Hindawi
Security and Communication Networks
Volume 2018, Article ID 7139391, 11 pages
https://doi.org/10.1155/2018/7139391

Research Article

A Relative Phase Based Audio Integrity Protection Method: Model and Strategy

Zhaozheng Li, Weimin Lei, Wei Zhang, and Kwanghyok Jo

School of Computer Science and Engineering, Northeastern University (NEU), Shenyang 110169, China

Correspondence should be addressed to Weimin Lei; leiweimin@ise.neu.edu.cn

Received 24 May 2018; Accepted 4 October 2018; Published 21 October 2018

Academic Editor: Emanuele Maiorana

Audio oriented integrity protection should consider the characteristics of audio signals based on the combination of audio application scenarios. However, in the current popular network interaction environment, traditional verification based solutions can no longer work. A kind of integrity protection scheme for audio business should be redesigned in these new scenarios. In this context, a method of audio integrity protection based on relative phase (RP-AIP) is proposed. Through the design of integrity object (I.O.), the integrity of audio can be abstracted as the completeness and accuracy of it. The I.O. is bound to the audio signal in a uniformly and randomly embedded manner, and the embedding rules are controlled by the relative phase characteristic of the audio itself. The model and strategy of RP-AIP are illustrated, and the corresponding process and algorithm are demonstrated as well. Simulation experiments illustrate the feasibility of the proposed solution and indicate the superiority of its performance.

1. Introduction

The rapid development of digital multimedia and Internet technology has facilitated the interaction of various multimedia services based on images, videos, and also audio. In particular, the widespread adoption of new generation instant messaging applications, such as “WhatsApp”, “WeChat”, and “LINE”, has made voice based communication more and more popular. Since voice messaging is more convenient and easier to use than text or other mediums, it has quickly become a widespread way of instant messaging products. Voice-class audio business has now become the mainstream of network interaction.

The voice messages can not only carry the speaker’s intention, but also identify the speaker. It is generally acknowledged that the voice-class interaction is more secure and reliable than the text-class. Indeed, it is even allowed to undertake high trust requirement functions. For example, voice authentication has been recommended in the process of transfer of accounts in “WeChat”. However, we should be aware that it is easy to change the original intention of the speaker by cutting or tampering operations on the voice messages. It must be recognized that malicious cutting or tampering attacks on voice-class audio business has become a new major issue troubling audio communication [1].

On this issue, audio integrity protection can effectively resist malicious tampering and cutting. However, currently available audio protection solutions are generally oriented towards copyright [2–4], and more details can be found in the survey done by R. D. Shelke et al. [5]. These studies are not only in time domain [6], but also in various transform domains to discuss how to protect the integrity of audio signals [7–10]. The issue of audio integrity protection for network interaction scenarios is rarely studied. When it comes to this kind of audio class business, it has some special features, such as the following.

(i) Streaming Characteristic

This is the most important and essential characteristic that is different from other interactive mediums. Compared to static texts and images, the streaming media characteristic of audio determines that the media should always accompany the process of integrity protection. In other words, integrity measurement based protection solutions (SHA-1, MD5, etc.) can no longer work. This characteristic poses new requirements and challenges for the integrity protection of audio.

How to design such a scheme that can accompany the media transmission and unlink the global association is the first essential issue to consider.

(ii) Imperceptibility

This characteristic is to ensure that the quality of experience (QoE) of users will not be degraded. Comparing to the human visual system (HVS), the human auditory system (HAS) is more sensitive and easy to perceive subtle changes in the audio signal. This is also why audio signal processing is more difficult than image and video signals. Time-domain-oriented signal processing method tends to have better synchronization, but at the same time, it is rarely used because of its serious signal distortion. This characteristic just contradicts the requirements of the previous one, which is another new difficulty we need to consider.

(iii) Self-Detectability

As streaming media, audio signal cannot stay: while listening while disappearing. Therefore, the integrity detection should be self-fulfilled, without relying on the original signal. That is to say, the integrity protection and detection process should be synchronized as well: while playing while detecting.

(iv) Underdetermination

Due to the unpredictability of transmission environment and noise effects, we do not emphasize the complete certainty of audio integrity protection. More specifically, we believe that the audio oriented integrity protection does not have to be completely determined. This is because our fundamental goal of audio integrity protection is to ensure the correctness and credibility of the audio information (e.g., the speaker's intention), rather than the audio signal medium itself.

For these reasons, a new method of audio integrity protection based on relative phase (RP-AIP) is proposed in this paper, which is used for protecting audio from being maliciously cut or tampered and ensuring the audio information can be correctly conveyed. Taking into account the above features, we first divided the audio into uniform segments and sampled them to get discrete audio signals. Then, we transformed them by DFT (Discrete Fourier Transform) and DCT (Discrete Cosine Transform), respectively. In the DFT domain, we can get the relative phase relation of the host segment and use this as a rule. In the DCT domain, we will embed a series of constructed eigenvalues as the integrity object (I.O.) into each segment furtherly. By this method, the integrity of the audio can be abstracted as the completeness and accuracy of the I.O. in each segment. This model has broken the direct relationship between integrity and audio media, so as to achieve the purpose of synchronous detection of streaming media.

The remainder of this paper is organized as follows. We introduced some related technologies about audio masking effect and the noise model in DCT. Then we illustrated the model and analysis of RP-AIP in Section 3. After that, we presented the process and strategy of RP-AIP and described the details of them in Section 4. In Section 5, we illustrated the feasibility of our model with simulation experiments. Finally, Section 6 concluded the paper and identified future directions.

2. Related Technologies

2.1. Audio Masking Effect. The HAS can be seen as a set of frequency analysis systems that contains about 26 band-pass filters, which can distinguish about 20Hz to 20kHz. However, HAS is difficult to distinguish between adjacent frequencies, that is, if a weak sound distributes in the adjacent frequency of a strong sound, the strong one will mask the weak one [11]. This is the audio masking effect, and the so-called adjacent frequency is called the critical band with "Bark" unit. The audio masking effect can be described by masking function (*MF*), which is related to the sound pressure (*SP*) of the audio and the distance (*d*) between the masked and the masker.

$$SP(k) = 10\lg\left[\frac{1}{N}\left\|\sum_{i=0}^{N-1} s(i)\, h(i) \exp\left(-j2\pi\frac{ik}{N}\right)\right\|^2\right] \quad (1)$$

where the $SP(k)$ is the sound pressure, $s(i)$ is a frame of audio signal, N is the number of samples per frame, and the $h(i)$ is a weighted *Hanning* window as shown in the following equation:

$$h(i) = \sqrt{\frac{8}{3}} \cdot \frac{1}{2}\left[1 - \cos\left(\frac{2\pi i}{N}\right)\right] \quad i \in [0, N-1] \quad (2)$$

Obviously, the closer to the masked and masker, the greater the masking effect. Otherwise, the weaker the masking effect is until the masked is out of the critical band of the masker. Furthermore, the masking function of the masker can be expressed as shown in (3). It can be seen that the masker has no effect on audio which is outside [-3, 8] Bark.

$$MF(i) = \begin{cases} 0 & d \in (-\infty, -3) \\ 17d - 0.4SP(i) + 11 & d \in [-3, -1) \\ [0.4SP(i) + 6]\, d & d \in [-1, 0) \\ -17d & d \in [0, 1) \\ -(d-1)[17 - 0.15SP(i)] & d \in [1, 8] \\ 0 & d \in (8, +\infty) \end{cases} \quad (3)$$

Audio masking effect is the basis of audio embedding. Through it, the embedding data can be embedded without the human ear being aware of it. Therefore, the embedding data should depend on the original audio, and the embedded distribution must be determined by the masking characteristic of the original audio signal.

