IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 July 2024, accepted 21 July 2024, date of publication 29 July 2024, date of current version 6 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434576

RESEARCH ARTICLE

A Novel Digital Audio Encryption and Forensics Watermarking Scheme

JUNJIE HE[1], PEI ZHU[2], ZHENGHUI LIU[3,4,5], (Member, IEEE), AND YI CAO[2]

[1]School of Mathematics and Statistics, Xinyang Normal University, Xinyang 464000, China
[2]School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China
[3]Guangdong Provincial Key Laboratory of Information Security Technology, Guangzhou 510275, China
[4]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060, China
[5]Shenzhen Key Laboratory of Media Security, Shenzhen 518060, China

Corresponding author: Zhenghui Liu (zhenghui.liu@163.com)

This work was supported in part by Henan Province Key Research and Development Project under Grant 241111212200, in part by the Science and Technology Research Key Project of the Education Department of Henan Province under Grant 23A510002, in part by the National Natural Science Foundation of Henan Province under Grant 232300420424, and in part by the Opening Project of Guangdong Provincial Key Laboratory of Information Security Technology under Grant 2020B1212060078-1.

ABSTRACT There are a large amount of audio signals stored in third-party storage centers. To improve the privacy and security of audio signals, a digital audio encryption and forensics watermarking scheme is proposed. We defined the signal energy ratio feature of audio signals and designed the embedding method by quantifying the feature, aiming to improve the security of watermarking system. We firstly encrypted the original audio by using scrambling and multiplication, to get the encrypted data. Secondly, we divided the encrypted data into frames and compressed each frame by sampling to get the compressed data. After that, we embedded the compressed data and frame number into the encrypted data to get the watermarked signal being uploaded to third-party storage centers. Authorized users download encrypted data and verify the authenticity of the data. For the intact data, they decrypt it directly to get the audio signal. If the downloaded data has been attacked, they locate the attacked frame and extract the compressed data to reconstruct the attacked signal approximately. Furthermore, we decrypted the reconstructed signal to obtain the expression meaning of the original audio. The experimental results demonstrate the effectiveness of the proposed scheme.

INDEX TERMS Digital watermark, encrypted audio, content authentication, tamper recovery.

I. INTRODUCTION

Currently the emergence of a large number of digital audio signal equipment makes it easier for people to create, record, and receive audio signals [1], [2]. Digital audio signals do provide convenience for information transmission and communication, but also providing opportunities for criminals to take advantage of them. Attackers may delete audio segments containing significant content and then transmit them to the receiver, resulting in the receiver cannot obtain the complete expression. An attacker may also use a phone to recapture copyrighted audio content and then they freely distribute the recaptured audio to obtain illegal benefits [3], [4], [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

In addition, for a large number of privacy-sensitive audio signals that are easy to cause disputes, we usually record the audio for convenient playback later. These large numbers of signals are often stored in third-party storage centers (TPSC), *e.g.*, cloud storage servers. If these signals are uploaded directly to TPSC, there will be a risk of privacy disclosure for sensitive audio signals. In order to protect the privacy and prevent the audio data from being stolen and misused, users need to find a way to protect the data from various types of existing threats. Encrypting audio signals is a common and effective method, which can distort audio signals to prevent unauthorized listeners from understanding them [6].

A. AUDIO ENCRYPTION

Encrypting audio data is more challenging than other types of data owing to the correlation between neighbouring samples.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Currently, encryption algorithms based on chaos theory are now widely used to protect digital audio and image data. Atul and Mohit proposed a novel chaos-based strategy for audio encryption [7]. The scheme used the Sine-Cosine Map, Logistic Sine Cosine map, and double DNA operation to encrypt audio signals. Some characteristics of the encrypted signal have changed greatly compared with the previous one, such as Peak Signal to Noise Ratio, Histogram Analysis, Number of Sample Change Rate, Correlation Coefficient, which indicates that the encryption method can hide the expression meaning of the original signal. Rahul et al. proposed a robust method for audio encryption based on chaos theory and user-biometric images [8]. The chaotic sequences generated by the logistic map were used to create different initial values for the Henon map and Lorenz Systems. The scheme is robust and efficient against some forms of cryptographic threats. Albahrani et al. presented an encryption method for two-channel audio files based on chaotic maps, which has an effective diffusion and confusion mechanism for audio communication with inside the area of telecommunication [9]. Alanazi et al. used Gensio-Tesi chaotic system to generate the substitution and permutation networks, and then designed the approach for audio encryption to securely store and transfer audio signals [10]. By using a particle swarm optimization algorithm and along with two transformations Multiwavelet and Arnold techniques, Hasan et al. proposed the audio encryption method, aiming to optimize the level of noise in the encrypted signal and improve the complexity and security [11].

In addition to the above encryption methods, there is another homomorphic encryption schemes for audio signals, for which, the result of addition or multiplication of encrypted data is equal to that of audio data [12]. So, the algorithm maintains the mathematical structure of original data and permits some operation on encrypted data. However, the major problem of audio homomorphic encryption is the huge data expansion of encrypted data. In order to address the issue, Shi et al. presented a speech homomorphic encryption scheme with less data expansion [13]. By comparing with the encryption method based on scrambling system, the encryption efficiency of the algorithm in [13] is obviously higher than other algorithms. Table 1 summarizes recent audio encryption algorithms.

Indeed, encryption can distort audio signals to prevent unauthorized listeners from understanding them. However, how to protect the authenticity and integrity of encrypted data has not been solved, such as the data stored in TPSC. It is possible that the data downloaded from the third-party servers have been tampered, occurred on these servers or during the process of transmission [14], [15], [16]. For digital audio signals, there is a risk of data loss if encrypted data is attacked. In addition, if users act accordingly based on the attacked version, it may cause serious consequences. So, for the downloaded data, users need verify the authenticity of the data downloaded before decryption [17], [18], [19], [20].

TABLE 1. Summary of recent audio encryption algorithms.

Methods	Year	Key innovations
Atul *et al*[7]	2023	Chaos-based strategy for audio encryption
Rahul *et al*[8]	2023	Robust method for audio encryption based on chaos theory and user-biometric images
Albahrani *et al*[9]	2023	Encryption method for two-channel audio files based on chaotic maps
Alanazi *et al*[10]	2023	Audio encryption method based on Gensio-Tesi chaotic system
Hasan *et al*[11]	2023	Audio encryption method by using a particle swarm optimization algorithm
Mahato *et al*[12]	2023	A comparative review on Homomorphic encryption
Shi *et al*[13]	2019	Speech homomorphic encryption scheme with less data expansion

B. AUDIO WATERMARKING

Audio watermarking is a solution for copyright protection and content forensics. It hides information with special purposes in the audio signal, which does not affect the auditory of the signal. When necessary, the users extract the hidden information to prove the ownership of the audio or to verify the authenticity of the signal [21], [22], [23].

Based on the application purpose, digital audio watermarking technology can be divided into fragile(semi-fragile) audio watermarking technology and robust audio watermarking technology. The fragile(semi-fragile) watermarking is mainly used for audio forensics [24], [25]. For such schemes, the watermark extracted from the attacked frame is usually different to the original one. Based on this, users can verify the watermarked audio intact or not. Hu and Lu [24] proposed a semi-fragile watermarking algorithm based on compressed sensing. Users got the compressed version of original audio signal by using compressed sensing technology. The compressed version and the tampering location data were as the watermark information and embedded in the region with low energy of high frequency coefficients and high energy of low frequency coefficients respectively, after 2-level Discrete Fourier Transform (DWT). The scheme has higher average signal-to-noise ratio (SNR) than others. However, due to the small quantization step size, the scheme cannot effectively resist some signal processing operations. Lai et al. [25] proposed a fragile privacy-preserving audio watermarking using homomorphic encryption the batching technique SIMD. Authors embedded the fragile watermark into the encrypted discrete wavelet transform domain. The scheme is very sensitive to common audio attacks. And non-malicious processing will cause the algorithm fail to work. Meanwhile, if the watermarked audio is attacked, the algorithm cannot recover the original signal.

The robust watermarking is mainly used for copyright protection [26], [27]. For such schemes, watermark can be extracted correctly after a certain degree of malicious attacks. Maha et al. [28] proposed digital watermarking scheme for security copyright protection applications,

in which authors involved neural network architecture in the insertion and detection processes. The scheme is imperceptibility and robust against some common audio attacks. Zhao et al. [29] designed the private data hiding system using state-switch DWT coefficients quantization on digital signal. The authors combined the quantization-embedding system and the weights of DWT coefficients with SNR to obtain an optimization model for watermarking. The scheme has a higher SNR value and a stronger robustness against re-sampling, amplitude scaling, and MP3 compression. While the schemes [28], [29] have a poor performance on the malicious attacks, such as the de-synchronization attacks.

In comparison, it is easier to design a digital audio watermarking scheme robust against common signal processing operations. However, it is a challenging work for the de-synchronization attacks. To solve this problem, Xiang et al. [30] presented robust digital audio watermarking scheme based on patchwork. Authors embedded the synchronization codes by quantifying the feature calculated by logarithmic DCT coefficients. At the decoding end, the extractor firstly analyzed the received audio to find the scaling factor imposed by an attack and then re-scaled the received audio to remove the scaling effect. This operation can make the scheme resistant to some de-synchronization attacks. Li and Xiang [31] proposed a robust watermarking scheme against de-synchronization attacks. The histogram shape in the DWT low-frequency component as a robust feature was mathematically invariant to TSM and robust to random cropping attacks. Based on the conclusion, authors embedded watermark into the histogram shape in the DWT low-frequency component. Experimental results shown the scheme provides strong robustness to TSM and random cropping attacks. Based on the segmental singular values summation (SSVS) and segmental singular values difference (SSVD), Zhao and Zong [32] proposed a robust watermarking method. They firstly extracted the SSVS feature and the SSVD feature and then embedded watermark bits via optimized embedding strategies based on these features. The scheme has higher SNR value and is robustness against de-synchronization attacks. While, the features used in the paper are public and easy to be obtained by attackers. Attackers can embed other copyright information by using the same embedding method, which may cause a fatal attack on the watermarked signal.

In addition, recently there have some audio watermarking scheme been proposed with different functions. Hwai-Tsu and Tung-Tsun [33] proposed a speech watermarking algorithm, by embedding the synchronization codes and information bits into separate DWT subbands. The scheme has a large capacity and the robustness to some attacks. Yamni et al. proposed a robust audio/speech watermarking, which is combined the discrete Tchebichef moment transform, the chaotic system of the mixed linear-nonlinear coupled map lattices, and DWT, aiming to resolve the conflicts between imperceptibility, payload capacity, and robustness. The schemes in [33] and [34] work for audio signals, but not for the encrypted version. Liu et al. [35] proposed a content authentication and tamper recovery scheme for digital speech signal. Authors discussed the compression method for speech signal based on discrete cosine transform (DCT) and embedded the compression data based on the block-based large capacity embedding method. The scheme not only can locate the attacked frames, but also can reconstruct the attacked signal. Salayani et al. [36] proposed a zero-watermarking scheme to improve the robustness to MP3 compression, high-pass filtering, and re-sampling attacks. The disadvantage of the scheme is that the watermark needs to be transmitted to the decoding end, which increases the transmission burden and risk. Table 2 summarizes recent audio watermarking schemes.

TABLE 2. Summary of recent audio watermarking methods.

Methods	Year	Key innovations
Hu *et al*[24]	2022	Semi-fragile watermarking algorithm based on compressed sensing
Lai *et al*[25]	2022	Fragile privacy-preserving audio watermarking using homomorphic encryption the batching technique SIMD
Maha *et al*[28]	2022	Watermarking scheme for security copyright protection applications using neural network architecture
Zhao *et al*[29]	2022	Watermarking scheme using state-switch DWT coefficients quantization
Xiang *et al*[30]	2014	Robust digital audio watermarking scheme based on patchwork
Li *et al*[31]	2022	Robust audio watermarking based on histogram shape in the DWT low-frequency component
Zhao *et al*[32]	2023	Robust watermarking method by using SSVS feature and the SSVD feature
Hu *et al*[33]	2019	Speech watermarking algorithm by embedding watermark into separate DWT subbands
Mehri *et al*[36]	2023	Zero-watermarking scheme

In the case of users uploading audio signals TPSC, there are two problems need to be addressed in order to protect the security of the data.

