EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH **Open Access**

# Music content authentication based on beat segmentation and fuzzy classification

Wei Li*, Xiu Zhang and Zhurong Wang

## Abstract

Digital audio has been ubiquitous over the past decade. Since it can be easily modified by editing tools, there has been a strong need to protect its content for secure multimedia applications. Previous audio authentication algorithms are mainly focused on either human speech or general audio with music as part of the test data, while special research on music authentication has been somewhat neglected. In this article, we propose a novel algorithm to protect the integrity and authenticity of music signals. Its main contributions include the following: (1) Music is segmented into beat-based frames, which not only endows the authentication units with more semantic meaning but also perfectly resolves the challenging synchronization problem. (2) Robust hashes are generated from chroma-based mid-level audio feature which can appropriately characterize the music content and integrated with an encryption procedure to ensure the security against malicious block-wise vector quantization attack. (3) Fuzzy logic is adopted to make the authentication decision in the light of three measures defined on bit errors, coinciding with the inherent blurred nature of authentication. The experiments exhibit good discriminative ability between admissible and malicious operations.

## 1. Introduction

Modern audio editing and processing tools make high-quality forgery pretty easy and convenient. For example, the semantic meaning of audio can be altered by simply reordering or dropping out a few small parts without introducing perceptible artifacts. Thus, judging the authenticity and integrity of audio data by human perception alone is far from enough, tamper detection is increasingly essential to secure audio applications. Traditional data authentication in cryptography does not permit any change of the binary bit stream; this is not suitable for audio data which can be equivalently represented in various formats without perceptible distinction. Therefore, audio authentication which is aimed at effectively protecting the perceptual authenticity and integrity of audio has become an emerging technique in recent years. It ensures that the received audio signal was not maliciously changed by a third party during the course of transmission, that is, the received and the original audio signals are the same in the sense of human auditory perception.

Based on the protection level, audio authentication can be classified into hard and soft authentication [1]. Hard authentication rejects any modifications except lossless compression or format conversion. Soft authentication passes certain incidental or admissible manipulations and rejects all the rest called malicious manipulations. Soft authentication can be further divided into quality-based authentication which rejects any manipulations that lower the perceptual quality below an acceptable level and content-based authentication which rejects any manipulations that change the semantic meaning of the content. Apparently, hard authentication has the minimum distortion endurance, while content-based soft authentication has the maximum capability.

Differentiating acceptable and unacceptable manipulations is the main research challenge in multimedia authentication techniques. In addition, it is dependent on a specific application, namely, an admissible operation in one application might be regarded as unacceptable in another situation. For example, MP3 compression is deemed as a content-preserving operation in most applications, whereas it must be excluded in the production of CD masters in recording studio because any quality loss should be avoided. Different from previous audio authentication algorithms which are specialized to speech or only

* Correspondence: weili-fudan@fudan.edu.cn
School of Computer Science and Technology, Fudan University, Shanghai, China

use music as part of the test data, this research is focused on music authentication, and in this circumstance we classify possible intermediate operations into two categories:

1. The first is content-preserving operations that only change the signal but not the content and typically include standard audio signal processing, such as MP3 compression, filtering, and resampling, and time-domain synchronization manipulations like time-scale modification (TSM) and jittering.
2. The second is malicious operations that substantially change the semantic meaning and commonly include three types of illegal tampering, i.e., cropping, adding, and replacing.

Soft authentication typically measures distortion in some metrics between a feature vector from the dubious signal and that from the original signal, and by comparing with a preset threshold, the final decision is made on the signal's authenticity. It is usually hard to distinguish distortions caused by incidental operations from that caused by malicious manipulations, namely, there is no sharp boundary between authentic and inauthentic signals. This intrinsic fuzziness makes the soft authentication design challenging and ad hoc in most cases [2]. Another bottleneck is how to resist time-domain synchronization distortions, like malicious cropping/adding, and content-preserving jittering and time-scale modifications. Because audio signals must be divided into many frames for the purpose of tamper localization, the above time-domain distortions will bring about ruinous results to most previous authentication algorithms.

In the literature, only a few audio authentication algorithms have been published. Note that algorithms of blind audio forensics summarized in [3,4] are out of the scope of this research. Most algorithms are focused on speech authentication or general audio authentication. In the latter case, some algorithms take music signals as part of the test data. In regard to speech authentication, Wu and Kuo started the earliest work in this field. In [5], they proposed a fragile speech watermarking scheme based on the modified odd/even modulation with exponential scale quantization and a localized frequency masking model. Malicious alterations can be distinguished from content preserving operations like resampling, white noise pollution, and G.711 and G.721 speech coding with very low error probabilities. In [6,7], they developed two robust hashing schemes integrated with CELP and ITU G.723.1 speech coders. Semantic-level speech features including pitch information, changing shape of the vocal tract, and energy envelope are extracted, encrypted, and attached as the header information. The speech signal could go through GSM-AMR speech coder, recompression, amplification, transcoding, resampling, D/A and A/D conversion, and minor white noise pollution without triggering the verification alarm. To gain resynchronization caused by content-preserving operations, a low-cost mechanism based on salient point detection is adopted in [6]. Besides, Jiao et al. designed a word-level robust speech hashing algorithm based on linear spectrum frequencies (LSFs) which can model the vocal tract [8]. Discrete cosine transform (DCT) is introduced to decorrelate the LSFs, and low-frequency DCT coefficients are taken to enhance the discriminative capacity. Owing to these global features, the algorithm is robust against speech transcoding, resampling, noise addition, random cropping, and slight time scaling. Park et al. proposed to detect speech forgery using curve-fitting-based watermark pattern recovery techniques [9]. The watermark pattern will be modified if some changes such as substitution, insertion, and removal have been made to the speech content; therefore, modification and forgery can be measured and detected by pattern recovery. This method uses cyclic pattern embedding to overcome the synchronization problems and enhance the robustness. With respect to general audio authentication, Radhakrishnan and Memon proposed a classical algorithm based on an invariant feature [10]. The core idea is that if two audio signals are perceptually similar, their psychoacoustic masking curves should also resemble each other. Accordingly, this property can be used to differentiate allowed signal processing like MP3 compression from certain malicious operations. Quan and Zhang designed a wavelet packet domain watermarking scheme that decomposes audio signals into subband structure close to the critical bands in psychoacoustic [11]. Not only it can authenticate the integrity but also locate time/frequency tampering. In [12], Steinebach and Dittmann used audio features including the root mean square, zero cross rate, and spectral information of frame-based audio samples to design a content-fragile authentication scheme. The error rates increase with the strength of attacks; accordingly, a threshold-based identification is adopted to differentiate content changes. Zmudzinski and Steinebach used a perception-based robust hash function adapted from the famous Philips audio fingerprint to verify the integrity of audio recordings [13]. Experiments show a high level of distinction between perceptually different audio data and high robustness against content-preserving signal transformations. In [14], Varodayan et al. developed a backward-compatible audio authentication scheme based on distributed source coding, which provides the desired robustness against legitimate encoding variations and at the same time detects illegitimate modifications. The key idea is to provide a Slepian-Wolf-encoded quantized perceptually significant audio projection as authentication data. Valenzise et al. combined compressive sensing and distributed source coding to generate compact hash signature and applied it to audio content protection [15]. Three kinds of tampering, i.e., time-localized tampering, frequency-localized tampering, and time-frequency-localized tampering, are classified and sparse tampering can be reconstructed. In summary, although the above algorithms have obtained certain achievements in different aspects of audio content authentication, they still exhibit some common weakness to be improved. First, audio signals are all segmented into fixed-length frames which may cause serious synchronization problems under cropping, adding, and time stretching; next, adopted features are not suitable enough to characterize the content of music signals; last, all algorithms take a yes/no decision instead of a fuzzy one.