2.2. Noise Model in DCT. The DCT transform can convert the frequency of the original audio signal to DCT coefficients, as shown in (4) [12]. The embedding of weak signals on strong signals can be regarded as changes in DCT coefficients, so the essence of the embedding data is equivalent to an additive noise.

$$S(k) = \delta(u) \sum_{n=0}^{N-1} s(n) \cos\left(\frac{(2n+1)k\pi}{2N}\right)$$
$$s(n) = \sum_{k=0}^{N-1} \delta(u) S(k) \cos\left(\frac{(2n+1)k\pi}{2N}\right) \qquad (4)$$
$$n, k \in [0, N-1]$$

where $s(n)$ is the discrete sequence of audio signal in time domain, $S(k)$ is the corresponding DCT coefficient sequence, N is the number of samples, and $\delta(u)$ is a weight factor as shown in the following equation;

$$\delta(u) = \begin{cases} \sqrt{\dfrac{1}{N}} & u = 0 \\ \sqrt{\dfrac{2}{N}} & otherwise \end{cases} \qquad (5)$$

Assuming that the data embedded on the i-th coefficient of $S(k)$ is $E(i)$, then the inverse transform signal $s'(n)$ can be calculated as shown in (6). In this way, we can find the changes between the original and transformed signals.

$$\begin{aligned} s'(n) &= \sum_{k=0}^{i-1} \delta(u) S(k) \cos\left(\frac{(2n+1)k\pi}{2N}\right) + \delta(i)[S(i) \\ &\quad + E(i)] \cos\left(\frac{(2n+1)i\pi}{2N}\right) + \sum_{k=i+1}^{N-1} \delta(u) S(k) \\ &\quad \cdot \cos\left(\frac{(2n+1)k\pi}{2N}\right) \xrightarrow{(2n+1)\pi/2N \triangleq \Psi} \\ &= \left[\sum_{k=0}^{i-1} \delta(u) S(k) \cos(k\Psi) + \delta(i) S(i) \cos(i\Psi)\right. \\ &\quad \left. + \sum_{k=i+1}^{N-1} \delta(u) S(k) \cos(k\Psi)\right] + \delta(i) E(i) \cos(i\Psi) \\ &= \sum_{k=0}^{N-1} \delta(u) S(k) \cos(k\Psi) + \delta(i) E(i) \cos(i\Psi) \\ &= s(n) + \delta(i) E(i) \cos(i\Psi) \end{aligned} \qquad (6)$$

Let $s'(n) - s(n) = \delta(i)E(i)\cos(i\Psi) \triangleq \varepsilon(i,n)$, then the noise effect in the time domain can be expressed as

$$\varepsilon(i,n) = \begin{cases} \sqrt{\dfrac{1}{N}} E(i) & i = 0 \\ \sqrt{\dfrac{2}{N}} E(i) \cos(i\Psi) & otherwise \end{cases} \qquad (7)$$

where $\Psi = (2n+1)\pi/2N$.

It can be seen from (6) and (7) that, in the DCT domain, the noise effect is only related to the embedded coefficient. Thus, we can evaluate the distortion of the original signal caused by the embedding property.

$$\begin{aligned} \sum_{n=0}^{N-1} \frac{\varepsilon(i,n)}{s(n)} &= \begin{cases} \displaystyle\sum_{n=0}^{N-1} \frac{\sqrt{1/N} E(i)}{s(n)} & i = 0 \\ \displaystyle\sum_{n=0}^{N-1} \frac{\sqrt{2/N} E(i) \cos(i\Psi)}{s(n)} & otherwise \end{cases} \\ &= \sqrt{\frac{2}{N}} \cdot \sum_{n=0}^{N-1} \frac{E(i)}{s(n)} \cdot \begin{cases} \dfrac{1}{\sqrt{2}} & i = 0 \\ \cos(i\Psi) & otherwise \end{cases} \end{aligned} \qquad (8)$$

In (8), it can be seen that $\sum_{n=0}^{N-1}(\varepsilon(i,n)/s(n)) \propto E(i)/s(n)$, and the scaling factor can be furtherly defined as $\rho(i)$ in the following equation:

$$\rho(i) = \begin{cases} \dfrac{1}{\sqrt{2}} & i = 0 \\ \displaystyle\sum_{n=0}^{N-1} \cos(i\Psi) & otherwise \end{cases} \qquad (9)$$

where $\Psi = (2n+1)\pi/2N$.

So far, we learn that $\rho(i)$ is the final effect caused by embedding on the audio signal and is also the sign of signal distortion.

3. Model and Analysis of RP-AIP

3.1. Mathematical Modeling. According to the audio noise model, we can establish the mathematical model as shown in (10). We expect to find such a binding relation that associates the I.O. with the audio signal. Then, the integrity of the host audio will be extracted and expressed accordingly.

$$\begin{aligned} &\exists \quad \otimes \\ &s.t. \quad S' = S \otimes E \\ &\qquad E \longleftarrow \text{integrity of } S \end{aligned} \qquad (10)$$

where S refers to the original audio, E refers to the I.O., and S' refers to the transformed audio. ⊗ here means a sort of binding relation we expected, which will be further introduced later. In addition, E should be unperceivable but robust to ensure that S will not be distorted.

3.2. Model Structure. The overall structure of the RP-AIP model is shown in Figure 1, and it consists of four parts: A: signal processing part, B: DFT part, C: DCT part, and D: I.O. generation part. Here, we would like to apply the random embedding approach as the binding relation. The audio signal is evenly divided into segments in Part A, and the segment is the basic unit for subsequent operations. Then perform DFT and DCT transforms on one audio segment in Parts B and C, respectively. The DFT transform is used to obtain the relative phase relation of the extreme points in its phase spectrum, and the DCT coefficient matrix is the final embedded target. The eigenvalues refer to the key encapsulated in the form of fuzzy vault, which are generated in Part D.