- The privacy of the audio content needs to be protected from being revealed.
- When a user downloads data from a third party, the user needs to verify the authenticity of the downloaded data.

Unfortunately, there are few research results that can hide the express meaning of the audio signal while at the same time being able to forensics the data. In order to protect the privacy and improve the security of audio data stored in TPSC, we proposed a digital audio encryption and authentication watermarking scheme. Firstly, we defined the signal energy ratio feature of digital audio and analyzed the properties of the feature. Based on the analysis, we designed a watermark embedding method by quantifying the feature. For the signal being uploaded to the third-party storage centers, we encrypted the signal by using scrambling and multiplication firstly. Secondly, we cut the encrypted data into

frames and compressed each frame to get the compressed data. Then we embedded the compressed data and frame number into the encrypted data to get the watermarked signal being uploaded to TPSC. At the decoding end, users verify the downloaded data is intact or not firstly.

If the downloaded data was attacked, users can locate the attacked frames by using the method. And then they extract the compressed data to reconstruct the attacked frames approximately. Furthermore, they decrypt the reconstructed signal to obtain the expression meaning of the original audio. Theoretical analysis and experimental results demonstrate that the scheme improves the security of digital audio stored in TPSC, and can authenticate the encrypted data and reconstruct the attacked frames approximately. The main contributions of our proposed method are summarized as follows:

- We designed an audio signals encryption method without increasing the data size of original audio, in order to improve the security and privacy of audio data stored in TPSC.
- We defined the security feature of encrypted audio and presented the watermark embedding and extraction method based on the feature, which improves the security of watermarking system.
- The proposed watermarking scheme can verify the authenticity of encrypted data downloaded from TPSC firstly. Secondly, it can locate the location of the attacked segments. Furthermore, for some malicious attacks, the scheme can approximately recover the attacked segments.

The organization of this paper is as follows. Section II analyzes the signal energy ratio of digital audio. In Section III, we design the audio encryption method to get the encrypted data. In Section IV, we propose the watermarking method based on the feature defined. We simulate the scheme and present the simulation experiment results in Section V. Finally, we summarize the conclusion in section VI.

II. THE FEATURE FOR WATERMARKING

A. THE DEFINITION OF SIGNAL ENERGY RATIO FEATURE

In order to improve the security and efficiency of watermarking system, we use the signal energy ratio (SER) feature for watermarking. For the N length audio $X = \{x(i), 1 \le i \le N\}$ and the N length signal $Y = \{y(i), 1 \le i \le N\}$, the SER X to Y is defined in Eq. (1).

$$SER_{X\_Y} = 10\lg\left(\sum_{i=1}^{N}\left(\frac{x(i)}{y(i)}\right)^2\right) \tag{1}$$

In Eq. (1), $y(i) \neq 0$. In this paper, we use the logistic chaotic mapping to generate the sequence Y. The logistic chaotic mapping is shown in Eq. (2).

$$y_{l+1} = \mu y_l(1 - y_l), y_0 = k \tag{2}$$

where k is the initial value of the logistic map, as the key of the watermarking system. μ is the logistic parameter, $0 \le \mu \le 4$.

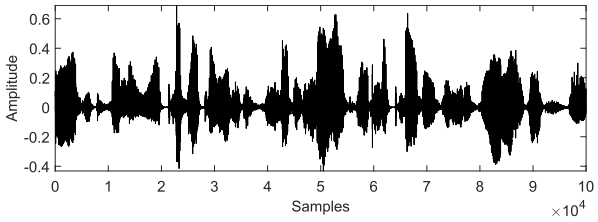


FIGURE 1. The audio signal selected randomly.

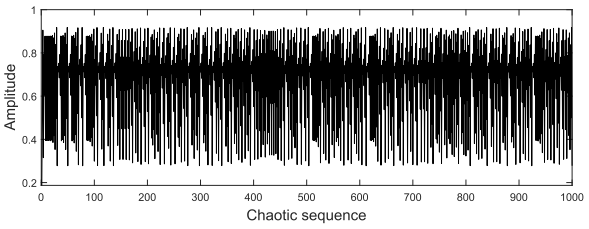


FIGURE 2. The 1000 length chaotic sequence Y1.

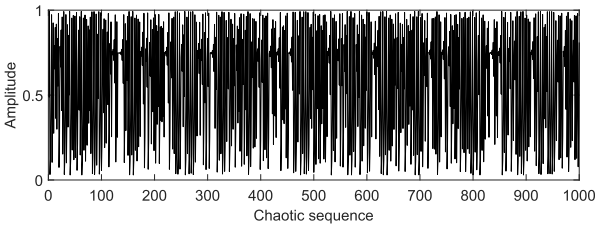


FIGURE 3. The 1000 length chaotic sequence Y2.

Under the condition that $3.5699 \le \mu \le 4$, the sequence generated by Eq. (2) is a pseudo-random distribution state, especially when the value of μ is close to 4. According to Eq. (1), we can get that

- *When the audio signal X is equal to Y,* $SER_{X\_Y} = 10\lg N$. So, the closer SER_{X_Y} is to $10\lg N$, the smaller the difference between X and Y is. Conversely the farther SER_{X_Y} is to $10\lg N$, the more the difference of the energies between X and Y is.
- SER_{X_Y} *is the relative energy of the audio signal X by comparing with the signal Y. That is a higher value of* $SER_{X\_Y} = 10\lg N$ indicates a larger energy of X for the same signal Y.

We select one audio signal randomly, which is shown in Fig.1. And we cut the signal into 100 frames. Each frame has 1000 samples. Then we select the 1000 length chaotic sequence, denoted by $Y1$, $Y2$, $Y3$ and $Y4$, which are shown in Fig. 2 to Fig. 5. We calculate the SER feature of X to $Y1$, $Y2$, $Y3$ and $Y4$, denoted by X-Y1, X-Y2, X-Y3 and X-Y4, respectively. The SER features are shown in Fig. 6. From the results, we can get the two conclusions. The one is, for the same chaotic sequence, the SER features of the 100 frames are almost different from each other. It demonstrates that there are significant differences in the SER features for different audio signals. The other is, the SER features of the same audio frame to the 4 different chaotic sequences

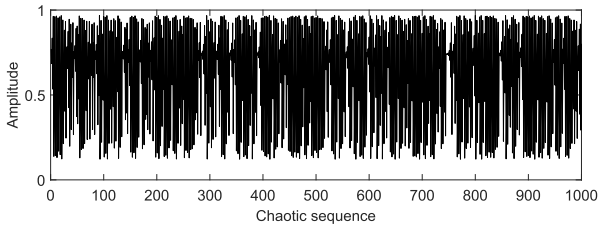


FIGURE 4. The 1000 length chaotic sequence Y3.

are almost different from each other. This is true for the 100 frames. It indicates that, if the chaotic sequence is unknown, it will be difficult to get the SER feature. So, in this paper, chaotic sequence is used as the secret key, aiming to improve the security of watermarking system.

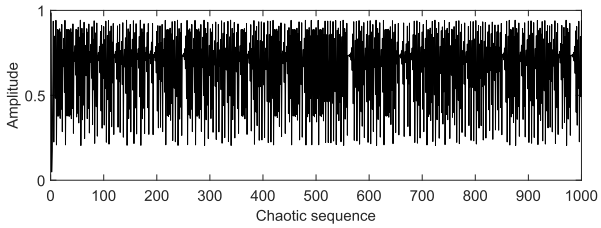


FIGURE 5. The 1000 length chaotic sequence Y4.

B. THE EMBEDDING METHOD BY QUANTIFYING THE SER FEATURE

Audio watermarking technology is to hide information called watermark into the host signal. The way to hide is to embed the information in the audio by quantifying some audio features usually. In this paper, we embed watermark bits into the SER feature. In the following, we analyze the embedding method by quantifying the SER feature. We denote X and Y as the host audio and the N length chaotic sequence, respectively. Y is generated by the logistic chaotic mapping shown in Eq. (2). Based on the Eq. (1), we can get

$$\sum_{i=1}^{N}\left(\frac{x\left(i\right)}{y\left(i\right)}\right)^{2}=\frac{SER_{X\_Y}}{10} \tag{3}$$

where $x\left(i\right)\in X$, $X=\left\{x\left(i\right),1\leq i\leq N\right\}$. We denote the watermarked audio as $X'=\left\{x'\left(i\right),1\leq i\leq N\right\}$, where $x'\left(i\right)$ is the i-th sample of X', And we denote $QSER_{X\_Y}$ as the SER of X' to Y. Similarly, based on the Eq. (1), the signal X' and it's SER feature $QSER_{X\_Y}$ should be satisfied

$$\sum_{i=1}^{N}\left(\frac{x'\left(i\right)}{y\left(i\right)}\right)^{2}=\frac{QSER_{X\_Y}}{10} \tag{4}$$

Combining the Eq. (3) and Eq. (4), we have

$$\frac{\sum_{i=1}^{N}\left(\frac{x'(i)}{y(i)}\right)^{2}}{\sum_{i=1}^{N}\left(\frac{x(i)}{y(i)}\right)^{2}}=10^{\frac{QSER_{X\_Y}-SER_{X\_Y}}{10}} \tag{5}$$

In the Eq. (5), the signal X and the SER feature SER_{X_Y} are known. Based on the watermark bits, we need quantify the feature SER_{X_Y} to $QSER_{X\_Y}$. Then the feature $QSER_{X\_Y}$ becomes a known parameter in Eq. (5). For the watermark embedding, the goal is to get the watermarked audio $X'=\left\{x'\left(i\right),1\leq i\leq N\right\}$. So, we need to find a solution for $x'\left(i\right)$ from the Eq. (5). In fact, there are many solutions. One of the solutions can be given by the Eq. (6).

$$x'\left(i\right)=x\left(i\right)\times\sqrt{10^{\frac{QSER_{X\_Y}-SER_{X\_Y}}{10}}} \tag{6}$$

So, based on Eq. (6), we can design the watermarking method by quantifying the SER feature.

III. THE DESIGNED AUDIO ENCRYPTION AND DECRYPTION

For the audio signals being upload to TPSC, in order to distort the signals to prevent unauthorized listeners from understanding them, we first encrypt the audio signals into encrypted data. In the following, we present the encryption and decryption method.

A. THE DESIGNED AUDIO ENCRYPTION METHOD

In this section, we denote the L length audio as $X=\{x\left(i\right),1\leq i\leq N\}$.

Step 1. According to the logistic chaotic mapping shown in Eq. (2), by using three different initial values k_1, k_2 and k_3 we generate the three L length different sequences, denoted by Y_1, Y_2 and Y_3, $Y_1=\left\{y_l^1,1\leq l\leq L\right\}$, $Y_2=\left\{y_l^2,1\leq l\leq L\right\}$ and $Y_3=\left\{y_l^3,1\leq l\leq L\right\}$.

Step 2. We multiply the audio signal X by the sequence Y_1 and denote the results as $E_1=\left\{e_l^1,1\leq l\leq L\right\}$, where $e_l^1=x_l\times y_l^1,1\leq l\leq L$.

