Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 5067313, 15 pages
http://dx.doi.org/10.1155/2016/5067313

*Research Article*

# Audio Watermarking Scheme Based on Singular Spectrum Analysis and Psychoacoustic Model with Self-Synchronization

**Jessada Karnjana,[1,2] Masashi Unoki,[1] Pakinee Aimmanee,[2] and Chai Wutiwiwatchai[3]**

[1]*School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*
[2]*Sirindhorn International Institute of Technology, Thammasat University, 131 Moo 5, Tiwanon Rd., Bangkadi, Muang, Pathum Thani 12120, Thailand*
[3]*National Electronics and Computer Technology Center, 112 Thailand Science Park, Phahonyothin Rd., Klong Luang, Pathum Thani 12120, Thailand*

Correspondence should be addressed to Jessada Karnjana; jessada@jaist.ac.jp

Received 28 June 2016; Revised 13 September 2016; Accepted 24 October 2016

Academic Editor: Kai Wang



This paper proposes a blind, inaudible, and robust audio watermarking scheme based on singular spectrum analysis (SSA) and the psychoacoustic model 1 (ISO/IEC 11172-3). In this work, SSA is used to analyze the host signals and to extract the singular spectra. A watermark is embedded into the host signals by modifying the singular spectra which are in the convex part of the singular spectrum curve so that this part becomes concave. This modification certainly affects the inaudibility and robustness properties of the watermarking scheme. To satisfy both properties, the modified part of the singular spectrum is determined by a novel parameter selection method based on the psychoacoustic model. The test results showed that the proposed scheme achieves not only inaudibility and robustness but also blindness. In addition, this work showed that the extraction process of a variant of the proposed scheme can extract the watermark without assuming to know the frame positions in advance and without embedding additional synchronization code into the audio content.

## 1. Introduction

Since the last decade, music sharing via the Internet has caused the music industry to lose annual sales of more than 3 billion US dollars [1] because the Internet is a good distribution system; that is, it distributes audio signals widely and very rapidly. In addition, all digital products have special characteristics; that is, they are expensive to produce for the first copy, but cheap to reproduce for duplicates [2]. One potential solution for protecting the digital content is audio watermarking [3]. Also, audio watermarking has been proposed as a solution for other purposes, such as ownership protection, content authentication, broadcast monitoring, information carrier, and covert communication [4–8].

The audio watermarking system consists of two processes: the embedding process and the extraction process, as illustrated in Figure 1. The first process embeds the watermark into the host audio signal. The second process extracts the watermark from the watermarked signal. Normally, the embedding process is frame-based. Therefore, the extraction process requires the frame positions in order to extract the watermark. The frame position requirement raises the frame synchronization problem. This problem is to be discussed in great detail in Section 3. Audio watermarking systems can be characterized by a number of properties [4]. Among them, there are five important properties [3, 9].

*(i) Inaudibility*. It is the property that the watermark does not affect the perceptual quality of the host signal.

*(ii) Robustness*. It is the ability to extract the watermark correctly when attacks are performed on the watermarked signals.

*(iii) Blindness*. It is the ability to be independent of the host signal in the extraction process. The system is blind when

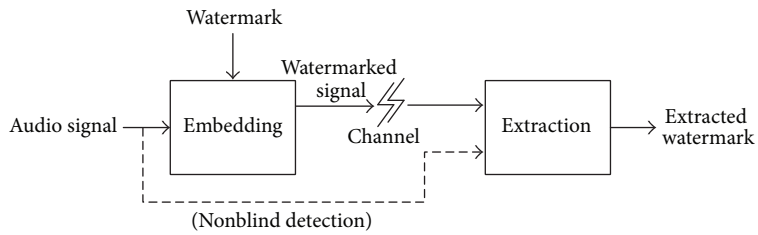


Figure 1: Audio watermarking system.

the extraction process does not require the host signal to be compared, in order to correctly extract the watermark. If the extraction process requires the host signal, as illustrated by the dashed line in Figure 1, it is nonblind.

*(iv) Confidentiality*. It is the property that keeps the watermark secret.

*(v) Capacity*. It is the quantity of the hidden information that is embedded into the host signals.

These required properties normally conflict with each other. Some techniques that obtain high robustness may suffer in inaudibility [10]. Some techniques are good at inaudibility but do not meet the blindness property [11]. Some with high capacity are not robust [12]. The method based on the least-significant-bit coding [13] obtains good inaudibility but loses on the robustness. The phase-coding method [14] achieves the inaudibility but fails the capacity. The phase modulation method [15] survives the inaudibility but does not pass the blindness property. The trade-off between the inaudibility and the robustness can be found in the methods based on adaptive phase modulation as well [16]. The method based on cochlear delay characteristics [17] is robust and inaudible. However, it has a significant trade-off between the inaudibility and the capacity. In addition, the blind cochlear-delay-based scheme reduces the sound quality of the watermarked signals, compared with nonblind ones. The methods based on echo hiding [18, 19] are blind and robust, but they perform poorly in inaudibility and confidentiality. The spread-spectrum-based technique is good in robustness but is poor in inaudibility and capacity [20]. These examples show that balancing among the required properties has always been a difficult task.

A literature review of audio watermarking has suggested that the schemes based on Singular-Value Decomposition (SVD) are robust [10, 11, 21–26]. In general, the SVD-based scheme extracts the singular values from the host signals and slightly changes some of those values with respect to the watermark bit. It is robust because the singular values are unchanged under common signal processing [27]. However, the balance between inaudibility and robustness for some audio signals needs to be further improved due to the fact that it has never taken the human perception into consideration.

The motivation for this work has started from the idea of exploiting the advantages of the SVD-based method and combining it with a human perceptual model. We turn to the SSA, which is SVD-based, and adopt it as the main analysis tool. We choose SSA because when a signal is analyzed, the singular values can be interpreted and have the physical meanings [28]. The physical meanings are of importance because they help us understand a relationship between the SSA and the perceptual model. Recently, we proposed the audio watermarking schemes based on the SSA [28, 29]. Also, we showed the benefits of using the SSA over the SVD. To verify the effectiveness of the SSA-based scheme, we used the *differential evolution* to adjust the balance. The results were quite successful [29]. However, as the search space was very large, therefore, the embedding process was time-consuming.

This work aims to show that SSA equipped with the perceptual model also gives the good balance between inaudibility and robustness. This work proposes a novel audio watermarking based on SSA and the human perceptual model. Also, it proposes a new method for automatic frame detection; that is, the frame positions are not required in the extraction process.

The rest of this paper is organized as follows. The proposed scheme and necessary background information are detailed in Section 2. Section 3 shows that we can slightly modify the proposed scheme to make it a self-synchronized one. The performance evaluation and experimental results are given in Section 4. The observations from the experiments are made and discussed in Section 5. Last, the whole work is summarized in Section 6.

## 2. Proposed Scheme

The proposed scheme is mainly based on the SSA-based audio watermarking scheme proposed by Karnjana et al. [28, 29]. The first two subsections are part of the embedding process, and the last two subsections are part of the extraction process. The proposed scheme with the self-synchronization is provided in Section 3.

*2.1. Embedding Process.* The embedding process consists of two major parts, as shown in Figure 2. The first part is a core structure. In this core structure, the basic SSA is mainly used to analyze the host signals and to extract the singular spectra. The basic SSA has experimentally proved to be useful for extracting meaningful information from signals [30, 31]. The second part, which is shown in the gray box, is the parameter selection method based on a psychoacoustic model. In this

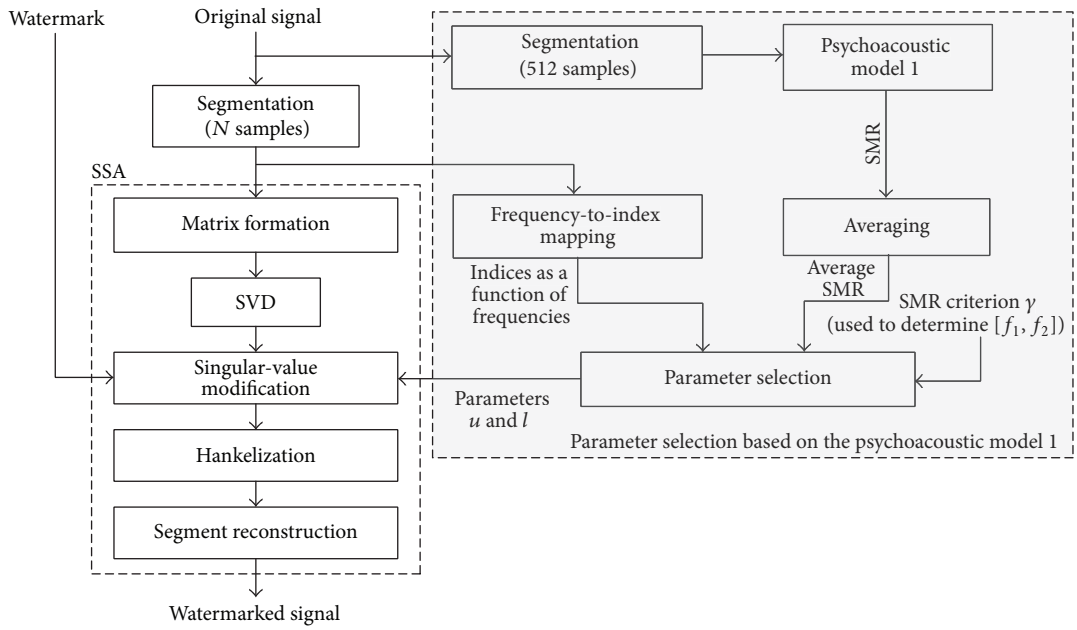


Figure 2: Embedding process.

work, we adopt the psychoacoustic model 1 (ISO/IEC 11172-3) [32] to the proposed scheme. The brief details of the psychoacoustic model and the parameter selection method are provided in the next subsection.

The core structure of the embedding process consists of six steps which are described as follows.