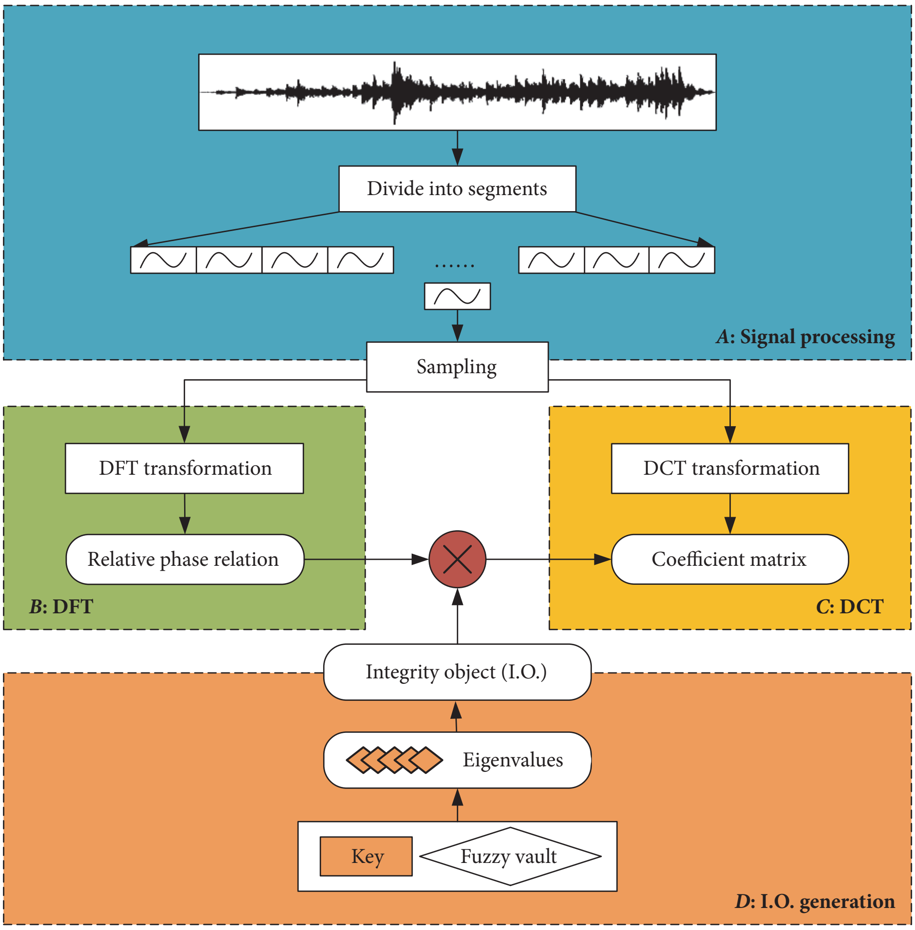


Figure 1: The overall structure of the RP-AIP model. The eigenvalues generated by the key and fuzzy vault are embedded into the DCT coefficient matrix, controlled by the relative phase relation indicated from DFT phase spectrum.

Since the segments are evenly divided, the I.O. can be equally distributed in the audio data, and the size of the fragments can be adjusted by actual requirements. In this way, the integrity of the audio can be abstracted as the completeness and accuracy of the I.O. by uniform embedding.

In addition, it is necessary to explain the function of Part *D*. Considering the issues of network congestion, accidental packet loss, and so forth caused by network anomalies, the received I.O. may not be exactly the same with the original. However, the loss and damage of I.O. caused by the transmission anomalies are generally accidental and random, but when it comes to malicious attack, it is generally deterministic and directional. Moreover, as stated earlier, the fundamental goal of audio integrity protection is to ensure the correctness and credibility of the audio information, rather than the audio signal medium itself. To this end, the use of fuzzy vault can solve this embarrassment, because the fuzzy vault can guarantee certain similarity rather than complete certainty.

3.3. Embedding Analysis. Through the introduction and analysis of audio integrity protection characteristics, we know that audio embedding should be transparent to the user. On the other hand, the distortion tolerance characteristic requires that the embedding process should have some robustness [13]. Interestingly, these two requirements are contradictory: to ensure transparency, it often means that the operation is usually based on the unimportant components of the audio signal, which are not sensitive to human ear, such as the high frequency regions. However, robustness is generally dependent on the important components of the audio signal, which are often sensitive to human ear, such as the low frequency regions [14].

After the audio signal sampled, it is transformed into a nonperiodic discrete signal in the time domain, and then a set of coefficient matrices composed of DCT coefficients can be obtained by the DCT transformation. Here, the low frequency signal energy is gathered in the upper left corner of the matrix, and the high frequency energy distributed in the lower right corner region is almost zero. Therefore, the eigenvalues can be embedded in these two regions: the low frequency region is robust because it carries a large amount of the main signal energy, while the high frequency region has good invisibility because of the absence of perception. Here comes the question, what kind of embedding strategy is the fairest? To this end, we propose to use the relative phase relation characteristic of the audio signal's frequency spectrum to control the embedding position.

4. Process and Strategy of RP-AIP

4.1. Audio Segmentation. According to the scheme design concepts, we need to embed the eigenvalues evenly in the audio signal, so this requires that the audio signal data should be divided evenly in advance.

The number of segments can be determined based on the elements of I.O. and the size of the audio. In general, assuming that there are n elements in the I.O., then we will divide the audio signal data into n segments as well. Thus, it can be ensured that the whole I.O. can be completely embedded and bound with the audio.

4.2. Relative Phase Invariance. Through the early part we learnt that the human ear is very sensitive to relative phase changes of audio signal but lacks the ability to perceive and distinguish the absolute phase. Therefore, for a certain audio signal, the phase component is more important than the amplitude component, and any change or damage to the phase component will cause unacceptable distortion of the audio quality or even completely destroy the audio information. At the same time, the communication theory also declares that the phase modulation has strong robustness to the noise signal.

More specifically, by DFT transform, the phase distribution of a segment of the audio signal is easy to obtain. There is only one maximum point and one minimum point in the phase spectrum as shown in Figure 2, and their relative positions are fixed (if there are multiple extreme points, the first one prevails). As mentioned above, the relative phase will maintain consistency regardless of any transformation. This characteristic offers the possibility to random embedding of the I.O.

4.3. Random Embedding Strategy. We will propose a strategy to determine the embedding position according to the position of the extreme points in the phase spectrum. Since the extreme points are random in the phase spectrum, the choice of the embedding position controlled by them is also random. Therefore, this strategy will guarantee the mutual balance between the transparency and robustness as previously described. Furthermore, the characteristic of relative phase invariance in frequency spectrum is unique because only the audio segment is involved, which guarantees the self- detectability as well.

An 8×8 coefficient matrix D can be obtained from a DCT transformation of an audio signal $s(n)$, as shown in Figure 3. The most robust embedding method recommended is embedding at Position (0,0), while the most transparent one is at Position (7,7). Embedding at Position (0,0) can improve robustness but adds more perceptual distortion, while embedding at Position (7,7) can improve imperceptibility but offers less robustness.