Step 3. The elements belong to the sequence Y_2 are sorted in ascending order by using Eq. (7), where $a(l)$ is the address index of the sorted chaotic sequence.

$$y_{a(l)}^2=ascend\left(y_l^2\right),l=1,2,\cdots,L \tag{7}$$

Based on the Eq. (7), we scramble the signal E_1 and denote the scrambled signal as E_2, $E_2=\left\{e_l^2,1\leq l\leq L\right\}$, where e_l^2 can be obtained by the Eq. (8).

$$e_l^2=e_{a(l)}^1 \tag{8}$$

Step 4. We divide the data E_2 into P nonoverlapping frames, which are denoted by $E_{2,i}=\left\{e_{(i-1)\times L/P+t}^2,1\leq t\leq L/P\right\}$, $1\leq i\leq P$. The length of the i-th frame A_i is L/P. Then we split each frame into two segments. The first and the second segment of the i-th frame are denoted by $E1_{2,i}$ and $E2_{2,i}$, $E1_{2,i}=\left\{e_{(i-1)\times L/P+t}^2,1\leq t\leq L/2P\right\}$, $E2_{2,i}=\left\{e_{(i-1)\times L/P+t}^2,L/2P+1\leq t\leq L/P\right\}$. We multiply the data in $E2_{2,i}$ by the data in $E1_{2,i}$. Then we can get the data $RE2_{2,i}=\left\{re_{(i-1)\times L/P+t}^2,L/2P+1\leq t\leq L/P\right\}$, where $re_{(i-1)\times L/P+t}^2$ can be obtained by the Eq. (9)

$$re_{(i-1)\times L/P+t}^2=e_{(i-1)\times L/P-L/2P+t}^2\times e_{(i-1)\times L/P+t}^2 \tag{9}$$

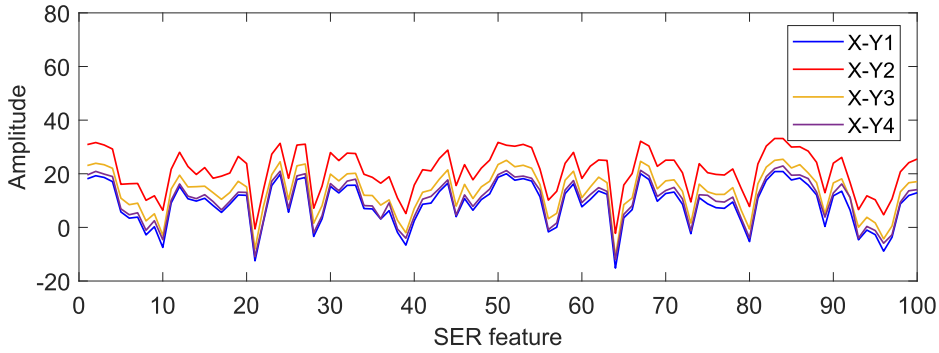


FIGURE 6. The SER features.

where $L/2P+1 \le t \le L/P$. We update the data in $E2_{2,i}$ as $RE2_{2,i}$, $1 \le i \le P$. For data E_2, the updated version is denoted by RE_2.

Step 5. Based on sequence $Y3$, the data in RE_2 is scrambled using the same scrambling method as the Step 3. The scrambled signal is denoted as $E_3 = \{e_l^3, 1 \le l \le L\}$, which is the encrypted version of the signal X. The encryption process is shown in Fig. 7.

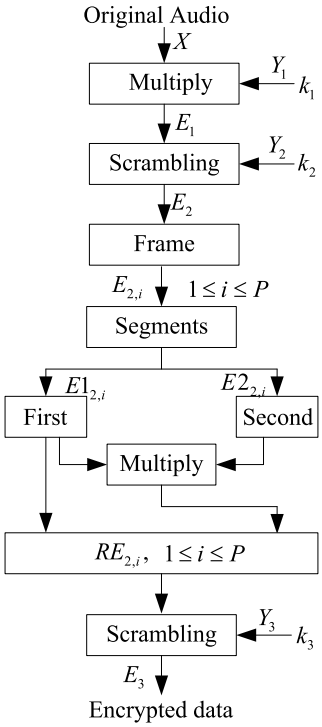


FIGURE 7. The encryption process.

B. THE DESIGNED AUDIO DECRYPTION METHOD

For the encrypted data, denoted by $E_3 = \{e_l^3, 1 \le l \le L\}$, the steps of decryption are presented in the following.

Step 1. By using the initial value k_3 and based on the Eq. (2), we generate the L length sequence $Y_3 = \{y_l^3, 1 \le l \le L\}$. Based on the scrambling system shown in Eq. (7) and Eq. (8), we perform anti scrambling operation on the encrypted data E_3. Then we can get the data denoted by RE_2, $RE_2 = \{re_l^2, 1 \le l \le L\}$.

Step 2. We cut the data RE_2 into P nonoverlapping frames, which are denoted by RE_2, $RE_{2,i} = \{re^2_{(i-1)\times L/P+t}, 1 \le t \le L/P\}$, $1 \le i \le P$. Then we split each frame into two segments. The first and the second segment of the i-th frame are denoted by $RE1_{2,i}$ and $RE2_{2,i}$, $RE1_{2,i} = \{re^2_{(i-1)\times L/P+t}, 1 \le t \le L/2P\}$, $RE2_{2,i} = \{e^2_{(i-1)\times L/P+t}, L/2P+1 \le t \le L/P\}$.

Step 3. Let's divide $RE2_{2,i}$ by $RE1_{2,i}$, and denote the result as $E2_{2,i} = \{e^2_{(i-1)\times L/P+t}, L/2P+1 \le t \le L/P\}$, where $e^2_{(i-1)\times L/P+t}$ can be obtained by the Eq. (10)

$$e^2_{(i-1)\times L/P+t} = re^2_{(i-1)\times L/P-L/2P+t} \times re^2_{(i-1)\times L/P+t} \quad (10)$$

where $L/2P+1 \le t \le L/P$. We update all the data in $RE2_{2,i}$ as $E2_{2,i}$, $1 \le i \le P$. For the data RE_2, the updated version is denoted by E_2.

Step 4. By using the initial value k_2 to generate the L length sequence Y_2, then we perform anti scrambling operation on the data E_2. We denote the result as E_1.

Step 5. Based on the chaotic system shown in Eq. (2), we generate the L length sequence Y_1 by using the initial value k_1. Then we divide E_1 by Y_1 to get the decrypted audio, denoted by X.

IV. WATERMARKING SCHEME

In this section, we firstly generate watermark by using frame number and the sampled data. Then we embed watermark into the encrypted data to generate the data that will be stored in TPSC. Secondly, we present the forensics method for the data downloaded from the TPSC. If the data is intact, the authorized users decrypt the data to get the audio signal. On the contrary, if the data is attacked, the users can locate the attacked frames and reconstruct the attacked signal approximately. The main work of the scheme proposed in this paper is shown in Fig. 8

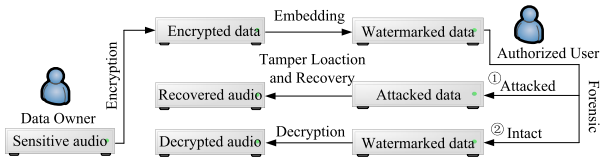


FIGURE 8. The main work of the scheme proposed.

A. WATERMARK EMBEDDING

By using the encryption method proposed in Section III, we first encrypt the L length original audio $A = \{a_l, 1 \le l \le L\}$ and denote the encrypted data as EA, $EA = \{ea_l, 1 \le l \le L\}$.

Step 1. We cut EA into P frames and denote the i-th frame as $EA_i = \left\{ea_{i,l}, 1 \le l \le L/P\right\}$. For the i-th frame EA_i, the frame number i is mapped to $FN_i = \{n_1, n_2, \cdots, n_M\}$, which can be obtained by the Eq. (11). FN_i is the identifier of the i-th frame EA_i and as a part of the watermark, which will be embedded in i-th frame EA_i.

$$i = n_1 \cdot 10^{M-1} + n_2 \cdot 10^{M-2} + \cdots + n_M \tag{11}$$

Step 2. For the i-th frame EA_i, We save one sample every T samples to get the sampled data, which is used as the compressed data of the i-th frame and denoted by $D_i = \left\{d_{i,j}, 1 \le j \le n\right\}$, $1 \le i \le P$, $n = L/T \cdot P$. Similarly, based on the initial value k_1, we can get P length sequence. And then, by using the same scrambling method shown in Eq. (7) and Eq. (8), we scramble the compressed data. After scrambling, the compressed data to be embedded into the i-th frame is denoted by C_i, $C_i = D_{a(i)}$, $1 \le i \le P$. C_i is as another part of the watermark and will be embedded in i-th frame. Therefore, in this paper, the watermark embedded in the i-frame includes FN_i and C_i.

Step 3. We split the i-th frame EA_i into three segments, denoted by $EA1_i$, $EA2_i$ and $EA3_i$, respectively. The first segment $EA1_i$ has n samples. For the second and the third segment $EA2_i$ and $EA3_i$, we split the two segments into $3M$ fragments, denoted by $EA2_{i,j}$ and $EA3_{i,j}$, respectively, $1 \le j \le 3M$. Each fragment has m elements. The segmentation method of the encrypted data EA is shown in Fig. 9.

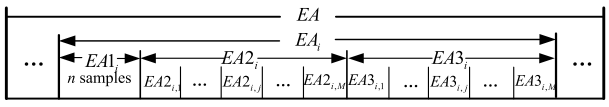


FIGURE 9. The segmentation method of the encrypted data.

Step 4. We denote $EA1_i = \left\{ea_{i,j}, 1 \le j \le n\right\}$. We embed C_i in $EA1_i$, by using the Eq. (12). In Eq. (12), $T1$ represents $c_{i,j} \ge 0$ and $\left\lfloor \left| ea_{i,j} \times 10 \right| \right\rfloor \bmod 2 = 0$, $T2$ represents $c_{i,j} \ge 0$ and $\left\lfloor \left| ea_{i,j} \times 10 \right| \right\rfloor \bmod 2 = 1$, $T3$ represents $c_{i,j} < 0$ and $\left\lfloor \left| ea_{i,j} \times 10 \right| \right\rfloor \bmod 2 = 1$, and $T4$ represents $c_{i,j} < 0$ and $\left\lfloor \left| ea_{i,j} \times 10 \right| \right\rfloor \bmod 2 = 0$.

$$wa_{i,j} = \begin{cases} s_{i,j} \times \dfrac{\left\lfloor \left| ea_{i,j} \times 10 \right| \right\rfloor}{10} + \dfrac{\left\lfloor \left| c_{i,j} \times 10 \right| \right\rfloor \bmod 10}{100} + \dfrac{\left\lfloor \left| c_{i,j} \times 100 \right| \right\rfloor \bmod 10}{1000}, & T1 or T3 \\ s_{i,j} \times \dfrac{\left\lfloor \left| ea_{i,j} \times 10 \right| \right\rfloor + 1}{10} + \dfrac{\left\lfloor \left| c_{i,j} \times 10 \right| \right\rfloor \bmod 10}{100} + \dfrac{\left\lfloor \left| c_{i,j} \times 100 \right| \right\rfloor \bmod 10}{1000}, & T2 or T4 \end{cases} \tag{12}$$

where $sign(\cdot)$ is a symbolic function and $s_{i,j} = sign\left(ea_{i,j}\right)$, $\lfloor \cdot \rfloor$ returns the largest integer less than the original value, and $|\cdot|$ represents the absolute value function, and $c_{i,j} \in C_i$. Then we denote the signal after being quantified as $WA1_i = \left\{wa_{i,j}, 1 \le j \le n\right\}$.

Step 5. We embed $FN_i = \{n_1, n_2, \cdots, n_M\}$ into $EA2_i$ and $EA3_i$, respectively. Here, we take the embedding of n_1 as an example to introduce the embedding method.

