Multimedia Tools and Applications (2023) 82:18599–18616
https://doi.org/10.1007/s11042-022-14197-w

A novel audio watermarking algorithm robust against recapturing attacks

Junjie He[1] · Zhenghui Liu[2,3,6] · Kejia Lin[4] · Qing Qian[5]

Received: 13 June 2022 / Revised: 13 September 2022 / Accepted: 27 October 2022 /
Published online: 23 November 2022

Abstract
Digital watermarking is a promising technology used in copyright protection for digital audio. However, in audio watermarking, it is still a challenge to achieve robustness against recapturing attacks, although much effort has been made in recent years. In this paper, a novel audio watermarking based on the frequency domain power spectrum (FDPS) feature is proposed to resist recapturing attacks. We divide a long audio signal into segments and split each segment into fragments. We embed synchronization codes in some fragments by quantifying the feature generated by discrete wavelet transform (DWT) approximate coefficients and copyright information in other fragments based on the FDPS feature. For the incoming audio, we extract synchronization codes to locate the watermarked fragments, further to extract copyright information. Experimental results demonstrate that the proposed method performs well under recapturing attacks and the bit error ratio (BER) value is decreased by more than 13% by comparing with the state-of-the-art methods.

Keywords Audio watermarking · Copyright protection · Robust feature · Recapturing attack · De-synchronization attacks

✉ Zhenghui Liu
zhenghui.liu@163.com

[1] School of Mathematics and Statistics, Xinyang Normal University, Xinyang 464000, China

[2] Guangdong Key Laboratory of Intelligent Information Processing and the Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China

[3] Guangdong Provincial Key Laboratory of Information Security Technology, Guangzhou 510275, China

[4] Xinyang Normal University Library, Xinyang 464000, China

[5] Guizhou University of Finance and Economics, Guiyang 550025, China

[6] Present address: School of Computer and Information Technology, Xinyang Normal University, Xinyang, China

1 Introduction

Digital watermarking is one of the assuring techniques to protect the copyright of digital multimedia data, such as digital audio, digital image and digital video. The method is to embed copyright information into a multimedia object. When necessary, copyright owners or authorized one can extract the embedded information to prove the ownership of the multimedia data in hand.

In recently, with the popularization of recording devices, recapturing has become a common method for people to get audio content. On the one hand, recapturing attacks can capture the meaning of watermarked audio. On the other hand, the attacks can cause devastating damage to the embedded information. To this day, recapturing attack is still a challenging issue for audio watermarking [5, 10, 18].

Many audio watermarking methods have been proposed over the past twenty years. The methods can be generally categorized into time domain [3, 4, 20] and transform domain methods [8, 12, 15]. In general, the transform domain methods have better robustness than the time domain methods. Therefore, transform domain methods are often used for copyright protection. The common transform domain methods are based on the techniques such as spread spectrum (SS) [11], patchwork [24].

Although many audio watermarking methods are robust to some attacks, it is difficult to extract accurate watermark information from recaptured signals. It is indeed hard to design an audio watermarking scheme robust against recapturing attacks. Figure 1 shows the rough process of recapturing attack. First, digital watermarked audio is converted into an analog signal by D/A conversion and played by the speaker. Second, the sound waves produced by speakers travel through the air, which will be disturbed by environmental noise and other disturbances. Third, recording equipment convert the received analog signal to digital signal and the compress it to get the recaptured vision. In short, the recapturing attack can result in a larger change in the time and frequency domains of the watermarked audio, which is a destructive attack to the embedded watermark. So it is difficult to extract the correct watermark bits from attacked signal.

In order to solve the above problems, we proposed the novel audio watermarking algorithm robust against recapturing attacks. In this paper, we divide a long audio signal into segments and split each segment into fragments. We embed synchronization codes in selected fragments in the discrete wavelet transform (DWT) domain to reduce computational load occurring in searching synchronization codes. We extract the feature frequency domain power spectrum (FDPS) from other fragments and embed copyright information by quantifying the FDPS feature. For the incoming audio, we cut the signal into many fragments and search for synchronization codes, to find the fragments with copyright information. After that we extract the copyright information. Experimental results show the superior

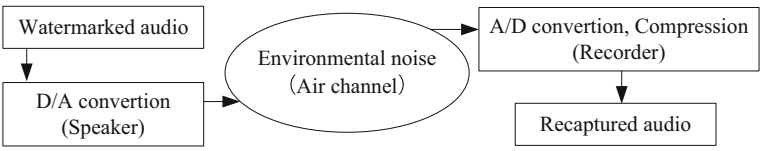

Fig. 1 The process of recapturing attack

performance of our proposed method. The main contributions of our proposed method are summarized as follows:

- The proposed FDPS feature and the embedding method are novel in audio watermarking. Compared to the existing methods, the FDPS feature has stronger resistance against recapturing attacks;
- The proposed novel synchronization codes embedding method can improve the robustness under the premise of not reducing the search efficiency.

The rest of the paper is organized as follows. Section 2 discusses the existing methods and their drawbacks. In Section 3, we present the fundamental theory of the scheme. In Section 4, we propose our watermarking scheme, including synchronization codes and copyright information embedding and extraction strategies. In Section 5, we discuss the results of the performance evaluation of the proposed method and further comparisons with other recently developed methods. Conclusions are given in Section 6.

2 Related work

In this section, popular audio watermarking methods are briefly discussed.

2.1 Patchwork based methods

In patchwork based methods, there are two groups, referred to as "patches". The watermark bits are embedded into the patches by modifying the feature extracted. Patchwork based methods have shown higher robustness against attacks.

Yeo et al. proposed a modified patchwork algorithm by using the enhanced version of the conventional patchwork algorithm [24], in which the transform domain used contains not only discrete cosine transform (DCT), but also discrete fourier transform (DFT) and discrete wavelet transform (DWT). In [7], based on the method called full index embedding, authors presented a patchwork algorithm for audio watermarking, in which the watermark embedding strength was psychoacoustically adapted. Regrettably, performance of the method relies on the assumption that the chosen patches have the same statistical property. Kalantari et al. [6] selected two patches having comparable statistical characteristics for watermark embedding and presented a multiplicative patchwork watermarking method. The audio segment suitable for watermark embedding must meet certain conditions, which limits the embedding capacity of watermarking methods.

