Twenty years of digital audio watermarking—a comprehensive review

Hua, Guang; Huang, Jiwu; Shi, Yun Q.; Goh, Jonathan; Thing, Vrizlynn L. L.

2016

Hua, G., Huang, J., Shi, Y. Q., Goh, J., & Thing, V. L. (2016). Twenty years of digital audio watermarking—a comprehensive review. Signal Processing, 128, 222-242.

https://hdl.handle.net/10356/80616

https://doi.org/10.1016/j.sigpro.2016.04.005

Twenty Years of Digital Audio Watermarking – A Comprehensive Review

Guang Hua[a,*], Jiwu Huang[b], Yun Q. Shi[c], Jonathan Goh[d], Vrizlynn L. L. Thing[d]

[a] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.*
[b] *College of Information Engineering, Shenzhen University, Shenzhen 518060, China.*
[c] *Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA.*
[d] *Cyber Security & Intelligence Department, Institute for Infocomm Research, Singapore 138632.*

Abstract

Digital audio watermarking is an important technique to secure and authenticate audio media. This paper provides a comprehensive review of the twenty years' research and development works for digital audio watermarking, based on an exhaustive literature survey and careful selections of representative solutions. We generally classify the existing designs into time domain and transform domain methods, and relate all the reviewed works using two generic watermark embedding equations in the two domains. The most important designing criteria, i.e., imperceptibility and robustness, are thoroughly reviewed. For imperceptibility, the existing measurement and control approaches are classified into heuristic and analytical types, followed by intensive analysis and discussions. Then, we investigate the robustness of the existing solutions against a wide range of critical attacks categorized into basic, desynchronization, and replacement attacks, respectively. This reveals current challenges in developing a global solution robust against all the attacks considered in this paper. Some remaining problems as well as research potentials for better system designs are also discussed. In addition, audio watermarking applications in terms of US patents and commercialized solutions are reviewed. This paper serves as a comprehensive tutorial for interested readers to gain a historical, technical, and also commercial view of digital audio watermarking.

Keywords: Audio watermarking, Robust watermark, Data hiding, Copyright protection, Media security.

[*] Corresponding author

Email addresses: ghua@ntu.edu.sg (Guang Hua), jwhuang@szu.edu.cn (Jiwu Huang), yun-qing.shi@njit.edu (Yun Q. Shi), jonathan-goh@i2r.a-star.edu.sg (Jonathan Goh), vriz@i2r.a-star.edu.sg (Vrizlynn L. L. Thing)

1. Introduction

Digital audio watermarking is an important research branch of multimedia data hiding [1–5], which involves embedding the watermarks into host audio data and when necessary, performing watermark extraction for copyright protection, authentication, and other digital rights management (DRM) purposes. The original work on information hiding, i.e., dirty paper writing, was carried out from a communication theory perspective, which dated back to 1983 [6], while the first reported work on digital audio watermarking was seen in 1996 [1] and the first systematic work was presented in 1997 [7]. Therefore, there has been a history of about twenty years for digital audio watermarking. Within the twenty years, the advanced signal processing techniques have been efficiently utilized for this topic, and numerous solutions have been proposed alongside [6–70]. A generic digital audio watermarking system is depicted in Fig. 1, where the terms in solid rectangles specify the general phases of the signal manipulations in an audio watermarking system and the terms in dashed rectangles represent possible users who manipulate the data. Note that normal users may "attack" the watermarked signal unintentionally during the processes of lossy compression, equalization, or adding effects, etc. Thus, we also call such "processing attacks" as unintentional attacks while deliberate attacks aiming at destroying or removing the watermarks are referred to intentional attacks. The watermark extraction phase in Fig. 1 is also termed as watermark detection in many existing works. While watermark detection and extraction could refer to a similar signal processing purpose at the receiver end, they are actually slightly different tasks. Specifically, if the receiver end implements a threshold-based correlation and detection scheme, and the copyright is claimed by true positive detection results, then such a process is usually termed watermark detection. In this case, the original watermarks need to be available in order to calculate the corresponding correlation functions. However, it is also quite often that the receiver aims at restoring the original watermark sequence from watermarked copies without the knowledge of the original watermarks. In this situation, such a process is more precisely termed as watermark extraction. For simplicity, in this paper, the two terms are treated as interchangeable and used where appropriate, and the original watermarks are assumed to be known at the receiver end unless otherwise mentioned. Also note that some literature uses encoding and decoding to describe the embedding and extraction processes, which can also be considered as equivalent descriptions.