- We select the first three fragments $EA2_{i,1}$, $EA2_{i,2}$ and $EA2_{i,3}$. Based on the Eq. (2) and the initial values k_1 we generate a m length sequence, denoted by $B = \{b_l, 1 \le l \le m\}$. Then we calculate the SER feature of $EA2_{i,1}$ to B, which is denoted by $SER2_{i,1}$. Similarly, we can get the SER features for other two fragments $EA2_{i,2}$ and $EA2_{i,3}$, which are denoted by $SER2_{i,2}$ and $SER2_{i,3}$.
- As to the three features $SER2_{i,1}$, $SER2_{i,2}$ and $SER2_{i,3}$, for the convenience of description, we denote $U0_t = \lfloor SER2_{i,t} \rfloor$, $U1_t = \lfloor SER2_{i,t} \times 10 \rfloor \bmod 10$, $U2_t = \lfloor SER2_{i,t} \times 100 \rfloor \bmod 10$, $U3_t = \lfloor SER2_{i,t} \times 1000 \rfloor \bmod 10$, $1 \le t \le 3$. Then, we calculate $Z = f\left(SER2_{i,1}, SER2_{i,2}, SER2_{i,3}\right)$ by using the Eq. (13).

$$f\left(SER2_{i,1}, SER2_{i,2}, SER2_{i,3}\right) = (U1_1 + U1_2 \times 2 + U1_3 \times 3) \bmod 10 \tag{13}$$

- Based on the value of $n_1 - Z$, we quantify the three features $SER2_{i,1}$, $SER2_{i,2}$ and $SER2_{i,3}$. The quantization method is shown in Table 3, where $QSER2_{i,1}$, $QSER2_{i,2}$ and $QSER2_{i,3}$ are the three quantified features, respectively. For example, if $n_1 - Z = 1$, we quantify the feature $SER2_{i,1}$ to $QSER2_{i,1}$, $QSER2_{i,1} = U0_1 + (U1_1 + 1)/10 + U2_1/100 + U3_1/1000$, and $QSER2_{i,2} = SER2_{i,2}$, $QSER2_{i,3} = SER2_{i,3}$.
- We denote $EA2_{i,1} = \left\{ea_{1,l}, 1 \le l \le T'\right\}$. If the feature $SER2_{i,1}$ is quantified to $QSER2_{i,1}$, the samples belonging to $EA2_{i,1}$ can be quantified by using the Eq. (14).

$$wa_{1,l} = ea_{1,l} \times \sqrt{10^{\frac{QSER2_{i,1} - SER2_{i,1}}{10}}}, 1 \le l \le T' \tag{14}$$

Then the watermarked signal can be denoted by $WA2_{i,1} = \{wa_{1,l}, 1 \le l \le T'\}$. Similarly, the quantified version of the signals $A2_{i,2}$ and $A2_{i,3}$ can be obtained and denoted as $WA2_{i,2}$ and $WA2_{i,3}$. By using the same method, we can embed FN_i into $A2_i$ and $A3_i$, and denote the watermarked signals as $WA2_i$ and $WA3_i$. Then we can get the i-th frame of watermarked audio WA_i, which can be represented as $WA_i = WA1_i \,|WA2_i\,|WA3_i$, $1 \le i \le P$. The process of embedding is shown in Fig. 10.

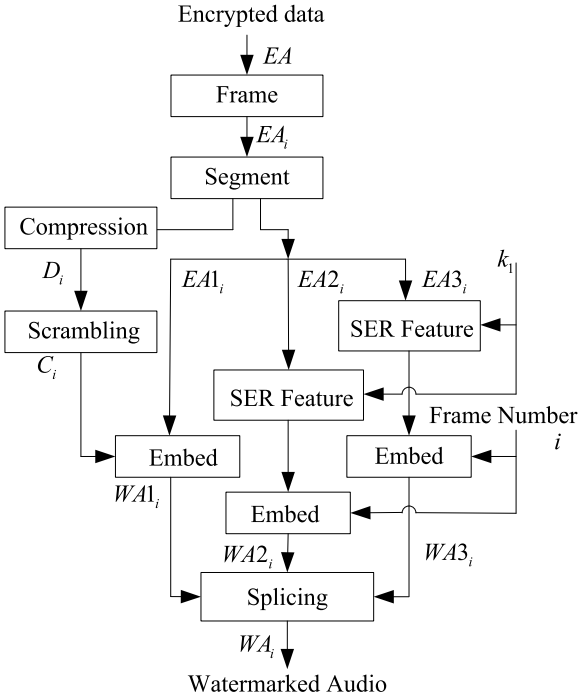


FIGURE 10. The process of the scheme proposed.

B. FORENSICS AND DECRYPTION

Suppose $WA = \{wa_l, 1 \le l \le WL\}$ is the WL length watermarked data downloaded from TPSC. The forensics steps for the watermarked data are shown as follows.

1) FORENSICS THE ENCRYPTED DATA

We select the first L/P samples from the watermarked data, denoted by $WA_1 = \{wa_l, 1 \le l \le L/P\}$. By using the method shown in Fig. 9, we cut WA_1 into three segments $WA1_1$, $WA2_1$, and $WA3_1$.

Step 1. We split $WA2_1$ and $WA3_1$ into $3M$ fragments, which are denoted by $WA2_{1,j}$ and $WA3_{1,j}$, $1 \le j \le 3M$, respectively.

- We calculate the SER feature of $WA2_{1,j}$ and $WA3_{1,j}$ to B, denoted by $WSER2_{1,j}$ and $WSER3_{1,j}$, $1 \le j \le 3M$.
- We denote $WU1_1 = \lfloor WSER2_{1,1} \times 10 \rfloor \bmod 10$, $WU1_2 = \lfloor WSER2_{1,2} \times 10 \rfloor \bmod 10$ and $WU1_3 = \lfloor WSER2_{1,3} \times 10 \rfloor \bmod 10$.
- Based on the Eq. (13), we calculate $w'_{1,1} = f(WU1_1, WU1_2, WU1_3)$. $w'_{1,1}$ is the watermark extracted from the first three fragments $WA2_{1,1}$, $WA2_{1,2}$ and $WA2_{1,3}$.
- Repeat the steps, we can extract all the watermark from the fragments belonging to $WA2_1$ and $WA3_1$, which are denoted by $W'_1 = \{w'_{1,j} \,|w'_{1,j} \in [0,9], 1 \le j \le M\}$ and $W^*_1 = \{w^*_{1,j} \,|w^*_{1,j} \in [0,9], 1 \le j \le M\}$.

Step 2. We identify the frame WA_1 is intact or not. If $W'_1 = W^*_1$, we regard the frame WA_1 is intact. And then we can reconstruct the frame number based on W'_1 or W^*_1, by using the Eq. (11). Otherwise, it indicates that the signal WA_1 has been tampered.

For the intact data, we decrypt it to get the audio signal by using the method in Section III. For the attacked encrypted data, we use the following methods to tamper recovery.

2) TAMPER RECOVERY

We suppose that the L/P samples belonging to WA_i have been tampered. Then we move the window and verify the next L/P samples in the same way, until we find the L/P continuous samples which can pass the verification, denoted by $WA_{i'}$. Because WA_i and $WA_{i'}$ are both intact, we can extract the frame number from WA_i and $WA_{i'}$, denoted by i and i'. Based on the scheme proposed, the difference between i and i' is the frame number of the attacked frames. The tamper location method is shown in Fig. 11.

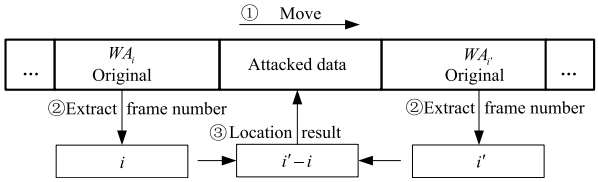


FIGURE 11. The method of tamper location identification.

For the attacked frame, denoted by WA^a, we reconstruct the signal approximately by using the following steps.

Step 1. We firstly obtain the frame number of the attacked signal WA^a by using the method shown in Fig. 11, which is denoted by i_a.

Step 2. According to the scrambling method based on the Eq. (8), we can obtain the embedding location of compressed data corresponding to the attacked signal WA^a. That is, the embedding position of the compressed data of the i_a-th frame.

Step 3. We suppose that the compressed data of the i_a-th frame is embedded into the i'_a-th frame $WA_{i'_a}$. We select the first n samples of $WA_{i'_a}$, denoted by $\{a_1, a_2, \cdots a_n\}$, from which we extract the compressed data denoted by $C_{i'_a} = \{c_1, c_2, \cdots c_n\}$ by using the following method.

$$c_l = sign(c_l) \times \left(\frac{\lfloor |a_l \times 100| \rfloor \bmod 10}{10} + \frac{\lfloor |a_l \times 1000| \rfloor \bmod 10}{100} \right) \tag{15}$$

where $sign(c_l) = 1$, when $\lfloor |a_l \times 10| \rfloor \bmod 2 = 0$; and $sign(c_l) = -1$, when $\lfloor |a_l \times 10| \rfloor \bmod 2 = 1$,

TABLE 3. The quantization method based on the watermark bit.

$n_1 - Z$	$QSER2_{i,1}$	$QSER2_{i,2}$	$QSER2_{i,3}$
0	$SER2_{i,1}$	$SER2_{i,2}$	$SER2_{i,3}$
1	$U0_1 + (U1_1 + 1)/10$ $+U2_1/100 + U3_1/1000$	$SER2_{i,2}$	$SER2_{i,3}$
2	$SER2_{i,1}$	$U0_2 + (U1_2 + 1)/10$ $+U2_2/100 + U3_2/1000$	$SER2_{i,3}$
3	$SER2_{i,1}$	$SER2_{i,2}$	$U0_3 + (U1_3 + 1)/10$ $+U2_3/100 + U3_3/1000$
4	$U0_1 + (U1_1 + 1)/10$ $+U2_1/100 + U3_1/1000$	$SER2_{i,2}$	$U0_3 + (U1_3 + 1)/10$ $+U2_3/100 + U3_3/1000$
5	$SER2_{i,1}$	$U0_2 + (U1_2 + 1)/10$ $+U2_2/100 + U3_2/1000$	$U0_3 + (U1_3 + 1)/10$ $+U2_3/100 + U3_3/1000$
6	$U0_1 + (U1_1 - 1)/10$ $+U2_1/100 + U3_1/1000$	$SER2_{i,2}$	$U0_3 + (U1_3 - 1)/10$ $+U2_3/100 + U3_3/1000$
7	$SER2_{i,1}$	$SER2_{i,2}$	$U0_3 + (U1_3 - 1)/10$ $+U2_3/100 + U3_3/1000$
8	$SER2_{i,1}$	$U0_2 + (U1_2 - 1)/10$ $+U2_2/100 + U3_2/1000$	$SER2_{i,3}$
9	$SER2_{i,1}$	$SER2_{i,2}$	$U0_3 + (U1_3 - 1)/10$ $+U2_3/100 + U3_3/1000$

$1 \leq l \leq n$. Based on the compressed data extracted, we reconstruct the L/P length signal. The method is to insert some 0 amplitude samples between two adjacent samples belonging to $C_{i'_a}$. After that we substitute the attacked frame by using the reconstructed frame. Then we can get the L length reconstructed data.

Step 4. We decrypt the L length reconstructed data by using the decryption method shown in III. So, the scheme proposed can approximately reconstruct the attacked signal, under the condition that the expression meaning of the recovered signal is the same as that of the original one.

V. PERFORMANCE ANALYSIS AND EXPERIMENTAL RESULTS

In this section, we select 500 WAVE format audio as the test signals, including 200 pieces of popular music and 300 recorded signals (150 telephone recordings and 150 live recordings). The reason is that the content of the recorded audio is usually the privacy of the user. In order to protect their privacy, users will adopt some feasible methods. Therefore, we chose the recorded audio as the test signal. All the audio are 16-bit quantified mono signals and sampled at 44.1 kHz. Each audio has a duration of 20 seconds. The parameters used are L=882000, P=20, M=6, n=1980, k_1=0.287, k_2=0.865, k_3=0.902, $\mu_1 = 3.97$, $\mu_2 = 3.89$, $\mu_3 = 3.73$,. The performance analysis and experimental results are shown as follows.

A. EMBEDDING CAPACITY

In this paper, we cut the host audio into P frames and split each frame into three segments. Then we embed the n length compressed data into the first segment and the M length decimal frame number into the second and the third segment, respectively.

For the frame number i, we mapped it to the sequence $FN_i = \{n_1, n_2, \cdots, n_M\}$, where n_l is a decimal integer and $n_l \in [0, 9]$, $1 \leq l \leq M$. If we convert n_l into a binary sequence, the sequence is 4 bits long. Based on the quantifying method shown in Table 3, we get the conclusion that the embedding of n_l requires quantifying at most two features. That is we can embed 4 binary bits by quantifying at most two SER features. So, for the decimal frame number embedding method proposed in this paper, the embedding capacity(EC) is $8M$ bps.