In this paper, we propose a novel content-based soft authentication algorithm for widespread music works. To overcome previous methods' fragility under time-domain synchronization distortions which are mainly caused by the fixed framing of audio signal, we instead adopt a beat tracking method to segment the bit stream. After post-processing, most reserved beat times can be roughly deemed as music edges like drums or onsets which are very important to human auditory perception and have been shown to be rather stable under various distortions. In other words, this is an implicit synchronization method which partitions the time axis into a series of authentication units with each unit bounded by a left and a right music beat. Combined with dynamic time warping (DTW) technique, synchronization between the original and the received music signal is perfectly achieved without loss of precision for tamper localization. By integrating an encryption procedure and chroma feature which is popularly used in music information retrieval to characterize the progression of melody and harmonics, we achieve the secure robust hash against various content-preserving distortions. To avoid the deficiency of previous audio authentication methods that give a definite classification between admissible and malicious operations, herein we perform fuzzy classification on three defined statistics between the original and the dubious hash sequence to make the final decision with an authentication degree.

## 2. System overview

To get an overall idea of this content-based music authentication algorithm, the general framework which is composed of two stages, i.e., the protection stage and the verification stage, is illustrated in Figure 1. In the protection stage, original music signal is first segmented into variable frames by an effective beat tracking algorithm and post-processing; then, chroma features that are commonly used in music analysis to characterize the content are extracted in every beat-based frame; next, they are nonuniformly quantized into binary sequences to form the final robust hash by integrating an encryption procedure; last, concatenated hashes are stored in a trusted authentication center for future use. In the verification stage, the same beat tracking and chroma-based hash calculation are first performed on the input music signal to be authenticated; then, the beat alignment is done to rectify the missegmented beats caused by distortions during the transmission, thus, the extracted hash sequence and its stored counterpart are resynchronized and compared in terms of normalized Hamming distance; finally, fuzzy classification is performed on three

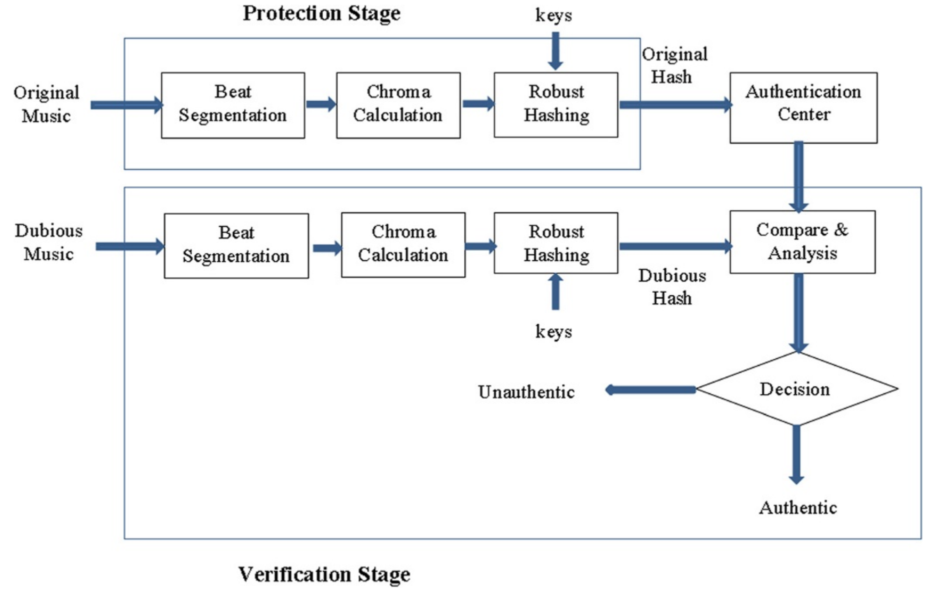

**Figure 1 General framework of the proposed music authentication scheme.**

statistical measures that are defined on the above distance to give the verification result with an authentication degree.

## 3. Protection stage

In this section, we describe the procedure of generating secure robust hashes for the music authenticity protection stage which consists of three steps, i.e., beat-based music segmentation, chroma-based feature extraction, and robust hash generation with security considerations.

### 3.1 Beat-based music segmentation

In order to fulfill the requirement of tamper localization, an authentication algorithm must divide the multimedia signal into many basic authentication units, e.g., frames for audio and blocks for images. Conventional audio authentication schemes usually use fixed-length framing, whereas this kind of signal partition will bring about two major problems. First, it breaks the natural continuity between adjacent audio segments and thus affects the semantic characterization of the content. Second, it will cause serious desynchronization problem due to time-domain distortions. It is known that music signals typically exhibit obvious rhythm. Therefore, fixed-length framing is inappropriate for music authentication; in this research, we instead adopt beat-based framing which is composed of a state-of-the-art beat tracking method and a post-processing module to partition authentication units. In the literature, similar ideas of music segmentation have been used in the scenario of robust watermarking for ownership protection [16,17]. In [16], music segmentation is also based on beat detection, while it uses a different beat tracking algorithm and a different mechanism for resynchronization between the original and the distorted beats. Moreover, it does not have the post-processing procedure. In [17], note onset detection instead of beat tracking is used for music segmentation. Audio feature extracted from a note duration is generally not so robust as within a beat duration, since one beat is normally several times as long as the smallest note and accordingly increase the feature's resistibility against various distortions. As expected, the experiments only show moderate robustness under some audio signal distortions and cropping, while results under other time-domain distortions were not reported.

As stated above, beat-based segmentation not only keeps the inherent relationship of audio samples per frame and hence endows semantic meaning to the generated hashes, but also provides a powerful mechanism to resist time-domain modifications since many beat times are perceptually important music edges and will be kept unchanged or only trivially changed under various distortions. In our implementation, we first resort to an existing beat tracking approach introduced in reference [18] and then perform post-processing to pick out those steadier beat times as frame boundaries. This algorithm includes the following steps:

1. Convert the input audio into an onset strength envelope by taking the first-order difference along time in each subband, throwing away negative values and convolving with a Gaussian envelope about 20 ms wide.
2. Estimate an approximate global tempo by calculating the autocorrelation of the onset strength and searching peaks in perceptual weighting windows. The period with the highest peak is identified as the target tempo.
3. Define an objective function to maximize both the onset strength at every hypothesized beat time and the consistency of the inter-beat interval with the estimated constant tempo.
4. The set of times that optimize the objective function are derived using dynamic programming and chosen as the beat times of the input music, denoted as $C = \{C_\mathrm{i}|\mathrm{i} = 1, 2, \ldots, M\}$, where $M$ is the total number of beats of a whole song.
5. In this step, we perform post-processing to pick out the steadier beats. Specifically, if the energy in a small local region centered at $C_\mathrm{i}$ is less than 1/4 of the average of all beats, then beat $C_\mathrm{i}$ is abandoned. The preserved beats are marked as $B = \{B_\mathrm{i}|\mathrm{i} = 1, 2,\ldots,\mathrm{N}\}$, they are used as frame boundaries and are generally very steady music edges.

Figure 2 is an illustration of the beat-based frames of a 5-s long music and its time-scale modified (−5%) version. It can be seen that most beats under this distortion are not obviously affected and are still able to be mapped to their original position.

### 3.2 Chroma-based feature extraction

In an effort to verify the semantic meaning of music content, selecting suitable features that can characterize the music plays a crucial role. On the basis of beat-based framing, we employ chroma which has been widely used in music content analysis as the key feature to characterize the progression of main melody and harmonics.

