Spread Spectrum Based High Embedding Capacity Watermarking Method for Audio Signals

Yong Xiang*, *Senior Member, IEEE*, Iynkaran Natgunanathan, Yue Rong, *Senior Member, IEEE*, and Song Guo, *Senior Member, IEEE*

***Abstract*— Audio watermarking is a promising technology for copyright protection of audio data. Built upon the concept of spread spectrum (SS), many SS-based audio watermarking methods have been developed, where a pseudonoise (PN) sequence is usually used to introduce security. A major drawback of the existing SS-based audio watermarking methods is their low embedding capacity. In this paper, we propose a new SS-based audio watermarking method which possesses much higher embedding capacity while ensuring satisfactory imperceptibility and robustness. The high embedding capacity is achieved through a set of mechanisms: embedding multiple watermark bits in one audio segment, reducing host signal interference on watermark extraction, and adaptively adjusting PN sequence amplitude in watermark embedding based on the property of audio segments. The effectiveness of the proposed audio watermarking method is demonstrated by simulation examples.**

***Index Terms*— Audio watermarking, spread spectrum, PN sequence, embedding capacity, copyright protection.**

EDICS: AUD-AUMM

I. Introduction

Recent advances in communication and multimedia technologies have made the reproduction, manipulation and distribution of digital multimedia data much easier than ever before. While these technologies have brought great benefits to our society and individuals, multimedia piracy is a serious problem and the financial loss caused by illegal multimedia data downloading and sharing is enormous. For example, in the global context, 95% of music downloads are illegal [1], which is worth many billions of dollars [2]. Therefore, there is a strong demand for preventing illegal use of copyrighted multimedia data. In this paper, we limit our attention to audio data such as music and speech.

In an open network environment, audio watermarking is a promising technology to tackle piracy for audio data. Technically speaking, audio watermarking aims to hide watermark data (such as publisher information, user identity, file transaction/downloading records, etc.) into the actual audio signal without affecting its normal usage. When necessary, the owner or law enforcement agencies can extract the watermark data, by using a secret key, to trace the source of illegal distribution. An effective and practical audio watermarking scheme should exhibit some important characteristics such as imperceptibility, robustness, security and high embedding capacity. Moreover, blind methods which extract watermarks without knowing the host audio signal are highly desirable as semi-blind and non-blind methods are not applicable to most practical applications [3]. So far, many audio watermarking methods have been developed, based on various schemes such as spread spectrum (SS) [4]-[8], echo-hiding [9]-[14], patchwork [3], [15]-[17], and others [18]-[22]. Among these audio watermarking methods, the SS-based audio watermarking methods have attracted much attention due to their simple watermark embedding and extraction structures and superior performance in terms of imperceptibility, security, and robustness against conventional attacks [6]. Besides, most of the SS-based audio watermarking methods are blind methods.

The concept of SS-based audio watermarking can be found in [4]. As shown in [4], a watermark bit is embedded into a host audio segment by using a spreading sequence. At the decoder, the embedded watermarks are extracted by correlating the watermarked signal with the spreading sequence. The watermark extraction mechanism used in [4] results in host signal interference. Large host signal interference could significantly degrade the accuracy of watermark extraction, and thus reduce the robustness of the audio watermarking method. To deal with this issue, two techniques are proposed in [5] and [6] to reduce host signal interference. Based on these two techniques, another two modified SS-based audio watermarking methods have been developed and reported in [7] and [8], respectively.

Since the SS-based watermarking methods exploit the second-order statistical property of the watermarked audio signal to extract watermarks, they usually require relatively long audio segments to achieve high watermark extraction accuracy or robustness. This inevitably leads to low embedding capacity. Although embedding capacity can be increased by using a spreading sequence with a larger amplitude and/or reducing the length of the audio segments, these measures will lower perceptual quality and/or robustness.

In this paper, a new SS-based audio watermarking method is proposed to increase the embedding capacity while maintaining high imperceptibility and robustness. In the embedding process of the proposed method, the discrete cosine transform (DCT) is first applied to the host audio signal to obtain the corresponding DCT coefficients. Then, those DCT coefficients vulnerable to filtering (e.g. low-pass and high-pass filtering) and compression attacks (e.g. MP3 and AAC) are discarded and the remaining DCT coefficients are selected

Y. Xiang and I. Natgunanathan are with the School of Information Technology, Deakin University, Burwood, VIC 3125, Australia (e-mail: yxiang@deakin.edu.au, iynkaran.n@research.deakin.edu.au).

Y. Rong is with the Department of Electrical and Computer Engineering, Curtin University, Bentley, WA 6102, Australia (e-mail: y.rong@curtin.edu.au).

S. Guo is with the Department of Computer Science and Engineering, The University of Aizu, Japan (e-mail: sguo@u-aizu.ac.jp).

for watermark embedding. Prior to embedding watermarks, these selected DCT coefficients are segmented, followed by further dividing each audio segment into a pair of fragments. On the other hand, a number of near-orthogonal pseudonoise (PN) sequences are formed by rotationally shifting a randomly generated seed PN sequence which is temporally white. Each of the PN sequences represents a group of watermark bits. These PN sequences not only act as the spreading sequences but also secret keys to introduce security into the proposed method. After that, one can embed a group of watermark bits into an audio segment by inserting the corresponding PN sequence into the designated pair of fragments in a multiplicative manner. While inserting the PN sequence, the amplitude of the PN sequence is adjusted by a scaling factor to maximize perceptual quality while ensure high robustness. The scaling factor is adaptively determined by an analysis-by-synthesis mechanism by exploiting the property of the audio segment. In the watermark extraction process, based on the near-orthogonality of the PN sequences, the correlations between the watermarked audio segments and the PN sequences are exploited to extract the embedded watermark bits. Moreover, the watermark extraction mechanism utilizes the similarity between the pair of fragments in an audio segment to reduce the host signal interference.

The proposed SS-based audio watermarking method has much higher embedding capacity than the other SS-based methods. It also has high imperceptibility and robustness against common attacks. Its superior performance results from the usage of a set of new techniques to embed multiple watermarks in one audio segment, to reduce host signal interference on watermark extraction, and to adaptively control the amplitude of the PN sequences in watermark embedding. Simulation results show the validity of our method. The remainder of the paper is organized as follows. The representative SS-based methods are reviewed and discussed in Section II. The proposed SS-based audio watermarking method is presented in section III. The simulation results are shown in Section IV and Section V concludes the paper.