To get a good compromise between these parameters, the mid-band coefficients such as (2,2) (L) and (4,4) (H) are

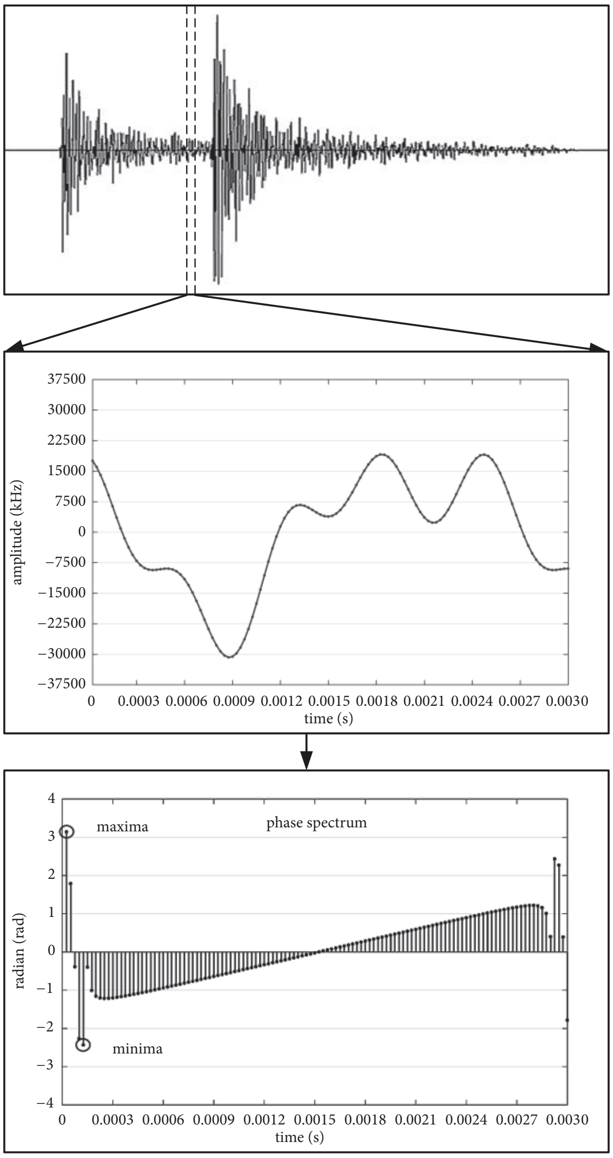


Figure 2: The phase spectrum of a segment of the audio signal. The audio sampling frequency is 44100kHz and the length is 0.5 seconds. The length of the intercepted fragment is 0.003 seconds, containing 120 sampling points.

	0	1	2	3	4	5	6	7
0								
1								
2			L					
3								
4					H			
5								
6								
7								

Figure 3: An 8×8 coefficient matrix of DCT. Position L represents the low frequency region, and position H represents the high frequency region, which cannot be perceived by human ear.

```
receiving the key vaults as $kv\ (i)$;
set the number of $kv\ (i)$ as $n$;
divide the audio signal equally into $n$ segments as $s\ (j)$;
sampling the $s\ (j)$ to $c\ (j)$;
$\forall c(j)$, initialize $j$=1;
while! end do
    if $j \leq n$ then
    calculate the phase spectrum $\varphi(\omega)$ of $s(j)$;
    determine the position of the maxima and minima in $\varphi(\omega)$;
    calculate the coefficient matrix $C\ (j)$ by $8 \times 8$ DCT of $c(j)$;
        if the maxima is in front of the minima then
            embedding the $kv\ (j)$ at $C\ (2,2)$;
        else
            embedding the $kv\ (j)$ at $C\ (4,4)$;
        end if
    end if
end while
```

Algorithm 1: The algorithm of random embedding strategy based on relative phase.

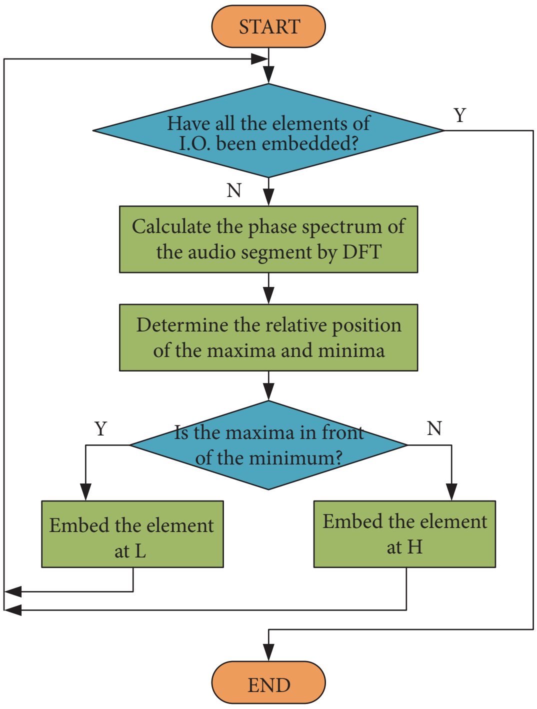

START
Have all the elements of I.O. been embedded?
Y
N
Calculate the phase spectrum of the audio segment by DFT
Determine the relative position of the maxima and minima
Y
Is the maxima in front of the minimum?
N
Embed the element at L
Embed the element at H
END

Figure 4: The flow chart of the random embedding strategy. The choice of the embedding position between L and H is controlled by the relative phase relation of the host audio segment.

selected as the embedding position instead. Position L or H is the Boolean choice for random embedding. The so-called Boolean choice means that a choice must be made from these two positions and the selection process is random. Based on the relative phase relation discussed above, a random embedding strategy is proposed as follows. The flow chart of this strategy is shown in Figure 4, and its algorithm is illustrated in Algorithm 1.

(i) If the maximum point of the audio segment appears before the minimum point, the corresponding eigenvalue will be embedded into the low frequency region (Position L).

(ii) Otherwise, the corresponding eigenvalue will be embedded into the high frequency region (Position H).

After determining the embedding position, the next step is to construct a fuzzy vault containing the secret sequence. On the finite field F, the secret sequence $key \in F$. By an encrypting set P ($P \in F$), the key can be packaged into the vault of P to generate the I.O. of the sender.

After embedding the I.O. into the audio, a new DCT coefficient matrix D' is generated. By making a DCT inverse transformation on D', we can obtain a new audio signal $s'(n)$ containing I.O. as well. Finally, $s'(n)$ is the transformed audio signal that we exactly expect.