Chroma, also called pitch class profile, is a frame-based representation of music signals where the full spectrum is projected into 12 semitone classes included in an octave to reflect the distribution of music notes [19]. Specifically, a 12-dimensional chroma feature of one frame is calculated as below:

$$X_{\mathrm{PCP}}\left(K', n\right) = \sum_{K:P(K)=K'} X_{\mathrm{STFT}}(K, n) \tag{1}$$

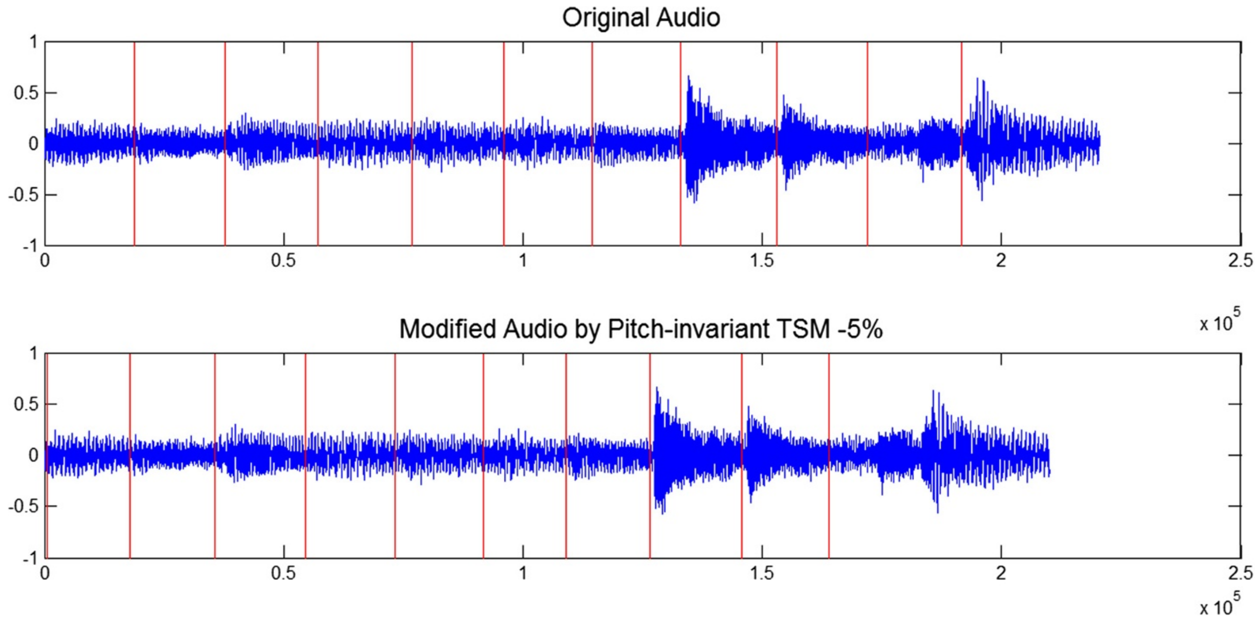

**Figure 2 Beat-based framing for a piece of music and its distorted copy under time-scale modification (−5%).**

where STFT means the short-time Fourier transform, $X_{\mathrm{PCP}}(K', n)$ and $X_{\mathrm{STFT}}(K, n)$ are the chroma feature and the magnitude spectrogram of music signal $x(n)$, respectively, $n$ is the time index and $K$, $K'$ are the frequency indices. The spectral warping between frequency index $K$ in STFT and $K'$ in chroma is described as:

$$P(K) = \mathrm{round}\left\{\left[12 \cdot \log_2\left(\frac{K}{\mathrm{NFFT}} \cdot \frac{f_s}{f_1}\right)\right]\right\} \bmod 12, \tag{2}$$

where NFFT is the FFT length, $f_s$ is the sampling rate, and $f_1$ is the reference frequency corresponding to a note in the standard tuning system.

Our goal is to reduce the music signal in a beat to a chroma-based feature vector. To accomplish this, each beat-based frame (usually several hundreds of ms long) is first subdivided into equal-length non-overlapping subframes (512 samples in our implementation); then a 12-dimensional chroma feature is calculated from each subframe, and all of them in the same beat are averaged to get the final feature vector. In chroma calculation, the frequency range is selected as 64 to 4,096 Hz covering six octaves from note C2 to B7. The reason is twofold. On one hand, most tonal instruments and vocals fall into this frequency band while much percussive noise produced by base drums, cymbals, and snare drums are filtered out accordingly, and the chroma calculation is

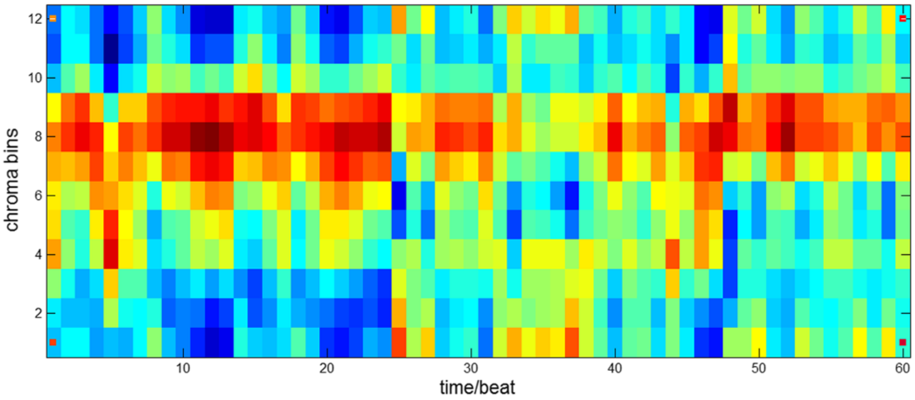

**Figure 3 Chroma-by-beat representation of a 15-s music signal.** Red color means stronger energy, and light blue means weaker energy.

greatly facilitated. On the other hand, middle frequency coefficients are usually less susceptible to various distortions than high-frequency coefficients and hence increase the robustness. Figure 3 shows an example of the chroma-by-beat representation of a 15-s long music signal.

### 3.3 Secured robust hashing

First, the chroma features are normalized so that all the components of a feature vector lie in between 0 and 1. Let $p(i)$ be the normalized chroma vector of the $i$th beat-based frame, non-uniform scalar quantization is then performed to get $\hat{p}(i)$ as below:

$$\hat{p}(i,j) = \begin{cases} floor(p(i,j) \times 10), 0 \leq p(i,j) < 0.7 \\ 7, \qquad 0.7 \leq p(i,j) \leq 1 \end{cases} \qquad j = 1, 2, \ldots, 12, \tag{3}$$

where $\hat{p}(i,j)$ and $p(i, j)$ are the $j$th element of $\hat{p}(i)$ and $p(i)$ respectively, floor($x$) denotes the largest integer less than or equal to $x$. Quantization of the feature values is not only necessary to reduce the data bits but also to increase the feature robustness against small disturbance. Next, each $\hat{p}(i,j)$ is converted from an integer into the form of three binary bits $\hat{p}(i,j) = [b_2 b_1 b_0]_2$, thus $\hat{p}(i)$ comprises 36 bits and is denoted as $h_1(i)$ hereafter.

In order to enhance the security of authentication, we adopt a two-layer encryption mechanism. In the first layer, we perform scrambling to 36 binary bits associated with each beat-based frame. Specifically, according to a secret key $k_1$, a random sequence $R = (r_1, r_2,\ldots, r_{36})$ is generated and rearranged so that $r_{a1} \leq r_{a2} \leq \ldots \leq r_{a36}$. In the light of Equation 4, the encrypted hash code $h_2(i)$ of the $i$th frame is obtained by rearranging the sequence of $h_1(i)$'s elements. Without the correct key, it would be very difficult for an attacker to forge the encrypted data.

$$h_2(i,j) = h_1(i, \alpha_j) \tag{4}$$

In the second layer, we have to avoid the vulnerability under vector quantization attack [20]. If the hash codes are all frame-wise independent, it would be possible for a hacker to make false authentication by substituting some small parts of the original music with other perceptually similar ones. Due to the high repeatability of music signals, this is always possible to be done that in some local regions, the hash values are kept almost unchanged even if the content has been substantially modified. To thwart this attack, an effective way is to make the hash code of one frame dependent not only on itself but also on its neighborhood. In this paper, we associate each beat-based frame with its two direct neighbors. By using another secret key $k_2$ to randomly select 14 bits from each neighbor in terms of Equation 5, a 64-bit chroma-based binary hash $h(i)$ is ultimately formed to represent the $i$th beat-based authentication unit, as follows:

$$h(i,j) = \begin{cases} h_2(i,j), 1 \leq j \leq 36 \\ h_2\left(i-1, c_{j-36}\right), 36 < j \leq 50) \\ h_2\left(i+1, c_{j-50}\right), 50 < j \leq 64 \end{cases}, \tag{5}$$
$$s_{c1} \leq s_{c2} \leq \ldots \leq s_{c14}$$

where the random sequence $S = (s_1, s_2,\ldots,s_{14})$ is generated by key $k_2$. Finally, $h(\cdot)$ for all beats of a music piece and secret keys $k_1$ and $k_2$ are stored in a credible data center for future verification.