II. Review of existing SS-based audio watermarking methods

Assume that $\mathbf{x}$ is a segment of the host audio signal in a given domain, $\mathbf{y}$ is the corresponding watermarked audio segment and $\mathbf{p}$ is a PN sequence. Here, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{p}$ are row vectors of equal length, the elements of $\mathbf{p}$ take values from $\{-1, +1\}$, and $\mathbf{p}$ is independent of $\mathbf{x}$. Given a watermark bit $w \in \{-1, +1\}$, the conventional SS-based watermarking method in [4] embeds w into the host audio segment $\mathbf{x}$ by

$$\mathbf{y} = \mathbf{x} + \alpha w \mathbf{p} \tag{1}$$

where α is a positive constant which controls the perceptual quality of the watermarked audio signal. At the watermark extraction stage, the embedded watermark bit is extracted from $\mathbf{y}$ by using $\mathbf{p}$ as a secret key. To proceed, we define

$$z = \frac{\mathbf{y}\mathbf{p}^T}{\mathbf{p}\mathbf{p}^T} \tag{2}$$

where the superscript ^{T} stands for transpose operation. By substituting (1) into (2), it yields

$$\begin{aligned} z &= \frac{(\mathbf{x} + \alpha w \mathbf{p})\mathbf{p}^T}{\mathbf{p}\mathbf{p}^T} \\ &= \alpha w + x' \end{aligned} \tag{3}$$

where

$$x' = \frac{\mathbf{x}\mathbf{p}^T}{\mathbf{p}\mathbf{p}^T}. \tag{4}$$

Then, the extracted watermark bit, denoted as w_e, is given by

$$w_e = \text{sign}(z) \tag{5}$$

where $\text{sign}(\cdot)$ is the sign function which gives $+1, -1$ and 0 if its argument is positive, negative and zero, respectively. Obviously, this audio watermarking method is a blind method as it does not require information of the host audio signal in the watermark extraction process.

In the context of watermark extraction, the term x' acts as an interference from the host audio signal, which is called host signal interference. From (3) and (5), one can see that if x' is sufficiently small and/or the constant α is large, then $w_e \approx \text{sign}(\alpha w) = w$, i.e., perfect watermark extraction can be achieved. Otherwise, extraction error could occur. However, the value of α must be small to ensure high perceptual quality. On the other hand, from the expression of x' in (4), a small x' is possible if the length of the audio segment is large, which leads to low embedding capacity. Consequently, the audio watermarking method in [4] cannot increase embedding capacity without significantly compromising imperceptibility and robustness.

Various efforts have been made to reduce the host signal interference effect via modifying the watermark embedding function in (1), while still using (5) for watermark extraction [5]-[8]. In [5], Malvar and Florencio proposed to use the following watermark embedding function:

$$\mathbf{y} = \mathbf{x} + (\alpha w - \lambda x')\mathbf{p} \tag{6}$$

where λ is a positive constant no greater than 1. Based on the watermark embedding function in (6), it follows from (2) and (4) that

$$\begin{aligned} z &= \frac{(\mathbf{x} + (\alpha w - \lambda x')\mathbf{p})\mathbf{p}^T}{\mathbf{p}\mathbf{p}^T} \\ &= \alpha w + (1 - \lambda)x'. \end{aligned} \tag{7}$$

Clearly, in contrast with (3), the host signal interference to the z value in (7) can be controlled by the parameter λ. By increasing λ towards 1, the impact of the host signal interference on watermark extraction can be reduced, which yields better robustness. However, as one can see from (6), a larger λ results in lower perceptual quality. To simultaneously obtain high imperceptibility and robustness, the audio segment length must be large enough such that x' is sufficiently small.

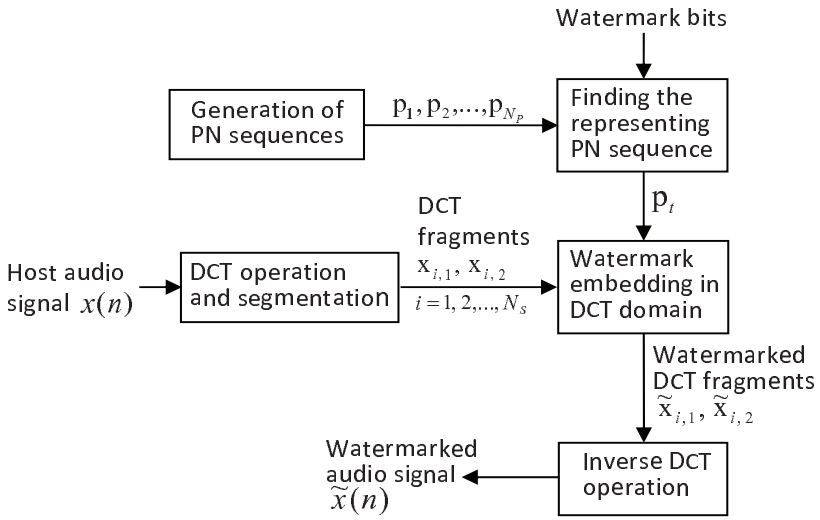

Fig. 1. Block diagram of the proposed watermark embedding scheme.

A different approach is proposed in [6] for watermark embedding:

$$\mathbf{y} = \begin{cases} \mathbf{x} + \alpha_1 \mathbf{p}, & \text{if } x' \geq 0,\ w = +1 \\ \mathbf{x} - \alpha_2 \mathbf{p} - \lambda' \mathbf{p}\mathbf{x}\mathbf{p}^T, & \text{if } x' \geq 0,\ w = -1 \\ \mathbf{x} - \alpha_1 \mathbf{p}, & \text{if } x' < 0,\ w = -1 \\ \mathbf{x} + \alpha_2 \mathbf{p} - \lambda' \mathbf{p}\mathbf{x}\mathbf{p}^T, & \text{if } x' < 0,\ w = +1 \end{cases} \tag{8}$$

where $0 < \alpha_1 < \alpha_2$ and $0 < \lambda' \leq 1$. It results from (2), (4) and (8) that

$$z = \begin{cases} x' + \alpha_1, & \text{if } x' \geq 0,\ w = +1 \\ x'(1 - \lambda' \mathbf{p}\mathbf{p}^T) - \alpha_2, & \text{if } x' \geq 0,\ w = -1 \\ x' - \alpha_1, & \text{if } x' < 0,\ w = -1 \\ x'(1 - \lambda' \mathbf{p}\mathbf{p}^T) + \alpha_2, & \text{if } x' < 0,\ w = +1. \end{cases} \tag{9}$$

From (9), it is clear that using large α_1 and α_2 and properly selected λ' will reduce the host signal interference and thus improve robustness. However, similar to the method in [5], a sufficiently long audio segment must be used to ensure high perceptual quality and robustness at the same time.

Other watermark embedding functions have also been used in the SS-based audio watermarking methods such as those recently developed in [7] and [8]. However, these methods still have the same problem, that is, how to achieve high embedding capacity while ensuring satisfactory imperceptibility and robustness. Next, we will propose a new SS-based audio watermarking method to tackle this problem.

III. Proposed SS-based audio watermarking method

In this section, the watermark embedding and extraction processes of the new method will be presented in detail. The selection of the key watermarking parameter, β, will be discussed. The proposed watermarking algorithm will be summarized at the end of the section.

A. Watermark embedding process

The watermark embedding process is composed of three major parts: generation of near-orthogonal PN sequences, DCT operation and segmentation, and embedding of watermark bits. Fig. 1 shows the block diagram of the proposed watermark embedding scheme.