4.4. Detection and Extraction. The detection and extraction of I.O. can be achieved by the difference between D and D'. When the receiver receives I.O., it is essentially a copy Q of the fuzzy vault P containing the secret sequence. On the finite field F, by matching the Q ($Q \in F$) and P, the key can be parsed from the vault to obtain the I.O. of the receiver. After the I.O. are recovered, the integrity of the audio signal can be determined according to the completeness and accuracy of them furtherly. Figure 5 shows the reconstruction and detection process.

In Figure 5, (a) shows the reconstruction process of the host audio and $s'(n)$ is the reconstructed one which contains the I.O. (b) is the detection and extraction process of I.O. $\tilde{e}_v$ is the eigenvalues and $\tilde{E}_v$ is its DCT transformation. $\tilde{e}'_v$ is the extracted eigenvalues and $\tilde{E}'_v$ is its DCT transformation. By matching the similarity between $\tilde{e}'_v$ and $\tilde{e}_v$, the similarity between the received and original audio can be evaluated. According to the features of the extracted I.O., it can be found

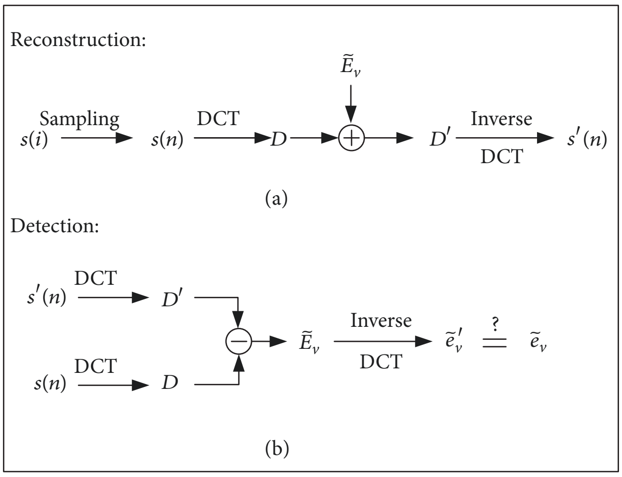


Figure 5: The process of the reconstruction and detection of the eigenvalues. (a) is the reconstruction process and (b) is the detection process.

whether the audio is cut or tampered. More details can be found in the experimental section.

4.5. Integrity Evaluation. After the I.O. is extracted, the similarity of $\widetilde{e}_v$ and $\widetilde{e}'_v$ can be calculated by their cosine distance d, and the integrity of the host audio I^{au} can be evaluated based on this, as shown in (11). It can be seen that when δ takes 1, I^{au} is a decimal within $(0, 1)$. The greater the value of I^{au}, the higher the similarity, also, the better the audio integrity.

$$I^{au} = \delta \cdot d = \delta \cdot \frac{\sum_{i=0}^{N-1} \sum_{i=0}^{N-1} \widetilde{e}(i) \widetilde{e}'(i)}{\sqrt{\sum_{i=0}^{N-1} \widetilde{e}^2(i)} \cdot \sqrt{\sum_{i=0}^{N-1} \widetilde{e}'^2(i)}} \tag{11}$$

where d is the cosine distance of $\widetilde{e}_v$ and $\widetilde{e}'_v$, δ is a weight coefficient, and the default value is 1. $\widetilde{e}(i)$ is the i-th element of $\widetilde{e}_v$ and $\widetilde{e}'(i)$ is the i-th element of $\widetilde{e}'_v$.

5. Experiments and Evaluations

5.1. Audio Quality Criterion. The signal to noise ratio (SNR) is the widely approved and used audio quality criterion for the evaluation of audio signal transformation, as shown in the following equation:

$$\mathrm{SNR} = 20 \log \left\{ \frac{\sqrt{\sum_{n=0}^{N-1} s^2(n)}}{\sqrt{\sum_{n=0}^{N-1} \left[s(n) - s'(n) \right]^2}} \right\} \tag{12}$$

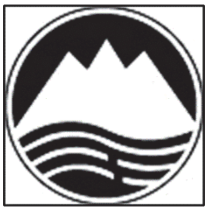


Figure 6: The black-and-white image of 100×100 pixels as I.O. in these experiments.

where $s'(n)$ is the transformation of $s(n)$.

$$\begin{aligned}
\mathrm{SNR} &= 20 \log \left\{ \frac{\sqrt{\sum_{n=0}^{N-1} s^2(n)}}{\sqrt{\sum_{n=0}^{N-1} \left[s(n) - s'(n) \right]^2}} \right\} \xrightarrow{\text{Eq.(6),Eq.(7)}} \\
&= 20 \log \left[\frac{\sqrt{\sum_{n=0}^{N-1} s^2(n)}}{\sqrt{\sum_{n=0}^{N-1} \varepsilon^2(i,n)}} \right] \\
&= 20 \log \sqrt{\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} \varepsilon^2(i,n)}} \\
&= 20 \log \sqrt{\sum_{n=0}^{N-1} \sum_{n=0}^{N-1} \frac{s^2(n)}{\varepsilon^2(i,n)}} \xrightarrow{\text{Eq.(9)}} \\
&= 20 \log \sqrt{\sum_{n=0}^{N-1} \frac{1}{\rho^2(i)}}
\end{aligned} \tag{13}$$

Therefore, after embedding the I.O. on the i-th coefficient of $s(n)$, the SNR can be calculated according to (6), (7), and (9), as shown in (13).

From (11), we can see that the SNR is only related to $\rho(i)$; that is to say, the audio quality after embedded eigenvalues is only related to the embedding position of its DCT coefficient i. This proves that the embedding position is crucial to the imperceptibility and robustness of the host audio in our RP-AIP model. Therefore, the embedding rule based on the relative phase of the audio signal itself is of extraordinary significance.

5.2. Embedding and Extraction of I.O. In this experiment, it is assumed that the I.O. is a black-and-white image as shown in Figure 6, and the host audio is a set of audio clips of the left channel as shown in Figure 7 and Table 1. Firstly, we uniformly embedded the I.O. in the host audio and then recovered it in the opposite way.

According to (12), the SNR of different audio types is calculated as shown in Table 1, which is in line with the requirements of robustness and imperceptibility.

5.3. Integrity Protection Performance. In the transmission process, the audio may suffer from network congestion, accidental packet loss, and other accidents caused by network

Table 1: The SNR of different audio types after embedding I.O.