## 4. Verification stage

This stage is aimed to verify the authenticity and integrity of a dubious music, namely, to check whether it has been maliciously modified during the transmission. Because time-domain distortions like TSM or jittering may occur before verification, the beat sequence of the susceptible music and that of its original version registered in the authentication center are not guaranteed to be the same. Therefore, beat alignment is firstly performed by virtue of dynamic time warping. Next, the chroma-based robust hash of the received music is calculated and compared with its original data stored in the authentication center. By using fuzzy logic, we calculate the hash difference's confidence belonging to acceptable operations and malicious modifications, respectively, thereby make the final decision of authentication.

### 4.1 Beat alignment

During the course of transmission, the original music signal might experience various acceptable signal distortions or malicious cropping, adding, replacing, etc. Therefore, at the verification end, the received music will not be segmented into exactly the same set of beats as the original ones in most cases. That is, let $B = \{B_i|\mathrm{i} = 1, 2,\ldots,\mathrm{N}\}$ and $\hat{B} = \left\{\hat{B}_j \,|\, \mathrm{j} = 1, 2, \ldots, \mathrm{N}'\right\}$ denote the segmented beat sets of the original and the dubious music, generally speaking $B \neq \hat{B}$. Since chroma-based feature and derived robust hash are calculated in a frame-wise manner, the beat alignment must be first performed to regain synchronization. Differently from [16], where a sophisticated beat normalization procedure composed of identifying the average beat period, locating each beat, and rescaling to the average beat period is used to recover the synchronization before watermark detection, we utilize dynamic time wrapping [21] to resynchronize possibly distorted beats.

It is known that DTW is an effective technique for measuring the similarity between two sequences which may vary in time or speed. Herein, it is applied to find the optimal matching between the original and the distorted beat sequences, using normalized Hamming distance of chroma feature per frame as the similarity metric. Ideally, the beat-pair map will be bijective and move along the main diagonal line of the DTW similarity matrix. However, due to the various acceptable and malicious operations during transmission, it is worth noting that a frame in the test music might be mapped to more than one frame in the original version and vice versa. In other words, some singular points that deviate from the diagonal trajectory will appear. For example, the yellow circle marked in Figure 4b gives an illustration that a specific $\hat{B}_j$ is mapped to both $B_i$ and $B_{i+1}$. In such cases, the average distance between a frame and its multiple mapped ones will be adopted.

### 4.2 Measures for fuzzy authentication

In the literature, most audio authentication algorithms make the final decision by comparing the distance, such as Hamming distance and Euclidean distance, between the hashes of the received and the original audio with a preset threshold. The main flaw of such measures is that they only reflect the global effect of errors while ignore the temporal distribution along the time axis. A malicious tampering and an acceptable signal processing often give rise to pretty much the same errors on the whole, whereas the former errors are generally located in a few small regions and the latter ones are evenly distributed in a much wider range (see Figure 5c,e for illustration). To solve this problem, we first introduce the concepts of possibly modified point (PMP), dense point (DP), and sparse point (SP), then based on which three statistical and temporal measures adapted from similar concepts of [22] in image authentication are redefined to characterize the error distribution and to differentiate acceptable operations from malicious manipulations.

#### *4.2.1 Possibly modified points*

As stated before, each beat-based frame of music signal is deemed as an authentication unit. For the *i*th beat, let $\mathrm{diff}(i) = \frac{1}{64}\sum_{j=1}^{64} h(i,j) \oplus \tilde{h}(i,j)$ be the normalized Hamming distance or bit error rate (BER) between the original hash *h*(i) and the extracted hash $\tilde{h}(i)$, where *j* means the *j*th bit of *h*(i). Then we define a set $D = \{\mathrm{diff}(i) | \mathrm{diff}(i) \in [0, 1], 1 \le i \le \mathrm{N}'\}$ to represent the beat-wise BERs between two hash sequences extracted from the original and the dubious music. A subset of points in *D* are identified as possibly modified points (PMPs) if their values are bigger than a threshold *T*, and their indexes in *D* are recorded in another array *pos*:

$$\mathrm{PMP} = \{\mathrm{diff}(i) | \mathrm{diff}(i) \ge T, 1 \le i \le \mathrm{N}'\} \tag{6}$$

$$\mathrm{POS} = \{\mathrm{pos}(j) | \mathrm{diff}(\mathrm{pos}(j)) = \mathrm{PMP}(j), j = 1, 2, \ldots, \mathrm{PMP}\}. \tag{7}$$

The threshold *T* is set as Equation 8, with acceptable and malicious operations both considered. Since the initial threshold $T_0$ is an experimentally determined value and set as 6/64 = 9.375% ≈ 10%, it can be roughly used to judge if malicious tampering has occurred. It is our observation that in most cases when malicious modifications occur, some locally continuous elements in set *D* are usually much bigger than $T_0$ like in Figure 5e; while when acceptable operations come up, most (if not all) elements are much smaller than $T_0$ and spread in a wide range like in Figure 5c. Therefore, after coarsely classifying the two cases, more appropriate thresholds

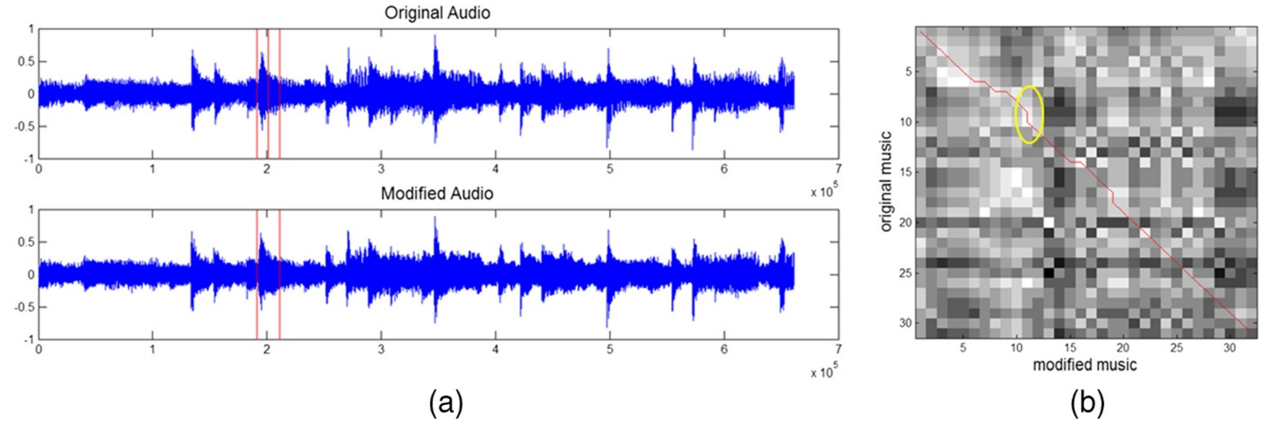

**Figure 4 Beat mapping between the original and the dubious music.** (**a**) One beat in the test music is mapped into two beats of the original music; (**b**) corresponding DTW representation of (**a**).