1) Generation of near-orthogonal PN sequences: In traditional SS-based watermarking methods [4]-[8], one PN sequence is used to embed one watermark bit, either "-1" or "+1", into one audio segment. In order to increase embedding capacity without sacrificing perceptual quality, we propose to use one PN sequence to embed multiple watermark bits into one audio segment. Assume that the number of watermark bits to be inserted into one audio segment is n_b. Obviously, n_b binary watermark bits can yield up to 2^{n_b} different watermark sequences. To represent each watermark sequence of length n_b by a different PN sequence, the number of PN sequences required is $N_p = 2^{n_b}$. For example, in the case of two watermark bits, the corresponding watermark sequences of length two are $\{-1,-1\}$, $\{-1,+1\}$, $\{+1,-1\}$ and $\{+1,+1\}$. So, four PN sequences are needed.

Now, we show how to generate the N_p PN sequences. Let

$$\mathbf{p}_1 = [p_1, p_2, \ldots, p_N] \tag{10}$$

be a temporally white PN sequence of length N, where $N > N_p$ and $p_i \in \{-1,+1\}$, $i = 1,2,...,N$. Based on $\mathbf{p}_1$, the other PN sequences are generated by shifting the elements of $\mathbf{p}_1$ in a circular manner as follows:

$$\begin{cases} \mathbf{p}_2 = [p_N, p_1, \ldots, p_{N-1}] \\ \mathbf{p}_3 = [p_{N-1}, p_N, p_1, \ldots, p_{N-2}] \\ \qquad \vdots \\ \mathbf{p}_{N_p} = [p_{N-N_p+2}, \ldots, p_N, p_1, \ldots, p_{N-N_p+1}] \end{cases} . \tag{11}$$

Since $\mathbf{p}_1$ is temporally white, the PN sequences $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N_p}$ are near-orthogonal. That is, if N is sufficiently large, the vector $\mathbf{p}_t \circ \mathbf{p}_j (t \neq j)$, where "$\circ$" stands

for the Hadamard product (i.e., the element-wise product), has almost half of its elements taking value "+1" and the other elements taking value "−1".

As will be shown in the watermark extraction process, the above near-orthogonality property of the PN sequences is essential to correctly extracting the watermark bits from the watermarked audio signal. In addition, since the PN sequences $\mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_{N_p}$ are generated from the PN sequence $\mathbf{p}_1$, one only needs to pass $\mathbf{p}_1$, instead of all N_p PN sequences, to the watermark extraction end. This greatly simplifies the proposed method.

2) DCT operation and segmentation: Let $x(n)$ be the host audio signal, which contains K samples. Similar to the SS-based audio watermarking methods in [4] and [8], we apply DCT to $x(n)$. The DCT coefficients corresponding to $x(n)$, denoted by $X(k)$, can be computed as follows [23]:

$$X(k) = l(k) \sum_{n=0}^{K-1} x(n) \cos\left\{\frac{\pi(2n+1)k}{2K}\right\} \quad (12)$$

where $k = 0, 1, ..., K-1$, and

$$l(k) = \begin{cases} \frac{1}{\sqrt{K}}, & \text{if } k = 0 \\ \sqrt{\frac{2}{K}}, & \text{if } 1 \le k < K \end{cases}. \quad (13)$$

It is known that very low and high frequency components are vulnerable to attacks such as filtering and compression. So, only the DCT coefficients corresponding to a certain frequency range $[f_l, f_h]$ should be used to embed watermarks, where f_l and f_h can be determined experimentally. Without loss of generality, assume that there are L such DCT coefficients. Then, these DCT coefficients are selected from $X(k)$.

To facilitate watermark embedding, we divide the selected DCT coefficients into N_s segments of length $2N$, and denote the ith segment as

$$X_i(k) = [X_i(0), X_i(1), \ldots, X_i(2N-1)] \quad (14)$$

where $i = 1, 2, \ldots, N_s$. Then, we further divide $X_i(k)$ into a pair of fragments $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$ as follows:

$$\mathbf{x}_{i,1} = [X_i(0), X_i(2), \ldots, X_i(2N-2)] \quad (15)$$
$$\mathbf{x}_{i,2} = [X_i(1), X_i(3), \ldots, X_i(2N-1)] \quad (16)$$

where the length of the fragments $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$ is N.

3) Embedding of watermark bits: Assume that the n_b watermark bits to be embedded into the ith segment $X_i(k)$ are represented by the PN sequence $\mathbf{p}_t$. Thus, we perform watermark embedding by inserting the PN sequence $\mathbf{p}_t$ into $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$ using the following embedding rule:

$$\tilde{\mathbf{x}}_{i,1} = (\mathbf{1} + \beta \mathbf{p}_t) \circ \mathbf{x}_{i,1} \quad (17)$$
$$\tilde{\mathbf{x}}_{i,2} = (\mathbf{1} - \beta \mathbf{p}_t) \circ \mathbf{x}_{i,2} \quad (18)$$

where $\tilde{\mathbf{x}}_{i,1}$ and $\tilde{\mathbf{x}}_{i,2}$ are the watermarked fragments, $\mathbf{1}$ is a length-N row vector whose elements are all one, and β is a constant satisfying $0 < \beta < 1$.

Define

$$\tilde{\mathbf{x}}_{i,1} = [\tilde{X}_{i,1}(0), \tilde{X}_{i,1}(1), \ldots, \tilde{X}_{i,1}(N-1)] \quad (19)$$

and

$$\tilde{\mathbf{x}}_{i,2} = [\tilde{X}_{i,2}(0), \tilde{X}_{i,2}(1), \ldots, \tilde{X}_{i,2}(N-1)]. \quad (20)$$

Once $\tilde{\mathbf{x}}_{i,1}$ and $\tilde{\mathbf{x}}_{i,2}$ are obtained from (17) and (18), the watermarked segment $\tilde{X}_i(k)$, which is the watermarked counterpart of $X_i(k)$, can be formed by

$$\tilde{X}_i(k) = [\tilde{X}_{i,1}(0), \tilde{X}_{i,2}(0), \tilde{X}_{i,1}(1), X_{i,2}(1), \ldots, \tilde{X}_{i,1}(N-1), \tilde{X}_{i,2}(N-1)]. \quad (21)$$

After obtaining all watermarked segments $\tilde{X}_i(k)$, $i = 1, 2, \ldots, N_s$, the watermarked signal $\tilde{x}(n)$ is constructed by applying inverse discrete cosine transform.

Remark 1: The watermarking parameter β in (17) and (18) is utilized to control the amplitude of $\mathbf{p}_t$ in watermark embedding with the purpose of maximizing perceptual quality while maintaining high robustness. The value of β needs to be properly selected, which will be discussed in the subsection III.C.

B. Watermark extraction process

This process aims to extract the embedded watermark bits from the received audio signal $y(n)$, with the help of the available PN sequence $\mathbf{p}_1$. Here, $y(n)$ is the post-attack counterpart of the watermarked audio signal $\tilde{x}(n)$. In the absence of attacks, $y(n) = \tilde{x}(n)$.