In order to improve the performance of patchwork based methods, Natgunanathan et al. [16] proposed a robust audio watermarking method. Authors cut the host audio into segments and divided each segment into tow sub-segments. They performed DCT on sub-segments and partitioned the DCT coefficients into a number of frame pairs according to a specified frequency region. Then, by a selection criterion, the suitable DCT frame pairs were selected for embedding. By comparing with some existing patchwork watermarking methods, the method did not require information of which frame pairs of the watermarked audio signal enclose watermarks and improved the robustness to conventional attacks. While de-synchronization attacks can disrupt the location of watermark and cause a considerable number of false watermarks being extracted for the method. Xiang et al. [21] presented a patchwork

based watermarking method to improve the robustness against de-synchronization attacks, in which synchronization codes were embedded in the logarithmic DCT (LDCT) domain of host audio. For the incoming signal being scaled, the method used the location information of the synchronization codes to find the scaling factor and then re-scaled the incoming signal to remove the scaling factor. The watermark was extracted from the modified version of the incoming signal. Natgunanathan et al. [17] presented a multilayer patchwork audio watermarking method, in which watermark was embedded to host signal repeatedly in an overlaying manner. In this paper, the watermark embedding algorithm was designed to ensure that the embedded watermarks in a certain layer did not affect the detection of watermarks in other layers. With an embedding error buffer, the method can withstand a wide range of common attacks. While, the added multiple layers of watermark bits inevitably reduced the perceptual quality of the method.

The most recent watermarking method was proposed in [25], in which the watermarking strategy is by unitizing the patchwork technique. One major problem with this method is that its performance against de-synchronization attacks is limited.

2.2 SS based methods

Watermark bit is embedded into a host audio via a spreading sequence (SS), the method of which is called SS based watermarking method.

Malvar et al. proposed a improved spread spectrum (ISS) based audio watermarking method [14], which achieved roughly the same noise robustness gain as quantization index modulation but without the amplitude scale sensitivity of QIM. For improving the embedding capacity, Xiang et al. proposed a new SS-based audio watermarking method [22], through a set of mechanisms: embedding multiple watermark bits in one audio segment, reducing host signal interference on watermark extraction, and adaptively adjusting PN sequence amplitude in watermark embedding based on the property of audio segments.

Although, the SS based method is robust against some attacks, it is vulnerable to de-synchronization attacks.

2.3 Other methods

In addition to the above methods shown in Sections 2.1 and 2.2, there are other methods designed to verify the authenticity of audio content. Based on Bessel-Fourier moments, Liu et al. [12] proposed a speech content authentication, which has the ability of tamper location for maliciously attacks. In order to verify the authenticity of compressed audio recordings, Korycki proposed an authentication scheme using for the detection of multiple compression and encoder's identification [9]. In [9], The compressed digital audio recordings are authenticated by evaluation of statistical features extracted from MDCT coefficients and other parameters obtained from compressed audio files, which are used for training selected machine learning algorithms. Chen et al., based on compression technique and codebook-excited linear prediction, proposed an authentication scheme for compressed speech signal [1]. Watermark bits are generated by the features extracted during compression process based on codebook-excite linear prediction and embedded based on lest significant bits (LSB). In the method, signal processing is considered a hostile attack, for the watermarking Strategy based on LSB is fragile.

3 Fundamental theory

The approximate way of the proposed scheme is to segment a long audio signal into some sections and then to embed one bit of synchronization code or copyright information into one section. Based on the time-frequency localization capability of DWT, we embed synchronization code into the approximate DWT coefficients of one section, to reduce computational load occurring in searching synchronization codes. In the following, we brief introduce the DWT.

3.1 Discrete wavelet transform

Discrete wavelet transform can be viewed as the multiresolution decomposition of a sequence [19]. It takes a length J sequence u (j), and generates an output sequence of length N. The output is the multiresolution representation of u (j). It has $J/2$ values at the highest resolution, $J/4$ values at the next resolution, and so on.

The structure of the DWT is due to the dyadic nature of its time-scale grid, shown in Fig. 2, in which AC_q and DC_q represent the approximate and detail coefficients from q-level DWT, respectively. The length of AC_q is $J/2^q$, equal to DC_q.

There are several advantages for applying DWT to audio watermarking. 1) DWT is known to have the time-frequency localization capability. And this characteristic can be used to improve computational efficiency greatly in searching synchronization codes. 2) Variable decomposition levels are available. 3) DWT itself needs a lower computation load compared with DCT and DFT [19].

3.2 The feature used for embedding

3.2.1 The feature definition

We denote the L length original audio as $A = \{a(l),\ 1 \le l \le L\}$, where $a(l)$ denotes the l-th sample. Then we perform DCT on the audio A and get the DCT coefficients, denoted by D, $D = \{d(l),\ 1 \le l \le L\}$. The audio feature frequency domain power spectrum (FDPS) can be defined by the Eq. (1).

$$F = \sqrt{\frac{\sum_{l=1}^{L} log_2\left(|d(l)|^2 + \lambda\right)}{L}} \tag{1}$$

where $d(l)$ denotes the l-th DCT coefficient, $\lambda > 0$.In this paper, we set $\lambda = 1.02$.

In this paper, we propose the embedding method by quantifying the FDPS feature. Denote QA as the L length watermarked audio, $QA = \{qa(l),\ 1 \le l \le L\}$, $qa(l)$ denotes the l-th sample. In the following, we give how to get the corresponding DCT coefficients from the quantified FDPS feature. According to the need of embedding, we quantify the FDPS feature F of original audio to QF. Based on the Eq. (2), we can get the quantified DCT coefficient $qd(l)$, $1 \le l \le L$.

$$qd(l) = sign(d(l)) \cdot \sqrt{\left(d(l)^2 + \lambda\right)^{QF/F} - \lambda} \tag{2}$$

where $sign(\cdot)$ is the symbolic function, F is the FDPS feature and $d(l)$ is the l-th DCT coefficient of original audio, QF is the quantified FDPS feature, and $qd(l)$ is the l-th quantified DCT coefficient.

After getting the quantified DCT coefficients $qd(l)$, we perform inverse DCT on $qd(l)$, and can get the samples $qa(l)$ of watermarked audio, $1 \le l \le L$.