In addition, we embed the n length compressed data into the first segment of all frames. Therefore, the scheme proposed in this paper has a larger embedding capacity greater than $8M$ bps. While, for the scheme in [4], one binary bit is embedded by quantifying two features. Besides, we also compared the embedding capacity of the proposed scheme with the state-of-the-art schemes [24], [37], [38]. The Comparison results are shown in Table 4. Based on the comparison results, we can conclude that the embedding capacity of scheme proposed has been significantly improved.

TABLE 4. The embedding capacity for different schemes.

Methods	EC(bps)
[4]	$3M$
[24]	$6M$
[37]	$2M$
[38]	$4M/9$
Our	$8M$

- In [4], Liu et al. proposed an audio watermarking scheme using patchwork framework in order to improve the robustness against recapturing and de-synchronization attacks. In the scheme, one watermark bit is embedded by quantifying the frequency-domain coefficients logarithmic mean features of the two adjacent segments.
- In [24], Hu and Lu proposed the fragile/semi fragile watermarking method used for tampering detection

and recovery. In process of embedding, watermark was embedded in the region with low energy of high frequency coefficients and high energy of low frequency coefficients respectively after 2-level DWT.

- In [37], Jiang et al. presented the audio watermarking algorithm against synchronization attacks by using the global characteristics and adaptive frame division. Experimental results revealed that the scheme exhibits effective robustness against de-synchronization attacks, such as jittering, time scale modification, and pitch shift modification.
- In [38], Wang et al. proposed an audio watermarking scheme robust against de-synchronization attacks based on pseudo-Zernike moment. In this paper, synchronization codes and watermark bits were embedded into the statistics average value of audio samples and the average value of modulus of the low-order pseudo-Zernike moments. The scheme is robust against the de-synchronization attacks. However, the embedding capacity of the proposed algorithm is low, for the embedding of a one bit watermark requires the quantification of nine consecutive segment features.

B. INAUDIBILITY AND RECONSTRUCT CAPABILITY

1) INAUDIBILITY OF THE DECRYPTED AUDIO

For audio watermarking technology, watermark embedding does not reduce the auditory quality of the audio signal. In this paper, we use subjective difference grades (SDG) [19] to test the inaudibility subjectively. The meaning of each score in SDG is listed in Table 5. The watermark bits are embedded into the encrypted data in this paper. In order to test the inaudibility, we firstly decrypt the data and then calculate the SDG values of the decrypted audio.

TABLE 5. Subjective difference grades.

SDG	Description of impairments	Quality
0	Imperceptible	Excellent
-1	Perceptible, but not annoying	Good
-2	Slightly annoying	Fair
-3	Annoying	Poor
-4	Very annoying	Bad

We provide all the original and decrypted signals to the 15 audience. They give SDG values according to the scoring criteria. The maximum, average and minimum SDG values are shown in Table 6. In addition, we evaluate the SDG values of watermarked signals generated by the state-of-the-art schemes in [4], [24], and [37], under the condition that the embedding capacity is 8Mbps. Then we compare our method with that in [4], [24], and [37] and show the results in Table 6. Based on the comparison results, we can get that the watermark embedding of our scheme does not affect the auditory quality of the original audio, and the scheme proposed outperforms the state-of-the-art watermarking schemes in [4], [24], and [37].

TABLE 6. The SDG values for different schemes.

Methods	SDG		
	Maximum	average	minimum
[4]	-0.87	-1.10	-1.27
[24]	-0.98	-1.15	-1.31
[37]	-1.09	-1.24	-1.42
Our	-0.65	-0.82	-0.94

2) RECOVERING CAPABILITY

If the encrypted data stored on TPSC is attacked, the scheme can reconstruct the attacked signal by using the compressed data. Then we substitute the attacked frame by using the reconstructed one. Furthermore, after decryption, we can get the same meaning as the original audio from the reconstructed signal. For convenience, let's use a deletion attack as an example. For other types of attacks, the results are similar. For the 500 test signals, we firstly embed watermark into encrypted data. Then we delete a different number of samples. After that, we reconstruct the attacked samples by using the compressed data and decrypt the signal after being reconstructed.

We use both subjective and objective methods to test the auditory quality of reconstructed signals. For the subjective method, audiences listen to the original signal and reconstructed signal, and score according to the SDG evaluation criteria. For the objective method, we calculate the SNR of the reconstructed signal. SNR is defined in Eq. (16). According to the SDG and SNR value, we test the auditory quality of the reconstructed signal subjectively and objectively.

$$SNR = 10\lg\left(\frac{\sum_{l=1}^{L} a(l)^2}{\sum_{l=1}^{L} (a(l) - ra(l))^2}\right) \tag{16}$$

where $a(l)$ and $ra(l)$ are the l-th sample of the original and reconstructed audio.

- We delete 1/8, 1/4, 3/8 and 1/2 of the samples of encrypted audio. After that we reconstruct the deleted samples. We test the maximum SDG (M_SDG) of all the reconstructed signals. In addition, under the same conditions, we also test the M_SDG of the reconstructed signals by using other schemes [39], [40]. The test results are shown in Table 7, in which NSD represents the number of samples deleted. The results shown in Table 7 indicate that, when the number of deleted samples is less than 3/8, the SDG value of the algorithm in this paper is the maximum. So, our scheme has better reconstruction performance for the attacked signal and we can more easily identify the expression meaning of the original signals from the reconstructed ones by using the scheme proposed. However, when the number of deleted samples is more than half, the SDG value of the reconstructed signal by using the proposed scheme is smaller than that in [39]. The reason is that the compressed signal generate by the proposed scheme is obtained by sampling the encrypted data,

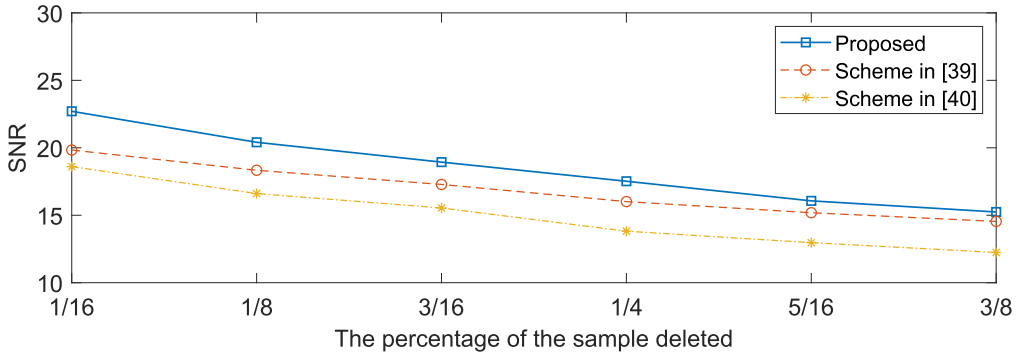


FIGURE 12. The SNR value of the reconstructed signals after different attacks.

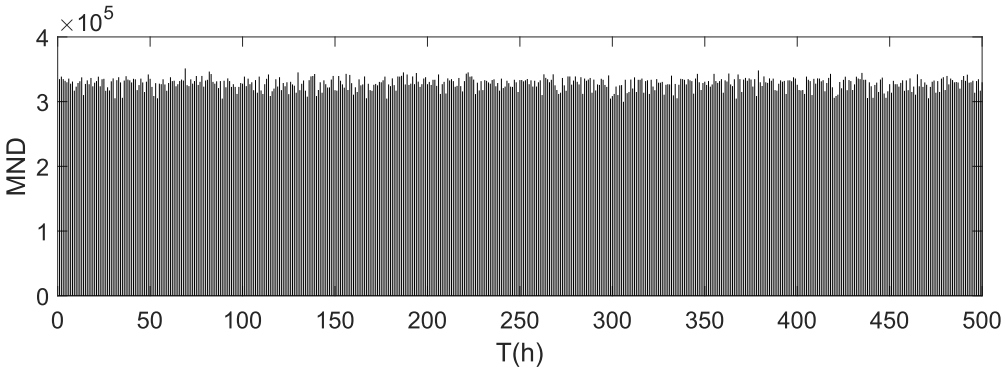


FIGURE 13. The maximum number of samples can be deleted.

and the quality of the reconstructed signal is not as good as that of the transform-based compression method [35].

- Similar to the subjective method, we delete different number of samples of encrypted data and reconstruct the attacked data, and then decrypt to obtain the reconstructed audio signal. After different attacks, we calculate the SNR values of all the reconstructed signals. Similarly, we also calculate the SNR values of the reconstructed signals by the schemes proposed in [39] and [40] under the same conditions. The mean value SNR is shown in Fig. 12. In audio watermarking schemes, SNR is often used to test the inaudibility of the embedded watermark. Moreover, when SNR is greater than 20, the watermark has good inaudibility. In this paper, we use SNR to objectively evaluate the auditory quality of reconstructed signals. The experimental results show that, if the only requirement is to be able to obtain the expressive meaning of the original audio from the reconstructed one, as long as the SNR value is greater than 15. Of course, in this case, there will be tolerable noise in the reconstructed signals. The results shown in Fig. 12 indicate that, when the proportion of samples deleted is less than 3/8, the SNR value of the reconstructed signals by using the proposed scheme is greater than 15. It also shows that, after different attacks, the SNR value of the reconstructed signals by the proposed scheme is greater than that of [39] and [40], especially [40].

We give the maximum number of samples that can be deleted for the 500 test signals, under the condition that the express meaning of the reconstructed signal is same as the original one(SDG>-1.5 and SNR >15). The SDG values are acquired from 15 listeners. It should be noted that, there may be tolerable noise in the reconstructed signals. The maximum number of samples that can be deleted for the 500 test signals is shown in Fig. 13, where MND represents the maximum number of samples can be deleted and $T(h)$ represents the h-th audio, $1 \leq h \leq 500$.

TABLE 7. The maximum SDG value of the reconstructed signals for different samples deleted.

Mthods	NSD	M_SDG
[39]	1/8 | 1/4 | 3/8 | 1/2	-0.73 | -0.96 | -1.38 | -2.37
[40]	1/8 | 1/4 | 3/8 | 1/2	-0.85 | -1.18 | -1.50 | -2.82
Our	1/8 | 1/4 | 3/8 | 1/2	-0.61 | -0.94 | -1.23 | -2.65

We denote R_c as the reconstruct capability of the scheme proposed, which is defined by the Eq. (17).

$$R_c = \frac{L_c}{L} \tag{17}$$

where L_c represents the maximum number of samples that can be deleted, under the condition that the listener can easily access the content from the reconstructed audio. Based on the results shown in Fig. 13, it indicates that the maximum number is about 328000 generally. So, the recovery capability of the proposed scheme is about $3/8$.

C. SECURITY

The main work of the scheme proposed includes two aspects. The one is that, we designed the encryption method to encrypt audio signal, aiming to protect the privacy of the audio uploaded to TPSC. The other is that, we presented the watermarking method to embed watermark bits into the encrypted data, used for forensics the data download from TPSC. So, we analyze the security of the scheme proposed from two aspects.

1) THE SECURITY OF THE ENCRYPTION METHOD

In this paper, we use multiple multiplication and scrambling operations to encrypt the original audio signal. For simplicity, we use different keys to generate different pseudo-random sequences. Firstly, we multiply the original signal by the one pseudo-random sequence and scramble the signal based another pseudo-random sequences. Then we cut the obtained data into frames and split each frame into two segments. The two segments of each frame are multiplied as the second segment in the frame. Finally, we scramble the obtained signal again to generate the final encrypted signal. For audio signals, on the one hand, scrambling can make audio signal close to noise signal. So scrambling operation can hide the expressed meaning of original audio. On the other hand, the multiplication operation can change the amplitude of the audio signal and further brings the samples closer to the random sequence, especially after multiplying many times.