1) Extraction of watermark bits from $y(n)$: Fig. 2 shows the block diagram of the proposed watermark extraction scheme. Firstly, from the PN sequence $\mathbf{p}_1$, the other $N_p - 1$ PN sequences $\mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_{N_p}$ are regenerated using (11). Then, we apply DCT to the received audio signal $y(n)$ to obtain the corresponding DCT coefficients $Y(k)$. The DCT coefficients corresponding to the frequency region $[f_l, f_h]$ are selected and partitioned into N_s length-$2N$ segments $Y_i(k)$, $i = 1, 2, \ldots, N_s$. Afterwards, these segments are further partitioned into N_s pairs of length-N fragments $\mathbf{y}_{i,1}$ and $\mathbf{y}_{i,2}$, $i = 1, 2, \ldots, N_s$. Clearly, if the received audio signal has not undergone any attacks, i.e., $y(n) = \tilde{x}(n)$, then it holds from (17) and (18) that

$$\mathbf{y}_{i,1} = \tilde{\mathbf{x}}_{i,1} = (\mathbf{1} + \beta \mathbf{p}_t) \circ \mathbf{x}_{i,1} \quad (22)$$

and

$$\mathbf{y}_{i,2} = \tilde{\mathbf{x}}_{i,2} = (\mathbf{1} - \beta \mathbf{p}_t) \circ \mathbf{x}_{i,2}. \quad (23)$$

Note that since $0 < \beta < 1$ and the elements of $\mathbf{p}_t$ take values from $\{-1, +1\}$, it is obvious that the elements of $(\mathbf{1} + \beta \mathbf{p}_t)$ and $(\mathbf{1} - \beta \mathbf{p}_t)$ are positive.

Define

$$\mathbf{y}_{i,d} = |\mathbf{y}_{i,1}| - |\mathbf{y}_{i,2}| \quad (24)$$

where $|\cdot|$ denotes the element-wise absolute value. From (22)-(24), it results in

$$\begin{aligned} \mathbf{y}_{i,d} &= |(\mathbf{1} + \beta \mathbf{p}_t) \circ \mathbf{x}_{i,1}| - |(\mathbf{1} - \beta \mathbf{p}_t) \circ \mathbf{x}_{i,2}| \\ &= (\mathbf{1} + \beta \mathbf{p}_t) \circ |\mathbf{x}_{i,1}| - (\mathbf{1} - \beta \mathbf{p}_t) \circ |\mathbf{x}_{i,2}| \\ &= (|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|) + \beta \mathbf{p}_t \circ (|\mathbf{x}_{i,1}| + |\mathbf{x}_{i,2}|). \end{aligned} \quad (25)$$

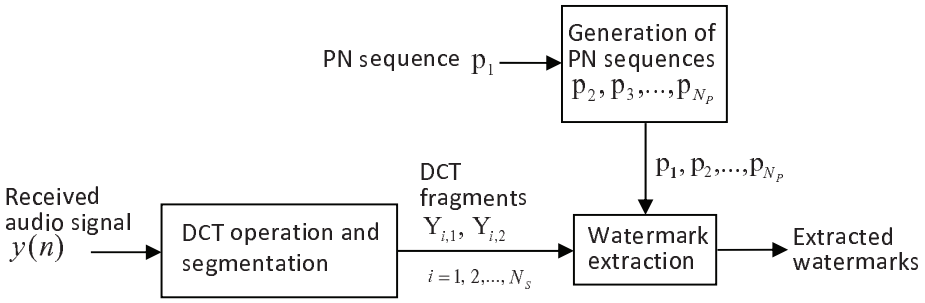

Fig. 2. Block diagram of the proposed watermark extraction scheme.

Recall that each PN sequence represents one and only one watermark sequence containing a unique set of watermark bits. So, the embedded watermark bits in the ith segment can be extracted through finding the corresponding PN sequence $\mathbf{p}_t$ from $\mathbf{y}_{i,d}$. This can be achieved by finding the PN sequence $\mathbf{p}_m$, where the index m is given by

$$m = \underset{j\in\{1,2,\ldots,N_p\}}{\operatorname{argmax}} \mathbf{y}_{i,d}\mathbf{p}_j^T. \quad (26)$$

Here, the function argmax$(\cdot)$ returns the j value with which $\mathbf{y}_{i,d}\mathbf{p}_j^T$ yields maximum.

2) Discussion about the rationale of (26): From (25), it holds that

$$\begin{aligned}\mathbf{y}_{i,d}\mathbf{p}_j^T &= (|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_j^T + \\ &\quad \beta(\mathbf{p}_t \circ (|\mathbf{x}_{i,1}| + |\mathbf{x}_{i,2}|))\,\mathbf{p}_j^T \end{aligned} \quad (27)$$

where $j = 1, 2, \ldots, t, \ldots, N_p$. In the above equation, the first term on the right-hand side acts as the host audio signal interference which has negative impact on watermark extraction. Hence, the smaller this term, the better the robustness. Obviously, the absolute value of each element in the vector $|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|$ is generally smaller than, at most equal to, the value of the corresponding element in either $|\mathbf{x}_{i,1}|$ or $|\mathbf{x}_{i,2}|$. As a result, $(|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_j^T$ is often much smaller than $|\mathbf{x}_{i,1}|\mathbf{p}_j^T$ or $|\mathbf{x}_{i,2}|\mathbf{p}_j^T$. The significant reduction of host audio signal interference is due to the properly designed watermark embedding rule.

Considering the values that j can take, it follows from (27) that

$$\begin{aligned}\mathbf{y}_{i,d}\mathbf{p}_{\bar{j}}^T &= (|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_{\bar{j}}^T + \\ &\quad \beta\,(\mathbf{p}_t \circ (|\mathbf{x}_{i,1}| + |\mathbf{x}_{i,2}|))\,\mathbf{p}_{\bar{j}}^T \\ &= (|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_{\bar{j}}^T + \\ &\quad \beta(\mathbf{p}_t \circ \mathbf{p}_{\bar{j}})\left(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|\right) \end{aligned} \quad (28)$$

where $\bar{j} = 1, 2, \ldots, t-1, t+1, \ldots, N_p$, and

$$\begin{aligned}\mathbf{y}_{i,d}\mathbf{p}_t^T &= (|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_t^T + \\ &\quad \beta(\mathbf{p}_t \circ \mathbf{p}_t)\left(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|\right) \\ &= (|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_t^T + \\ &\quad \beta\mathbf{1}\left(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|\right) \end{aligned} \quad (29)$$

with $\mathbf{1}$ being the length-N row vector with all-one elements. Since the PN sequences $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N_p}$ are independent of $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$, thus $(|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_{\bar{j}}^T$ and $(|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_t^T$ are statistically comparable. That is, the first term on the right-hand side of (28) is comparable to the corresponding term in (29). On the other hand, it is clear that all elements of the vector $|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|$ are positive and the vector $\mathbf{p}_t \circ \mathbf{p}_{\bar{j}},\ \forall \bar{j} \neq t$ has almost equal number of elements taking values "+1" and "−1", respectively. Consequently, the value of $\beta\mathbf{1}\left(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|\right)$ which is the second term on the right-hand side of (29) is usually much greater than that of $\beta(\mathbf{p}_t \circ \mathbf{p}_{\bar{j}})\left(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|\right)$ which is the corresponding term in (28). Therefore, $\mathbf{y}_{i,d}\mathbf{p}_t^T$ is likely to be greater than $\mathbf{y}_{i,d}\mathbf{p}_{\bar{j}}^T$ for $\bar{j}$ not equal to t. In other words, $\mathbf{y}_{i,d}\mathbf{p}_j^T$ reaches maximum at $j = t$. This verifies the validity of (26).