3.2.2 Robustness of the FDPS feature

In order to test the robustness of the FDPS feature, we experimentally give the features before and after the recapture attack and signal processing. We select one audio randomly and recapture the signal, shown in Figs. 3 and 4, respectively. And recapture the audio. Then we cut the two audio into 50 frames, and calculate the FDPS feature of the frames. The features are shown in Fig. 5.

It can be seen from the results shown in Fig. 5, if we take three frames as a group, the size relationship of the three features is almost unchanged. Such as, ① for the first three frames, the feature of the 2nd frame is greater than the 1st frame and less than the 3rd frame; ② for the 10th to 12th frames, the 11th frame has the largest feature. The size relationship of the features remains unchanged before and after the attacks. Besides the signal shown in Fig. 3, we select 100 audio signals (each signal has the duration of 30 seconds) and recapture the signals by using different recording devices containing mobile phone (Apple iPhone 13 Pro, HUAWEI P50, SAMSUNG Galaxy S21) and voice recorder (SONY ICD-UX560F). For each signal, we cut the original and recaptured one into 300 segments and calculate the FDPS feature of each segment. Similarly, we take the three segments as one group and count all the groups, for which the three features size relationship is almost unchanged

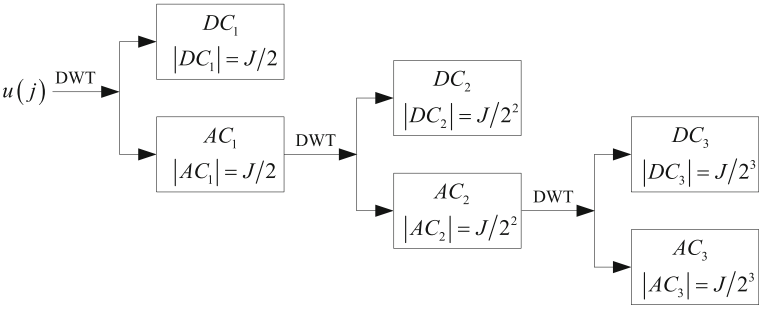

Fig. 2 The structure of the DWT

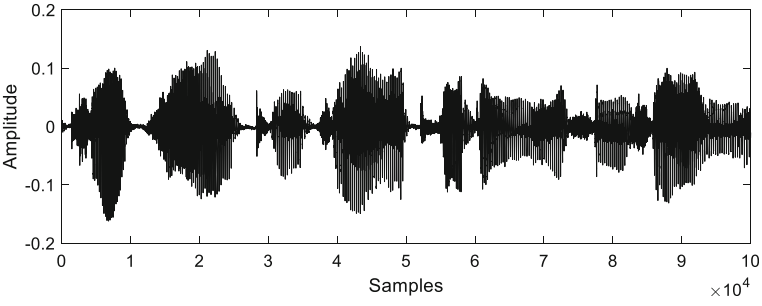

Fig. 3 The original audio selected

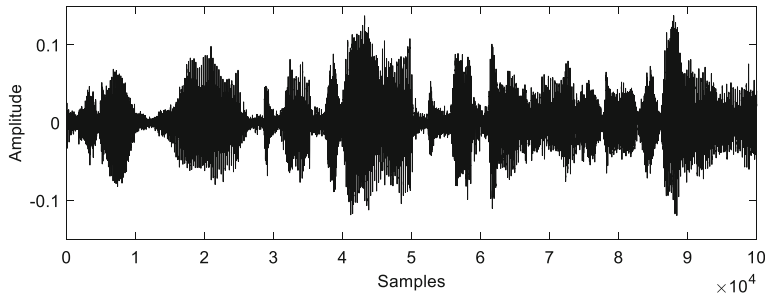

Fig. 4 The recaptured audio

before and after recapturing attacks. For every recording device, we show the number of groups in the Fig. 6.

From the results shown in Fig. 6, we can get the conclusion that the FDPS features of audio signal have the characteristic that the feature size relationship between the selected segment and the two adjacent segments remains almost unchanged after recapturing attacks. Based on the analysis, we propose the watermark embedding method. In short, for embedding, we select three consecutive frames and quantify the FDPS feature of one frame, making the quantified features satisfying a certain relationship.

4 The Scheme

In practical applications, audio signals are often subject to de-synchronization attacks, such as deletion attacks, etc.. In this paper, we embed synchronization codes and copyright information. Synchronization codes are used for the resistance of de-synchronization attacks, and the copyright information is used for copyright protection and traceability.

We denote the L length original audio as $A = \{a(l),\ 1 \le l \le L\}$, where $a(l)$ is the l-th sample. Then we cut the signal A into N length frames, and denote the i-th frame as $A_i = \{a_i(l),\ 1 \le i \le L/N,\ 1 \le l \le N\}$. We denote the binary bits that will be embedded into the i-th frame A_i as $W_i = \{w_m|\ w_m \in \{0,\ 1\},\ 1 \le m \le M\}$.$W_i$ consists of two parts, $W1_i$ and $W2_i$. $W1_i = \{w_m|\ 1 \le m \le M_1\}$ is generated by the synchronization codes, and $W2_i = \{w_m|\ M_1 + 1 \le m \le M\}$ is the whole watermark (generated by copyright information) or a part of the watermark. Then W_i can be re-expressed as$W_i = [W1_i,\ W2_i]$.

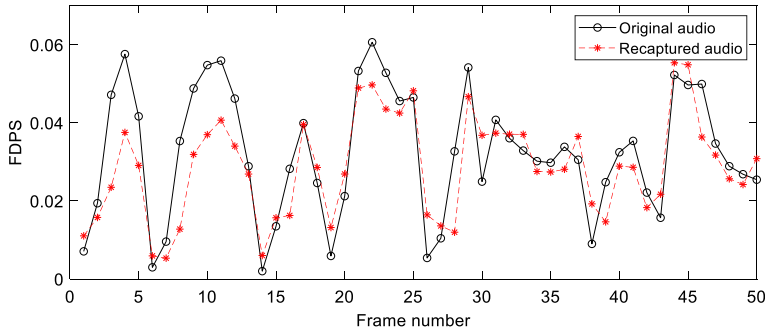

Fig. 5 The feature of the 50 frames of original and recaptured audio

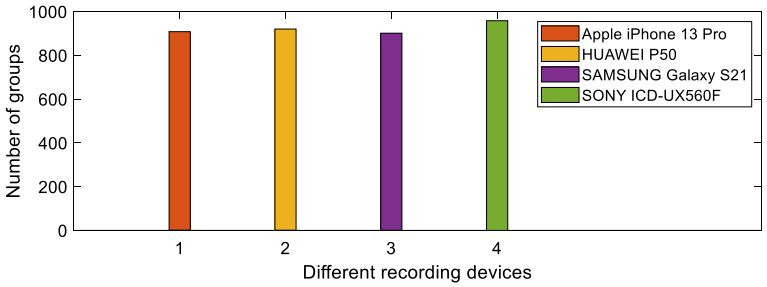

Fig. 6 The number of groups for which the three features size relationship unchanged before and after recapturing attacks