In order to more clearly show the encryption method, in the following, we select an audio signal (shown in Fig. 14) and encrypt it by using the method. Firstly, we can generate three different pseudo-random sequences by using different keys. We multiply the signal by the first pseudo-random sequence, the result is shown in Fig. 15. After that, we scramble the signal in Fig. 15 and show the result in Fig. 16. The data in Fig. 16 is cut into frames and each frame is split into two segments. The two segments of each frame are multiplied as the second segment in the frame. The result is shown in Fig. 17. Finally, we scramble the obtained signal (shown in Fig. 17) again to generate the final encrypted signal, which is shown in Fig. 18. By comparing the original signal (shown in Fig. 14) and the encrypted data (shown in Fig. 18), it can be seen that the encrypted data is closer to a noise signal. In addition to that, we give the histograms of the original and the encrypted signals, which are shown in Fig. 19 and Fig. 20. It is obvious that the histograms of the two signals are quite different and the statistical characteristics of original signal are changed after encryption.

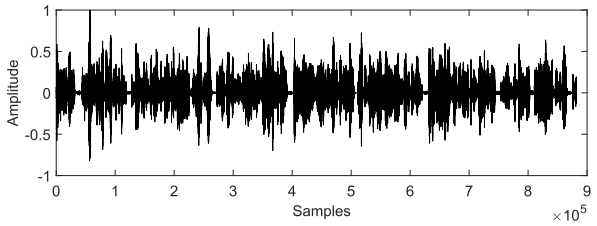


FIGURE 14. One audio signal selected randomly.

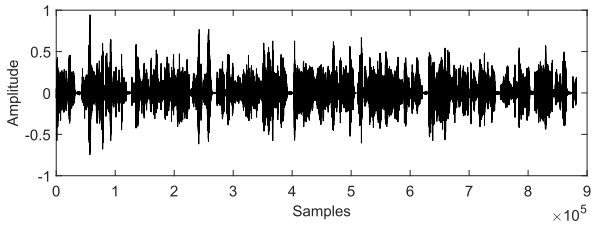


FIGURE 15. The audio multiplication.

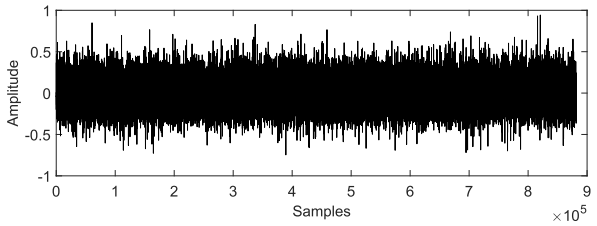


FIGURE 16. The signal after scrambling.

Based on the analysis above, we can get the conclusion that the encrypted signal is different from the original one and it's very hard for attackers to get the expressed meaning effectively from the encrypted data. So, the encryption method not only can hide the expression meaning of digital

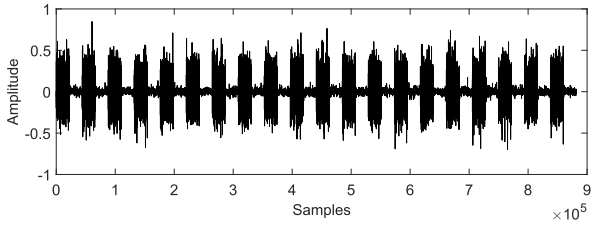


FIGURE 17. The result two segments of each frame multiplied.

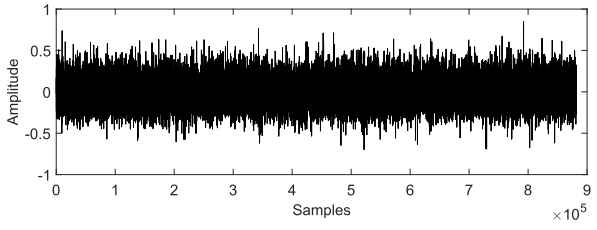


FIGURE 18. The encrypted data.

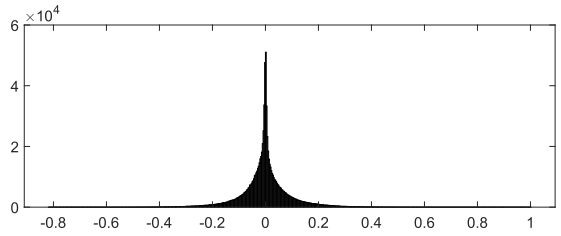


FIGURE 19. The histogram of the original audio.

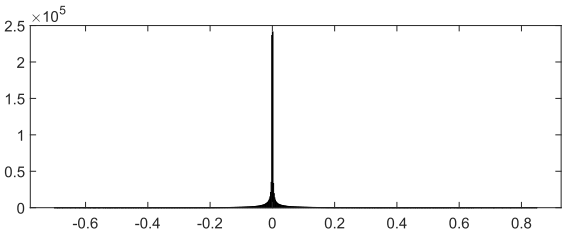


FIGURE 20. The histogram of the encrypted data.

audio and protect the audio privacy, but also can improve the attack resistance of the encrypted signal.

2) EMBEDDING METHOD

In this paper, we embed watermark bits into the SER features of the encrypted signal. To get the SER feature of one frame, we need to the encrypted signal and the pseudo-random sequence simultaneously. The pseudo-random sequence is generated based on the logistic chaotic mapping, in which the initial value is the secret key. So, without the secret key, it's difficult to get the pseudo-random sequence, and further to get the SER feature. If attackers generate the pseudo-random sequence by using other initial value, the probability that the watermark can be correctly extracted is only $1/10^M$. So, if one frame of watermarked signal is attacked, the probability that it can be detect is

$$A_s = 1 - \frac{1}{10^M} \tag{18}$$

where A_s represents the ability against attack, and M represents the length of watermark bits embedded into one frame.

D. EFFICIENCY AND DATA EXPANSION

Efficiency is critical when dealing with large amounts of data, such as for a large amounts of audio data uploaded to TPSC. For encryption and decryption of audio signals, the efficiency of the algorithm determines the length of time required for encryption and decryption. The complexity is a common method to measure the efficiency of an algorithm. The more complex the algorithm, the less efficient it is, and the longer it takes to complete the same task. Therefore, the schemes with high complexity are not suitable for processing large number of audio signals.

In addition, for some encryption schemes, the data size of the encrypted signal will be greatly increased compared with the original one, which is often referred to as data expansion (DE). The increase of data size will inevitably increase the burden of data transmission. In the following, we analyze the complexity and the data expansion of related schemes, which are shown in Table 8, in which, L represents the length of the original audio, g is the public key, R and P are the two random prime number. From the results shown in Table 8, we can get that the complexity of our scheme is $O(n)$, much lower than other schemes. In [8], for logistic map, authors generated the hash values of the plain audio signals and biometric images using SHA-256 hash algorithm to construct the initial values, which increases the time complexity. In [9], original audio data was encoded into other data range. It increases the burden of data transmission. Additionally, the DE of the encrypted data generated by the schemes in [13] and [41] are $g^L \times R^P$ and L^2. While, it is the same as the original one for our scheme. Based on the results in Table 8, it indicates that, the method proposed has the lowest complexity. And at the same time, compared with the original data, the amount of encrypted data does not increase, except for a few secret keys used to generate pseudo-random sequences. Therefore, in terms of improving time efficiency and controlling data expansion, the method proposed in this paper has certain advantages.

TABLE 8. The complexity and data expansion for different schemes.

Methods	Complexity	DE
[8]	$O(n^2)$	L
[9]	$O(n^{3/2})$	$2L$
[13]	$O(n^3)$	L^2
[41]	$O(g^n)$	$g^L \times R^P$
[42]	$O(n^2)$	$2L$
Our	$O(n)$	L

- In [8], Rahul *et al*. proposed a method for audio encryption based on chaos theory and user-biometric images. In the scheme, the chaotic sequences generated by the logistic map were used to create different initial values, which are constructed using the hash values of the plain audio signals and biometric images produced by the SHA-256 hash algorithm. The complexity of the SHA-256 hash algorithm is relatively high. And it increases the time complexity of the encryption algorithm.
- In [9], Albahrani et al. presented a substitution and permutation method for the encryption/decryption of two-channel audio files based on chaotic maps. The original audio data was encoded into other data range and translated to a binary sequence. In addition, to produce the keys, the authors designed the key generation algorithm based on the properties of the square root of large prime numbers and a hyper chaotic system. Although the proposed scheme can encrypt two-channel audio, the algorithm complexity of key generation is high, which is not conducive to the application of the algorithm.

- In [13], Shi et al. presented a speech homomorphic encryption scheme. In the scheme, the original digital speech with some random numbers selected was firstly grouped to form a series of speech matrix, which was encrypted by the encryption method proposed. The disadvantage of the algorithm is that the size of the encrypted data has grown too much. The increase of data size will inevitably increase the burden of data transmission.
- In [41], Xiang and Luo proposed the homomorphic encryption method for digital images, which has the ability of reversible data hiding. In addition, the scheme has lower computation complexity, higher security performance and better embedding performance. Unfortunately, the data size of the watermarked signal obtained by the scheme was increased greatly.
- In [42], Cao and Liu proposed a chaos-based dual-channel one-time one-key audio encryption scheme. A non-degenerate 2D integer domain hyper chaotic map was constructed, based on which authors constructed two keyed strong S-Boxes. In addition to audio data, this encryption method requires a lot of additional data and increases the transmission bandwidth.

E. TAMPER LOCATION AND TAMPER RECOVERY

There is a risk of being attacked for the encrypted data stored in TPSC. So, for safety's sake, it is necessary for users to forensics the data downloaded from TPSC. If the encrypted data is intact, the user can decrypt it directly. If the encrypted data is attacked, the user needs to locate the attacked frame and then reconstruct the attacked signal by using the compressed data. Furthermore, after decryption, we can get the same meaning as the original audio from the reconstructed signal.

The attack model of encrypted data stored in TPSC is shown in Figure 21. For audio signals, the common types of attacks include deletion attack, substitution attack and insertion attack. We take these three attack methods as examples to give the tamper location and tamper recovery methods in the following.

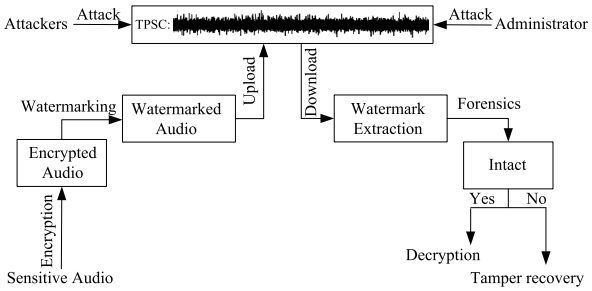


FIGURE 21. The attack model of encrypted data stored in TPSC.

1) DELETION ATTACK

We select one audio signal randomly, which is shown in Fig. 22. In addition, the encrypted version is shown in Fig. 23. Then we perform deletion attack on the encrypted data. The attacked signal is shown in Fig. 24. By using the forensic method, we select the first L/P samples and cut the samples into three segments. We divide the second and the third segment into $3M$ fragments, respectively. Then we calculate the SER feature of the fragments and verify the L/P samples are intact or not by using the method in Section IV. If the L/P samples are intact, we reconstruct the frame number of the L/P samples by using the Eq. (11). We extract all the frame number of the intact frame and show them in Fig. 25, from which we can get that the 5-th to 7-th frame have been attacked.

According to the selected parameter in this section, the compressed data of the 5-th to 7-th frames of the encrypted signal is embedded into the 1-th, 8-th and 18-th frame. Then we extract the compressed data and reconstruct the attacked signals by using the method in Section IV. The reconstructed signal is shown in Fig. 26. After that we divide the encrypted data by the pseudo-random sequence Y and perform anti-scrambling transformation to decrypt the reconstructed signal. Then we can get the decrypted version, which is shown in Fig. 27.

2) SUBSTITUTION ATTACK

We select 6000 samples randomly to substitute the samples belonging to the encrypted data shown in Fig. 23. The attacked signal is shown in Fig. 28. By using the forensic method, we can construct all the frame number of intact frame

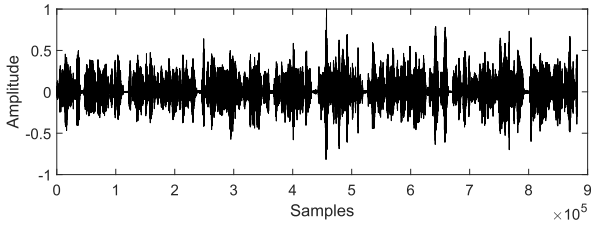


FIGURE 22. The audio selected randomly from the database.