(1) The host audio signal is segmented into nonoverlapping frames. The number of frames is equal to the number of the watermark bits since one bit is embedded into one frame. Let $F = [f_0 \ f_1 \ f_2 \ \cdots \ f_{N-1}]^T$ denote a frame of size $N$, where $N$ is greater than 2. Remark that the embedding capacity is the sampling frequency of the host signal divided by $N$.

(2) The trajectory matrix $\mathbf{X}$ which represents each frame $F$ is constructed. The construction of $\mathbf{X}$ is done as follows:

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{bmatrix}, \quad (1)$$

where $L$, called a *window length* of the matrix formation, is the only parameter of the basic SSA and not greater than $N$, and $K = N - L + 1$.

(3) SVD is performed on each trajectory matrix $\mathbf{X}$ to obtain the singular spectrum $\{\sqrt{\lambda_0}, \sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{L-1}}\}$, where $\lambda_0, \lambda_1, \ldots$, and $\lambda_{L-1}$ denote the eigenvalues of $\mathbf{X}\mathbf{X}^T$.

(4) When the watermark bit is 1, we modify singular values $\sqrt{\lambda_i}$, for $i = u+1, u+2, \ldots, l-1$, given that $\sqrt{\lambda_u}$ is greater than $\sqrt{\lambda_l}$, as follows.

$$\sqrt{\lambda_i} = \sqrt{\lambda_l} + 0.9 \times \left(\sqrt{\lambda_u} - \sqrt{\lambda_l}\right). \quad (2)$$

When the watermark bit is 0, the singular values are left unchanged. In this step, there are two parameters, $u$ and $l$, and these parameters are determined by the parameter selection algorithm based on the psychoacoustic model.

(5) The modified trajectory matrix is constructed by SVD reversion, and then it is hankelized. The hankelization of a modified trajectory matrix $\mathbf{Y}$ to a signal $G = [g_0 \ g_1 \ g_2 \ \cdots \ g_{N-1}]^T$ is defined as follows:

$$g_k = \begin{cases} \dfrac{1}{k+1} \displaystyle\sum_{m=1}^{k+1} y^*_{m,k-m+2}, & \text{for } 0 \le k < L^* - 1, \\ \dfrac{1}{L^*} \displaystyle\sum_{m=1}^{L^*} y^*_{m,k-m+2}, & \text{for } L^* - 1 \le k < K^*, \\ \dfrac{1}{N-k} \displaystyle\sum_{m=k-K^*+2}^{N-K^*+1} y^*_{m,k-m+2}, & \text{for } K^* \le k < N, \end{cases} \quad (3)$$

where $y_{ij}$ is an element at the row $i$ and column $j$ of the matrix $\mathbf{Y}$, $L^* = \min(L, K)$, $K^* = \max(L, K)$, $y^*_{ij} = y_{ij}$ if $L < K$, and $y^*_{ij} = y_{ji}$ if $L \ge K$.

(6) Finally, the watermarked signal is reconstructed by stacking those hankelized frames.

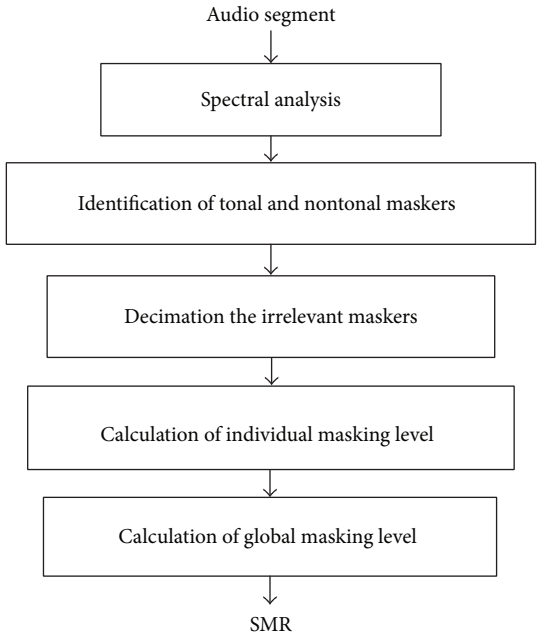


Figure 3: Psychoacoustic model 1.

*2.2. Parameter Selection Based on Psychoacoustic Model.* The block diagram of the parameter selection method based on the psychoacoustic model 1 is shown in the gray box of Figure 2. The psychoacoustic model 1, which is deployed in the MPEG-1 Layer 1, is adopted in order to deliver a signal-to-mask ratio (SMR) of the analyzed signal, and then the SMR is used as a criterion for selecting the parameters $u$ and $l$.

Basically, the psychoacoustic model 1 is built based on three psychoacoustic principles: the absolute threshold of hearing, the simultaneous masking, and the upward spread of masking [33]. It consists of five steps [33, 34], as shown in Figure 3. According to the standard ISO/IEC: 11172-3, the overview of the process is summarized as follows. First, the FFT and the power spectral density (PSD) of the signal are calculated, and then the PSD is normalized with the maximum sound pressure level (SPL) of 96 dB. Next, the PSD is used to identify the tonal (more sinusoid-like) and nontonal (more noise-like) components of the signal. This identification is used for the calculation of the masking levels due to the tonal and nontonal maskers. Then, the irrelevant maskers are removed by applying two psychoacoustic principles in the following manner. The maskers which are lower than the absolute threshold of hearing are removed, and only the strongest masker within a distance of 0.5 Bark is kept. Subsequently, these survival maskers are used to calculate the individual masking levels. Finally, we combine all masking levels to calculate the global masking level. The output of the psychoacoustic model is SMR. The SMR is defined as the difference between the SPL of the global masking level and the PSD of the analyzed signal. Figure 4 shows an example of the SMR (red line) of one frame.

In perceptual audio codings, such as MP3 compression, the SMR is used to allocate the quantization bits. The frequency components with the lower SMR are assigned

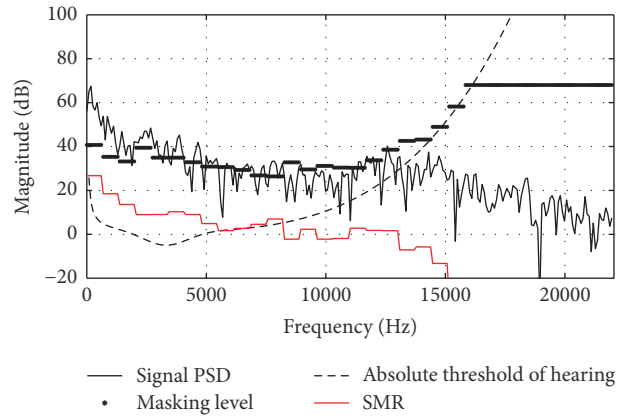


Figure 4: Signal PSD, global masking level, and SMR.

with smaller numbers of bits since the human auditory system is less sensitive to those frequency components. In this work, the SMR is used as guidance to determine the appropriate parameters because embedding the watermark into the components with the low SMRs helps to improve the inaudibility. The algorithm that we use to deliver the parameters $u$ and $l$ consists of the following five steps:

(1) We first calculate the SMR of each frame. According to the standard, the frame size used for this calculation is 512 samples. Note that the frame size from the segmentation of the core structure is not necessary to be the same as that of this psychoacoustic model.

(2) We use the SMRs obtained from the previous step to calculate the average SMR of the host signal.

(3) We identify the frequency band $[f_1, f_2]$ with the average SMR lower than a predefined value, $\gamma$. If there are more than one band, the band with the lowest frequency is selected. If the frequency bandwidth $f_2 - f_1$ is wider than a predefined bandwidth, it is limited to the predefined bandwidth. In our simulation, the predefined bandwidth is 10 kHz. An example is shown in Figure 5.

(4) For each frame, we map the selected band to a singular-value interval. In this step, we have to find the relationship between the frequencies and the singular-value indices because the output of the psychoacoustic model, the SMR, is expressed as a function of frequency. When the basic SSA is used to decompose a signal, the singular values of the matrix representing the signal can be interpreted as the scale factors of the oscillatory components of the signal [28]. After analyzing each oscillatory component by the Fourier transform, we found that a frequency band of each oscillatory component is quite narrow compared with the signal bandwidth, as shown in Figure 6. We associate the index of the singular value with the peak frequency of its oscillatory components. Figure 7 shows an example of the relationship between the frequencies and the singular-value indices. Thus, to map the frequency

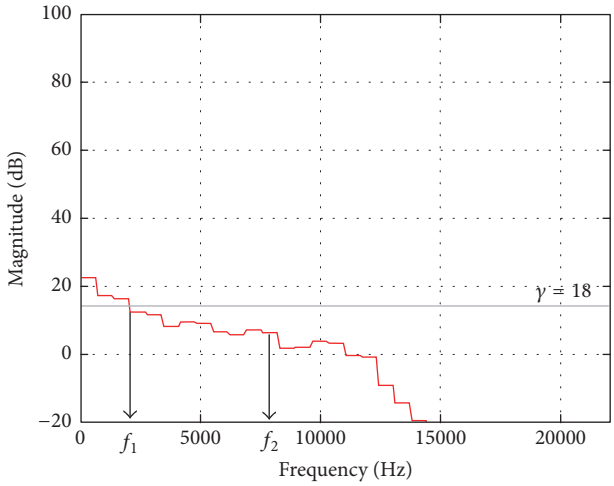


FIGURE 5: Average SMR (red) and an example of selecting the frequency band $[f_1, f_2]$ given the SMR criterion $\gamma = 18$ dB. If the frequency bandwidth $f_2 - f_1$ is wider than 10 kHz, it is limited to 10 kHz.

range $[f_1, f_2]$ to the interval $[u, l]$, we first find the local minimum which is closest to $f_1$, and set $u$ the index of this local minimum. Then, we find the local maximum that is closest to $f_2$ which must be on the right side of $u$, and then set $l$ the index of this local maximum. An example of this mapping is shown in Figure 8. Note that different frames may have different intervals from one another, and the word *frame* in this step means the frame from the segmentation process.

(5) Finally, the parameters $u$ and $l$ for embedding the watermark are selected using the arithmetic mean of boundaries of all intervals.