We cut the frame of original audio A_i into M segments, and denote the j-th segment as $A_{i,j}$, $A_{i,j} = \{a_{i,j}(l),\ 1 \le l \le N/M,\ 1 \le j \le M\}$. Then the watermark bits W_i can be embedded into the the i-th A_i by using the following method.

4.1 Synchronization codes embedding

We embed synchronization codes ($W1_i = \{w_m | 1 \le m \le M_1\}$) into the first M_1 segments of $A_i(A_{i,1}, A_{i,2}, \ldots, A_{i,M_1})$. The process of synchronization codes embedding is shown in Fig. 7.

Step 1. We select the first segment of A_i and perform D-level DWT on the signal $A_{i,1}$. We get the approximate coefficient denoted by $C_{i,1}$.

Step 2. We divide $C_{i,1}$ into 2 segments and denote the two segments as $C_{i,1}^1 = \{ac_l^1 | 1 \le l \le N/2M\}$ and $C_{i,1}^2 = \{ac_l^2 | N/2M + 1 \le l \le N/M\}$. Then we calculate the sum of coefficients of first segment based on the Eq. (3).

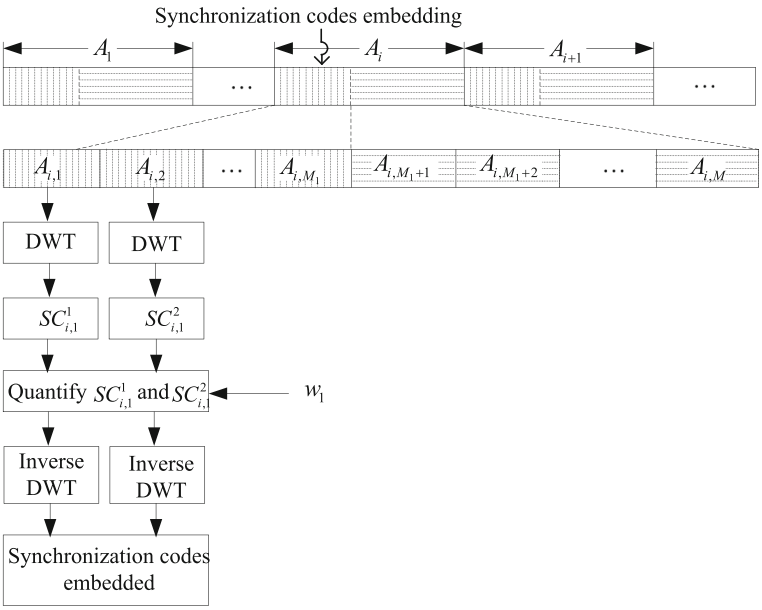

Fig. 7 The process of synchronization codes embedding

$$SC_{i,1}^{1} = \sum_{l=1}^{N/2M} \left| ac_{l}^{1} \right| \tag{3}$$

where ac_l^1 represents the l-th approximate coefficient, $1 \le l \le N/2M$. Similarly, we can get the sum of coefficients of second segment $SC_{i,1}^2$.

Step 3. If $SC_{i,1}^1 > SC_{i,1}^2$, $w_1 = 0$, we quantify the coefficient $C_{i,1}^1$ by using the Eq. (4). If $SC_{i,1}^1 < SC_{i,1}^2$, $w_1 = 1$, we quantify the coefficient $C_{i,1}^2$ by using the Eq. (5), where $\lambda_1 > 0$ and it is the parameter.

$$qac_{l}^{1} = \lambda_1 \times \frac{SC_{i,1}^{2}}{SC_{i,1}^{1}} \times ac_{l}^{1}, 1 \le l \le N/2M \tag{4}$$

$$qac_{l}^{2} = \lambda_1 \times \frac{SC_{i,1}^{1}}{SC_{i,1}^{2}} \times ac_{l}^{2}, N/2M + 1 \le l \le N/M \tag{5}$$

Step 4. Perform D-level inverse DWT on the quantified approximate coefficient and other detail coefficients, we can embed the bit w_1 into $A_{i,1}$. Repeat the steps above, we can get the synchronization codes embedded signal.

4.2 Copyright information embedding

We embed copyright information ($W2_i = \{w_m | M_1 + 1 \le m \le M\}$) into the second M_2 segments of $A_i(A_{i,M_1+1}, A_{i,M_1+2}, \ldots, A_{i,M})$.

Step 1. We divide the segment A_{i,M_1+1} into 3 sub-segments, $A_{i,M_1+1}^1, A_{i,M_1+1}^2, A_{i,M_1+1}^3$. Based on Eq. (1), we calculate the FDPS feature of these sub-segments, denoted by FA_{i,M_1+1}^1, FA_{i,M_1+1}^2, FA_{i,M_1+1}^3., respectively.

Step 2. By using the following method, we quantified the feature FA_{i,M_1+1}^2 to QFA_{i,M_1+1}^2. Here we denote $S = sign\left(FA_{i,M_1+1}^2 - FA_{i,M_1+1}^1\right) \times sign\left(FA_{i,M_1+1}^2 - FA_{i,M_1+1}^3\right)$, where $sign(\cdot)$ is the symbolic function.

