotonically with the increase of the block size and reach a stable status when block size is equal to 90 granules. Therefore, the parameter block size is set as 90 in the experiment.

4.1.3 BER thresholding

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 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 a perceptual sense if the bit error rate is below the threshold T. Theoretical and semi-theoretical analysis on bit error rate have been studied in the literature (for example, [8,38]). However, these approaches rely on the assumption that fingerprint bits are random independent and identically distributed, and error bits can be further modeled by normal distribution. This is unfortunately not the case in reality due to relevance incurred by large overlap between contiguous frames, and a theoretical FPR value is usually much smaller than its corresponding experimental value. Consequently, we prefer to adopt the experimental method to estimate the false positive rate. First, a set of fingerprint pairs combined from different songs are constructed, then the BER of each pair is calculated. All BER values exhibit a bell-shaped distribution around 0.5. Given a specific threshold T, false positive rate is determined by dividing the number of falsely matched queries by that of all fingerprint pairs. We further observe that experimental FPRs corresponding to most thresholds, for example from 0.2 to 0.4, are acceptable in practice. Then, which threshold is most appropriate? To help make this selection, we did some experiments from another point of view to investigate the relationship between top-1 identification hit rate and the BER threshold T as shown in Figure 13. It can be seen that when T increases from 0.30 to 0.40, the hit rate lines under common audio signal distortions, pitching shifting, and time scale modifications first successively go upwards monotonously and then keep steady after 0.34; in other words, bigger thresholds do not significantly contribute to the identification hit rate any more. In conclusion, 0.34 is adopted as the BER threshold in the experiment.

4.2 Retrieval results under distortions

To simulate the real-world interference, we apply various audio signal operations on the compressed

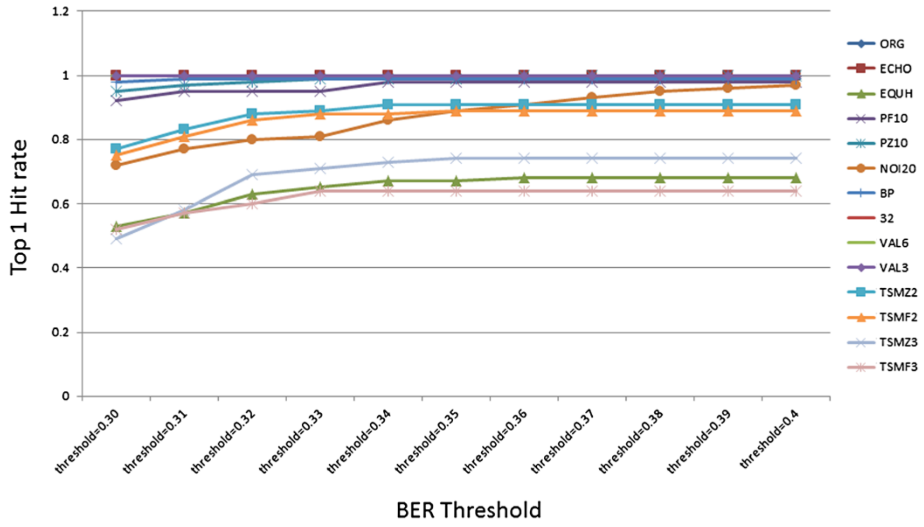

Figure 13 Relationship between BER threshold and top-1 hit rate.

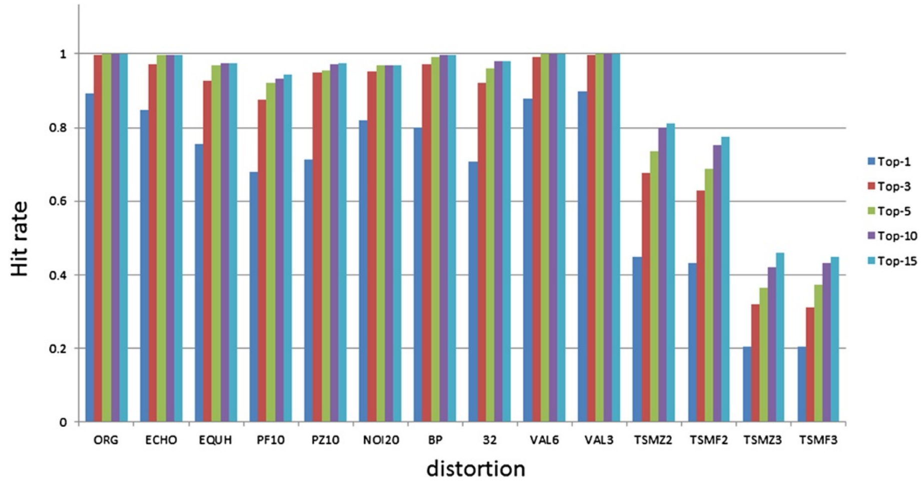

Figure 14 Retrieval performance under various time-frequency distortions.