*2.3. Extraction Process.* The plot of singular spectra is normally convex, as shown in Figure 9. However, after the watermark bit 1 is embedded into an interval $[u + 1, l - 1]$ of the singular spectrum of a host frame, the embedding process causes the concave part on the interval $[u + 1, l - 1]$ of the singular spectrum of the reconstructed, watermarked frame [29], as shown in Figure 10. We exploit this property to extract the watermark bit. The extraction process consists of five steps, as shown in Figure 11. The details of each step are as follows.

(1) We segment the watermarked signal into nonoverlapping frames. At this stage, we assume that we know the frame positions and the frame size.

(2) We construct the trajectory matrix in the same way we do in the embedding process.

(3) We perform the SVD operation on the trajectory matrix to obtain the singular spectrum.

(4) If the parameters $u$ and $l$ are not provided, automatic parameter estimation, which is illustrated in the gray box of Figure 11, is used to estimate them. The details of this parameter estimation are given in the upcoming subsection.

(5) We approximate all singular values of $[u + 1, l - 1]$ using the quadratic equation, $y(x) = ax^2 + bx + c$, where $y$ is the singular value and $x$ is the index of the singular value. Since the coefficient $a$ of the quadratic formula indicates the rate of change of the singular values, the sign of the coefficient $a$ is used to determine the watermark bit. A minus sign indicates concavity or the watermark bit 1, and a plus sign indicates convexity or the watermark bit 0.

*2.4. Automatic Parameter Estimation.* To automatically estimate the parameters $u$ and $l$, we use the fact that when watermark bit 1 is embedded into a frame, there exists a concave part in the singular spectrum plot. In other words, when watermark bit 1 is embedded into the frame, we can find some pairs of indices $m$ and $n$, where $m < n$, such that the singular values of the interval $[m, n]$ are mostly above the line segment connecting two singular values $\sqrt{\lambda_m}$ and $\sqrt{\lambda_n}$. Thus, this automatic parameter estimation estimates the parameters from the width of the concave part.

We first define the concavity density as a measurement of degree of the concavity. Given a singular spectrum $\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_d}\}$, the concavity density $D_{m,n}$ of singular values from $\sqrt{\lambda_m}$ to $\sqrt{\lambda_n}$ is defined as follows.

$$D_{m,n} = \sum_{i=m+1}^{n-1} \left( \sqrt{\lambda_i} - L(i) \right),$$
$$L(i) = \sqrt{\lambda_m} + \frac{\sqrt{\lambda_m} - \sqrt{\lambda_n}}{m - n} (i - m), \quad (4)$$

where $L(i)$ is the function defining the line connecting $\sqrt{\lambda_m}$ and $\sqrt{\lambda_n}$.

Starting from the first singular value ($m = 1$), a sequence of the singular values that is used to calculate the concavity density is shifted to the right by one singular-value point at a time to determine the set of the concavity density $\{D_{1,n}, D_{2,n+1}, \ldots, D_{i,n+i-1}, \ldots, D_{d-n+1,d}\}$. An example of the positive and the negative concavity density of two sequences of the singular values is shown in Figure 12.

Figure 13 shows an example of the concavity density curve of the singular spectrum in Figure 12 when a sequence of the singular values used to calculate the concavity density has a length of 30. It can be seen that the positive density roughly corresponds to the concave part of the singular spectrum. However, the concavity density depends upon the choice of the length of the sequence used to calculate the concavity density. In this work, we get around the problem by using the average density at the different lengths. Then, the average-density curve is refined as follows. First, any negative-density value is ignored because it implies convexity. Second, any positive density curve that is narrower than $\Gamma \times (l - u)$, where $\Gamma$ is a user-defined real number around 1, is neglected because, practically, we can set the minimum value of $l - u$ in advance.

Subsequently, the indices at the rising and falling edges of the consequent density curve, together with an offsetting

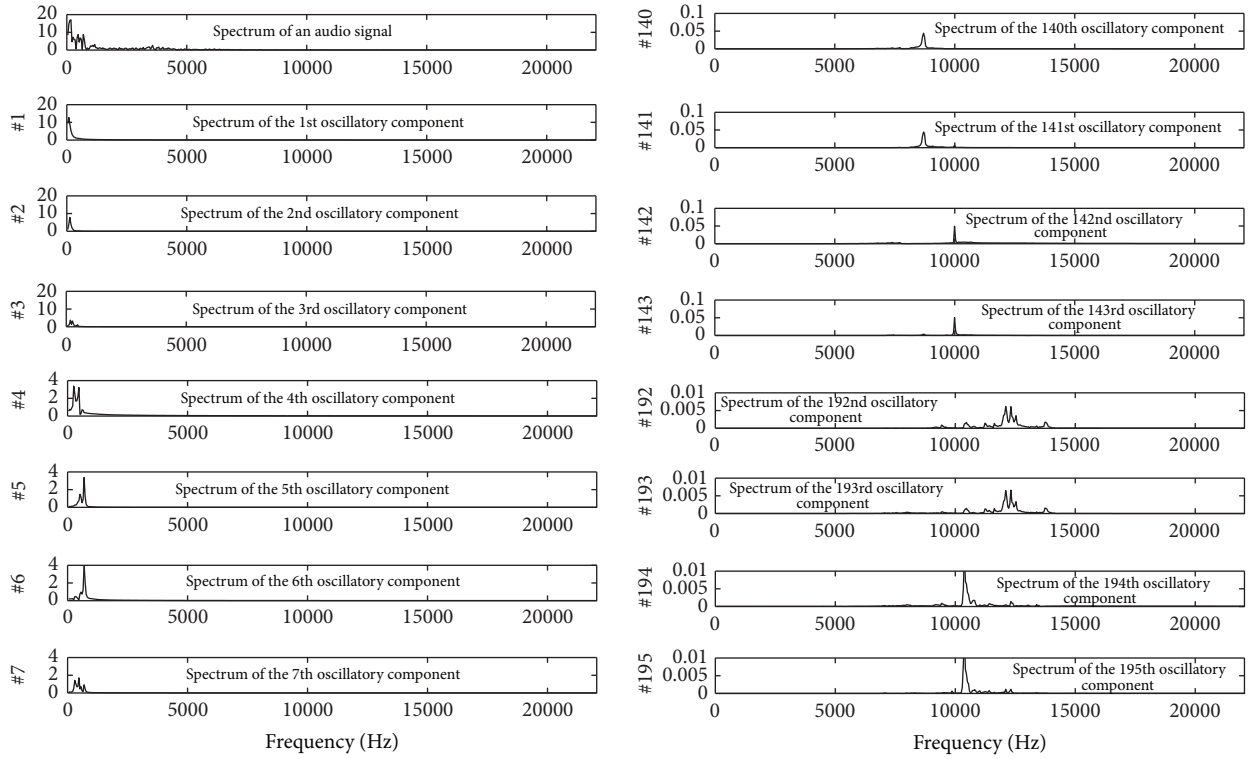


Figure 6: Examples of spectra of oscillatory components: the top-left panel is a spectrum of an audio signal, and the other panels show spectral of some oscillatory components. The numbers labeled on vertical axes are the orders of the components, which are indices of their associated singular values. Compared with the spectrum of the audio signal, the spectral of oscillatory components has narrower bandwidths.

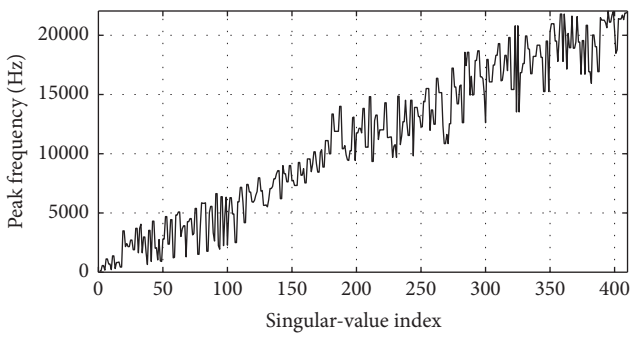


Figure 7: Relationship between frequency and singular-value index (frame size, $N = 1024$).

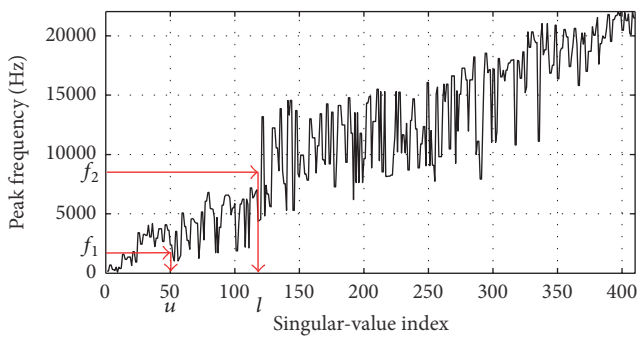


Figure 8: Example of parameter selection: frequency range $[f_1, f_2]$ is mapped to the interval $[u, l]$.

constant, are used to estimate the parameters $u$ and $l$ for the given frame. Finally, the parameters $u$ and $l$ for the watermarked signal are calculated by averaging the estimated parameters $u$ and $l$ from all frames. The averaging algorithm is depicted in Figure 14 and detailed as follows.

Let $\widehat{u}_{i,j}$ and $\widehat{l}_{i,j}$ for $j = 1, 2, \ldots, k_i$ denote the estimated parameters of the frame $i$. The subscripts $j$ indicate that there can be more than one concave interval detected within one frame. The maximum number of intervals detected within the frame $i$ is denoted by $k_i$.

The general idea of the averaging algorithm is as follows. Given two integral intervals $[\mu_1, \iota_1]$ and $[\mu_2, \iota_2]$, where $\mu_1$, $\iota_1$,

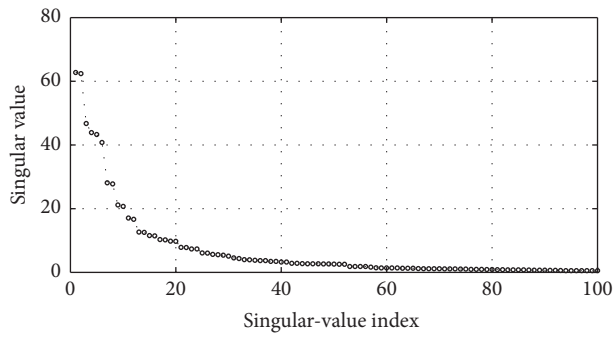


Figure 9: Singular spectrum.