If $w_{M_1+1} = 0$, we use Eq. (6) to quantify the feature FA_{i,M_1+1}^2. While, if $w_{M_1+1} == 1$, we use Eq. (7) to quantify the feature FA_{i,M_1+1}^2.

$$QFA_{i,M_1+1}^{2} = \begin{cases} \min\left(FA_{i,M_1+1}^{1}, FA_{i,M_1+1}^{3}\right) + \lambda_2 \times \left| FA_{i,M_1+1}^{1} - FA_{i,M_1+1}^{3} \right|, S = 1 \\ FA_{i,M_1+1}^{2}, \qquad S = -1 \end{cases} \tag{6}$$

$$QFA_{i,M_1+1}^{2} = \begin{cases} FA_{i,M_1+1}^{2}, \qquad S = 1 \\ \max\left(FA_{i,M_1+1}^{1}, FA_{i,M_1+1}^{3}\right) + \lambda_3, S = -1 \end{cases} \tag{7}$$

where QFA_{i,M_1+1}^2 is the quantified feature, $0 < \lambda_2 < 1$, $\lambda_3 > 0$. The approximate quantified feature size relationship of the 3 sub-segments is shown in Figs. 8 and 9.

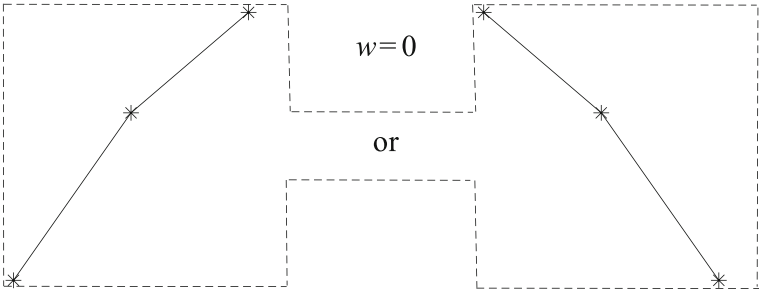

Fig. 8 The approximate quantified feature size relationship of the 3 sub-segments for embedding of 0

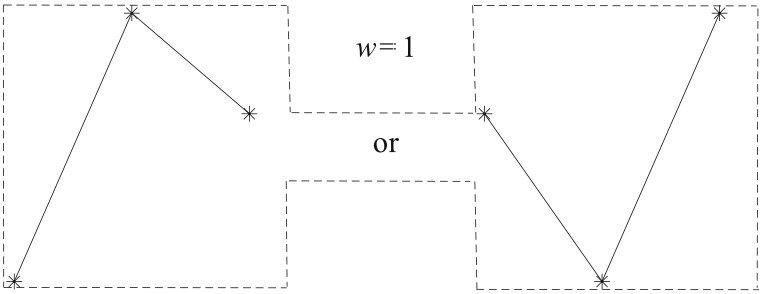

Fig. 9 The approximate quantified feature size relationship of the 3 sub-segments for embedding of 1

Step 3. Based on the quantified feature QFA^2_{i,M_1+1} and the Eq. (2), we can get the corresponding DCT coefficients. Then we perform inverse DCT on the coefficients to get the embed the bit w_{M_1+1} into A_{i,M_1+1}. Repeat the steps, we can embed the copyright information in host signal. The process of embedding is shown in Fig. 10.

Combining the two embedding methods, synchronization codes embedding and copyright information embedding, we can get the watermarked audio. The detailed process is summarized in Algorithm 1.

Algorithm 1: Copyright information embedding

Input: A_i

Output: Copyright information embedded

1: **for** $m = M_1 + 1 : M$ **do**

2: Calculate the FDPS feature of the three sub-segments of each segment.

3: **if** the bit is 0, quantify the FDPS feature satisfying the relationship shown in Fig. 8.

4: **else**
Quantify the FDPS feature satisfying the relationship shown in Fig. 9.

5: Get the DCT coefficients of the quantified feature.

6: Perform inverse DCT on the coefficients.

7: **end if**

8: **end for**

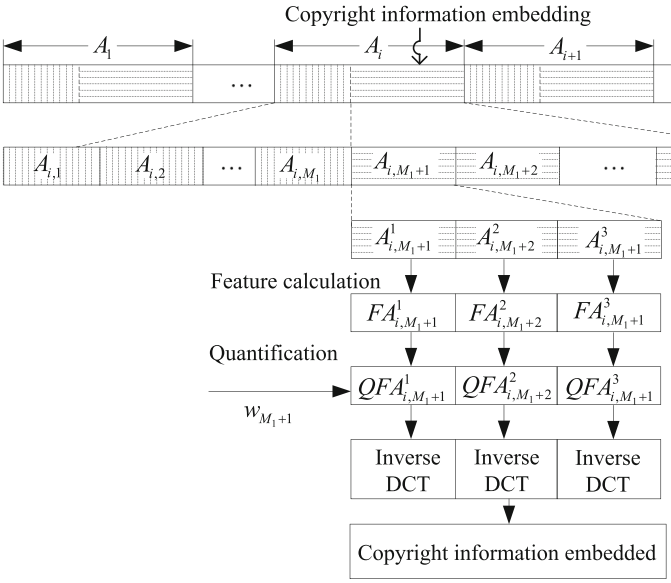

Fig. 10 The process of copyright information embedding

4.3 Watermark extraction

4.3.1 Search for synchronization codes

We first segment the incoming audio into many sections. Then we perform D-leve DWT on the segment and extract binary bits from the approximate coefficients. Secondly we search for synchronization codes in the extracted bits and repeat the procedure by shifting the selected segment one sample at a time until a synchronization code is found. After determining the location of the synchronization code, we can locate the segment with copyright information. And then we extract the binary bits from the located segment as the extracted copyright information. The detailed process is described as follows.

Let denote $A^{'} = \left\{a^{'}(l) \mid 1 \le l \le L_A^{'}\right\}$ as the watermarked signal, containing $L_A^{'}$ samples. We search synchronization codes firstly to locate the signals watermark embedded. We select the $N \cdot M_1/M$ consecutive samples from the first sample (denoted by A_1^s). Then we extract binary bits as follows.

Step 1. We cut A_1^s into M_1 segments, and denote the first segment as $A_{1,1}^s$.

Step 2. We perform D-level DWT on the segment $A_{1,1}^s$ and can get the approximate coefficient denoted by $C_{1,1}^s$.

Step 3. We divide $C_{1,1}^s$ into 2 sub-segments and calculate the sum of coefficients of the 2 sub-segments, based on the Eq. (3), denoted by $SC_{1,1}^{s1}$ and $SC_{1,1}^{s2}$.