	Audio clips		
	A_1	A_2	A_3
Audio types	A segment of voice message	A prompt tone of Windows 10	A piece of music
SNR	30.96	42.33	46.40

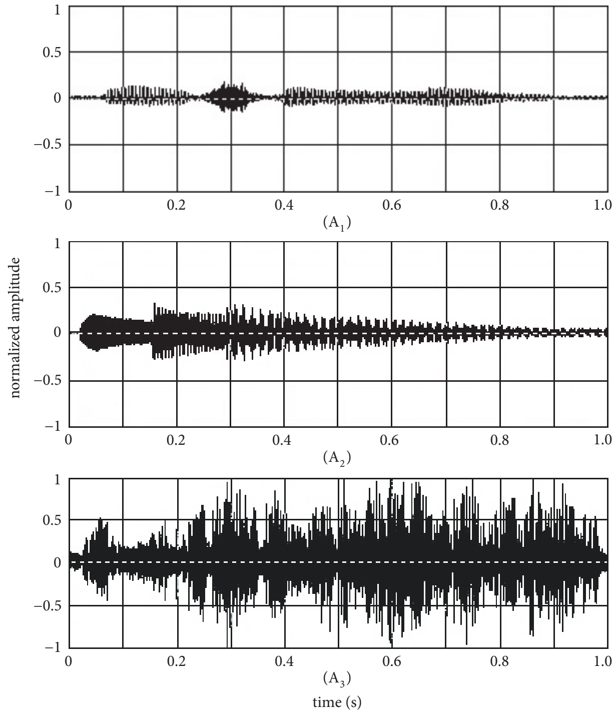


Figure 7: The host audio clips in the experiment. The audio sampling frequency is 22050 kHz and the length is 1.0 seconds. (A_1): a segment of voice message; (A_2): a prompt tone of Windows 10; (A_3): a piece of music.

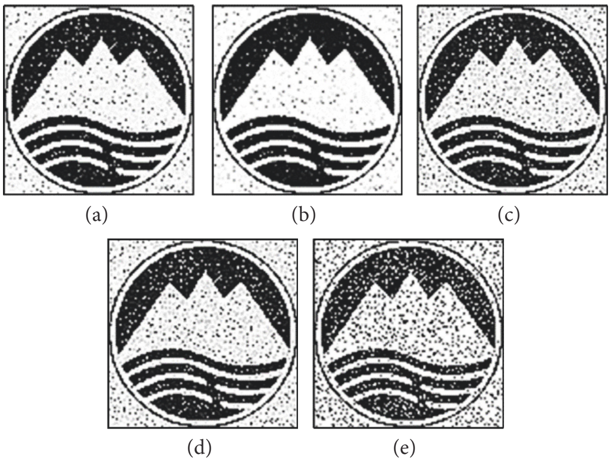


Figure 8: The simulation results of the recovered I.O. of different audio processing methods.

anomalies, even filtering, compression, and A/D conversion. Therefore, the first thing to consider is whether the model can effectively distinguish between such accidents and malicious attacks, such as cutting and tampering. As is stated above, the loss and damage of I.O. caused by the transmission anomalies are generally accidental and random, but deterministic and directional caused by malicious attacks. Thus, this issue can be judged from the recovered I.O.

5.3.1. Audio Processing. In this experiment, we simulated five common audio processing methods: smoothing/low-pass filtering (with 4kHz cutoff frequency), band-pass filtering (with 200-2kHz cutoff frequency), MPEG-1 (with compression ratio of 10.5:1 and 12:1) compression, and A/D conversion. The evaluation of audio signal distortion is indicated by the similarity between the original and recovered I.O. as well.

Figure 8 shows the simulation results. Among them, (a) is the recovered I.O. after a smoothing/low-pass filtering, (b) is the one after a band-pass filtering, (c) is the one after a MPEG-1 (with compression ratio of 10.5:1) compression, (d) is the one after a MPEG-1 (with compression ratio of 12:1) compression, and (e) is the one after an A/D conversion. The similarity between the original and recovered I.O. and the comparison with the other two common audio embedding methods (embedding the I.O. on 0×0 or $n \times n$ coefficient, which is called 0-coefficient or n-coefficient embedding for short) can be found in Table 2 and presented by Figure 9.

It can be seen from Figure 8 and Table 2 that the RP-AIP model can well adapt to the common audio processing; that is to say, the general audio processing method will not lead to the destruction and loss of I.O. It can be seen from Figure 9 that the RP-AIP model has similar robustness against the common audio processing compared with the 0-coefficient embedding model. In addition, as far as the smoothing/low-pass and band-pass filtering methods are concerned, the performance of band-pass filtering of RP-AIP is superior to other models significantly, thanks to the method of random embedding in the mid-band coefficient. In general, RP-AIP exhibits sufficient robustness for general audio processing methods.

5.3.2. Malicious Attack. In this experiment, we simulated three kinds of malicious attacks as cutting, tampering, and resampling. Moreover, in the type of tampering, it can be divided into self-media tampering and non-self-media tampering: self-media tampering refers to using the media itself for shifting, swapping, and so forth, while non-self-media tampering refers to using other media for replacement, coverage, and so forth. Resampling can also be divided into upsampling and downsampling: upsampling refers to

Table 2: The similarity ρ between the original and recovered I.O. in audio processing experiments.

No.	Processing methods	Similarity ρ		
		Random embedding (RP-AIP)	0-coefficient embedding	n-coefficient embedding
(a)	Original	1.00	1.00	1.00
(b)	Without processing	0.97	0.99	0.95
(c)	Smoothing/low-pass filtering	0.88	0.98	0.18
(d)	Band-pass filtering	0.90	0.85	0.84
(e)	MPEG-1 (with 10.5:1 ratio)	0.79	0.88	0.68
(f)	MPEG-1 (with 12:1 ratio)	0.81	0.85	0.64
(g)	A/D conversion	0.75	0.74	0.76

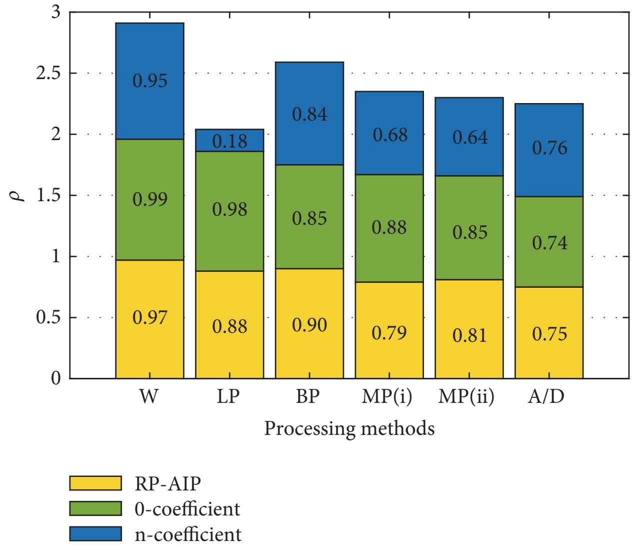
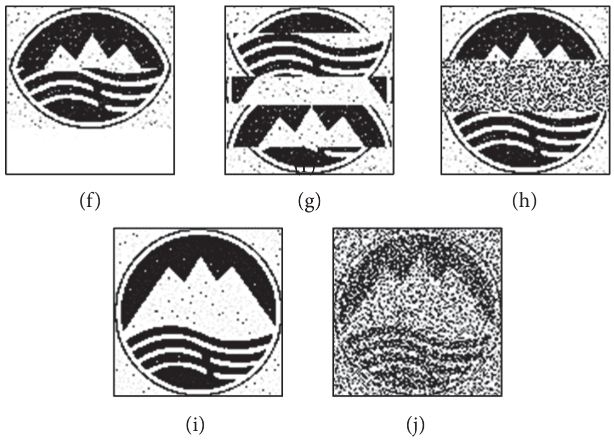