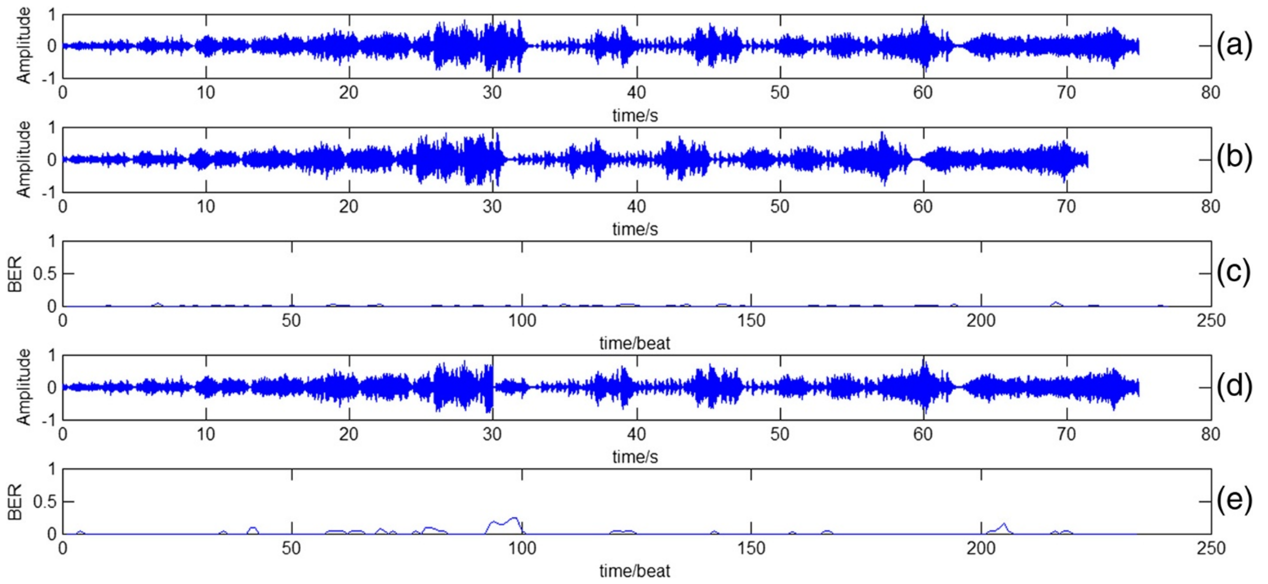

**Figure 5 Beat-wise hash error distribution caused by acceptable and malicious modification along the time axis.** (**a**) Original music. (**b**) Modified music after TSM (−5%). (**c**) Error distribution between hashes of (**a**) and (**b**). (**d**) Tampered music (0:30 ~ 0:32 s is modified). (**e**) Error distribution between hashes of (**a**) and (**d**).

for malicious and admissible operations are defined as below:

$$T = \begin{cases} max(\mathrm{PMP}) \times 0.5, max(\mathrm{PMP}) \geq T_0 \\ median(PMP), \max(\mathrm{PMP}) < T_0 \end{cases} \tag{8}$$

### *4.2.2 Dense points and sparse points*

For a particular point $i$ in PMP, it is defined as a dense point (DP) if at least one of its eight neighbors in the region $N_8(i) = [i - 4, i - 1] \cup [i + 1, i + 4]$ is a PMP. Otherwise, it is called a sparse point (SP).

### *4.2.3 Statistical and temporal measures*

First, based on the above concepts, we herein define three statistical and temporal measures that exhibit distinct properties under admissible and malicious operations for latter authenticity judgment.

#### Average distortion

The average distortion of a dubious music signal is measured by the mean BER of all PMPs.

$$\mathrm{AD} = \frac{1}{L_{\mathrm{pmp}}} \sum_{j=1}^{L_{\mathrm{pmp}}} \mathrm{PMP}(j) \tag{9}$$

where $L_{\mathrm{pmp}}$ = PMP is the length of set PMP. Average distortion describes the degree of modification to the original music's content. Malicious manipulations typically result in larger average distortion (AD), while acceptable operations result in a smaller one.

#### Uniformity degree

The uniformity degree aims at assessing the uniformity of modifications to the original music at the time axis. Let $DIS = \{dis(j)|dis(j) = pos(j+1) - pos(j), j = 1, 2, \ldots, PMP - 1\}$ denote all the beat intervals between every two adjacent PMPs, uniformity degree (UD) which is defined as the standard variance of DIS and calculated as below,

$$\mathrm{UD} = \left[ \frac{1}{N_{\mathrm{dis}}} \sum_{j=1}^{N_{\mathrm{dis}}} \left( \mathrm{dis}(j) - \frac{1}{N_{\mathrm{dis}}} \sum_{j=1}^{N_{\mathrm{dis}}} \mathrm{dis}(j) \right)^2 \right]^{\frac{1}{2}}, \tag{10}$$

where $N_{\mathrm{dis}} = \mathbf{DIS} = \mathbf{PMP} - 1$ is the length of set DIS. Obviously, larger UD indicates uneven distribution of the PMPs which is more likely caused by malicious operations, whereas smaller UD means a relatively even distribution that is more possibly induced by acceptable processing.

#### Maximum connected area size

A connected area is made up of a group of consecutive dense points (DPs), with its size defined as the total number of points included. Of all the connected areas, maximum connected area size (MC) denotes the maximum size. In general, the MC values caused by malicious manipulations are much larger than those caused by acceptable operations, because in the former case affected points tend to tightly concentrate around some local areas while in the latter, circumstance tend to scatteredly spread out on the time axis.

### 4.3 Content authenticity verification

Multimedia authentication is by nature a gradually changed procedure, without an unambiguous boundary between authentic and inauthentic status [2]. Although each of the above three measures exhibits certain potential to differentiate malicious manipulations from acceptable ones, we here combine them together to further reinforce this ability. In accordance with the intrinsic fuzziness of music content authentication, fuzzy classification [23] on the combination is performed to judge whether the received music has been maliciously modified or not. For the purpose of parameter tuning, a small dataset composed of 16 pop songs are collected. For each song, 54 content-preserving operations and 20 malicious modifications are performed. Altogether, 1,184 distorted copies are used for training.

#### *4.3.1 Membership function selection*

In a fuzzy set, a membership degree between 0 and 1 is assigned to each element according to its fitness to certain criterion. The mathematical relationship is modeled by a membership function. On the basis of the above three metrics AD, UD, and MC, it can be qualitatively concluded that when they tend to be small (large), the possibility that acceptable (malicious) modifications have occurred increases. Therefore, we need to choose a suitable membership function for each metric so that given a specific value it can be quantitatively described to what extent it is deemed as small or large.

For the AD, an informal agreement is that it should have an upper bound $\mathrm{th}_2$ of 20%. Namely, if AD is bigger than 0.2, the membership considered as large (small) should be 1 (0). On the other hand, due to unavoidable interference from the environment, a lower bound $\mathrm{th}_1$ which is a little bigger than 0 should also be set. If AD is smaller than $\mathrm{th}_1$, the membership degree considered as small (large) should be 1 (0). Besides, when AD is between $\mathrm{th}_1$ and $\mathrm{th}_2$, the membership degree is approximately linearly changed based on our observation. Therefore, in accordance with these requirements, a trapezoidal membership function is chosen to model AD as shown in Equations 11 and 12):

$$X_{\mathrm{As}}(\mathrm{AD}) = \begin{cases} 1, 0 \le AD \le \mathrm{th}_1 \\ \dfrac{1}{\mathrm{th}_1 - \mathrm{th}_2}(\mathrm{AD} - \mathrm{th}_2), \mathrm{th}_1 < AD \le \mathrm{th}_2 \\ 0, AD > \mathrm{th}_2 \end{cases} \tag{11}$$

$$X_{\mathrm{Al}}(\mathrm{AD}) = \begin{cases} 0, 0 \le AD \le \mathrm{th}_1 \\ \dfrac{1}{\mathrm{th}_2 - \mathrm{th}_1}(\mathrm{AD} - \mathrm{th}_1), \quad \mathrm{th}_1 < AD \le \mathrm{th}_2 \, , \\ 1, AD > \mathrm{th}_2 \end{cases} \tag{12}$$

where $X_{\mathrm{As}}(\mathrm{AD})$ and $X_{\mathrm{Al}}(\mathrm{AD})$ are, respectively, the membership degree that AD is deemed as small or large. The parameter $\mathrm{th}_1$ defines the threshold below which AD is completely small and $\mathrm{th}_2$ defines the threshold above which AD is completely large. In our experiment, they are set to 0.04 and 0.2, respectively. The shapes of these two membership functions are shown in the top subgraph of Figure 6.