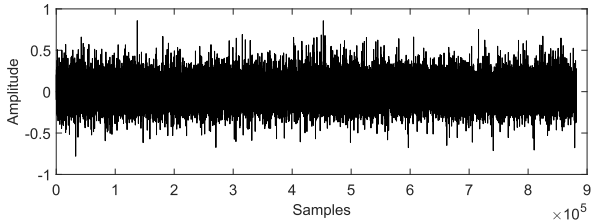


FIGURE 23. The encrypted version.

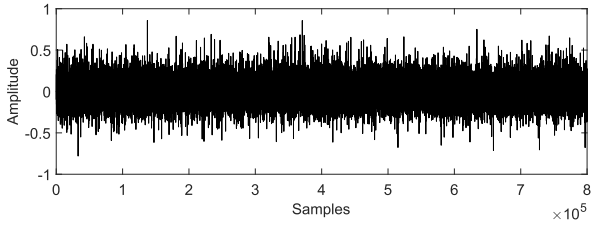


FIGURE 24. The deletion attacked data.

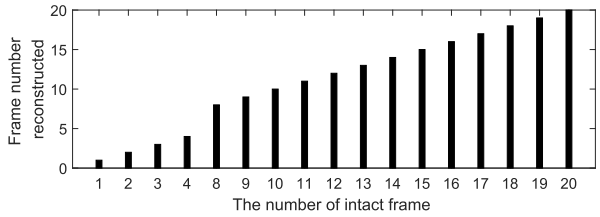


FIGURE 25. The frame number extracted from the data after being deletion attacked.

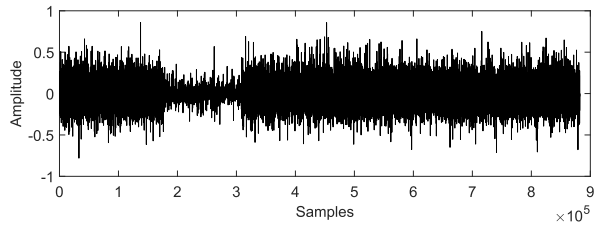


FIGURE 26. The reconstructed data for deletion attack.

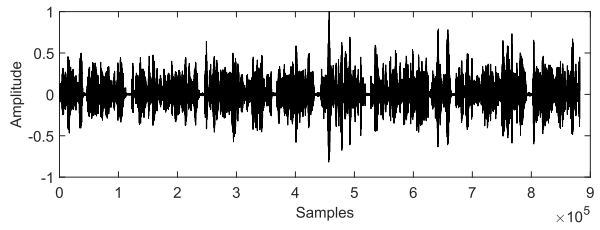


FIGURE 27. The decrypted signal for deletion attack.

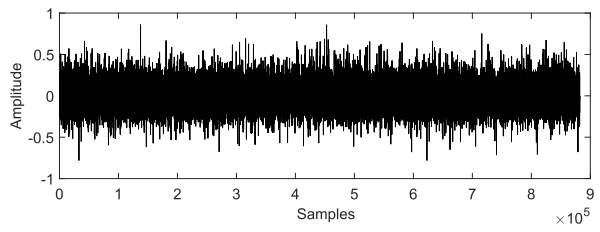


FIGURE 28. The audio substitution attacked.

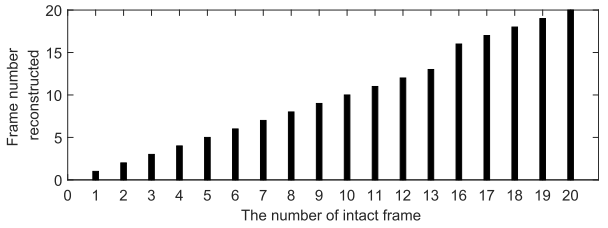


FIGURE 29. The frame number extracted from the substitution attacked data.

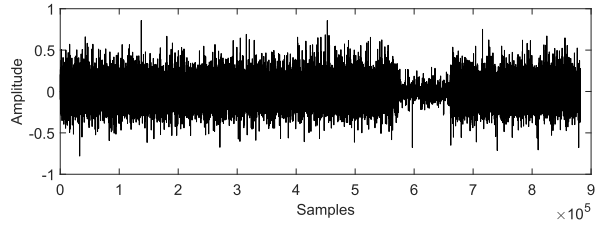


FIGURE 30. The reconstructed data for substitution attack.

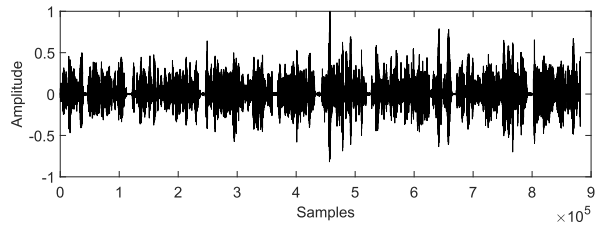


FIGURE 31. The decrypted signal for substitution attack.

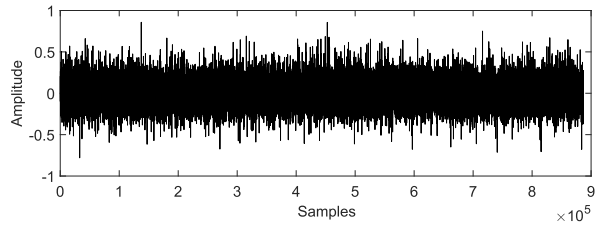


FIGURE 32. The audio insertion attacked.

and show the results in Fig. 29. From the result we can get that the 14-th and 15-th frame are attacked. Based on the scrambling system in this paper, the compress data of the 14-th and 15-th frame have been embedded into the 12-th and 10-th frame respectively. We extract the compressed data from the two frames, and then we reconstruct the attacked signal by using the compressed data. After that we can get the reconstructed signal shown in Fig. 30. Furthermore, after decryption, we can get decrypted signal, shown in Fig. 31.

3) INSERTION ATTACK

We randomly select some samples from other encrypted audio and insert them into the current signal(shown in Fig. 23). The attacked signal is shown in Fig. 32. As with deletion and substitution attacks, by using the forensics method proposed in this paper, we extract the frame number of all the intact frames and shown them in Fig. 33. The results show that the 19-th frame has been attacked. Based on the scrambling system in this paper, the compress data of the 19-th frame has been embedded into the 6-th frame. We extract the compress data and reconstruct the attacked frame and shown the signal in Fig. 34. Then we can get the decrypted signal after decryption, which is shown in Fig. 35.

For the deletion and insertion attack, they firstly destroys the synchronization of the encrypted data. And then it removes some meaningful samples. To get the meaning of the original signal, we extract the compressed data and then reconstruct the attacked frame approximately. After that we replace the attacked signal with the reconstructed one. The approximate reconstructed signal can be spreaded to different parts after anti-scrambling, which do not change the meaning of the original signal. For the reconstructed signals, shown in

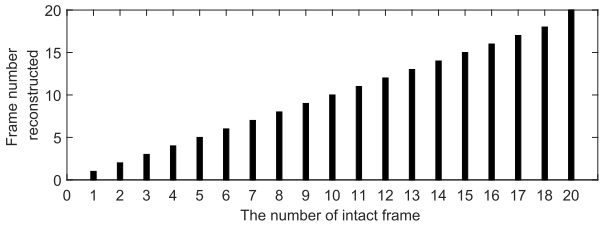


FIGURE 33. The frame number extracted from the insertion attacked data.

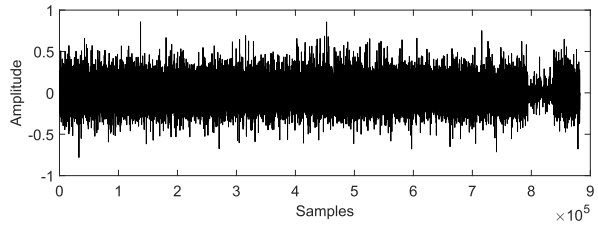


FIGURE 34. The reconstructed data for insertion attack.

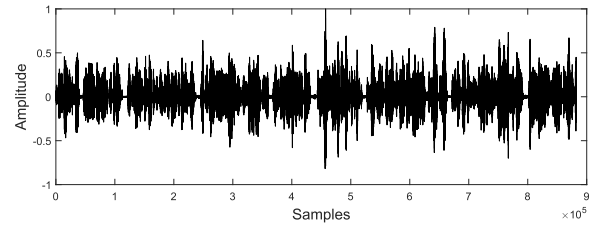


FIGURE 35. The decrypted signal for insertion attack.

Fig. 27, Fig. 31 and Fig. 35, we compared the auditory quality of the original signal with these reconstructed signals. The results, which are obtained from 15 listeners, show that the expression meaning obtained from the reconstructed signals is the same as the original audio.

The objective of this paper is to encrypt the audio signal and forensics the encrypted data. Based on this purpose and according to the above experimental results, we compare our scheme with others in three aspects: the ability to encrypt signals (AES); forensics on encrypted signals (FES) and the ability to reconstruct the attacked signal (ARAS). The comparison results are shown in Table 9.

TABLE 9. Comprehensive performance comparison results with other schemes.

Schemes	AES	FES	ARAS
[6]	Yes	No	Yes
[9]	Yes	No	Yes
[24]	No	No	Yes
[25]	Yes	Yes	No
[29]	No	No	No
[35]	No	No	Yes
[42]	Yes	No	No
Our	Yes	Yes	Yes

- In [6], authors proposed a chaos based dual-channel audio block encryption algorithm,which mainly adopts confusion and diffusion. The process is similar to other schemes, except for the scrambling system.
- In [9], the encryption and decryption method for two-channel audio signal based on chaotic maps was proposed. The scheme encoded the original audio into other data range. Each value in the resulted range was translated to a binary sequence and both substitution and permutation operations were accomplished using the chaotic state and chaotic parameters. Although it can encrypt the audio data, the algorithm complexity is high. In addition, if the encrypted data is attacked, it will be powerless.
- In [24], authors get the compressed version of original signal and embed the compressed signal into host audio. For the attacked frame, they extract watermark in areas which are not damaged and get the recovered signal after decompression. For the scheme, the extraction of compressed signal plays a crucial role in tamper recovery. But for encrypted data, it's hard to get the effective compressed signal by using the method. So, the method performs well for audio signals, but cannot be used for tampering recovery of encrypted audio.
- In [25], the homomorphic encryption and the batching technique SIMD in cloud computing were used for encrypted audio watermarking. Indeed, the scheme can also encrypt the audio signal and forensics the encrypted data. However, if the encrypted data is attacked, this algorithm cannot reconstruct the expression meaning of the original audio.
- In [29], the authors combined the quantization-embedding system and the weights of DWT coefficients with SNR to obtain an optimization model for watermarking. If we perform DWT on the encrypted data, the DWT coefficients can not reflect the characteristics of the encrypted data well. So, for the method, on the one hand, it cannot be used to encrypt digital audio and for forensics the encrypted data. On the other hand, the embedding domain is public, causing the scheme to be easily attacked.
- In [35], authors firstly get the compressed data of original signal based on DCT coefficients. Then they embed the compressed data by using the block-based method. If watermarked signal is attacked, users can extract the compressed data to reconstruct the attacked frames. Similarly, the method cannot be used for the encryption of audio signals and the forensics of the encrypted data.
- In [42], authors proposed a chaos-based dual-channel one-time one-key audio encryption scheme. The method can only encrypt audio signals, unable to verify the encrypted data intact or not, and cannot reconstruct the attacked data.

Based on the above analysis, we conclude that the proposed scheme can effectively encrypt the audio signal and embed the watermark bits into the encrypted data. The authorized

users can download the watermarked data and verify the data is intact or not. For the attacked data, they can reconstruct the audio signal approximately. The scheme proposed can effectively protect the audio signal stored in third-party storage centers and has a better comprehensive performance than the state-of-the-art watermarking schemes.