query examples using audio editing tools Cool Edit (Adobe Systems Inc., CA, USA) and Gold Wave (GoldWave® Inc., Canada). Since music identification is done in a fragmental way, the processing procedure is actually equivalent to a mixture of random cut plus signal processing. For each song in the testing database, where 21,185 distinct songs are collected all together, a 10-s segment is first randomly cut and then manipulated by 13 various audio signal distortions. Accordingly, the query set amounts to 296,590 audio excerpts. With the parameters set as above (i.e., block size = 90, hop size = 1, and BER threshold = 0.34), the top-1, 5, and 10 identification rates of the queries within the testing dataset are averaged and illustrated in Figure 14. The horizontal axis lists the abbreviation of audio signal distortions adopted in the experiment. ORG means original audio signal which is not distorted. ECHO means echo addition with 100-ms delay and 50% decay. EQUH means 10-band equalization. PF10 and PZ10 mean pitch shifting by −10% and +10%, respectively. NOI20 means noise addition at SNR of 20 dB. BP means band-pass filtering from 100 to 6,000 Hz. 32 means MP3 recompression under 32 kbps. VAL6 and VAL3 mean volume change under −6.02 and +3.52 dB, respectively. TSMZ2, TSMF2, TSMZ3, and TSMF3 mean time scale modification under +2%, −2%, +3%, and −3%, respectively.

It can be seen that this proposed MDCT Zernike moment-based fingerprint shows satisfying identification results, even under severe audio signal processing like heavy lossy recompression, volume modulation, echo adding, noise interference, and various frequency wrappings such as band-pass filtering, equalization, and pitch shifting (±10%). To be more specific, when the queries are original or only distorted by echo adding, band-pass filtering, and volume modulation, the top-5 hit rates (green bars) are almost not influenced and all come close to 100%. Under other more severe signal manipulations such as equalization, pitch shifting, noise addition, and MP3 compression, the top-5 hit rates are pretty good and still above 90%. The only deficiency is that under pitch-reserved time scale modifications, which can be modeled as a kind of cropping/pasting to relatively smooth local parts in between music edges [37], the identification results drop quickly with the increase of scaling factors and become unacceptable when ±3% time scale modifications are performed.

This weakness 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 cannot be applied. The only way to resist serious time-domain desynchronization is to increase the overlap between consecutive blocks and design more steady fingerprints; however, the overlap has an upper limit of 100% (98% has been used in this method), and discovering more powerful features is not an easy work.

At present, it is difficult to quantitatively compare with other algorithms because different datasets or even different evaluation measures are used. It is also unrealistic to precisely reimplement these methods due to the lack of adequate details. In fact, one important task of our work is to collect enough songs (21,185 songs in the experiment) and queries (296.590

Table 2 False statistics of identification results

Actual	Predicted	
	Positive	Negative
Positive (27,378)	23,336	4,042
Negative (29,256)	106	29,150

distorted queries) under various distortions (13 kinds of audio signal distortions) so that the experimental results are more convincing. Since this dataset is much larger and more comprehensive than those used in the cited references, the source codes will be published to the research community and serve as a baseline system.

4.3 False analysis

In a practical identification system, two important false statistics must be taken into account to thoroughly evaluate the overall performance. The first is called false negative rate, which fails to detect the correct songs even though the query is included in the database. The second is false positive rate, which returns wrong matched results for a query that does not belong to the database and is more annoying in commercial applications. Below, a confusion matrix is adopted to analyze the two types of mistakes. To achieve this aim, we prepare 27,378 queries that exist in the testing database and 29,256 queries that come from outside the database. In all the true queries, 4,042 of them are not successfully retrieved from the database (i.e., the false negative rate is 14.7%), while for all the false queries, 106 of them are falsely judged to be within the database and get wrong results (i.e., the false positive rate is 3.6×10^{-3}), as shown in Table 2. The false positive rate is acceptable in practical application, while the false negative rate is relatively big. The reasons are twofold, one is that the above numbers are top-1 identification results, the other is that many database songs of a same singer have quite similar musical aspects in rhythm, harmonic progression, instrument arrangement, etc. so that the queries are confused.

5 Conclusion

In this paper, a novel music identification algorithm is proposed, which directly works on the MP3-encoded bit stream by constructing the MDCT-granule auditory images and then calculating the auditory Zernike moments. By virtue of the short-time stationary characteristics of such feature and large overlap, 10-s long query excerpts are shown to have achieved promising retrieval hit rates from the large-scale database containing intact MP3 songs and distorted copies under various audio signal operations including the challenging pitch shifting and time scale modification. For future work, combining the MDCT Zernike moments with other powerful compressed domain features using information fusion will be our main approach to improve the retrieval performance and robustness against large time domain misalignment and stretching. Cover song identification performed right on the compressed domain is our final aim to be accomplished.