C. Selection of parameter β

The selection of the parameter β in (17) and (18) is vital as it affects both perceptual quality and robustness. On one hand, the value of β should not be too small in order to achieve satisfactory robustness. For the watermarks to be detectable, $\mathbf{y}_{i,d}\mathbf{p}_t^T$ should be greater than $\mathbf{y}_{i,d}\mathbf{p}_{\bar{j}}^T$. From (28) and (29), it follows

$$\begin{aligned}&(|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_t^T + \beta\mathbf{1}\left(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|\right) > \\ &(|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)\,\mathbf{p}_{\bar{j}}^T + \beta(\mathbf{p}_t \circ \mathbf{p}_{\bar{j}})\left(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|\right)\end{aligned} \quad (30)$$

which leads to

$$\beta > \frac{(|\mathbf{x}_{i,1}| - |\mathbf{x}_{i,2}|)(\mathbf{p}_{\bar{j}}^T - \mathbf{p}_t^T)}{(\mathbf{1} - (\mathbf{p}_t \circ \mathbf{p}_{\bar{j}}))(|\mathbf{x}_{i,1}^T| + |\mathbf{x}_{i,2}^T|)}. \quad (31)$$

On the other hand, a smaller β value yields higher perceptual quality. Hence, we set a range $[\beta_{min},\ \beta_{max}]$ for β, where β_{min} satisfies (31). Next, we present an analysis-by-synthesis approach to selecting a suitable β within the range $[\beta_{min},\ \beta_{max}]$, which has the smallest value but ensures correct watermark extraction.

Let $\triangle\beta$, γ_1 and γ_2 are three constants satisfying $0 < \triangle\beta \ll \beta_{max}$, $0 < \gamma_1 < 1$ and $\gamma_2 > \gamma_1$. Given $\mathbf{x}_{i,1}$, $\mathbf{x}_{i,2}$, $\mathbf{p}_t$, and $\bar{\mathbf{p}}$ constructed by

$$\bar{\mathbf{p}}^T = \left[\mathbf{p}_1^T, \ldots, \mathbf{p}_{t-1}^T, \mathbf{p}_{t+1}^T, \ldots, \mathbf{p}_{N_p}^T\right] \quad (32)$$

the suitable β value can be determined by using the selection approach shown in Table 1. It can be seen from Table 1 that the value of β is dependent on $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$, as well as $\mathbf{p}_t^T$ and $\bar{\mathbf{p}}^T$.

TABLE I

AN APPROACH TO SELECTING β VALUE

Step 1: Set initial β to $\beta = \beta_{min}$ and construct $\bar{\mathbf{p}}$ by (32).
Step 2: Compute
$\mathbf{d} = (\lvert\mathbf{x}_{i,1}\rvert - \lvert\mathbf{x}_{i,2}\rvert) + \beta\mathbf{p}_t \circ (\lvert\mathbf{x}_{i,1}\rvert + \lvert\mathbf{x}_{i,2}\rvert)$
$u_1 = \mathbf{d}\mathbf{p}_t^T$
$u_2 = \max(\mathbf{d}\bar{\mathbf{p}}^T)$
where $\max(\mathbf{u})$ returns the largest element of $\mathbf{u}$.
Step 3: If $\gamma_1 u_1 > \gamma_2$, set $v = \gamma_1 u_1$.
Otherwise, set $v = \gamma_2$.
Step 4: If $u_2 \geq u_1 - v$, go to Step 5.
Otherwise, end.
Step 5: Increase β to $\beta + \triangle\beta$. If $\beta \leq \beta_{max} - \triangle\beta$, go to Step 2.
Otherwise, end.

That is, the selection method jointly exploits the properties of the host audio signal and the PN sequences. It should be noted that in the absence of attacks, successful watermark extraction can be attained by ensuring $u_2 < u_1$. However, in order to enhance robustness, we choose β in such a way that results in $u_2 < u_1 - v$ (see Step 4 in Table 1). The parameter v is introduced to create an error buffer, which is set to be $\gamma_1 u_1$ with the lower limit of γ_2. If the negative impact of attacks on watermark extraction is within this error buffer, correct watermark extraction can be guaranteed. On the other hand, when the value of β needs to be increased, $\triangle\beta$ serves as the step size of the increment. The values of γ_1, γ_2 and $\triangle\beta$ can be chosen experimentally.

In summary, the proposed SS-based audio watermarking algorithm is formulated as follows.

Watermark embedding

- Step 1: Randomly generate the temporally white PN sequence $\mathbf{p}_1$ and then construct $\mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_{N_p}$ by (11).
- Step 2: Apply DCT to $x(n)$ by (12) and (13), select those DCT coefficients corresponding to the frequency range $[f_l, f_h]$, and segment them to obtain $X_i(k),\ i = 1, 2, \ldots, N_s$.
- Step 3: Given the ith DCT segment, construct the fragments $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$ by (15) and (16), respectively.
- Step 4: Use the approach shown in Table I to select β value.
- Step 5: Insert the PN sequence $\mathbf{p}_t$ into $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$ by (17) and (18).
- Step 6: Form the watermarked DCT segment $\tilde{X}_i(k)$ by (21).
- Step 7: After obtaining all $\tilde{X}_i(k),\ i = 1, 2, \ldots, N_s$, construct the watermarked signal $\tilde{x}(n)$ by applying inverse DCT.

Watermark extraction

- Step 1: Construct the PN sequences $\mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_{N_p}$ from $\mathbf{p}_1$ by (11).
- Step 2: Similar to Steps 2 and 3 in the watermark embedding part, construct $\mathbf{y}_{i,1}$ and $\mathbf{y}_{i,2}$ from the received audio signal $y(n)$, where $i = 1, 2, \ldots, N_s$.
- Compute $\mathbf{y}_{i,d}$ by (24) and then find the index m by (26). Thus, the embedded PN sequence $\mathbf{p}_m$ is extracted.
- From the one-to-one mapping relationship between the PN sequences and the watermark sequences, the embedded watermark bits can be recovered.

Remark 2: In the proposed audio watermarking method, the embedding capacity is increased mainly due to three reasons: i) Multiple watermark bits are embedded into an audio segment using one PN sequence, while traditional SS-based methods only embed one watermark bit into an audio segment. ii) The host signal interference encountered at the watermark extraction end is eased significantly thanks to the properly designed embedding rule. iii) While ensuring satisfactory robustness, the β value chosen in our method is the smallest, which leads to possibly the highest imperceptibility. This provides an opportunity for further increasing embedding capacity by slightly compromising perceptual quality (but still maintain perceptual quality at a high level).

IV. SIMULATION RESULTS

In this section, simulation examples are provided to illustrate the performance of the proposed SS-based audio watermarking method and compare it with three latest SS-based methods in [6]-[8]. In the simulations, we use 40 randomly selected audio clips belonging to four different genres as host audio signals, as detailed in Table II. All these audio clips have the duration of 10 seconds. They are sampled at the rate of 44.1 kHz and quantized with 16 bits. All of the obtained samples are used in the DCT operation.