VI. CONCLUSION

In order to protect the audio signals stored in third-party storage centers, we proposed the encryption and forensics watermarking scheme. We defined a security feature (signal energy ratio) of digital audio and analyzed the properties of the feature. At the same time, we designed a watermark embedding method by quantifying the signal energy ratio feature. For the signal being uploaded to the third-party storage centers, we encrypted the signal by using scrambling and multiplication. For the encrypted data, we divided it into frames. And then, we compressed each frame to get the compressed data. After that we embedded the compressed data and frame number into each frame to get the watermarked data being uploaded to the third-party storage centers. For the authorized users, they can download the data from the third-party storage centers. Then they verify the authenticity of the downloaded data. If the downloaded data is intact, the users decrypted the data to get the audio signal. On the contrary, if the downloaded data has been attacked, the users can reconstruct the audio signal. In addition, the scheme proposed improves the security of digital audio stored on local servers and in TPSC.

The scheme proposed can indeed reconstruct the attacked audio signal. However, when a large number of samples are attacked (such as more than half of the sample points), it is difficult to reconstruct the expression content of the original audio signal. In the future, we will try to compress the audio signal into less data and embed the compressed data into the encrypted signal, in order to design the watermarking scheme with stronger recovery ability.

REFERENCES

[1] R. Giampiccolo, A. Bernardini, and A. Sarti, "Wave digital models of piezoelectric transducers for audio applications," *IEEE Sensors J.*, vol. 23, no. 1, pp. 389–400, Jan. 2023.

[2] F. S. Mobley, G. Bowers, M. Ugolini, E. Fox, and N. Gillespie, "Modeling aircraft similarity with musical auditory feature extraction," *Appl. Acoust.*, vol. 214, Nov. 2023, Art. no. 109689.

[3] H.-T. Hu, H.-H. Chou, and T.-T. Lee, "Robust blind speech watermarking via FFT-based perceptual vector norm modulation with frame self-synchronization," *IEEE Access*, vol. 9, pp. 9916–9925, 2021.

[4] Z. Liu, Y. Huang, and J. Huang, "Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1171–1180, May 2019.

[5] G. Zhang, L. Zheng, Z. Su, Y. Zeng, and G. Wang, "M-sequences and sliding window based audio watermarking robust against large-scale cropping attacks," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1182–1195, 2023.

[6] H. Liu, "Audio block encryption using 3D chaotic system with adaptive parameter perturbation," *Multimedia Tools Appl.*, vol. 82, no. 18, pp. 27973–27987, Jul. 2023.

[7] A. Kumar and M. Dua, "Audio encryption using two chaotic map based dynamic diffusion and double DNA encoding," *Appl. Acoust.*, vol. 203, Feb. 2023, Art. no. 109196.

[8] B. Rahul, K. Kuppusamy, and A. Senthilrajan, "Chaos-based audio encryption algorithm using biometric image and SHA-256 hash algorithm," *Multimedia Tools Appl.*, vol. 82, no. 28, pp. 43729–43758, Nov. 2023.

[9] E. A. Albahrani, T. K. Alshekly, and S. H. Lafta, "New secure and efficient substitution and permutation method for audio encryption algorithm," *J. Supercomput.*, vol. 79, no. 15, pp. 16616–16646, Oct. 2023.

[10] A. S. Alanazi, N. Munir, M. Khan, and I. Hussain, "A novel design of audio signals encryption with substitution permutation network based on the genesio-tesi chaotic system," *Multimedia Tools Appl.*, vol. 82, no. 17, pp. 26577–26593, Jul. 2023.

[11] Z. A. Hasan, S. M. Hadi, and W. A. Mahmoud, "Speech scrambler with multiwavelet, Arnold transform and particle swarm optimization," *Pollack Periodica*, vol. 18, no. 3, pp. 125–131, Sep. 2023.

[12] G. K. Mahato and S. K. Chakraborty, "A comparative review on homomorphic encryption for cloud security," *IETE J. Res.*, vol. 69, no. 8, pp. 5124–5133, Sep. 2023.

[13] C. Shi, H. Wang, Y. Hu, Q. Qian, and H. Zhao, "A speech homomorphic encryption scheme with less data expansion in cloud computing," *KSII Trans. Int. Inf. Syst.*, vol. 13, no. 5, pp. 2588–2609, 2019.

[14] M. R. Sudha, "A novel study of cloud computing and its security," *Int. J. Comput. Inf. Technol.*, vol. 15, no. 1, pp. 1–11, 2023.

[15] Y. Al-Issa, M. A. Ottom, and A. Tamrawi, "Survey on eHealth cloud security challenges," *Int. J. Appl. Eng. Res.*, vol. 13, no. 1, pp. 456–470, 2023.

[16] C. Wang, S. S. M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *IEEE Trans. Comput.*, vol. 62, no. 2, pp. 362–375, Feb. 2013.

[17] C. Wang, Q. Wang, K. Ren, N. Cao, and W. Lou, "Toward secure and dependable storage services in cloud computing," *IEEE Trans. Services Comput.*, vol. 5, no. 2, pp. 220–232, Apr. 2012.

[18] S. Gadde, J. Amutharaj, and S. Usha, "A security model to protect the isolation of medical data in the cloud using hybrid cryptography," *J. Inf. Secur. Appl.*, vol. 73, Mar. 2023, Art. no. 103412.

[19] J. He, Z. Liu, K. Lin, and Q. Qian, "A novel audio watermarking algorithm robust against recapturing attacks," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 18599–18616, May 2023.

[20] E. Salah, K. Amine, K. Redouane, and K. Fares, "A Fourier transform based audio watermarking algorithm," *Appl. Acoust.*, vol. 172, Jan. 2021, Art. no. 107652.

[21] C.-J. Chen, H.-N. Huang, S.-Y. Tu, C.-H. Lin, and S.-T. Chen, "Digital audio watermarking using minimum-amplitude scaling on optimized DWT low-frequency coefficients," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2413–2439, Jan. 2021.

[22] Z. Su, L. Chang, G. Zhang, J. Jiang, and F. Yue, "Window switching strategy based semi-fragile watermarking for MP3 tamper detection," *Multimedia Tools Appl.*, vol. 76, no. 7, pp. 9363–9386, Apr. 2017.

[23] J. Zhang, "Audio dual watermarking scheme for copyright protection and content authentication," *Int. J. Speech Technol.*, vol. 18, no. 3, pp. 443–448, Sep. 2015.

[24] Y. Hu, W. Lu, M. Ma, Q. Sun, and J. Wei, "A semi fragile watermarking algorithm based on compressed sensing applied for audio tampering detection and recovery," *Multimedia Tools Appl.*, vol. 81, no. 13, pp. 17729–17746, May 2022.

[25] R. Lai, X. Fang, P. Zheng, H. Liu, W. Lu, and W. Luo, "Efficient fragile privacy-preserving audio watermarking using homomorphic encryption," in *Proc. 8th Int. Conf. Artif. Intell. Secur. (ICAIS)*, Qinghai, China, Jul. 2022, pp. 373–385.

[26] K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski, "Robust speech watermarking by a jointly trained embedder and detector using a DNN," *Digit. Signal Process.*, vol. 122, Apr. 2022, Art. no. 103381.

[27] A. Kanhe and A. Gnanasekaran, "A blind audio watermarking scheme employing DCT-HT-SD technique," *Circuits, Syst., Signal Process.*, vol. 38, no. 8, pp. 3697–3714, Aug. 2019.

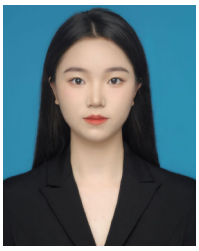
[28] M. Charfeddine, E. Mezghani, S. Masmoudi, C. B. Amar, and H. Alhumyani, "Audio watermarking for security and non-security applications," *IEEE Access*, vol. 10, pp. 12654–12677, 2022.

[29] M. Zhao, M. Li, X. Tong, J. Li, and S.-T. Chen, "Private data hiding system using state-switch DWT coefficients quantization on digital signal," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 36, no. 2, Feb. 2022, Art. no. 2258002.

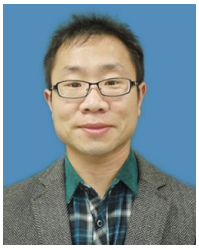
[30] Y. Xiang, I. Natgunanathan, S. Guo, W. Zhou, and S. Nahavandi, "Patchwork-based audio watermarking method robust to de-synchronization attacks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 1413–1423, Sep. 2014.

[31] J. Li and S. Xiang, "Audio-lossless robust watermarking against desynchronization attacks," *Signal Process.*, vol. 198, Sep. 2022, Art. no. 108561.
[32] J. Zhao, T. Zong, Y. Xiang, L. Gao, G. Hua, K. Sood, and Y. Zhang, "SSVS-SSVD based desynchronization attacks resilient watermarking method for stereo signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 448–461, 2023.
[33] H.-T. Hu and T.-T. Lee, "Frame-synchronized blind speech watermarking via improved adaptive mean modulation and perceptual-based additive modulation in DWT domain," *Digit. Signal Process.*, vol. 87, pp. 75–85, Apr. 2019.
[34] M. Yamni, H. Karmouni, M. Sayyouri, and H. Qjidaa, "Efficient watermarking algorithm for digital audio/speech signal," *Digit. Signal Process.*, vol. 120, Jan. 2022, Art. no. 103251, doi: 10.1016/J.DSP.2021.103251.
[35] Z. Liu, F. Zhang, J. Wang, H. Wang, and J. Huang, "Authentication and recovery algorithm for speech signal based on digital watermarking," *Signal Process.*, vol. 123, pp. 157–166, Jun. 2016.
[36] M. Salayani, B. Bakhtiari, and S. H. Ghafarian, "A robust zero-watermarking for audio signal using supervised learning," *Circuits, Syst., Signal Process.*, vol. 42, no. 6, pp. 3668–3705, Jun. 2023.
[37] W. Jiang, X. Huang, and Y. Quan, "Audio watermarking algorithm against synchronization attacks using global characteristics and adaptive frame division," *Signal Process.*, vol. 162, pp. 153–160, Sep. 2019.
[38] X.-Y. Wang, T.-X. Ma, and P.-P. Niu, "A pseudo-zernike moment based audio watermarking scheme robust against desynchronization attacks," *Comput. Electr. Eng.*, vol. 37, no. 4, pp. 425–443, Jul. 2011.
[39] S. Wang, W. Yuan, Z. Zhang, and L. Wang, "Speech watermarking based tamper detection and recovery scheme with high tolerable tamper rate," *Multimedia Tools Appl.*, vol. 83, no. 3, pp. 6711–6729, Jan. 2024.
[40] Q.-Y. Zhang and F.-J. Xu, "Encrypted speech authentication and recovery scheme based on fragile watermarking," *Telecommun. Syst.*, vol. 82, no. 1, pp. 125–140, Jan. 2023.
[41] S. Xiang and X. Luo, "Reversible data hiding in homomorphic encrypted domain by mirroring ciphertext group," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3099–3110, Nov. 2018.
[42] Y. Cao and H. Liu, "An audio encryption algorithm based on a non-degenerate 2D integer domain hyper chaotic map over GF(2^n)," *Multimedia Tools Appl.*, Mar. 2024, doi: 10.1007/S11042-024-18746-3.

JUNJIE HE was born in 1981. He received the B.S. degree from Beijing Institute of Technology, Luoyang, in 2003, and the M.S. degree from Xinyang Normal University, Xinyang, in 2011. His current research interests include cryptography and information security.



PEI ZHU was born in 2002. She received the bachelor's degree from Xinyang Normal University, where she is currently pursuing the master's degree. Her current research interests include multimedia information security and audio content authentication.



ZHENGHUI LIU (Member, IEEE) was born in 1983. He received the B.S. degree from Luoyang Normal University, Luoyang, in 2005, the M.S. degree from Xinyang Normal University, Xinyang, in 2010, and the Ph.D. degree from the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, in 2014. His current research interests include multimedia information security and audio content authentication.

YI CAO was born in 1998. She received the bachelor's degree from Xinyang Normal University, where she is currently pursuing the master's degree. Her current research interests include multimedia information security and audio content authentication.