In regard to UD, the increase of its value monotonically makes its membership degree of being large (small) go up (down) without absolute upper bound and lower bound. Therefore, using conventional sigmoidal membership function of Equations 13 and 14) to depict UD will be an appropriate choice.

$$X_{\mathrm{Us}}(\mathrm{UD}) = 1 - \frac{1}{1 + e^{-\alpha(\mathrm{UD} - \beta)}} \tag{13}$$

$$X_{\mathrm{Ul}}(\mathrm{UD}) = \frac{1}{1 + e^{-\alpha(\mathrm{UD} - \beta)}} \tag{14}$$

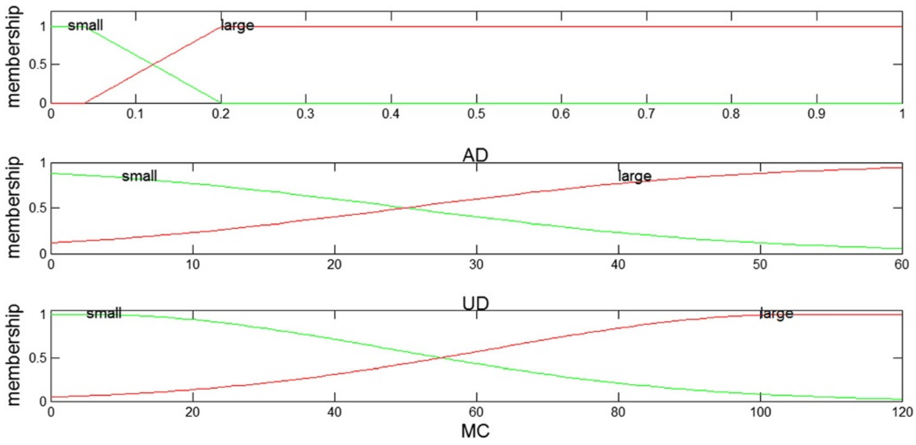

**Figure 6 Membership functions.** Membership functions for the average distortion (AD), the uniformity degree (UD), and the maximum connected area size (MC).

**Table 1 Eight fuzzy classes combined from the three measures**

| Class | Measure | | | Class | Measure | | |
|---|---|---|---|---|---|---|---|
| | AD, UD, MC | | | | AD, UD, MC | | |
| $C_1$ | As | Us | Ms | $C_5$ | Al | Us | Ms |
| $C_2$ | As | Us | Ml | $C_6$ | Al | Us | Ml |
| $C_3$ | As | Ul | Ms | $C_7$ | Al | Ul | Ms |
| $C_4$ | As | Ul | Ml | $C_8$ | Al | Ul | Ml |

As, Us, and Ms mean that AD, UD, and MC are small; Al, Ul, and Ml mean that AD, UD, and MC are large.

where $X_{\mathrm{Us}}(\mathrm{UD})$ and $X_{\mathrm{Ul}}(\mathrm{UD})$ mean the membership degree that UD is small or large individually. $\beta$ is an average UD value acquired from a set of training signals modified by both acceptable and malicious manipulations and $\alpha$ controls the changing speed especially at the point UD = $\beta$, they are experimentally set to 25 and 0.08 in our implementation. Shapes of these two membership functions are shown in the middle subgraph of Figure 6.

With respect to *MC*, we observed from experiment that when it increases (decreases), the membership degree of being large (small) also becomes larger continuously and smoothly. Since this is a gradually changed procedure, we select commonly used Gaussian membership function defined in Equations 15 and 16 to model MC.

$$X_{\mathrm{Ms}}(\mathrm{MC}) = \begin{cases} 1, MC \le \mu_1 \\ e^{-\frac{(\mathrm{MC}-\mu_1)^2}{2\sigma^2}}, MC > \mu_1 \end{cases} \qquad (15)$$

$$X_{\mathrm{Ml}}(\mathrm{MC}) = \begin{cases} 1, MC \ge \mu_2 \\ e^{-\frac{(\mathrm{MC}-\mu_2)^2}{2\sigma^2}}, MC < \mu_2 \end{cases}, \qquad (16)$$

where $X_{\mathrm{Ms}}(\mathrm{MC})$ and $X_{\mathrm{Ml}}(\mathrm{MC})$ are the membership degrees that MC is small or large respectively, with $\mu_1$ = 5, $\mu_2$ = 105, and $\sigma^2 = (\mu_2 - \mu_1)^2/8\ln 2$ in experiment. The shapes of these two membership functions are shown in the bottom subgraph of Figure 6.

### *4.3.2 Authenticity verification*

After introducing fuzziness to the three measures as above, a specific value is no longer definitely treated as large or small but simultaneously belongs to both states with distinct membership degree. The combination of the three fuzzy measures falls into eight fuzzy classes listed in Table 1. Given an arbitrary measure vector $m$ = ($m$1, $m$2, $m$3) = (AD, UD, MC), its membership degree pertaining to a particular class $C_i$ is calculated as follows, according to the theory of fuzzy classification:

$$X_{C_i}(m) = \sum_{j=1}^{3} w_j X_{C_{ij}}(m_j), i = 1, 2, \ldots, 8, \qquad (17)$$

where $X_{\mathrm{C}i}(m)$ means the membership degree of $m$ belonging to class $\mathrm{C}_i$, $\mathrm{X}_{\mathrm{C}ij}(m_j)$ is the membership degree that $m_j$ fits the status denoted as $\mathrm{C}_{ij}$ according to Equations 11 to 16), and $w$ = [0.3, 0.45, 0.25] is an empirical weight vector that describes the relative significance of each measure. Experiments show that $X_{C_1}(m) > X_{C_2}(m) > \ldots > X_{C_8}(m)$ for acceptable manipulations and, on the contrary, $X_{C_1}(m) < X_{C_2}(m) < \ldots < X_{C_8}(m)$ for malicious ones. In light of such regularity, the degree of authenticity ($D_y$) and that of inauthenticity ($D_n$) are derived as follows:

$$D_y = \sum_{i=1}^{8} w_{y_i} X_{C_i}(m), w_y = [1, 0.9, 0.5, 0.2, 0.2, 0.1, 0, 0] \qquad (18)$$

$$D_n = \sum_{i=1}^{8} w_{n_i} X_{C_i}(m), w_n = [0, 0, 0.5, 0.8, 0.8, 0.8, 0.9, 1], \qquad (19)$$

where $w_y$ and $w_n$ are experimentally determined weight vectors assigned to each class depending on its contribution to the authentication result. The final decision measure and rules are defined in Equations 20 and 21. If *authRatio* > 1, the dubious music is judged as authentic and otherwise inauthentic. Note that *authRatio* is also a measure with fuzziness. Namely, in case 1 bigger *authRatio* means higher confidence to the authenticity, while in case 2 smaller *authRatio* brings more reliability to inauthenticity.

**Table 2 Average authentication results under acceptable audio manipulations**

| Manipulation | authRatio | Result | Manipulation (kbps) | authRatio | Result |
|---|---|---|---|---|---|
| LP (8 kHz) | 1.4584 | ✓ | MP3 (128) | 1.4563 | ✓ |
| LP (4 kHz) | 1.4496 | ✓ | MP3 (96) | 1.4550 | ✓ |
| RS (22,050) | 1.4023 | ✓ | MP3 (64) | 1.4506 | ✓ |
| RS (16,000) | 1.3876 | ✓ | MP3 (48) | 1.4450 | ✓ |
| - | - | - | MP3 (32) | 1.4396 | ✓ |

LP means low-pass filtering; RS means resampling.