Competing interests

The authors declare that they have no competing interests.

Authors' information

WL received his PhD degree in Computer Science from Fudan University, Shanghai, China in 2004. He is now a professor in the School of Computer Science and Technology, Fudan University, leading the multimedia security and audio information processing laboratory. He has published 40 refereed papers so far, including international leading journals and key conferences, such as IEEE Transactions on Audio, Speech, and Language Processing, IEEE Transactions on Multimedia, Computer Music Journal, IWDW, ACM SIGIR, and ACM Multimedia. He is a reviewer of international journals like IEEE Transactions on Signal Processing, IEEE Transactions on Multimedia, IEEE Transactions on Audio, Speech & Language Processing, IEEE Transactions on Information Forensics and Security, and Signal Processing, and conferences such as ICME, ACM MM, and IEEE GlobalCom.

Acknowledgments

This work is supported by NSFC (61171128), 973 Program (2010CB327900), and 985 Project (EZH2301600/026).

Author details

[1]School of Computer Science and Technology, Fudan University, Shanghai 201203, China. [2]China Electric Power Research Institute, Beijing 100192, China.

Received: 11 March 2013 Accepted: 11 July 2013
Published: 1 August 2013

References

1. China Internet Network Information Center, *he 29th statistical report on internet development in China, 2012.* (CNNIC, 2013), http://www1.cnnic.cn/IDR/ReportDownloads/. Accessed January 2012
2. P Cano, E Batlle, T Kalker, J Haitsma, A review of audio fingerprinting. J. VLSI Signal Process. **41**(3), 271–284 (2005)
3. CC Liu, PJ Tsai, *Content-based retrieval of MP3 music objects. Paper presented at the ACM international conference on information and knowledge management (CIKM 2001)* (Atlanta, Georgia, USA, 2001)
4. WN Lie, CK Su, *Content-based retrieval of MP3 songs based on query by singing. Paper presented at the IEEE international conference on acoustics, speech, and signal processing (ICASSP 2004)* (Montreal, Quebec, Canada, 2004)
5. TH Tsai, JH Hung, Content-based retrieval of MP3 songs for one singer using quantization tree indexing and melody-line tracking method, in *Proceeding of The IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, ed. by IEEE (IEEE, New York, 2006), pp. 505–508
6. TH Tsai, YT Wang, *Content-based retrieval of audio example on MP3 compression domain. Paper presented at the IEEE workshop on multimedia signal processing (MMSP 2004)* (Siena, Italy, 2004)
7. D Pye, *Content-based methods for the management of digital music. Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP 2000)* (Istanbul, Turkey, 2000)
8. YH Jiao, B Yang, MY Li, XM Niu, *MDCT-based perceptual hashing for compressed audio content identification. Paper presented at the IEEE workshop on multimedia signal processing (MMSP 2007)* (Chania, Crete, Greece, 2007)
9. R Zhou, Y Zhu, A robust audio fingerprinting algorithm in MP3 compressed domain. WASET Journal **55**, 715–719 (2011)
10. CC Liu, PF Chang, *An efficient audio fingerprint design for MP3 music. Paper presented at the ACM international conference on advances in mobile computing and multimedia (MoMM 2011)* (Ho Chi Minh City, Vietnam, 2011)
11. A Khotanzad, YH Hong, Invariant image recognition by Zernike moments. IEEE Trans Pattern Anal Mach Intell **12**(5), 489–497 (1990)
12. HS Kim, HK Lee, Invariant image watermark using Zernike moments. IEEE Trans. Circuits Syst. Video Technol. **13**(8), 766–775 (2003)
13. J Haddadnia, M Ahmadi, K Faez, An efficient feature extraction method with pseudo-Zernike moment in RBF neural network-based human face recognition system. EURASIP J Adv. Signal Process. **9**, 890–901 (2003)
14. NK Kamilaa, S Mahapatrab, S Nanda, RETRACTED: Invariance image analysis using modified Zernike moments. Pattern Recogn. Lett. **26**(6), 747–753 (2005)