TABLE II

HOST AUDIO SIGNALS USED IN SIMULATIONS

Host audio signals	Genres
$S_{01} \sim S_{10}$	Western pop music
$S_{11} \sim S_{20}$	Eastern folk music
$S_{21} \sim S_{30}$	Eastern classical music
$S_{31} \sim S_{40}$	Speeches

A practically applicable watermarking method should be robust to conventional attacks while maintaining high imperceptibility. Same as [3] and [24], the perceptual evaluation of audio quality (PEAQ) algorithm [25] is used to evaluate the perceptual quality of the proposed methods. The PEAQ algorithm compares the quality of the host audio signal with its watermarked counterpart and returns a parameter called objective difference grade (ODG) ranging between -4 and 0. The perceptual quality improves with the increase of the ODG value. Also, we employ the detection rate (DR) as a measure to assess the robustness of our method, which is defined as

$$\mathrm{DR} = \left(\frac{\text{Number of watermarks correctly extracted}}{\text{Number of watermarks embedded}}\right) \times 100\%.$$

The following common attacks are used in the evaluation of robustness:

- *Closed-loop attack* : The watermarks are extracted from the watermarked signals without any attacks.
- *Re-quantization attack* : Each sample of the watermarked signals is re-quantized from 16 bits to 8 bits.
- *Noise attack*: Random noise is added to the watermarked signals, where the ratio of the watermarked signal to noise is 20 dB.
- *Amplitude attack* : The amplitudes of the watermarked signals are enlarged by 1.2 times and 1.8 times.

TABLE III

DETECTION RATES OF THE PROPOSED METHOD AND THE METHODS IN [6] AND [8] , WHERE ODG=-0.7 AND EMBEDDING RATE=84 BPS FOR ALL THREE METHODS

Attacks	Host signals	DR (%)		
		Method in [6]	Method in [8]	Proposed method
Closed-loop	$S_{01} \sim S_{10}$	84.1	76.7	100
	$S_{11} \sim S_{20}$	91.9	80.6	100
	$S_{21} \sim S_{30}$	88.6	79.3	100
	$S_{31} \sim S_{40}$	97.9	76.0	100
Re-quantization	$S_{01} \sim S_{10}$	84.1	76.6	100
	$S_{11} \sim S_{20}$	91.8	80.4	100
	$S_{21} \sim S_{30}$	88.4	78.5	100
	$S_{31} \sim S_{40}$	97.7	74.4	100
Noise	$S_{01} \sim S_{10}$	84.0	74.9	99.6
	$S_{11} \sim S_{20}$	91.5	78.7	99.7
	$S_{21} \sim S_{30}$	86.4	75.4	99.6
	$S_{31} \sim S_{40}$	97.8	71.2	99.9
Amplitude (1.2)	$S_{01} \sim S_{10}$	84.1	76.7	100
	$S_{11} \sim S_{20}$	91.9	80.6	100
	$S_{21} \sim S_{30}$	88.6	79.3	100
	$S_{31} \sim S_{40}$	97.9	76.0	100
Amplitude (1.8)	$S_{01} \sim S_{10}$	84.1	76.7	100
	$S_{11} \sim S_{20}$	91.9	80.6	100
	$S_{21} \sim S_{30}$	88.6	79.3	100
	$S_{31} \sim S_{40}$	97.9	76.0	100
MP3 (128 kbps)	$S_{01} \sim S_{10}$	76.8	70.9	100
	$S_{11} \sim S_{20}$	84.4	70.4	100
	$S_{21} \sim S_{30}$	82.0	70.5	100
	$S_{31} \sim S_{40}$	93.3	68.0	100
MP3 (96 kbps)	$S_{01} \sim S_{10}$	76.4	69.7	99.8
	$S_{11} \sim S_{20}$	82.7	66.6	99.3
	$S_{21} \sim S_{30}$	81.6	67.0	99.6
	$S_{31} \sim S_{40}$	93.2	67.4	100
AAC (128 kbps)	$S_{01} \sim S_{10}$	79.4	71.7	100
	$S_{11} \sim S_{20}$	87.7	74.1	100
	$S_{21} \sim S_{30}$	86.7	74.3	100
	$S_{31} \sim S_{40}$	95.4	69.0	100
AAC (96 kbps)	$S_{01} \sim S_{10}$	78.5	70.2	99.8
	$S_{11} \sim S_{20}$	86.4	73.0	98.4
	$S_{21} \sim S_{30}$	84.3	72.8	99.1
	$S_{31} \sim S_{40}$	94.7	66.6	99.8
HPF (50 Hz)	$S_{01} \sim S_{10}$	87.9	76.4	100
	$S_{11} \sim S_{20}$	93.4	80.2	100
	$S_{21} \sim S_{30}$	87.5	77.9	100
	$S_{31} \sim S_{40}$	97.6	76.4	100
HPF (100 Hz)	$S_{01} \sim S_{10}$	85.3	76.2	100
	$S_{11} \sim S_{20}$	94.9	80.1	100
	$S_{21} \sim S_{30}$	84.1	77.6	100
	$S_{31} \sim S_{40}$	97.6	76.3	100
LPF (12 kHz)	$S_{01} \sim S_{10}$	69.3	66.1	100
	$S_{11} \sim S_{20}$	78.4	68.5	100
	$S_{21} \sim S_{30}$	77.1	67.6	100
	$S_{31} \sim S_{40}$	88.7	62.1	100
LPF (8 kHz)	$S_{01} \sim S_{10}$	62.9	59.3	100
	$S_{11} \sim S_{20}$	67.5	60.6	100
	$S_{21} \sim S_{30}$	65.3	59.9	100
	$S_{31} \sim S_{40}$	82.3	58.6	100

- MP3 *attack*: MPEG 1 Layer III compression is performed on the watermarked signals, where the compression bit rates are 128 kbps and 96 kbps.
- AAC *attack*: MPEG 4 advanced audio coding based compression is performed on the watermarked signals, where the compression bit rates are 128 kbps and 96 kbps.
- *High-pass filtering (HPF)*: High-pass filters with 50 Hz and 100 Hz cut-off frequencies are applied to the watermarked signals.
- *Low-pass filtering (LPF)*: Low-pass filters with 12 kHz and 8 kHz cut-off frequencies are applied to the watermarked signals.

Firstly, we compare the robustness of the proposed method with those in [6] and [8] at the same embedding rate of 84 bps and under the same perceptual quality with ODG $= -0.7$. This ODG value guarantees that the watermarked signals produced by these watermarking methods are of high perceptual quality. For our method, the simulation parameters are: $N_p = 64$, $N = 750$, $f_s = 100$ Hz, $f_e = 8$ kHz, $\beta_{min} = 0.001$, $\beta_{max} = 0.2$, $\triangle\beta = 0.005$, $\gamma_1 = 0.1$ and $\gamma_2 = 2$. Since $N_p = 64$, it means that 64 PN sequences $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{64}$

will be used to perform watermark embedding and extraction. We first randomly generate the temporally white PN sequence $\mathbf{P}_1$ and then form the other PN sequences $\mathbf{P}_2, \mathbf{P}_3, \ldots, \mathbf{P}_{64}$ by shifting the elements of $\mathbf{P}_1$ in a circular manner. Since $\mathbf{P}_2, \mathbf{P}_3, \ldots, \mathbf{P}_{64}$ can be obtained from $\mathbf{P}_1$, one only needs to pass $\mathbf{P}_1$ to the watermark extraction end, which is similar to the methods in [6]-[8].