Step 4. If $SC_{1,1}^{s1} < SC_{1,1}^{s2}$, we extract the first binary bit $w_1^{'} = 0$. And if $SC_{1,1}^{s1} \ge SC_{1,1}^{s2}$, we extract the binary bit $w_1^{'} = 1$.

Step 5. Repeat above steps, we can extract the M_1 length binary bits ($w_1^{'}$, $w_2^{'}$, ⋯, $w_{M_1}^{'}$) from the M_1 segments.

In this paper, we suppose $\{w_1, w_2, \cdots, w_{M_1}\}$ is the original synchronization codes, and $\left\{w_1^{'}, w_2^{'}, \cdots, w_{M_1}^{'}\right\}$ is the extracted bits. If $\sum_{m=1}^{M_1} w_m \oplus w_m^{'} \leq T$($T$ is the predefined threshold and set to 21 in this paper), the bits $\left\{w_1^{'}, w_2^{'}, \cdots, w_{M_1}^{'}\right\}$ can be determined as the synchronization codes. Otherwise, we shift one sample and repeat the procedure until a synchronization codes are found. With the position of a synchronization code determined, we can locate the segment with copyright information and extract the hidden information bits from it.

4.3.2 The extraction of copyright information

With the position of a synchronization code determined, we extract the binary bits by using the following method. Suppose WA' as the located audio segment with copyright information.

Step 1. We cut WA' into 3 sub-segments, $WA_1^{'}$, $WA_2^{'}$, $WA_3^{'}$. By using the Eq. (1), we calculate the FDPS feature of the three segments, denoted by $WF_1^{'}$, $WF_2^{'}$, $WF_2^{'}$, respectively.

Step 2. We denote cw as the bit extract from WA', which can be obtained based on Eq. (8).

$$cw = \begin{cases} 0, SW = -1 \\ 1, SW = 1 \end{cases} \tag{8}$$

where $SW = sign\left(WF_2^{'} - WF_1^{'}\right) \times sign\left(WF_2^{'} - WF_3^{'}\right)$.

Repeat the steps, we extract other binary bits from the located segments as the extracted copyright information.

5 Experimental results

In this paper, we used MATLAB software to simulate and analyze the performance of the proposed scheme. The test database includes 500 host audio signals, belonging to four different genres, marching music, light music, pop music and human voice. The watermark bits are generated by m-sequence (with a length of 44 binary bits) and copyright information (with a length of 56 binary bits). The m-sequence as synchronization code is used to locate the position of copyright information, in order to resist the cropping and shifting attacks. We embedded the watermark bits into the host signals, and then recaptured the watermarked signals with different recording devices, mobile phone (Apple iPhone 13 Pro) and voice recorder (SONY ICD-UX560F). We also tested the signals recaptured by other devices (such as HUAWEI P50, SAMSUNG Galaxy S21, and etc.), and the simulation results are similar. The recording environment is all in a relatively quiet room. All of the signals and the recaptured version are 16-bit quantified mono signals, with 44.1 kHz sample rate, and have the duration of 10 seconds.

5.1 Embedding capacity

In this paper, we define embedding capacity as the number of watermark bits embedded in per second. Denote V as the embedding capacity of the selected audio. We embed M length binary

bits (m-sequence and copyright information) in the fragment, which has the duration of 4 seconds. So the embedding capacity of the scheme is $M/4$. According to the value of the parameters in this paper, we can get the embedding capacity is about 25 bits/s.

5.2 Quality of watermarked signal

For a practically feasible watermarking method robust against recapturing attacks, in addition to the relatively strong robustness, watermark embedding cannot degrade the auditory quality of audio signals. That is to say the watermark embedded should be inaudible. In the following, we test the inaudibility of our algorithm, using both objective and subjective methods. We use signal-to-noise ratio (SNR) to measure the objective quality and subjective difference grade (SDG) to evaluate the subjective quality of the watermarked signals. The SNR is defined in Eq. (9), and the scores of SDG is listed in Table 1.

$$\mathrm{SNR} = 10\lg\left(\frac{\sum_{l=1}^{L} a(l)^2}{\sum_{l=1}^{L}(a(l)-a'(l))^2}\right) \tag{9}$$

where $a(l)$ and $d(l)$ are the l-th original and watermarked signal. The larger the SNR value is, the better the watermark is imperceptible.

We show the SNR mean values in different quantization steps in Fig. 11. As the results shown in Fig. 11, we can get the conclusion that the SNR values decrease with the increase of quantization steps. It is true that the increase of quantization step size can reduce the inaudibility of watermarked audio, but it can increase the robustness of the algorithm. Therefore, many algorithms increased the robustness of the algorithm by increasing the quantization step size. For the algorithm robust against de-synchronization and recapturing attacks proposed in this paper, under the premise that the watermark is inaudible (SNR value

Table 1 Subjective difference grades

SDG	Description of impairments	Quality
0.0	Imperceptible	Excellent
−1.0	Perceptible, but not annoying	Good
−2.0	Slightly annoying	Fair
−3.0	Annoying	Poor
−4.0	Very annoying	Bad

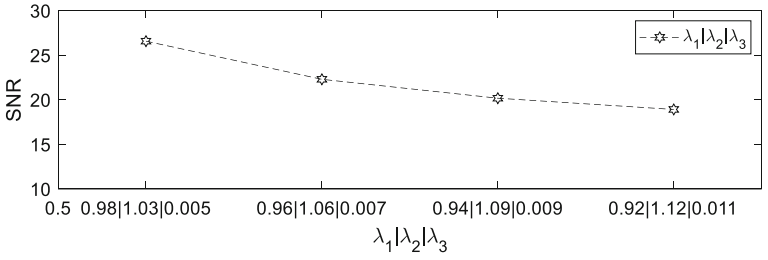

Fig. 11 The SNR values in different quantization steps

great than 20) [13], we try to improve the robustness. Based on the principle that watermarking method applied for copyright protection should be robust to attacks while maintaining high perceptual quality, we set $\lambda_1 = 0.94$, $\lambda_2 = 1.09$, $\lambda_3 = 0.09$to ensure high imperceptibility and robustness. The following simulation results show that the quantization steps are suitable for the proposed scheme with high imperceptibility and robustness against attacks.

In the following, we test the inaudibility of the proposed algorithm subjectively. Table 2 gives the SDG mean values for the different genres signals with the quantization step $\lambda_1 = 0.94$, $\lambda_2 = 1.09$, $\lambda_3 = 0.09$, which are acquired by 18 people. All the SDG values in Table 2 are larger than −1, which demonstrates that the watermark embedded is inaudible.