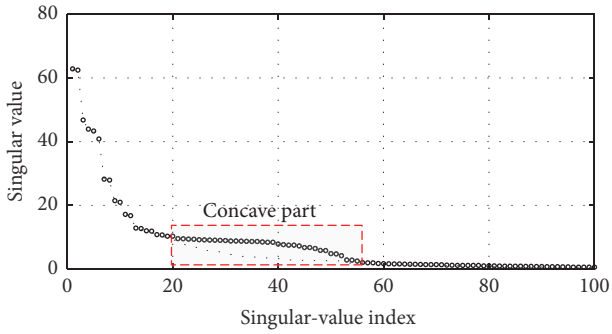


FIGURE 10: Singular spectrum with a concave part due to embedding watermark bit 1.

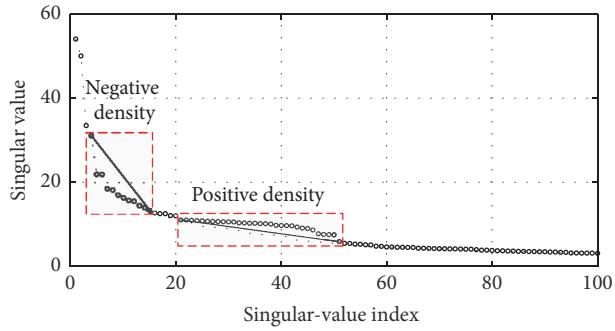


FIGURE 12: Examples of regions with negative and positive concavity density.

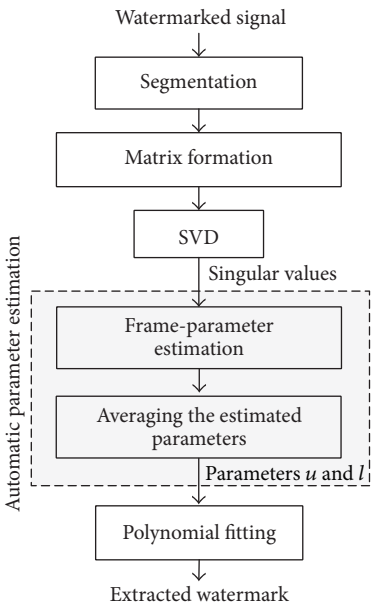


FIGURE 11: Extraction process.

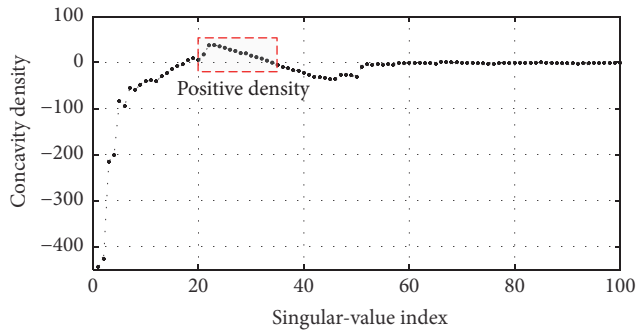


FIGURE 13: Concavity density set $\{D_{1,31}, D_{2,32}, \ldots, D_{100,130}\}$. Notice that there is a strong relationship between the region of positive density and the indices of modified singular values.

$\mu_2$, and $\iota_2$ are integers and $\iota_1 - \mu_1 \leq \iota_2 - \mu_2$, we say that there is an overlap between those two intervals if $\mu_2 < \iota_1 \leq \iota_2$ or $\mu_2 \leq \mu_1 < \iota_2$. For a pair of overlapping intervals, $([\mu_1, \iota_1], [\mu_2, \iota_2])$, we define the overlap degree $\eta$ as

$$\eta = \frac{\max(\iota_1, \iota_2) - \min(\mu_1, \mu_2)}{\min(\iota_2 - \mu_1, \iota_1 - \mu_2)}, \tag{5}$$

where $\max(\cdot)$ and $\min(\cdot)$ are the maximum and minimum functions, respectively.

Given a set $E$ of the estimated parameter-interval $[\widehat{u}_{i,j}, \widehat{l}_{i,j}]$, we can expect that the set $E$ must contain many overlapping intervals $[\widehat{u}_{i,j}, \widehat{l}_{i,j}]$. By the same token, we know that there is no overlap between intervals $[\widehat{u}_{i,j_1}, \widehat{l}_{i,j_1}]$ and $[\widehat{u}_{i,j_2}, \widehat{l}_{i,j_2}]$ when $j_1 \neq j_2$. Then, the averaging algorithm is just the process of recursively grouping the overlapping members of the set $E$. The following is the procedure used in the averaging algorithm.

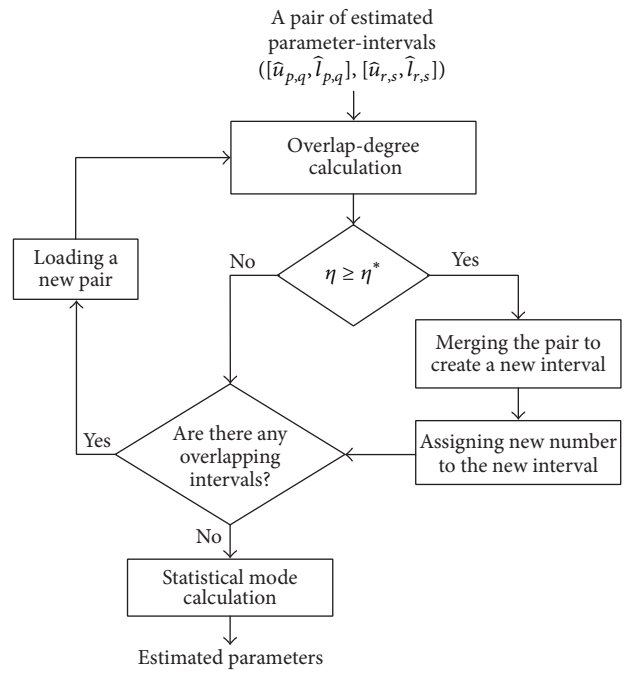


FIGURE 14: Averaging algorithm used in the automatic parameter estimation.

(1) We assign the frequency weight to each interval $[\widehat{u}_{i,j_2}, \widehat{l}_{i,j_2}]$ in the set $E$. Initially, the frequency weight is set to 1.

(2) We calculate the overlap degree $\eta$ of a pair of estimated parameter-intervals. If $\eta$ is greater than a predefined value $\eta^*$, the two intervals are merged to create a new interval. Then, the two old intervals are removed. The frequency weight of the new interval is the sum of the frequency weights of the two old intervals. The average of the lower bounds and that of the upper bounds of the old intervals are used for the new interval.

(3) Step (2) is repeated until set $E$ has no overlapping members.

(4) The interval with the highest frequency is chosen as the estimated parameters $\widehat{u}$ and $\widehat{l}$. If there are multiple intervals with the highest frequency, the estimated parameters $\widehat{u}$ and $\widehat{l}$ are randomly chosen from them.

## 3. Self-Synchronization

The embedding and extraction processes as described in previous section are frame-based. That means that the host signal is divided into frames, and one watermark bit is embedded into one frame. Thus, to correctly extract the watermark, the extraction process must know the frame positions. The assumption that the extraction process knows the frame positions in advance may not be practical in some situations. For example, an attacker can attack watermarked signals by cutting a few audio samples. This causes the extraction process to work improperly. This is known as a cropping attack. How the frame positions are acquired is the frame synchronization problem.

There are two solutions to solve the frame synchronization problem [11]. The first solution is by binding the watermark with some invariant audio features of the host signal [35] or performing self-synchronization [36–38]. The second solution is by embedding the frame synchronization code into the host signal [39, 40].

From experiments, we found that the proposed scheme can automatically detect the watermarked frames. In order to do that, we need to modify the scheme slightly. To fully grasp the idea behind the new rules, let us start with the basic findings from this work.

Consider an audio signal with three frames of equal length $N$, where the watermark bit 1 is embedded in its middle frame by the method described in Section 2.1, as illustrated in Figure 15. The starting and the last indices of audio samples of the middle frame $G_{n,N}$ are denoted by $n$ and $n+N-1$, respectively. According to the embedding and extraction processes, if we use the frame $G_{n,N}$ to construct the trajectory matrix, then we can detect the concave pattern in the singular spectrum plot. If $\tau$ is an integer which is less than $N$, then the frames $G_{n,N}$ and $G_{n-\tau,N}$ are overlapping. We discovered that the singular spectrum curve of the trajectory matrix constructed by the frames $G_{n-\tau,N}$ also has a concave part if the overlapping region is large enough. A similar

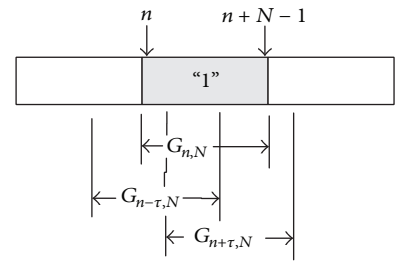


Figure 15: Example of an audio clip with 3 frames and three segments from which trajectory matrices are constructed.

effect occurs to the frame $G_{n+\tau,N}$ as well. In general, if we construct matrices from frames $G_{i,N}$ for $i = 0$ to $2N$, there are many matrices that we can interpret as having the watermark bit 1 embedded. Those matrices are the ones in which $i$ is in the same neighborhood with $n$. This overlapping effect of embedding the watermark bit 1 is utilized in our automatic frame detection. We perform a scanning operation by first constructing the frames $G_{i,N}$ for $i = 0$ to the last possible frame and then extracting the watermark bit from those frames. This effect implies that we can localize the watermarked frame where watermark bit 1 is embedded by performing a scan operation. This is the reason why we need to modify the proposed scheme if we want to make it self-synchronizing. The modification is as follows.