**Table 3 Average authentication results under time-domain desynchronization distortions**

| Manipulation | authRatio | Result | Manipulation | authRatio | Result |
|---|---|---|---|---|---|
| TSM (−1%) | 1.4588 | ✓ | TSM (+1%) | 1.4640 | ✓ |
| TSM (−2%) | 1.4505 | ✓ | TSM (+2%) | 1.4645 | ✓ |
| TSM (−3%) | 1.4468 | ✓ | TSM (+3%) | 1.4578 | ✓ |
| TSM (−4%) | 1.4398 | ✓ | TSM (+4%) | 1.4524 | ✓ |
| TSM (−5%) | 1.4460 | ✓ | TSM (+5%) | 1.4482 | ✓ |
| TSM (−6%) | 1.4354 | ✓ | TSM (+6%) | 1.4465 | ✓ |
| TSM (−7%) | 1.4268 | ✓ | TSM (+7%) | 1.4324 | ✓ |
| TSM (−8%) | 1.4324 | ✓ | TSM (+8%) | 1.4285 | ✓ |
| TSM (−9%) | 1.4205 | ✓ | TSM (+9%) | 1.4263 | ✓ |
| TSM (−10%) | 1.4154 | ✓ | TSM (+10%) | 1.4186 | ✓ |
| TSM (−11%) | 1.4106 | ✓ | TSM (+11%) | 1.4120 | ✓ |
| TSM (−12%) | 1.4045 | ✓ | TSM (+12%) | 1.4082 | ✓ |
| TSM (−13%) | 1.3940 | ✓ | TSM (+13%) | 1.4002 | ✓ |
| TSM (−14%) | 1.3768 | ✓ | TSM (+14%) | 1.3967 | ✓ |
| TSM (−15%) | 1.3428 | ✓ | TSM (+15%) | 1.3658 | ✓ |
| TSM (−16%) | 1.3198 | ✓ | TSM (+16%) | 1.3218 | ✓ |
| TSM (−17%) | 1.2956 | ✓ | TSM (+17%) | 1.3145 | ✓ |
| TSM (−18%) | 1.3014 | ✓ | TSM (+18%) | 1.3025 | ✓ |
| TSM (−19%) | 1.2543 | ✓ | TSM (+19%) | 1.2845 | ✓ |
| TSM (−20%) | 1.2745 | ✓ | TSM (+20%) | 1.2649 | ✓ |
| Jittering (1/2,000) | 1.4527 | ✓ | Jittering (1/500) | 1.4393 | ✓ |
| Jittering (1/1,500) | 1.4546 | ✓ | Jittering (1/100) | 1.4420 | ✓ |
| Jittering (1/1,000) | 1.4488 | ✓ | - | - | - |

TSM means time-scale modification.

$$\mathrm{authRatio} = \frac{D_y}{D_n} \tag{20}$$

$$\begin{cases} Case\ 1 : authRatio > 1 \Rightarrow authentication\, passed \\ Case\ 2 : authRatio \leq 1 \Rightarrow authentication\ failed \end{cases}. \tag{21}$$

### 4.4 Tamper localization

If a music signal is judged as inauthentic, all authentication units in the set of $\hat{B} = \{\hat{B}_j | \mathrm{diff}(j) \geq T,\ \ \mathrm{j} = 1, 2, \ldots, \mathrm{N}'\}$ are marked as tampered regions. Remember that in the procedure of robust hashing, each beat is associated with its two near neighbors. In light of Equation 5, the tampered regions are located in the current beat $\hat{B}_j$ only when the different bits between the original hash $h$(i) and the extracted hash $\tilde{h}(i)$ are within the first 36 bits; otherwise, the possibly tampered regions will actually be extended to $\tilde{B}_{j-1}$ and/or $\tilde{B}_{j+1}$, i.e., three beats (generally 1 ~ 2 s) in the worst case.

## 5. Experimental results

In this section, we perform robustness and fragility experiments to investigate this algorithm's capability of differentiating admissible operations from malicious modifications. The test dataset is composed of 344 Chinese popular songs (the 16 songs for training are not included), and their 25,456 legitimately and maliciously modified copies are used for testing the authentication performance. Each music piece is WAVE format, 2 to 5 min long, 44.1 kHz sampled, 16 bits/sample quantized, and monophonic. The audio editing and manipulating tools are Adobe Audition (Adobe Systems Inc., San Jose, CA, USA) and Gold Wave (GoldWave Inc., St. John's, Newfoundland and Labrador, Canada).

### 5.1 Authentication tests and false statistics

As this algorithm is aimed at music content-based soft authentication, we first check its authentication results

**Table 4 Average authentication results under malicious manipulations**

| Manipulation | authRatio | Result |
|---|---|---|
| Cropping | 0.8598 | x |
| Adding | 0.8402 | x |
| Replacing | 0.8554 | x |

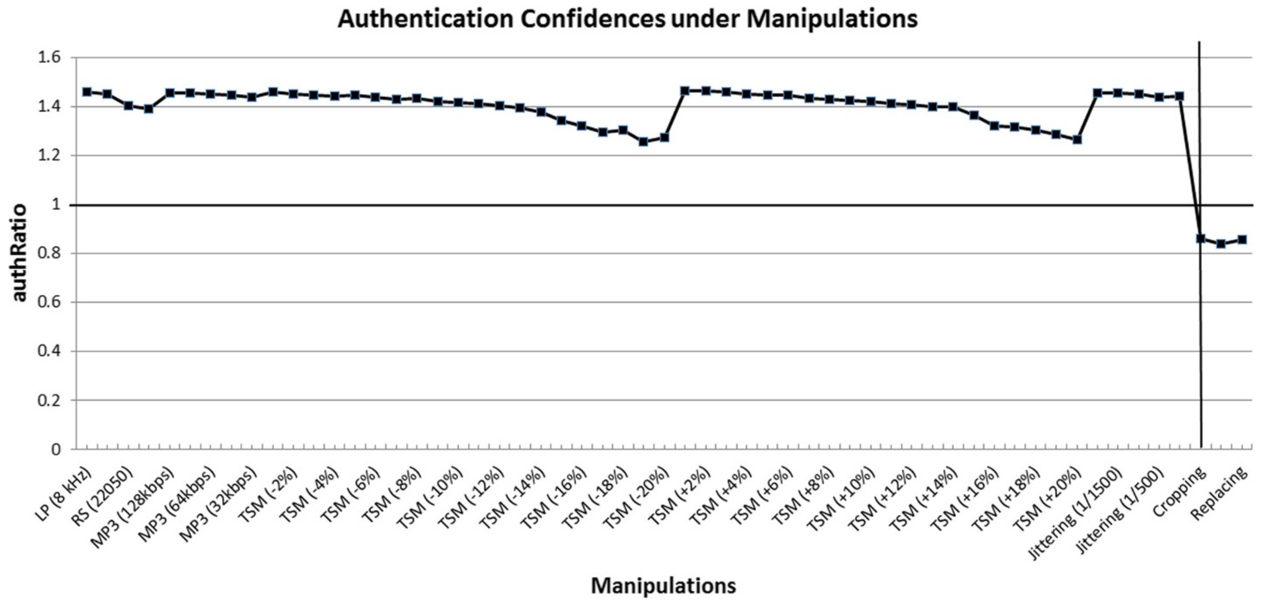

**Figure 7 Average authentication results under acceptable and malicious manipulations.**

under various acceptable audio operations. As stated above, *authRatio* is adopted as the authentication measure in accordance with the aforementioned fuzzy classification methodology. If it is bigger than 1, the algorithm is said to be able to sustain certain admissible operations. The average results performed on the above test dataset are summarized in Tables 2 and 3. It can be seen that in virtue of the power of beat segmentation/alignment and invariant chroma features, this algorithm is robust enough under common content-preserving distortions, like MP3 lossy compression, resampling, and low-pass filtering, and time-domain desynchronization distortions like time-scale modification and jittering. In most cases under the above admissible manipulations, authentication are correctly passed with average *authRatio*s higher than 1.2543, namely, only the perceptual quality is degraded and the semantic meaning is preserved.