Figure 9: The comparison of the simulation results of different audio processing and embedding methods. W: without processing; LP: smoothing/low-pass filtering; BP: band-pass filtering; MP (i): MPEG-1 compression (with compression ratio of 10.5:1); MP (ii): MPEG-1 compression (with compression ratio of 12:1); A/D: A/D conversion.

interpolation expansion of the original audio, while downsampling refers to the extraction of the original audio. Cutting and tampering are undoubtedly malicious attacks on audio and attempt to change the speaker's intention. However, resampling attacks often only affect the length or quality of the audio, especially the upsampling, but cannot change the speaker's intention.

Figure 10 shows the simulation results. Among them, (f) is the recovered I.O. after a cutting attack, (g) is the one after a self-media tampering attack, (h) is the one after a non-self-media tampering attack, (i) is the one after an upsampling attack, and (j) is the one after a downsampling attack. The similarity between the original and recovered I.O. and the comparison with the other two common audio embedding methods as mentioned above can be found in Table 3 and presented by Figure 11.

As can be seen from Figure 10 and Table 3, in addition to the upsampling attack, the RP-AIP model can effectively


(f) (g) (h)

(i) (j)

Figure 10: The simulation results of the recovered I.O. of different malicious attacks.

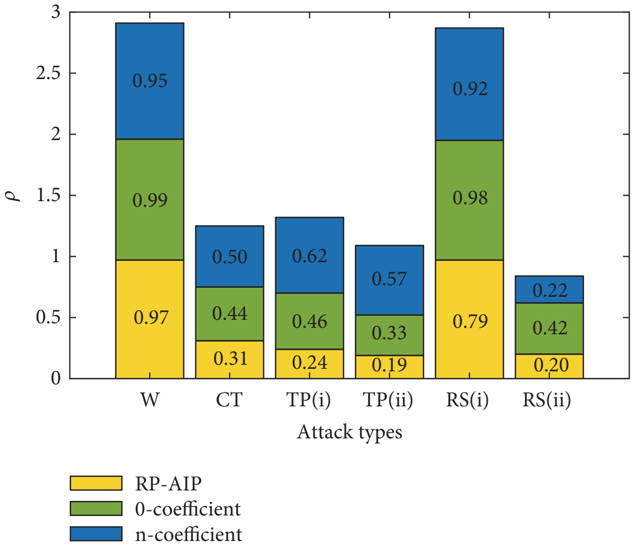


Figure 11: The comparison of the simulation results of different malicious attacks and embedding methods. W: without processing; CT: cutting attack; TP (i): self-media tampering attack; TP (ii): non-self-media tampering attack; RS (i): upsampling attack; RS (ii): downsampling attack.

Table 3: The similarity ρ between the original and recovered I.O. in malicious attack experiments.

No.	Attack types	Similarity ρ		
		Random embedding (RP-AIP)	0-coefficient embedding	n-coefficient embedding
(a)	Original	1.00	1.00	1.00
(b)	Without attacks	0.97	0.99	0.95
(h)	Cutting	0.31	0.44	0.50
(i)	Tampering (self-media)	0.24	0.46	0.62
(j)	Tampering (non-self-media)	0.19	0.33	0.57
(k)	Resampling (upsampling)	0.97	0.98	0.92
(l)	Resampling (downsampling)	0.20	0.42	0.22

detect and reflect malicious attacks. As stated earlier, the I.O. is the tangible representation form of the audio integrity; that is, changes in I.O. directly reflect changes in audio media. Therefore, we can evaluate the type and degree of the attack on the audio by observing I.O. In the case of (f), since cutting will directly lead to the loss of I.O., when the I.O. is reconstructed by 100×100 pixels, it will become an incomplete image. The missing part of I.O. indicates that the corresponding audio part has been cut off. In the same way, the tampered part appears as random noise or disordered as shown in (g) and (h), because the audio has suffered the corresponding tampering attack. However, for resampling attacks, the RP-AIP shows different results as shown in (i) and (j). This is because the upsampling tends to interpolate the original audio signal without causing loss of I.O., while downsampling is the extraction of the original audio signal, which directly leads to the loss and destruction of I.O. As stated in the Introduction, we are more concerned about whether the attack has tampered with the speaker's intention, but resampling is generally not.

It can be seen from Figure 11 that the RP-AIP model has definite better performance than the other two models. This is because attacks such as cutting and tampering are generally oriented to the audio signal's time domain rather than the frequency domain, so the damage to the low and high frequency parts is random. The embedding method with only consideration of low or high frequency can only guarantee one aspect, while the random embedding method in RP-AIP model can synthesize these two frequency parts. This is also the innovation and highlight of our work.

6. Conclusion

Audio integrity protection for network interaction environment is not only an effective way to protect audio information, but also an important link in building trusted communication. Due to real-time constraints of interaction scenarios, this kind of audio protection needs to reconsider some new characteristics, such as streaming characteristic, imperceptibility, self-detectability, and distortion tolerance, which are not covered by the traditional programs. This poses new challenges to the existing audio protection solutions.

To this end, a method of audio integrity protection based on relative phase (RP-AIP) is proposed in this work. In the RP-AIP model, we established the concept of integrity object (I.O.) in order to propose and abstract the integrity of the audio signal and then transform it into a tangible representation form. In addition, we also fully considered the characteristics of the audio signal in the DFT and DCT transform domain and used the frequency characteristics of the audio itself to guide the random embedding of I.O., thereby achieving blind detection and extraction of it. The tangible expression of the audio integrity and the relative phase based random embedding scheme are the highlight of our work.

The simulation experiments show that the RP-AIP model can guarantee the SNR of the audio signal, and the embedding of I.O. will not cause signal distortion. In addition, it has good adaptability to some common audio processing, such as smoothing/low-pass filtering, band-pass filtering, MPEG-1 compression with different ratios, and A/D conversion. However, aiming at malicious attacks such as cutting, tampering, and resampling, the RP-AIP model shows obvious superiority to other similar protection solutions. Of course, the research about audio integrity protection is still in its infancy, and more work will be done to make it more complete.

Data Availability

Some of the data is uploaded to the public server: https://doi.org/10.6084/m9.figshare.6382709.v1. In the file, you can find the demo audio and its spectrum data, the I.O. data, and the audio data used in the experiment.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China [No. 61671141, No. 61401081], the Liaoning Provincial Natural Science Foundation of China [No. 20180551007], and the Ministry of Education-China Mobile Scientific Research Funds [No. MCM20150103].