We first divide the frame into 4 equal subframes, where each subframe has a length of $M$. Each watermark bit is represented by the four-bit strings of either "0100" or "0110" depending upon the watermark bit. If the watermark bit is 0, four bits of "0100" are embedded into the 4 subframes. If the watermark bit is 1, "0110" are embedded into those subframes, as illustrated in Figure 16. For example, if the watermark bits are "001", then the subframe-embedding bits are "010001000110".

Given a frame $G = [g_0 \;\; g_1 \;\; \cdots \;\; g_{4M-1}]^T$ of length $4M$, we define the subframe-scan operator on $G$ as follows.

$$\text{Scan}\,[G] = (b_0, b_1, \ldots, b_{\lfloor 3M/\delta \rfloor}), \tag{6}$$

$$b_i = \xi\,(u, l, G_{i\delta,M}), \tag{7}$$

where $\delta$ is a scan-step size, $G_{i\delta,M}$ is the subframes $[g_{i\delta} \;\; g_{i\delta+1} \;\; \cdots \;\; g_{i\delta+M-1}]^T$, for $i = 0$ to $3M$, of the frame $G$, and $\xi(u, l, G_{i\delta,M})$ is 1 if the singular spectrum curve of the matrix constructed from the subframe $G_{i\delta,M}$ on the interval $[u+1, l-1]$ is concave; otherwise, $\xi(u, l, G_{i\delta,M})$ is 0.

The meaning of this operator is that the scanner Scan[·], which operates on $M$ samples, scans through the frame $G$ with step size $\delta$ and returns to 0 or 1, depending upon the characteristics of the singular spectra of the scanned subframes.

We use the first appearance of "1" in "0100" and "0110" as the synchronization point of watermark bit 0 and 1, respectively. If we can detect the next concavity, we interpret it as the watermark bit 1; otherwise, it is 0. Since the first detected concavity is used as the synchronization point, to ensure that all concavities are surrounded by convexities and

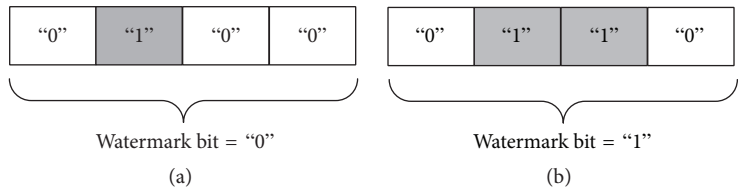


Figure 16: Four bits of "0100" are embedded into 4 subsegments of a frame, which represents embedding "0" (a), and 4 bits of "0110" are embedded into 4 subsegments of a frame, which represents embedding "1" (b).

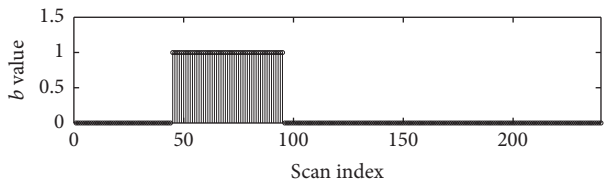


Figure 17: Illustration of performing the subframe-scan operation on a 3200-sample frame; that is, $M = 800$, with $\delta = 10$, $u = 30$, $l = 80$, and "1" being embedded into the second subframe of the frame.

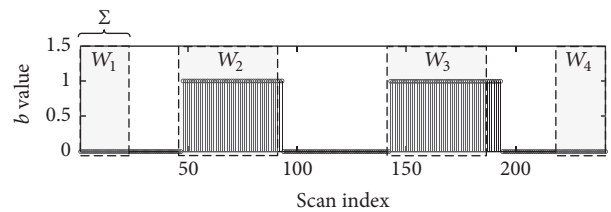


Figure 18: Example of performing the four windows on Scan$[Y_i]$.

that the distance between two concavities is far enough, "0" is added at the first and the last of the four-bit patterns. This is the concept behind our new proposed self-synchronization. An example of performing the subframe-scan operation according to (6) is shown in Figure 17.

To detect a watermarked frame, we define another scanner, which operates on $4M$ samples, called the *frame-scan operation*. Given a watermarked audio signal $Y = [y_0 \ y_1 \ \cdots \ y_{H-1}]^T$ of length $H$ greater than $4M$, the frame-scan operation $F[Y]$ scans from $y_0$ with a scan step of $\Delta$ until it detects the first watermarked frame.

Given $Y_i = [y_{i\Delta} \ y_{i\Delta+1} \ \cdots \ y_{i\Delta+4M-1}]^T$ is a frame scanned at step $i$, we first perform Scan$[Y_i] = (b_0, b_1, \ldots, b_{\lfloor 3M/\delta \rfloor})$.

Let four rectangular windows $W_j = (w_0^j, w_1^j, \ldots, w_{\lfloor 3M/\delta \rfloor}^j)$, for $j = 1$ to 4, where

$$w_k^j = \begin{cases} 1 & \text{if } k \in [\sigma, \sigma + \Sigma], \text{ for } j = 1, 2, 3, \\ 1 & \text{if } k \in [\sigma - \Sigma, \sigma], \text{ for } j = 2, 3, 4, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

and $\sigma = \lfloor 3M/4\delta \rfloor (j-1)$, and $\Sigma$ is a positive integer, called the *overlap margin*. Then, each of these windows is element-wisely multiplied with Scan$[Y_i]$; that is, $W_j * \text{Scan}[Y_i] = (b_0^j, b_1^j, \ldots, b_{\lfloor 3M/\delta \rfloor}^j)$, where the value of $b_k^j$ is calculated by the following equation.

$$b_k^j = w_k^j \times b_k, \tag{9}$$

for $k = 0, 1, \ldots, \lfloor 3M/\delta \rfloor$.

If $\sum_{k=0}^{\lfloor 3M/\delta \rfloor} b_k^j$ is greater than $\lceil \Sigma/2 \rceil$ for $j = 1$ or 4, or if $\sum_{k=0}^{\lfloor 3M/\delta \rfloor} b_k^j$ is greater than $\Sigma + 1$ for $j = 2$ or 3, we say that, by looking through the window $W_j$, the concavity of the singular spectrum is detected.

Table 1: Conditions for stopping frame-scan operation. Note that "∘" is "The concavity of singular spectrum can be detected," and "×" is "The concavity of singular spectrum cannot be detected."

| | $W_1$ | $W_2$ | $W_3$ | $W_4$ |
|---|---|---|---|---|
| Watermark bit $w = 0$ | × | ∘ | × | × |
| Watermark bit $w = 1$ | × | ∘ | ∘ | × |

The scanner $F[Y]$ stops scanning and declares a watermarked frame only when the conditions described in Table 1 are satisfied. The extracted watermark bit 0 is detected if and only if the concavity of singular spectrum cannot be detected through the windows $W_1, W_3$, and $W_4$ but can be detected through window $W_2$. In comparison, the extracted watermark bit 1 is detected if and only if the concavity of the singular spectrum curve cannot be detected through the windows $W_1$ and $W_4$ but can be detected through windows $W_2$ and $W_3$. Otherwise, it continues scanning with a step size of $\Delta$. The frame-scan operation is restarted repeatedly until it reaches the end of the watermarked signal $Y$.

An example of performing the four windows on one frame is shown in Figure 18. In this figure, the second and third subframes are embedded so that the frame-scan operation can decode the pattern of $b_i$ as the watermarked bit 1.

## 4. Evaluation

Twelve host signals from the RWC music-genre database (Track numbers 01, 07, 13, 28, 37, 49, 54, 57, 64, 85, 91, and 100) [41] were used in our experiments. All have a sampling rate of 44.1 kHz, 16-bit quantization, and two channels. Unless stated otherwise, the hidden information was embedded in one channel, starting from the initial segment of host signals. The frame size $N$ was set to 2450 samples. The embedding capacity was 18 bit per second (bps). We chose this capacity

Table 2: Parameters $u$ and $l$ obtained from the parameter selection based on the psychoacoustic model 1.

| | 01 | 07 | 13 | 28 | 37 | 49 | 54 | 57 | 64 | 85 | 91 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter $u$ | 30 | 40 | 20 | 45 | 40 | 40 | 30 | 75 | 30 | 60 | 40 | 35 |
| Parameter $l$ | 90 | 90 | 60 | 110 | 120 | 100 | 90 | 160 | 90 | 140 | 100 | 93 |

Table 3: Estimated parameters $u$ and $l$ obtained from the automatic parameter estimation in the extraction process.

| | 01 | 07 | 13 | 28 | 37 | 49 | 54 | 57 | 64 | 85 | 91 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimated parameter $\hat{u}$ | 26 | 39 | 23 | 43 | 44 | 38 | 30 | 81 | 28 | 64 | 37 | 38 |
| Estimated parameter $\hat{l}$ | 90 | 93 | 66 | 110 | 120 | 99 | 90 | 155 | 90 | 140 | 100 | 95 |

because the number is not too low or not too high, and it seems reasonable for general applications. The window length $L$ for the matrix formation was 980. One hundred and fifty bits of the watermark were embedded in total. The audio duration of each signal was about 8.33 seconds.

The parameters $u$ and $l$, obtained from the parameter selection based on the psychoacoustic model 1, are shown in Table 2. The estimated parameters, obtained from the automatic parameter estimation, are shown in Table 3. We implemented the proposed scheme using an adaptive criterion for the predefined SMR level $\gamma$ as follows. If the maximum SMR is greater than 25 dB, $\gamma = 18$. If the maximum SMR is less than 20 dB, $\gamma = 12$. Otherwise, $\gamma = 15$.

The proposed schemes were compared with the previously proposed schemes [28, 29] and the conventional SVD-based scheme [23]. There are three reasons for comparing with the conventional SVD-based scheme. First, it is one of a few blind SVD-based techniques. Second, its published results are promising. Last, both the SSA-based and SVD-based schemes belong to the same family of audio watermarking schemes; that is, they extract singular values from the host signals and embed the information into the signals by modifying those singular values. The following subsections report evaluations of the performance in the aspects of sound quality, robustness, and self-synchronization.