Table III shows the DRs of these methods under conventional attacks. From Table III, it can be seen that the proposed method achieves 100% detection rate under closed-loop attack, re-quantization attack, amplitude attack, MP3 and AAC attacks with 128 kbps compression bit rate, and filtering attacks. It also achieves high detection rates under other attacks. Overall, our method outperforms the methods in [6] and [8] by large margins under all attacks considered. Note that the method in [7] implements a psychoacoustics masking module in MPEG-I Layer I and thus it uses block-wise sample processing. Due to this reason, the segment length can only be changed to certain values. Consequently, its embedding rate cannot be adjusted to a value near 84 bps and thus this method is not compared in this simulation.

Secondly, we compare the proposed method with the method in [7]. In the simulation, we keep the embedding rate of our method at 84 bps while the embedding rate of the one in [7] is at 57 bps. Table IV shows the DRs of both methods under conventional attacks. We can see that although the proposed method uses an embedding rate which is about 47% greater than the embedding rate utilized by the method in [7], our method still yields higher DRs than the latter does under all considered attacks. The performance margins are particularly large under re-quantization, noise, MP3, AAC, and LPF attacks.

In the third simulation, we evaluate the DRs of the proposed method and the methods in [6] and [8] versus different embedding rates, in the scenario of closed-loop attack. For each method, at a given embedding rate, the DR values obtained from different types of audio clips are averaged. Fig. 3 shows the simulation results. As expected, DRs decrease with the rise of embedding rates for all three methods. However, compared with the methods in [6] and [8], our method has a much smaller DR descent rate. Moreover, the DR of the proposed method is much higher than those of the other two methods at all embedding rates.

Furthermore, we compare the ODG values of the proposed method and the methods in [6] and [8] under different embedding rates. In this simulation, the parameters of these watermarking methods are adjusted such that they can achieve 100% DR under closed-loop attack. As shown in Fig. 4, the ODG values of all methods decrease when increasing embedding rate, which is expected. Nevertheless, under all considered embedding rates, the proposed method has much better perceptual quality than the methods in [6] and [8].

TABLE IV

DETECTION RATES OF THE PROPOSED METHOD AND THE METHOD IN [7], WHERE ODG=-0.7 FOR BOTH METHODS, EMBEDDING RATE=84 BPS FOR THE PROPOSED METHOD, AND EMBEDDING RATE=57 BPS FOR THE METHOD IN [7]

Attacks	Host signals	DR (%)	
		Method in [7]	Proposed method
Closed-loop	$S_{01} \sim S_{10}$	98.1	100
	$S_{11} \sim S_{20}$	99.8	100
	$S_{21} \sim S_{30}$	98.7	100
	$S_{31} \sim S_{40}$	99.4	100
Re-quantization	$S_{01} \sim S_{10}$	84.3	99.8
	$S_{11} \sim S_{20}$	89.4	100
	$S_{21} \sim S_{30}$	87.1	100
	$S_{31} \sim S_{40}$	82.4	100
Noise	$S_{01} \sim S_{10}$	60.2	99.6
	$S_{11} \sim S_{20}$	61.5	99.7
	$S_{21} \sim S_{30}$	60.5	99.6
	$S_{31} \sim S_{40}$	53.8	99.9
Amplitude (1.2)	$S_{01} \sim S_{10}$	98.1	100
	$S_{11} \sim S_{20}$	99.8	100
	$S_{21} \sim S_{30}$	98.7	100
	$S_{31} \sim S_{40}$	99.4	100
Amplitude (1.8)	$S_{01} \sim S_{10}$	98.1	100
	$S_{11} \sim S_{20}$	99.8	100
	$S_{21} \sim S_{30}$	98.7	100
	$S_{31} \sim S_{40}$	99.4	100
MP3 (128 kbps)	$S_{01} \sim S_{10}$	78.2	100
	$S_{11} \sim S_{20}$	80.8	100
	$S_{21} \sim S_{30}$	79.4	100
	$S_{31} \sim S_{40}$	76.3	100
MP3 (96 kbps)	$S_{01} \sim S_{10}$	72.1	99.8
	$S_{11} \sim S_{20}$	68.9	99.3
	$S_{21} \sim S_{30}$	70.2	99.6
	$S_{31} \sim S_{40}$	68.4	100
AAC (128 kbps)	$S_{01} \sim S_{10}$	80.8	100
	$S_{11} \sim S_{20}$	86.0	100
	$S_{21} \sim S_{30}$	84.1	100
	$S_{31} \sim S_{40}$	78.9	100
AAC (96 kbps)	$S_{01} \sim S_{10}$	74.2	99.8
	$S_{11} \sim S_{20}$	75.8	98.4
	$S_{21} \sim S_{30}$	74.9	99.1
	$S_{31} \sim S_{40}$	68.8	99.8
HPF (50 Hz)	$S_{01} \sim S_{10}$	98.0	100
	$S_{11} \sim S_{20}$	99.8	100
	$S_{21} \sim S_{30}$	98.5	100
	$S_{31} \sim S_{40}$	99.4	100
HPF (100 Hz)	$S_{01} \sim S_{10}$	97.2	100
	$S_{11} \sim S_{20}$	99.1	100
	$S_{21} \sim S_{30}$	97.4	100
	$S_{31} \sim S_{40}$	98.1	100
LPF (12 kHz)	$S_{01} \sim S_{10}$	75.9	100
	$S_{11} \sim S_{20}$	69.7	100
	$S_{21} \sim S_{30}$	74.2	100
	$S_{31} \sim S_{40}$	58	100
LPF (8kHz)	$S_{01} \sim S_{10}$	60.1	100
	$S_{11} \sim S_{20}$	58.4	100
	$S_{21} \sim S_{30}$	58.9	100
	$S_{31} \sim S_{40}$	55.2	100

V. CONCLUSION

In this paper, we propose a novel SS-based watermarking method for audio signals. Unlike the existing SS-based audio watermarking methods, the proposed method can embed multiple watermark bits into one audio segment, which increases

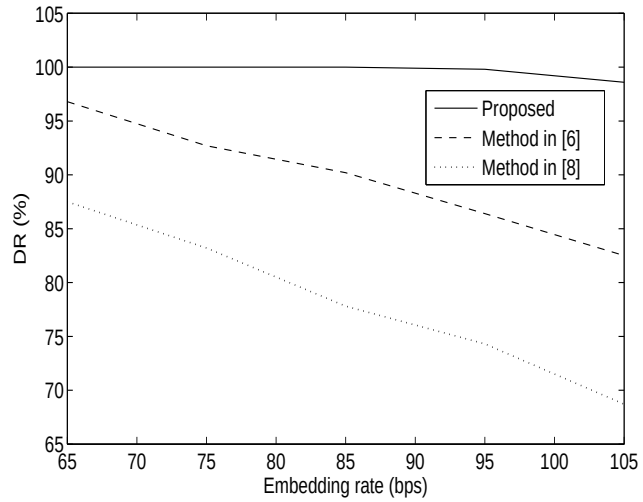

Fig. 3. DRs versus embedding rates under closed-loop attack, where ODG=-0.7 for all three methods.