References

[1] R. F. Olanrewaju and O. Khalifa, “Digital audio watermarking; techniques and applications,” in *Proceedings of the 2012 International Conference on Computer and Communication Engineering, ICCCE 2012*, pp. 830–835, Malaysia, July 2012.

[2] J. Seok, J. Hong, and J. Kim, “A novel audio watermarking algorithm for copyright protection of digital audio,” *ETRI Journal*, vol. 24, no. 3, pp. 181–189, 2002.

[3] H. Yassine, B. Bachir, and K. Aziz, “A Secure and High Robust Audio Watermarking System for Copyright Protection,” *International Journal of Computer Applications*, vol. 53, no. 17, pp. 33–39, 2012.

[4] J. Haitsma, M. van der Veen, T. Kalker, and F. Bruekers, “Audio watermarking for monitoring and copy protection,” in *Proceedings of the ACM Workshops on Multimedia ACM*, pp. 119–122, 2000.

[5] R. D. Shelke and M. U. Nemade, “Audio watermarking techniques for copyright protection: A review,” in *Proceedings of the 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, ICGT-SPICC 2016*, pp. 634–640, December 2016.

[6] S. V. Dhavale, R. S. Deodhar, and L. M. Patnaik, “High Capacity Lossless Semi-fragile Audio Watermarking in the Time Domain,” in *A hierarchical CPN model for mobility analysis in zone based MANET*, pp. 843–852, 2012.

[7] S. V. Dhavale, R. S. Deodhar, D. Pradhan, and L. M. Patnaik, “State Transition Based Embedding in Cepstrum Domain for Audio Copyright Protection,” *IETE Journal of Research*, vol. 61, no. 1, pp. 41–55, 2015.

[8] H. H. Tsai, J. S. Cheng, and P. T. Yu, “Audio Watermarking Based on HAS and Neural Networks in DCT Domain,” *Eurasip Journal on Advances in Signal Processing*, vol. 3, no. 2, pp. 1–12, 2003.

[9] Y. Xiang, I. Natgunanathan, Y. Rong, and S. Guo, “Spread spectrum-based high embedding capacity watermarking method for audio signals,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2228–2237, 2015.

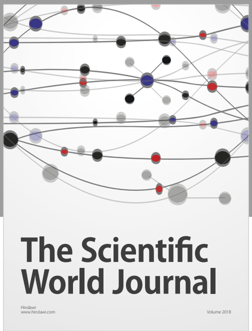
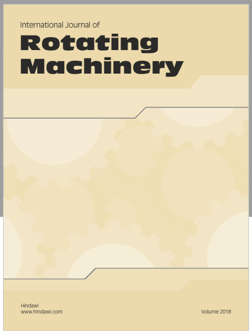
[10] X. Huang, A. Nishimura, and I. Echizen, “A Reversible Acoustic Steganography for Integrity Verification,” in *Digital Watermarking*, p. 305, Springer, Berlin, Heidelberg, Germany, 2011.

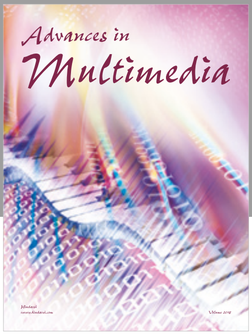
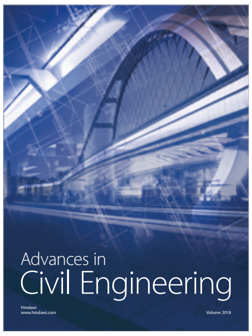
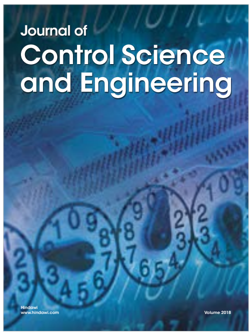
[11] H. Cao, Y. Wang, J. Li et al., “Audio data hiding using perceptual masking effect of HAS,” in *Proceedings of the International Symposium on Multispectral Image Processing and Pattern Recognition, International Society for Optics and Photonics*, 2007.

[12] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–512, 2000.

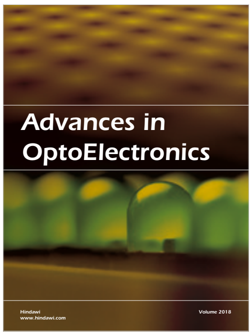
[13] S. V. Dhavale, R. S. Deodhar, D. Pradhan et al., “Robust Multiple Stereo Audio Watermarking for Copyright Protection and Integrity Checking,” in *Proceedings of the International Conference on Computational Intelligence and Information Technology IET*, pp. 9–16, 2014.

[14] L. Boney, A. H. Tewfik, and K. N. Hamdy, “Digital watermarks for audio signals,” in *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, pp. 473–480, IEEE, 2002.

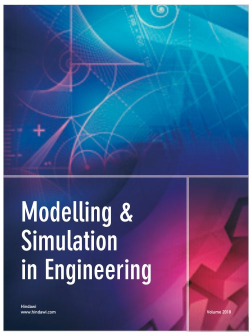
Journal of
Engineering

The Scientific World Journal

International Journal of
Rotating Machinery

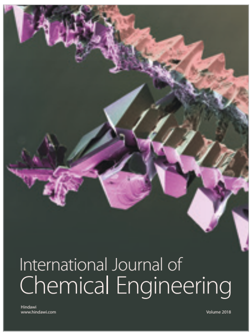
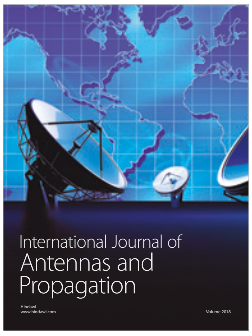
Journal of
Sensors

Advances in
Multimedia

Advances in
Civil Engineering

Journal of
Control Science and Engineering

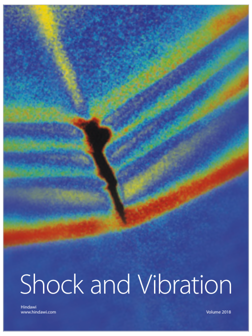
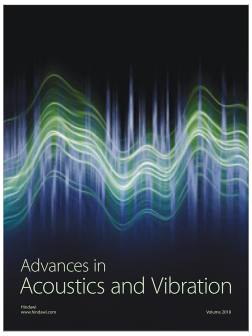
Journal of
Robotics

Journal of
Electrical and Computer Engineering

Advances in
OptoElectronics

VLSI Design

International Journal of
Navigation and Observation

Modelling & Simulation in Engineering

International Journal of
Aerospace Engineering

International Journal of
Chemical Engineering

International Journal of
Antennas and Propagation

Active and Passive Electronic Components

Shock and Vibration

Advances in
Acoustics and Vibration