*4.1. Sound-Quality Evaluation.* Three distance measures were chosen to evaluate the sound quality of watermarked signals: the evaluation of audio quality (EAQUAL) [42], log-spectral distance (LSD), and the signal-to-distortion ratio (SDR). The EAQUAL measures the degradation of the watermarked signal, compared with the original, and covers a scale, called the objective difference grade (ODG), from −4 (very annoying) to 0 (imperceptible).

The LSD is a distance measure between two spectra. Given $P(\omega)$ and $\widehat{P}(\omega)$ are power spectra of the original and the watermarked signals, respectively, the LSD is defined as the following formula:

$$\mathrm{LSD} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log\frac{P(\omega)}{\widehat{P}(\omega)}\right]^2 d\omega}. \tag{10}$$

Table 4: ODGs, LSDs, and SDRs: comparison of the conventional SVD-based method (SVD), SSA-based scheme (SSA), SSA-based scheme with the differential evolution (SSA.DE), proposed scheme without the automatic parameter estimation (Prop.), and proposed one with the automatic parameter estimation (Prop.APE).

| | ODG | LSD | SDR |
|---|---|---|---|
| SVD [23] | −0.23 | 0.11 | 26.82 |
| SSA [28] | −0.70 | 0.36 | 20.96 |
| SSA.DE [29] | −0.10 | 0.16 | 35.25 |
| Prop. | −0.52 | 0.25 | 25.61 |
| Prop.APE | −0.52 | 0.25 | 25.61 |

The SDR is a power ratio between the signal and the distortion. Given the amplitudes of original and watermarked signals, $A_{\mathrm{org}}(n)$ and $A_{\mathrm{wmk}}(n)$, the SDR is defined as follows.

$$\mathrm{SDR} = 10\log\frac{\sum_n \left[A_{\mathrm{org}}(n)\right]^2}{\sum_n \left[A_{\mathrm{org}}(n) - A_{\mathrm{wmk}}(n)\right]^2}. \tag{11}$$

The evaluation criteria for good sound quality are as follows. The ODG must be greater than −1 (not annoying), the LSD must be less than 0.4 dB, and the SDR must be greater than 25 dB. An ODG of −1 indicates that the noise perceived in the watermarked signal is perceptible but not annoying. Based on our simulations, ODG values between 0 and −1 mean excellent in sound quality. We set the criteria for LSD and SDR to 0.4 dB and 25 dB, respectively, because we found from our preliminary experiments that either an LSD greater than 0.4 dB or an SDR lower than 25 dB can cause an annoying perception.

The comparison of the average ODGs, average LSDs, and average SDRs is shown in Table 4. The proposed scheme satisfies the inaudibility criteria and is considerably improved when it is compared with the SSA-based method [28]. Compared with the conventional SVD-based method and the SSA-based method with differential evolution [29], the proposed method is less inaudible. However, the difference in the inaudibility among them is nonsignificant. Based on our listening-test experiment [29], we found that the signals that satisfy all conditions ODG $>$ −1, LSD $<$ 0.4 dB, and SDR $>$ 25 dB are hardly distinguishable in terms of the sound quality. Therefore, these results show that we can use the

psychoacoustic model to deliver the parameters $u$ and $l$, in order to improve the sound quality of the watermarked signal obtained from the previously proposed SSA-based method [28]. However, the parameters determined by the differential evolution give the best performance in terms of sound quality.

*4.2. Robustness Evaluation.* The effectiveness of the proposed schemes in terms of robustness is measured by the watermark extraction precision. We use the bit-error rate (BER) to represent the watermark extraction precision. Given the embedded watermark bit-string $w(i)$ and the extracted watermark bit-string $\widehat{w}(i)$ for $i = 1$ to the frame length $N$,

$$\text{BER} = \frac{1}{N}\sum_{i=1}^{N} w(i) \oplus \widehat{w}(i), \tag{12}$$

where $\oplus$ is the bitwise XOR operator. The criterion for the robust scheme is that the BER must be less than 0.1 or 10%. At this level of BER, it is possible to reduce the BER further to close to 0 by adding error correction code. Furthermore, at this level, the BER can be reduced practically and effectively by the embedding-repetition scheme. That is, a frame is segmented into several subframes, and a watermark bit is embedded repeatedly into those subframes. Then the majority rule is applied in the extraction process to decode the extracted watermark bit.

Five attacks were performed on watermarked signals: Gaussian-noise addition with average signal-to-noise ratio (SNR) of 36 dB, resampling with 16 and 22.05 kHz, band-pass filtering with 100–6000 Hz and −12 dB/Oct, MP3 compression with 128 kbps joint stereo, and MP4 compression with 96 kbps.

The results from the robustness evaluation are shown in Table 5. The average BERs of the proposed schemes are less than 10% on almost all evaluation attacks except MP3 compression and the band-pass filtering (BPF). For MP3 and BPF, the average BERs are slightly above 10%. If we consider the overall average BERs, which is the average of BERs from all types of attacks, our proposed methods are still below 10% and less than that of the conventional SVD-based method. Table 6 shows the overall average of all methods.

Compared with the conventional SVD-based method, the proposed schemes are slightly less robust in the case of "no attack," "MP4," "AWGN," "RES16," and "BPF." However, the overall average BERs of the proposed schemes are better than that of the conventional SVD-based one. In general, when the BER is low enough (e.g., 10%), it can be reduced further by applying error correction code or by employing embedding repetitions. On the other hand, the proposed schemes outperform the conventional SVD-based method in the case of "MP3" and "BPF." Since the average BERs of the conventional SVD-based method in both cases are close to the chance level, they are hard to be improved further by those techniques.

Compared with the previously proposed SSA-based methods, it is less robust to some degree. Therefore, the overall performance of the proposed scheme seems to be slightly poorer than that of the SSA-based one with the differential evolution. The explanation concerning this issue is discussed in Section 5.

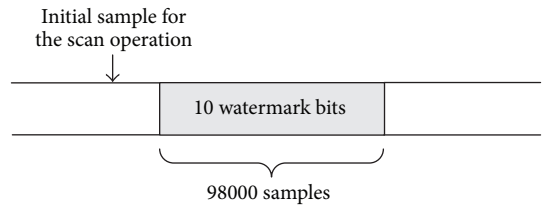


Figure 19: The initial sample, which is before the watermarked segment, is randomly chosen for the scan operation.

When the extraction process does not assume to know the parameters $u$ and $l$ in advance, the average BER increases about 2%. The root-mean-square deviation of the difference between the estimated values and actual values is about 2.83. Thus, the extraction process is sensitive to the correctness of the parameter values to some degree. When it extracts the watermark with less information, the BER increases.

*4.3. Self-Synchronization Evaluation.* To test the selfSynchronization, we implemented the scheme with settings shown in Table 7. Each test signal is randomly chosen as a segment of 98000 samples (about 2.2 seconds), and 10 bits of the watermark were embedded into the segment.

To detect the watermarked frame and to extract the watermark, we randomly choose the initial sample, which is before the embedded segment, for the scan operation, as depicted in Figure 19. The accuracy of the frame detection and the watermark extraction is defined as the number of correctly extracted watermark bits divided by the total number of embedded watermark bits. Since there is naturally the concavity on singular spectra, it is possible that our proposed method identifies an unwatermarked segment as a watermarked frame. In this case we will have a misidentified frame. The false positive rate is defined as the number of misidentified frames divided by the total number of frames identified by the algorithm. The test results show that the accuracy of the frame detection and the watermark extraction is 80%. The false positive detection rate is 6.42%.

## 5. Discussion

Even though the proposed scheme satisfies the robustness and inaudibility criteria, there are other aspects that need to be improved. In this section, five issues concerning the performance and limitation of the proposed scheme are discussed. The first two issues are about the performance of the proposed scheme. The next two issues are about the limitation of the currently proposed self-synchronized scheme. The last one is a general problem in terms of the confidentiality property.

First, we have shown that the psychoacoustic model can be used to determine the parameters $u$ and $l$. These parameters are host-signal-dependent and of importance because their values determine the balance between the inaudibility and the robustness. In our previously proposed method,

Table 5: Robustness evaluation using BERs (%): comparison of the conventional SVD-based method (SVD) [23], SSA-based scheme (SSA) [28], SSA-based scheme with the differential evolution (SSA.DE) [29], proposed scheme without the automatic parameter estimation (Prop.), and proposed one with the automatic parameter estimation (Prop.APE) when attacks (i.e., MP3 and MP4 compression, white-Gaussian-noise addition (AWGN), resampling with 16 and 22.05 kHz (RES 16 and RES 22.05, resp.), and band-pass filtering (BPF)) were performed. AV and SD are average and standard deviation, respectively.