5.3 Robustness

In this paper, we use the synchronization codes to locate the signal segment with copyright information. So, for the incoming signals, we extract binary bits and searching for synchronization codes from the bits extracted firstly. The accuracy of the location will directly affect the extraction quality of copyright information. We use the Eq. (10) to calculate the probability of correctly finding the synchronization codes from the incoming signals.

$$P = \frac{N_s^{'}}{N_s} \tag{10}$$

where $N_s^{'}$ denotes the number of segments located correctly based on synchronization codes, N_s denotes the total number of the segments with copyright information. The higher the probability of correctly extracting the synchronization codes, the more accurate it is to locate the audio segment containing copyright information, aiming to better extract copyright information. On the basis of the correctly searching for the synchronization codes, we extract the copyright information from the located segment, and evaluate the robustness by using BER, defined as follows.

$$\mathrm{BER} = \frac{1}{M} \sum_{m=1}^{M} w_m \oplus w_m^{'}, \tag{11}$$

where $\oplus$ is the exclusive (XOR) operator, w_m represents the original watermark, $w_m^{'}$ represents the extracted binary bit, and M is the number of watermarks embedded.

5.3.1 Robustness against de-synchronization attacks

We extract binary bits from the attacked signals to searching for synchronization codes. The probability of correctly finding the synchronization codes P of our method is shown in Table 3. Besides, we simulate the methods in [2, 5, 23, 25] under the same experimental environment and list the results in Table 3 too. The embedding rates are 25 bits/s for all methods.

In Table 3, signal processing operations include low-pass filtering with cutoff frequencies of 8 kHz, MP3 compression (with compression bit rates of 64 and 128 kbps), AAC compression (128 kbps). And the de-synchronization attacks contain jittering attack (1/10, 1/100 and 1/1000 mean that random removal of one sample out every 10, 100, and 1000 samples from the watermarked signals, respectively), time-scaling and pitch-scaling attacks, with the four scaling factors 80%, 90%, 110%, 120%. $P = 100\%$ means that all synchronization codes can be correctly searched.

Table 2 SDG values of different genres under the quantization step $\lambda_1 = 0.94, \lambda_2 = 1.09, \lambda_3 = 0.09$

Different genres	SDG
Marching music	−0.764
Light music	−0.876
Pop music	−0.692
Human voice	−0.658

Based on the results, it can be conclude that, most of the methods are robust against signal processing operations. While for the de-synchronization attacks, some of the methods have poor performance. For the method [2, 23], as to jittering attack (1/10), the P values are 93% and 89%, lower than our method 100%. And for the method in [5, 23], as to time-scaling attack with the four scaling factors 80%, the P values are 88% and 86%, less than our method 5% and 7%. For attacks, the P values are all lower than our method. It demonstrates that the proposed method has better performance against de-synchronization attacks than the methods in [2, 5, 23, 25].

5.3.2 Robustness against recapturing attacks

In the following, we perform comprehensive attacks on the watermarked audio. We play the watermarked audio through the speaker and then recapture the sound waves propagating in the air, to obtain the recaptured signal. Then we perform signal processing operations and de-synchronization attacks on the recaptured signal. Firstly, we extract binary bits from the attacked signal to search the synchronization codes and locate the segment with copyright information. After that we extract binary bits from the located segment and calculate the BER value. Following, under the same experimental environment, we also simulate the methods [2, 5, 23, 25] and show the test results in Table 4.

Table 3 Comparison results of synchronization codes extracted correctly of our method and the methods in [2, 5, 23, 25]

Attacks		P(%)				
		[5]	[25]	[2]	[23]	Our
Low pass (8 kHz)		96	100	99	98	**100**
MP3 (64 kbps)		95	96	97	94	**99**
MP3 (128 kbps)		98	99	99	97	**100**
AAC (128 kbps)		91	97	96	92	**98**
Jittering	1/10	94	95	93	89	**100**
	1/100	96	98	97	94	**100**
	1/1000	100	100	100	98	**100**
Time-scaling	80%	88	90	89	86	**93**
	90%	94	94	92	91	**97**
	110%	95	93	93	94	**98**
	120%	93	92	91	92	**95**
Pitch-scaling	80%	89	91	91	90	**94**
	90%	92	94	93	91	**94**
	110%	94	92	93	93	**97**
	120%	90	91	92	91	**95**

The embedding rates are 25 bits/s for all methods

Table 4 Comparison results of the BER values for comprehensive attacks of the proposed method and the methods in [2, 5, 23, 25]

Attacks			BER(%)				
			[5]	[25]	[2]	[23]	Our
Recapturing attack			24	21	26	17	**4**
Recapturing attacks and	Low pass (8 kHz)		25	23	28	19	**4**
	MP3 (64 kbps)		25	24	28	19	**5**
	MP3 (128 kbps)		24	22	27	18	**4**
	AAC (128 kbps)		27	23	29	21	**6**
	Jittering	1/10	29	31	33	24	**4**
		1/100	27	25	27	21	**4**
		1/1000	25	22	27	18	**4**
	Time-scaling	80%	31	29	33	28	**14**
		90%	29	27	29	26	**10**
		110%	29	28	30	23	**8**
		120%	28	26	31	25	**11**
	Pitch-scaling	80%	29	29	36	26	**12**
		90%	28	27	33	24	**9**
		110%	28	26	29	23	**8**
		120%	27	27	34	25	**9**

The embedding rates are 25 bits/s for all methods

The results shown in Table 4 are quite different from that in Table 3. It demonstrates that the performance of the methods against recapturing attack varies significantly. Indeed, after recapturing attacks, the performance of all methods degrades to some extent, but our method degrades slightly.

As to recapturing attacks, the BER values of copyright information extraction of the proposed method is 4, much lower than the methods [2, 5] (which are 24 and 26). For comprehensive attacks (recapturing attacks, post-processed with common signal processing operations and de-synchronization attacks), the BER values are lower than the methods [2, 5, 23, 25] for all the attacks. In particular for the pitch-scaling attacks (scaling factors 80%), compared with the method [2], the BER value of proposed method is decreased by 24%. In [2], the adaptive mean modulation is used for embedding and watermarks extraction is more sensitive to the changes audio signal. The recapturing attacks can result in a larger change of watermarked audio in the time and frequency domains, which is a devastating attack for watermarked signal. So, it is difficult to extract correct watermark for the method in [2].