15. R Jarina, N O'Connor, S Marlow, N Murphy, *Rhythm detection for speech-music discrimination in compressed-domain. Paper presented at the IEEE international conference on digital signal processing (DSP 2002)* (Pine Mountain, Georgia, USA, 2002)
16. K Takagi, S Sakazawa, *Light weight MP3 watermarking method for mobile terminals. Paper presented at the ACM international conference on multimedia (ACM Multimedia 2005)* (Hilton, Singapore, 2005)
17. Y Wang, M Vilermo, *A compressed domain beat detector using MP3 audio bit streams. Paper presented at the ACM international conference on multimedia (ACM Multimedia 2001)* (Ottawa, Ontario, Canada, 2001)
18. A D'Aguanno, G Vercellesim, *Tempo induction algorithm in MP3 compressed-domain. Paper presented at the ACM international conference on multimedia information retrieval (ACM MIR 2007)* (University of Augsburg, Germany, 2007)
19. G Tzanetakis, P Cook, *Sound analysis using MPEG compressed audio. Paper presented at the IEEE international conference on acoustics, speech, and signal processing (ICASSP 2000)* (Istanbul, Turkey, 2000)
20. CC Liu, CS Huang, *A singer identification technique for content-based classification of MP3 music objects. Paper presented at the ACM international conference on information and knowledge management (CIKM 2002)* (McLean, Virginia, USA, 2002)
21. CC Liu, PC Yao, *Automatic summarization of MP3 music objects. Paper presented at the IEEE international conference on speech, acoustics, and signal Processing (ICASSP 2004)* (Montreal, Quebec, Canada, 2004)
22. X Shao, CS Xu, Y Wang, M Kankanhalli, *Automatic music summarization in compressed-domain. Paper presented at the IEEE international conference on speech, acoustics, and signal Processing (ICASSP 2004)* (Montreal, Quebec, Canada, 2004)
23. R Jarina, N O'Connor, N Murphy, S Marlow, *An experiment in audio classification from compressed data. Paper presented at the international workshop on systems, signals and image processing (IWSSIP 2004)* (Poznan, Poland, 2004)
24. A Rizzi, NM Buccino, M Panella, *A Uncini, Genre classification of compressed audio data. Paper presented at the IEEE workshop on multimedia signal processing (MMSP 2008)* (Cairns, Queensland, Australia, 2008)
25. T Painter, A Sanias, Perceptual coding of digital audio. Proc IEEE **88**(4), 451–513 (2000)
26. K Salomonsen, S Søgaard, EP Larsen, *Design and implementation of an MPEG/audio layer III bit stream processor, Thesis* (Department of Communication Technology, Aalborg University, Denmark, 1997)
27. S Pfeiffer, T Vincent, *Formalisation of MPEG-1 compressed-domain audio features, in technical report number 01/196, CSIRO Mathematical and Information, Sciences, Australia*, 2001
28. K Lagerster, *Design and implementation of an MPEG-1 layer III audio decoder, Thesis* (Chalmers University of Technology, 2001)
29. RJ Prokop, AP Reeves, A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP-Graph. Model. Im. **54**(5), 438–460 (1992)
30. TY Chang, *Research and implementation of MP3 encoding algorithm, Thesis* (National Chiao Tung University, Hsinchu, Taiwan, ROC, 2002)
31. R Rifkin, J Bouvrie, K Schutte, S Chikkerur, M Kouh, T Ezzat, T Poggio, Phonetic classification using hierarchical, feed-forward, spectro-temporal patch-based architectures, in *Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2007-019*, 2007
32. Y Ke, D Hoiem, R Sukthankar, *Computer vision for music identification. Paper presented at the IEEE computer society conference on computer vision and pattern recognition (CVPR 2005)* (San Diego, CA, USA, 2005)
33. R Sukthankar, Y Ke, D Hoiem, *Semantic learning for audio applications: a computer vision approach. Paper presented at the IEEE computer society conference on computer vision and pattern recognition workshop (CVPRW 2006)*, 2006
34. S Baluja, M Covell, *Audio fingerprinting: combining computer vision & data stream processing. Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP 2007)* (Honolulu, Hawaii, USA, 2007)
35. S Baluja, M Covell, Waveprint: efficient wavelet-based audio fingerprinting. Pattern Recogn. **41**(11), 3467–3480 (2008)
36. I Cox, M Miller, J Bloom, J Fridrich, T Kalker, *Digital Watermarking and Steganography*, 2nd edn. (Morgan Kaufmann, Burlington, MA, 2007)
37. W Li, XY Xue, PZ Lu, Localized audio watermarking technique robust against time-scale modification. IEEE Trans. Multimedia **8**(1), 60–69 (2006)
38. J Haitsma, T Kalker, *A highly robust audio fingerprinting system. Paper presented at the international conference on music information retrieval (ISMIR 2002)* (Paris, France, 2002)

doi:10.1186/1687-6180-2013-132
Cite this article as: Li *et al.*: Low-order auditory Zernike moment: a novel approach for robust music identification in the compressed domain. *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:132.