| | | #01 | #07 | #13 | #28 | #37 | #49 | #54 | #57 | #64 | #85 | #91 | #100 | AV | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No Attack | SVD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.00** | 0.00 |
| | SSA | 0.00 | 0.00 | 4.00 | 0.00 | 0.67 | 0.00 | 2.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.67 | **0.67** | 1.30 |
| | SSA.DE | 0.00 | 0.83 | 0.83 | 0.00 | 0.00 | 0.83 | 1.67 | 6.67 | 1.67 | 0.83 | 0.83 | 0.83 | **1.25** | 1.79 |
| | Prop. | 0.00 | 0.00 | 2.00 | 0.00 | 0.00 | 0.67 | 0.00 | 12.67 | 1.33 | 0.67 | 1.33 | 0.67 | **1.61** | 3.54 |
| | Prop.APE | 5.33 | 0.00 | 0.67 | 0.00 | 0.00 | 4.00 | 0.00 | 12.67 | 2.67 | 0.00 | 3.33 | 1.33 | **2.50** | 3.70 |
| MP3 | SVD | 47.22 | 98.33 | 70.56 | 87.78 | 82.50 | 33.06 | 41.94 | 33.61 | 95.83 | 18.89 | 74.72 | 23.61 | **59.00** | 29.03 |
| | SSA | 1.33 | 2.00 | 26.67 | 0.67 | 2.67 | 2.67 | 5.33 | 0.67 | 4.67 | 0.67 | 2.67 | 1.33 | **4.28** | 7.21 |
| | SSA.DE | 0.00 | 2.50 | 15.00 | 3.33 | 3.33 | 2.50 | 8.33 | 5.00 | 1.67 | 6.67 | 9.17 | 4.17 | **5.14** | 4.11 |
| | Prop. | 2.00 | 12.00 | 26.67 | 2.00 | 6.67 | 7.33 | 12.00 | 36.67 | 3.33 | 8.67 | 10.00 | 4.00 | **10.94** | 10.51 |
| | Prop.APE | 12.00 | 6.00 | 17.33 | 4.67 | 4.67 | 7.33 | 8.67 | 21.33 | 0.67 | 4.67 | 24.67 | 1.33 | **9.44** | 7.80 |
| MP4 | SVD | 0.28 | 0.00 | 13.89 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | **1.20** | 4.00 |
| | SSA | 0.67 | 1.33 | 16.00 | 0.67 | 4.00 | 0.67 | 3.33 | 0.67 | 2.67 | 0.00 | 2.67 | 1.33 | **2.83** | 4.33 |
| | SSA.DE | 0.83 | 1.67 | 25.83 | 2.50 | 3.33 | 0.00 | 12.50 | 6.67 | 4.17 | 2.50 | 10.83 | 7.50 | **6.53** | 7.23 |
| | Prop. | 0.00 | 2.00 | 16.00 | 0.67 | 3.33 | 2.00 | 4.00 | 20.00 | 5.33 | 2.00 | 5.33 | 2.00 | **5.22** | 6.25 |
| | Prop.APE | 2.67 | 6.00 | 10.00 | 3.33 | 6.67 | 8.00 | 0.00 | 21.33 | 12.67 | 0.00 | 8.67 | 0.00 | **6.61** | 6.24 |
| BPF | SVD | 25.83 | 48.61 | 47.78 | 43.61 | 56.67 | 21.39 | 40.56 | 28.89 | 62.22 | 0.83 | 50.56 | 30.00 | **38.08** | 17.30 |
| | SSA | 0.00 | 0.67 | 40.00 | 0.00 | 12.67 | 0.67 | 1.33 | 2.67 | 0.67 | 0.67 | 0.67 | 0.00 | **5.00** | 11.57 |
| | SSA.DE | 5.00 | 1.67 | 40.00 | 2.50 | 5.00 | 2.50 | 10.83 | 25.00 | 3.33 | 2.50 | 15.00 | 5.83 | **9.93** | 11.68 |
| | Prop. | 0.67 | 0.67 | 40.00 | 0.00 | 14.00 | 0.00 | 0.00 | 15.33 | 2.67 | 0.00 | 2.67 | 0.67 | **6.39** | 11.90 |
| | Prop.APE | 0.67 | 6.00 | 52.00 | 3.33 | 15.33 | 10.67 | 3.33 | 24.67 | 12.67 | 0.00 | 8.00 | 1.33 | **11.50** | 14.64 |
| AWGN | SVD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.00** | 0.00 |
| | SSA | 0.00 | 0.00 | 4.00 | 0.00 | 0.67 | 0.00 | 2.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.67 | **0.67** | 1.30 |
| | SSA.DE | 0.00 | 0.83 | 0.83 | 0.00 | 0.00 | 0.83 | 1.67 | 6.67 | 1.67 | 0.83 | 0.83 | 0.83 | **1.25** | 1.79 |
| | Prop. | 0.00 | 0.00 | 2.00 | 0.00 | 0.00 | 0.67 | 0.00 | 12.67 | 1.33 | 0.67 | 1.33 | 0.67 | **1.61** | 3.54 |
| | Prop.APE | 5.33 | 0.00 | 0.67 | 0.00 | 0.00 | 4.00 | 0.00 | 12.67 | 2.67 | 0.00 | 3.33 | 1.33 | **2.50** | 3.70 |
| RES 16 | SVD | 5.56 | 0.83 | 39.44 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.56 | 1.67 | 0.28 | 1.67 | **4.17** | 11.22 |
| | SSA | 1.33 | 0.00 | 26.67 | 0.00 | 8.00 | 0.00 | 3.33 | 0.00 | 0.67 | 0.00 | 0.00 | 0.67 | **3.39** | 7.69 |
| | SSA.DE | 1.67 | 0.00 | 37.50 | 0.83 | 3.33 | 1.67 | 3.33 | 7.50 | 1.67 | 0.83 | 5.83 | 2.50 | **5.56** | 10.29 |
| | Prop. | 0.67 | 0.00 | 27.33 | 0.00 | 15.33 | 0.67 | 0.00 | 13.33 | 2.67 | 0.67 | 0.67 | 0.67 | **5.17** | 8.79 |
| | Prop.APE | 6.67 | 0.00 | 30.67 | 3.33 | 36.00 | 2.00 | 0.00 | 17.33 | 6.67 | 0.00 | 2.00 | 1.33 | **8.83** | 12.47 |
| RES 2205 | SVD | 1.67 | 0.00 | 12.83 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.67 | **1.35** | 3.67 |
| | SSA | 0.67 | 0.00 | 5.33 | 0.00 | 4.67 | 0.00 | 2.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.67 | **1.17** | 1.95 |
| | SSA.DE | 0.83 | 0.83 | 2.50 | 0.00 | 2.50 | 0.00 | 3.33 | 7.50 | 1.67 | 0.83 | 0.83 | 1.67 | **1.87** | 2.05 |
| | Prop. | 0.00 | 0.00 | 4.00 | 0.00 | 6.67 | 0.67 | 0.00 | 12.67 | 2.00 | 0.67 | 1.33 | 0.67 | **2.39** | 3.81 |
| | Prop.APE | 1.33 | 0.00 | 2.00 | 0.00 | 42.00 | 4.67 | 0.00 | 14.00 | 6.00 | 0.00 | 1.33 | 1.33 | **6.06** | 12.01 |

differential evolution was used to determine the parameters. Compared with using differential evolution, there are two advantages of using the psychoacoustic model.

(i) The computational time is reduced considerably because the differential evolution optimization has a large search space. The comparison of the computational time is shown in Table 8. To determine the parameters $u$ and $l$ for one signal, differential evolution takes about 13 hours, whereas the psychoacoustic-model-based method takes about 4.3 seconds.

(ii) The optimal parameters from the differential evolution depends on many factors, such as the simulations included in the optimizer [29]. Moreover, the cost function has two additional parameters. In this sense, using the psychoacoustic model reduces the number of scheme parameters.

Table 6: Overall average BERs (%): comparison of the conventional SVD-based method (SVD), SSA-based scheme (SSA), SSA-based scheme with the differential evolution (SSA.DE), proposed scheme without the automatic parameter estimation (Prop.), and proposed one with the automatic parameter estimation (Prop.APE).

| | Overall average BER |
|---|---|
| SVD [23] | 14.83 |
| SSA [28] | 2.57 |
| SSA.DE [29] | 4.50 |
| Prop. | 4.76 |
| Prop.APE | 6.78 |

Table 7: Simulation conditions for the self-synchronized SSA-based audio watermarking scheme.

| Conditions | |
|---|---|
| Subframe length $M$ | 2450 |
| Window length $L$ | 980 |
| Subframe-scan step $\delta$ | 10 |
| Frame-scan step $\Delta$ | 10 |
| Overlap margin $\Sigma$ | 20 |
| Embedding capacity | 4.5 bps |
| Payload | 10 bits |
| Total duration | 2.22 seconds |

However, the robustness of the proposed scheme is slightly poorer than that of the previously proposed methods. This is because only the SMR is used as the guidance for the parameter determination. The low SMR can gain inaudibility but may lose robustness because the lower SMRs associates with the lower singular-value indices. In addition, the components with the lower SMRs are more likely to be destroyed by the perceptual codings. To improve the robustness of the proposed scheme, we may include the other masking phenomena, such as the nonlinear excitatory masking, to the psychoacoustic model. This is one of our future work.

Second, different from the previously proposed scheme, this scheme does not modify the singular spectrum when the watermark bit 0 is embedded. We found that the effectiveness in terms of robustness is the same, but in terms of inaudibility, the objective scores improve slightly, as shown in Table 9. The previously proposed schemes, especially the one with the differential evolution optimization, can benefit from this fact because the optimization function directly handles the trade-off between inaudibility and robustness.

Third, the proposed self-synchronization is time-consuming. The extraction process with the self-synchronization takes up to $(\lfloor 3M/\delta \rfloor + 1)$ times that of one without. Based on our simulation, the extraction process without the self-synchronization took about 1.6 seconds to extract one watermark bit, whereas the one with the self-synchronized process took about 20 minutes. This explains why we separately simulated and evaluated the self-synchronization.

Fourth, although the synchronization rate of 80% of the proposed scheme with self-synchronization does not satisfy the criterion of BER being less than 10%, it can confirm the fundamental concepts on which the self-synchronized, proposed scheme is based. From our analysis, we found that the detection rate is determined by the algorithm that interprets the bit-string $b_i$. In the proposed scheme, our algorithm uses the simplest rectangular windows to find the pattern of $b_i$. Even in the case that the algorithm could not detect a watermark bit, we found that the string $b_i$ correctly presented the concavity on the singular spectra. Therefore, some effective pattern recognition techniques could be helpful to improve the situation.

Also, the false positive detection rate indicates that the algorithm sometimes detects a watermark bit when no hidden information is embedded there. We investigated this problem by analyzing unwatermarked signals with the proposed automatic frame detection. We found that in those false positive detection cases, there is some concavities on the singular spectra. If the false positive detection is a serious concern, we can solve this problem by first detecting the natural concavity and then hiding the watermark only in the no-concavity frames. Otherwise, good pattern recognition is required due to our findings that the patterns of the string $b_i$ of the natural concavity are different from those of the embedded watermark. This problem will be further investigated in the future.

Fifth, since this work has shown that we can completely blindly scan and analyze the watermarked signals to detect and extract the watermark, there is a question on the confidentiality of the watermark. As a result, if the secrecy of the watermark is a concern, we may need to encrypt the watermark with an encryption key before it is embedded into the host signals. Later, in the extraction process, a decryption key is required to decrypt the extracted, encrypted watermark to obtain the original one.