Based on the analysis above, we can get the conclusion the performance of this algorithm is better than the state-of-the-art methods [2, 5, 23, 25].

6 Conclusion

In this paper, we proposed a novel robust audio watermarking method against recapturing attack. We defined the FDPS feature of digital audio and analyzed the robustness of this feature. Then we divided a long audio signal into segments and split each segment into fragments. We embedded synchronization codes in some fragments and copyright information in other fragments by quantifying the FDPS feature. For the incoming audio, we extract synchronization codes to locate the watermarked fragments, further to extract copyright

information. Experimental results show that our method has better performance against recapturing attack than the state-of-the-art methods.

In the future, we will try to extend the applications of the proposed FDPS feature. And we will consider more noise environments and further improve the performance of the method.

Acknowledgments This paper is supported by the National Natural Science Foundation of China (Grant No. 61902085), Nanhu Scholars Program for Young Scholars of XYNU, the Science and Technology Project of Henan Province (No. 212102310993), Natural Science Foundation of Henan Province (No. 222300420274), and the Opening Project of Guangdong Provincial Key Laboratory of Information Security Technology (No.2020B1212060078-1). We would like to thank the anonymous reviewers for their constructive suggestions.

Declarations

Conflict of interest The authors confirm that the manuscript has been submitted solely to this journal and is not published, in press, or submitted elsewhere. There are no financial interests or benefits dependent on this manuscript. There is also no conflict of interest with any other research.

References

1. Chen OTC, Liu CH (2007) Content-dependent watermarking scheme in compressed speech with identifying manner and location of attacks. IEEE Trans Audio Speech Lang Process 15:1605–1616
2. Hu HT, Lee TT (2019) Frame-synchronized blind speech watermarking via improved adaptive mean modulation and perceptual-based additive modulation in DWT domain. Digit Signal Process 87:75–85
3. Hu P, Peng D, Yi Z, Xiang Y (2016) Robust time-spread echo watermarking using characteristics of host signals. Electron Lett 52:5–6
4. Hua G, Goh J, Thing VLL (2015) Time-spread echo-based audio watermarking with optimized imperceptibility and robustness. IEEE Trans Audio Speech Lang Process 23:227–239
5. Jiang W, Huang X, Quan Y (2019) Audio watermarking algorithm against synchronization attacks using global characteristics and adaptive frame division. Signal Process 162:153–160
6. Kalantari NK, Akhaee MA, Ahadi SM, Amindavar H (2009) Robust multiplicative patchwork method for audio watermarking. IEEE Trans Audio Speech Lang Process 17:1133–1141
7. Kang H, Yamaguchi K, Kurkoski B, Yamaguchi K, Kobayashi K (2008) Full-index-embedding patchwork algorithm for audio watermarking. IEICE Trans Inf Syst E91-D:2731–2734
8. Kang X, Yang R, Huang J (2011) Geometric invariant audio watermarking based on an LCM feature. IEEE Trans Multimed 13:181–190
9. Korycki R (2014) Authenticity examination of compressed audio recordings using detection of multiple compression and encoders' identification. Forensic Sci Int 238:33–46
10. Kosta P, Slavko K, Igor D, Adam W (2022) Robust speech watermarking by a jointly trained embedder and detector using a DNN. Digit Signal Process 122:103381
11. Li R, Yu S, Yang H (2016) Spread spectrum audio watermarking based on perceptual characteristic aware extraction. IET Signal Process 10:266–273
12. Liu ZH, Wang HX (2014) A novel speech content authentication algorithm based on Bessel–Fourier moments. Digit Signal Process 24:197–208
13. Liu Z, Zhang F, Wang J, Wang H, Huang J (2016) Authentication and recovery algorithm for speech signal based on digital watermarking. Signal Process 123:157–166
14. Malvar HS, Florencio DAF (2003) Improved spread spectrum: a new modulation technique for robust watermarking. IEEE Trans Signal Process 51:898–905
15. Nadeau A, Sharma G (2017) An audio watermark designed for efficient and robust resynchronization after analog playback. IEEE Trans Inf Forensics Secur 12:1393–1405
16. Natgunanathan I, Xiang Y, Rong Y, Zhou W, Guo S (2012) Robust patchwork-based embedding and decoding scheme for digital audio watermarking. IEEE Trans Audio Speech Lang Process 20:2232–2239
17. Natgunanathan I, Yong X, Hua G, Beliakov G, Yearwood J (2017) Patchwork-based multilayer audio watermarking. IEEE/ACM Trans Audio Speech Lang Process 25:2176–2187

18. Salah E, Amine K, Redouane K, Fares K (2021) A Fourier transform based audio watermarking algorithm. Appl Acoust 172:107652
19. Wu S, Huang J, Huang D (2005) Efficiently self-synchronized audio watermarking for assured audio data transmission. IEEE Trans Broadcast 51:69–76
20. Xiang Y, Peng D, Natgunanathan I, Zhou W (2011) Effective pseudonoise s denoted uence and decoding function for imperceptibility and robustness enhancement in time-spread echo-based audio watermarking. IEEE Trans Multimed 13:2–13
21. Xiang Y, Natgunanathan I, Guo S, Zhou W, Nahavandi S (2014) Patchwork-based audio watermarking method robust to de-synchronization attacks. IEEE/ACM Trans Audio Speech Lang Process 22:1413–1423
22. Xiang Y, Natgunanathan I, Rong Y, Guo S (2015) Spread spectrumbased high embedding capacity watermarking method for audio signals. IEEE/ACM Trans Audio Speech Lang Process 23:2228–2237
23. Yamni M, Karmouni H, Sayyouri M, Qjidaa H (2022) Efficient watermarking algorithm for digital audio/speech signal. Digit Signal Process. 120:103251
24. Yeo IK, Kim HJ (2003) Modified patchwork algorithm: a novel audio watermarking scheme. IEEE Trans Speech Audio Process 11:381–386
25. Zhao J, Zong T, Xiang Y, Hua G, Lei X, Gao L, Beliakov G (2022) Frequency spectrum modification process-based anti-collusion mechanism for audio signals. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2022.3156973