Specifically, for MP3 lossy compression, resampling, low-pass filtering, and jittering, the authentication confidences are rather high (slightly fluctuating around 1.4) and stable, which means that these operations can be correctly classified as admissible so that the authentication successfully passes with high confidence. With regard to time-scale modifications, the authentication confidences decline from around 1.46 to 1.25 when scaled from 1% to 20%. It shows that slight TSMs are definitely deemed as admissible so that the authentication passes with high confidence, while more serious TSMs gradually move towards the boundary between admissible and malicious with smaller and smaller confidences. This phenomenon verifies the fuzzy nature of audio authentication, namely, it is a gradually changed procedure rather than a sharp transition between legitimate and malicious modifications.

To test the fragility of this algorithm under malicious operations, we investigate the performance under three typical content-changing manipulations, i.e., cropping, adding, and replacing. The authentication results performed on the test dataset are averaged and shown in Table 4; the most malicious modifications are correctly judged as inauthentic with average *authRatio* lower than 0.86. The overall results are combined together and illustrated in Figure 7. It can be clearly seen that on average, *authRatio*s are above 1 for acceptable manipulations and below 1 for malicious operations.

In a practical authentication system, two important false statistics must be taken into consideration. The first is the false positive rate, which is the rate of considering a music signal as authentic when it has been maliciously modified. The second is the false negative rate, which is defined as the rate of judging a piece of music as inauthentic when it has actually undergone content-preserving operations. Below, we adopt confusion matrix to demonstrate the overall system performance (see Table 5). In the 18,576 admissible operations, 82 of them are falsely judged as malicious so the authentication fails,

**Table 5 Confusion matrix of the false statistics**

| **Confusion matrix** | | **Predicted** | |
|---|---|---|---|
| | | **Malicious** | **Admissible** |
| Actual | Malicious | *a* = 6,854 | *b* = 26 |
| | Admissible | *c* = 82 | *d* = 18,494 |

thus the false negative rate is 0.0044; in the 6,880 malicious modifications, 26 of them are falsely judged as admissible so the authentication passes, therefore, the false positive rate is 0.0038. It can be seen from the matrix that the authentication system is able to make distinction between admissible operations and malicious modifications pretty well.

At present, it is difficult to quantitatively compare this algorithm with other audio authentication methods. One reason is that different algorithms use different test datasets and evaluation measures. The other is that since authentication experiments are indeed a rather subjective test, malicious tampering that might occur in reality are inexhaustible and can only be exemplified in an article.

## Conclusions

In this paper, we propose an algorithm on music content authentication which has been somewhat ignored by the research community. By integrating beat-based segmentation, mid-level chroma feature, and fuzzy authentication, we obtain high robustness against acceptable operations and fragility under malicious modifications at the same time. Results are given in the form of authenticity degree to fit the intrinsic fuzzy nature. Overall, beat-based segmentation is a radical step for music authentication. Therefore, the proposed method is only suitable for music genres with perceptible rhythm, e.g., pop and rock, but does not work with classical music. A more precise beat mapping mechanism has to be designed in the future. This will not only further improve the robustness under admissible operations and the classification precision of malicious modifications but also be a solution for fragment authentication of audio that has never been scarcely touched in the research community.

### Competing interests

The authors declare that they have no competing interests.

### Authors' information

WL received the Ph. D. 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 more than 30 refereed papers so far, including international leading journals and key conferences, such as *IEEE Transactions on Multimedia*, *Computer Music Journal*, IWDW, ACM SIGIR, and *ACM Multimedia*. He is a reviewer for international journals like *IEEE Transactions on Signal Processing*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Audio, Speech & Language Processing*, *IEEE Transactions on Inform,ation Forensics and Security*, *Signal Processing*, and conferences, such as ICME, ACM MM, and IEEE Globalcom.

### Acknowledgment

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

Received: 11 March 2013 Accepted: 22 May 2013
Published: 18 June 2013

### References

1. BB Zhu, MD Swanson, AH Tewfik, When seeing isn't believing. IEEE Sig. Proc. Mag. **21**(2), 40–49 (2004)
2. CW Wu, On the design of content-based multimedia authentication systems. IEEE T. Multimedia **4**(3), 385–393 (2002)
3. S Gupta, S Cho, CC Kuo, Current developments and future trends in audio authentication. IEEE. Multimedia **19**(1), 50–59 (2012)
4. DPN Rodriguez, JA Apolinrio, LWP Biscainho, Audio authenticity: detecting ENF discontinuity with high precision phase analysis. IEEE T. In. For. Security **5**(3), 534–543 (2010)
5. CP Wu, *CC Kuo, Fragile speech watermarking based on exponential scale quantization for tamper detection, in Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)* (IEEE, Orlando, 2002), pp. 3305–3308
6. CP Wu, CC Kuo, Speech content authentication integrated with CELP speech codes, in *IEEE International Conference on Multimedia and Expo (ICME)* (Tokyo, 2001)
7. CP Wu, CC Kuo, IEEE International Conference on Information Technology, *Speech content integrity verification integrated with ITU G.723.1 speech coding* (Coding and Computing (ITCC), IEEE, Las Vegas, 2001, 2001), pp. 680–684
8. YH Jiao, LP Ji, XM Niu, Robust speech hashing for content authentication. IEEE Signal Proc. Let. **16**(9), 818–821 (2009)
9. CM Park, D Thapa, GN Wang, Speech authentication system using digital watermarking and pattern recovery. Pattern Recogn. Lett. **28**(8), 931–938 (2007)
10. R Radhakrishnan, N Memon, Audio content authentication based on psycho-acoustic model. Proc. SPIE. **4675**, 110–117 (2002)
11. X Quan, H Zhang, *Perceptual criterion based fragile audio watermarking using adaptive wavelet packets, in Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, vol. 2 (IEEE, Cambridge, 2004), pp. 867–870
12. M Steinebach, J Dittmann, Watermarking-based digital audio data authentication. EURASIP J. Appl. Signal Proc. **10**, 1001–1015 (2003)
13. S Zmudzinski, M Steinebach, *Perception-based audio authentication watermarking in the time-frequency domain, in Proceedings of the International Workshop on Information Hiding (IH)* (Springer, Berlin, Heidelberg, 2009), pp. 146–160
14. D Varodayan, *YC Lin, B Girod, Audio authentication based on distributed source coding, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Piscataway, IEEE, 2008), pp. 225–228
15. G Valenzise, G Prandi, M Tagliasacchi, A Sarti, Identification of sparse audio tampering using distributed source coding and compressive sensing techniques. EURASIP J. Image. Vid. Proc. **2009**, 1–12 (2009)
16. H Attias, D Kirovski, *Audio watermark robustness to desynchronization via beat detection, in Proceedings of the International Workshop on Information Hiding (IH)* (Springer, Berlin, Heidelberg, 2002), pp. 160–175
17. C Xu, N Maddage, X Shao, Q Tian, Content-adaptive digital music watermarking based on music structure analysis. ACM Trans. Multimed. Comput. Commun. Appl. **3**(1), 1–16 (2007)
18. DPW Ellis, Beat tracking by dynamic programming. J. New Mus Res. **36**, 51–60 (2007)
19. DPW Ellis, GE Poliner, *Identifying cover songs with chroma features and dynamic programming beat tracking, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4 (Piscataway, IEEE, 2007). p. 1429–1432
20. M Holliman, N Memon, Counterfeiting attacks on oblivious blockwise independent invisible watermarking schemes. IEEE Trans. Image Process. **9**(3), 432–441 (2000)
21. V Ewert, M Muller, *High resolution audio synchronization using chroma onset features, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Piscataway, IEEE, 2009), pp. 1869–1872
22. S Ye, Q Sun, E Chang, Statistics- and spatiality-based feature distance measure for error resilient image authentication. LNCS Tran. Mul. Sec. **4499**, 48–67 (2007)
23. M Friedman, A Kandel, *Introduction to Pattern Recognition - Statistical, Structural, Neural and Fuzzy Logic Approaches* (World Scientific, London, 1999)

doi:10.1186/1687-4722-2013-11
**Cite this article as:** Li *et al.*: Music content authentication based on beat segmentation and fuzzy classification. *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:11.