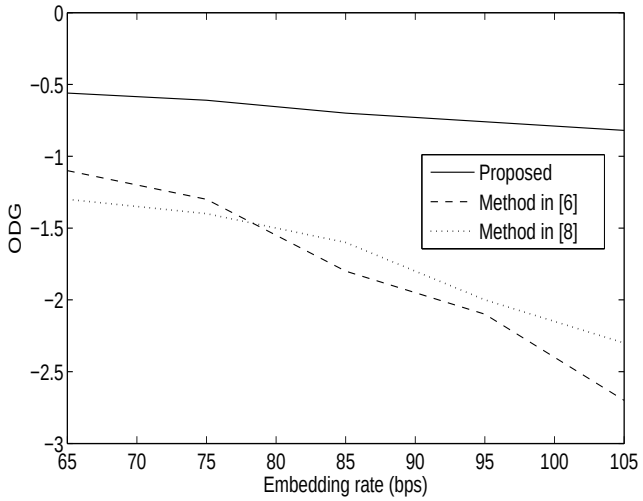

Fig. 4. ODGs versus embedding rates, where 100% DR under closed-loop attack is considered for all three methods.

embedding capacity dramatically. The watermark embedding is implemented by inserting a corresponding PN sequence into a pair of fragments in the segment, which have similar property. This embedding scheme can significantly reduce the host signal interference occurred in the process of watermark extraction and thus enhances robustness. Moreover, our method can adaptively control the amplitude of PN sequences, which improves perceptual quality. So, the proposed method exhibits superior performance in terms of high imperceptibility, robustness and embedding capacity. Compared with the latest SS-based audio watermarking methods, the proposed method can achieve much higher embedding capacity, while ensuring high level of imperceptibility and robustness. Its effectiveness is demonstrated by simulation results.

REFERENCES

[1] *Digital Music Report 2009*, International Federation of the Phonographic Industry, 2009.

[2] *Digital Music Report 2011*, International Federation of the Phonographic Industry, 2011.

[3] N.K. Kalantari, M.A. Akhaee, S.M. Ahadi, and H. Amindavar, "Robust multiplicative patchwork method for audio watermarking," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 6, pp. 1133-1141, Aug. 2009.

[4] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shanon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673-1687, Dec. 1997.

[5] H. S. Malvar and D. A. Florencio, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 898-905, Apr. 2003.

[6] A. Valizadeh and Z. J. Wang, "Correlation-and-bit-aware spread spectrum embedding for data hiding," *IEEE Trans. Information Forensics and Security*, vol. 6, no. 2, pp. 267-282, Jun. 2011.

[7] P. Zhang, S. Xu, and H. Yang, "Robust audio watermarking based on extended improved spread spectrum with perceptual masking," *Int. Journal of Fuzzy Syst.*, vol. 14, no. 2, pp. 289-295, Jun. 2012.

[8] X. Zhang and Z.J. Wang, "Correlation-and-bit-aware multiplicative spread spectrum embedding for data hiding," in *Proc. 2013 IEEE Int. Workshop on Information Forensics and Security*, 2013, pp. 186-190.

[9] H. Wang, R. Nishimura, Y. Suzuki, and L. Miao, "Fuzzy self-adaptive digital audio watermarking based on time-spread echo hiding," *Applied*

Acoustics, vol. 69, no. 10, pp. 868-874, Oct. 2008.

[10] O. T.-C. Chen and W.-C. Wu, "Highly robust, secure, and perceptual-quality echo hiding scheme," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 3, pp. 629-638, Mar. 2008.

[11] B.-S. Ko, R. Nishimura, and Y. Suzuki, "Time-spread echo method for digital audio watermarking," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 212-221, Apr. 2005.

[12] Y. Xiang, D. Peng, I. Natgunanathan, and W. Zhou, "Effective pseudonoise sequence and decoding function for imperceptibility and robustness enhancement in time-spread echo based audio watermarking," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 2-13, Feb. 2011.

[13] Y. Xiang, I. Natgunanathan, D. Peng, W. Zhou, and S. Yu, "A dual-channel time-spread echo method for audio watermarking," *IEEE Trans. Information Forensics and Security*, vol. 7, no. 2, pp. 383-392, Apr. 2012.

[14] H. J. Kim and Y. H. Choi, "A novel echo-hiding scheme with backward and forward kernels," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 13, no. 8, pp. 885-889, Aug. 2003.

[15] H. Kang, K. Yamaguchi, B. Kurkoski, K. Yamaguchi, and K. Kobayashi "Full-index-embedding patchwork algorithm for audio watermarking," *IEICE Trans. Inf. and Syst.*, vol. E91-D, no. 11, pp. 2731-2734, Nov. 2008.

[16] 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, and Language Process.*, vol. 22, no. 9, pp. 1413-1423, Sept. 2014.

[17] I. Natgunanathan, Y. Xiang, Y. Rong, W. Zhou, and S. Guo, "Robust patchwork-based embedding and decoding scheme for digital audio watermarking," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 8, pp. 2232-2239, Oct. 2012.

[18] S. Kirbiz and B. Gunsel, "Robust audio watermark decoding by supervised learning," in *Proc. 2006 IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2006, pp. 761-764.

[19] D. Lakshmi, R. Ganesh, R. Marni, R. Prakash, and P. Arulmozhivarman, "SVM based effective watermarking scheme for embedding binary logo and audio signals in images," in *Proc. 2008 IEEE Region 10 Conf.*, 2008, pp. 1-5.

[20] C. M. Pun and X. C. Yuan, "Robust segments detector for desynchronization resilient audio watermarking," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 11, pp. 2412-2424, Nov. 2013.

[21] K. Khaldi and A. O. Boudraa, "Audio watermarking via emd," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 3, pp. 675-680, Mar. 2013.

[22] A. Abrardo and M. Barni, "A new watermarking scheme based on antipodal binary dirty paper coding," *IEEE Trans. Information Forensics and Security*, vol. 9, no. 9, pp. 1380-1393, Sept. 2014.

[23] J. L. Wu and J. Shin, "Discrete cosine transform in error control coding," *IEEE Trans. Communications*, vol. 43, no. 5, pp. 1857-1861, May 1995.

[24] C. Baras, N. Moreau, and P. Dymarski, "Controlling the inaudibility and maximazing the robustness in an audio annotation watermarking system," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 5, pp. 1772-1782, Sep. 2006.

[25] *Rec.B. S. 1387: Methods for Objective Measurements of Perceived Audio Quality*, Rec. B.S. 1387, Int. Telecomm. Union, Geneva, Switzerland, 2001.