## 6. Conclusion

The main objective of this work is to show that SSA, equipped with the psychoacoustic model, can give a good balance between inaudibility and robustness, so that it can overcome the problems in the previously proposed SSA-based method [28] and the SVD-based method. Even though the overall performance of the currently proposed schemes is poorer than that of the SSA-based one with differential evolution, the processing time is reduced considerably. Integrated with the psychoacoustic model, the SSA-based audio watermarking scheme achieves three required properties of the audio watermarking system: inaudibility, robustness, and blindness. Also, this paper presented a novel method for self-synchronization. The synchronization rate of the proposed self-synchronized scheme was about 80%. Improving the synchronization rate and reducing the computational time of the self-synchronized scheme are our future work.

Table 8: Comparison of the computational times for determining the parameters of an host signal when the automatic parameterization is based on the differential evolution [29] and when it is based on the psychoacoustic model. The simulations were operated on the Fujitsu CX250 Cluster (JAIST parallel computers), where each node is equipped with Two Intel Xeon E5-2680v2 2.80 GHz (10 cores each) and 64 GB memory.

| | Function | Computational time of the function | Search space/No. of operations | Approximated total computational time |
|---|---|---|---|---|
| Differential Evolution | Cost function evalution | 3 minutes 44 seconds | 31815 possible vectors | 13 hours 9 minutes |
| Psychoacoustic Model | SMR calculation | 0.36 seconds | 717 times | 4.3 minutes |

Table 9: ODGs, LSDs, and SDRs: comparison of the proposed scheme when the singular spectrum is modified and when it is not modified to embed the watermark bit 0.

| | ODG | LSD | SDR |
|---|---|---|---|
| The singular spectrum is modified. | 0.18 | 0.34 | 24.30 |
| The singular spectrum is not modified. | 0.18 | 0.25 | 25.61 |

## Competing Interests

The authors declare that they have no competing interests.

## Acknowledgments

This work was supported by an A3 foresight program made available by the Japan Society for the Promotion of Science and partially supported by a Grant-in-Aid for Scientific Research (A) (no. 25240026). It was also under a grant in the SIIT-JAIST-NECTEC Dual Doctoral Degree Program and the National Research University Funding from Thailand.

## References

[1] S. Bhattacharjee, R. D. Gopal, and G. L. Sanders, "Digital music and online sharing: software piracy 2.0?" *Communications of the ACM*, vol. 46, no. 7, pp. 107–111, 2003.

[2] R. D. Gopal and G. L. Sanders, "Global software piracy: you can't get blood out of a turnip," *Communications of the ACM*, vol. 43, no. 9, pp. 83–89, 2000.

[3] A. M. Al-Haj, *Advanced Techniques in Multimedia Watermarking: Image, Video and Audio Applications*, IGI Global, 2010.

[4] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, Morgan Kaufman, 2007.

[5] X. Quan and H. Zhang, "Perceptual criterion based fragile audio watermarking using adaptive wavelet packets," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 867–870, IEEE, Cambridge, UK, August 2004.

[6] T. Kalker and J. Haitsma, "Efficient detection of a spatial spread-spectrum watermark in MPEG video streams," in *Proceedings of the International Conference on Image Processing (ICIP '00)*, pp. 434–437, September 2000.

[7] I. J. Cox and M. L. Miller, "The first 50 years of electronic watermarking," *EURASIP Journal on Advances in Signal Processing*, vol. 2002, no. 2, pp. 1–7, 2002.

[8] I. J. Cox, "Watermarking, steganography and content forensics," in *Proceedings of the ACM WMsec*, pp. 1–2, 2008.

[9] S. Craver, M. Wu, and B. Liu, "What can we reasonably expect from watermarks?" in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '01)*, pp. 223–226, 2001.

[10] A. Al-Haj, C. Twal, and A. Mohammad, "Hybrid DWT-SVD audio watermarking," in *Proceedings of the 5th International Conference on Digital Information Management (ICDIM '10)*, pp. 525–529, IEEE, Ontario, Canada, July 2010.

[11] B. Lei, I. Y. Soon, and E.-L. Tan, "Robust SVD-based audio watermarking scheme with differential evolution optimization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 11, pp. 2368–2378, 2013.

[12] M. Fallahpour and D. Megías, "High capacity audio watermarking using the high frequency band of the wavelet domain," *Multimedia Tools and Applications*, vol. 52, no. 2-3, pp. 485–498, 2011.

[13] C. H. Yeh and C. J. Kuo, "Digital watermarking through quasi m-arrays," in *Proceedings of the IEEE Workshop on Signal Processing Systems (SiPS '99)*, pp. 456–461, 1999.

[14] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3-4, pp. 313–335, 1996.

[15] S.-S. Kuo, J. D. Johnston, W. Turin, and S. R. Quackenbush, "Covert audio watermarking using perceptually tuned signal independent multiband phase modulation," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '02)*, pp. 1753–1756, Orlando, Fla, USA, May 2002.

[16] N. M. Ngo and M. Unoki, "Watermarking for digital audio based on adaptive phase modulation," in *Digital Forensics and Watermarking*, Lecture Notes in Computer Science 9023, pp. 105–119, Springer, 2015.

[17] M. Unoki and R. Miyauchi, "Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay characteristics," *IEICE Transactions on Information and Systems*, vol. 98, no. 1, pp. 38–48, 2015.

[18] H. O. Oh, J. W. Seok, J. W. Hong, and D. H. Youn, "New echo embedding technique for robust and imperceptible audio watermarking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, pp. 1341–1344, Salt Lake City, Utah, USA, May 2001.

[19] H. J. Kim and Y. H. Choi, "A novel echo-hiding scheme with backward and forward kernels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 885–889, 2003.

[20] P. Bassia, I. Pitas, and N. Nikolaidis, "Robust audio watermarking in the time domain," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 232–241, 2001.

[21] H. Ozer, B. Sankur, and N. Memon, "An SVD-based Audio Watermarking Technique," in *Proceedings of the 1st ACM Workshop on Information Hiding and Multimedia Security*, pp. 51–56, IEEE Press, 2005.

[22] A. Al-Haj and A. Mohammad, "Digital audio watermarking based on the discrete wavelets transform and singular value decomposition," *European Journal of Scientific Research*, vol. 39, no. 1, pp. 6–21, 2010.

[23] V. Bhat K, I. Sengupta, and A. Das, "A new audio watermarking scheme based on singular value decomposition and quantization," *Circuits, Systems, and Signal Processing*, vol. 30, no. 5, pp. 915–927, 2011.

[24] P. K. Dhar and T. Shimamura, "A DWT-DCT-based audio watermarking method using singular value decomposition and quantization," *Journal of Signal Processing*, vol. 17, no. 3, pp. 69–79, 2013.

[25] F. E. Abd El-Samie, "An efficient singular value decomposition algorithm for digital audio watermarking," *International Journal of Speech Technology*, vol. 12, no. 1, pp. 27–45, 2009.

[26] K. V. Bhat, I. Sengupta, and A. Das, "An audio watermarking scheme using singular value decomposition and dither-modulation quantization," *Multimedia Tools and Applications*, vol. 52, no. 2-3, pp. 369–383, 2011.

[27] L. Lamarche, Y. Liu, and J. Zhao, "Flaw in SVD-based watermarking," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE '06)*, pp. 2082–2085, Ottawa, Canada, May 2006.

[28] J. Karnjana, M. Unoki, P. Aimmanee, and C. Wutiwiwatchai, "An audio watermarking scheme based on singular-spectrum analysis," in *Digital Forensics and Watermarking*, vol. 9023 of *Lecture Notes in Computer Science*, pp. 145–159, 2015.

[29] J. Karnjana, P. Aimmanee, M. Unoki, and C. Wutiwiwatchai, "An audio watermarking scheme based on automatic parameterized singular-spectrum analysis using differential evolution," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA '15)*, pp. 543–551, Hong Kong, December 2015.

[30] H. Hassani, "Singular spectrum analysis: methodology and comparison," *Journal of Data Science*, vol. 5, no. 2, pp. 239–257, 2007.

[31] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman and Hall/CRC, Boca Raton, Fla, USA, 2001.

[32] K. Bradenburg and G. Stoll, "Iso/mpeg-1 audio: a generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792, 1994.

[33] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*, John Wiley & Sons, 2006.

[34] Y. You, *Audio Coding: Theory and Applications*, Springer Science & Business Media, 2010.

[35] W. Li, X. Xue, and P. Lu, "Localized audio watermarking technique robust against time-scale modification," *IEEE Transactions on Multimedia*, vol. 8, no. 1, pp. 60–69, 2006.

[36] S. Wu, J. Huang, D. Huang, and Y. Q. Shi, "Efficiently self-synchronized audio watermarking for assured audio data transmission," *IEEE Transactions on Broadcasting*, vol. 51, no. 1, pp. 69–76, 2005.

[37] S. Shuifa and K. Sam, "A self-synchronization blind audio watermarking algorithm," in *Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS '05)*, pp. 133–136, Hong Kong, December 2005.

[38] S. Wu, J. Huang, D. Huang, and Y. Q. Shi, "Self-synchronized audio watermark in DWT domain," in *Proceedings of the 2004 IEEE International Symposium on Cirquits and Systems (ISCAS '04)*, pp. V-712–V-715, Vancouver, Canada, May 2004.

[39] X.-Y. Wang and H. Zhao, "A novel synchronization invariant audio watermarking scheme based on DWT and DCT," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4835–4840, 2006.

[40] K. Hiratsuka, K. Kondo, and K. Nakagawa, "On the accuracy of estimated synchronization positions for audio digital watermarks using the modified patchwork algorithm on analog channels," in *Proceedings of the 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP '08)*, pp. 628–631, Harbin, China, August 2008.

[41] https://staff.aist.go.jp/m.goto/RWC-MDB/.

[42] A. Learch, *Software: EAQUAL—Evaluation of Audio Quality, v.0.1.3alpha ed*, 2002.