The effectiveness of an audio watermarking system is characterized by several performance

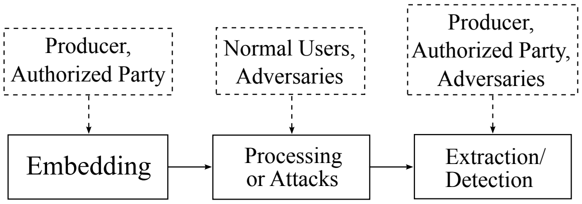

Figure 1: A generic block diagram of a digital audio watermarking system.

criteria [2], i.e., *imperceptibility*, *robustness*, *security*, *capacity*, and *computational complexity*, etc. First, imperceptibility characterizes the fidelity of watermarked audio data, indicating that the embedded watermarks should introduce perceptually indistinguishable changes to the host signal. Therefore, sometimes, fidelity, transparency, and inaudibility are used equivalently as imperceptibility. Second, robustness refers to the availability of successful watermark extraction when the watermarked signal has been attacked intentionally or unintentionally. It is the most complicated feature for an audio watermarking system because of the variety of attacks. Third, security means that the system should be designed in such a way that only authorized parties are able to extract the watermarks. Fourth, the amount of information that can be embedded into the given host data is called capacity. At last, the designed system is preferred to be computationally efficient.

For conciseness and efficiency, the comprehensive review reported in this paper mainly considers imperceptibility and robustness among the criteria, because they determine the key performance of most existing audio watermarking systems. In contrast, security is usually achieved via the use of random keys, which is widely incorporated in most of the existing solutions. However, we would like to note here that although not being in the scope of this paper, watermarking security is also an active research area and interested readers can refer to [71, 72] and references therein for more information. For the criterion of embedding capacity, it is usually optional in those systems with typical objectives of successful extraction of the watermarks and declaring the ownership and copyright of the audio files. In this situations, watermarks are in the forms of random sequences (simply a mark, as can be seen from most of the reviewed works in this paper), and whether the watermarks correspond to meaningful information or how much it contains are less important. A different situation would be seen in another research area, i.e., steganography [73], where the hidden information (now it becomes a message rather than a mark) itself becomes important to establish covert communications in, e.g., military and health applications. Another important difference between watermarking and steganography is that at the extraction phase, the former has the knowledge of the watermarks and focuses mainly on matching it with the extracted version, while the latter does not have any clues of the hidden message. Lastly, based on current literature, none of the existing solutions suffer from severe computational issues, and the embedding and extraction of watermarks are usually performed “off-line” (there do exist several “on-line” applications, which are not the major focus of this review). Thus, we also exclude computational complexity. To further clarify the scope of this paper, we incorporate the classification of information hiding topics from [3], where the digital audio watermarking systems considered in this paper belong to the branch of “Imperceptible watermarking”. Therefore, audio fingerprinting [74] and fragile watermarking [75] are not considered here either. In general, emphasizing on which performance criterion is indeed an application-oriented choice. However, imperceptibility and robustness have been considered in the majority of the existing works, according to which we set the scope of this paper. Readers can refer to [3] for more information about the similarities and differences among different applications.

Within the last twenty years, a tremendous amount of research works have been carried out for digital audio watermarking systems, and numerous solutions have been proposed [6–70], thanks to the maturely developed digital signal processing techniques. The variety of the existing solutions calls for a systematic review for researchers and engineers in related fields or general interested peers to obtain a big picture and a historic point of view for this topic. However, current available review works are highly insufficient. Early review results, [3], [4], and [5], date back to more than ten years ago, thus recent works could not be taken into consideration. In recent years, no systematic review work has been seen in the literature, either. Therefore, we are motivated to carry out a comprehensive and rigorous review to reveal current research and development status of audio watermarking, clarify existing problems and difficulties, and discuss about potential research directions and strategies towards more advanced solutions. In particular, the contributions of this paper are as follows. i) We systematically categorize all the existing watermark embedding schemes in a concise and effective way based on two generic embedding functions. Then, the similarities, differences, and key features of the solutions for each category are discussed. ii) The measurement and control approaches to ensure the imperceptibility of audio watermarking systems are exhausted and categorized into heuristic and analytical groups. After that, we compare and discuss the choices of perceptual regions for watermark embedding and the corresponding effects on imperceptibility and robustness. iii) Existing attacks to audio watermarking systems are comprehensively studied and evaluated against a series of representative audio watermarking systems. The attacks include basic signal processing attacks, advanced desynchronization attacks, and the most challenging

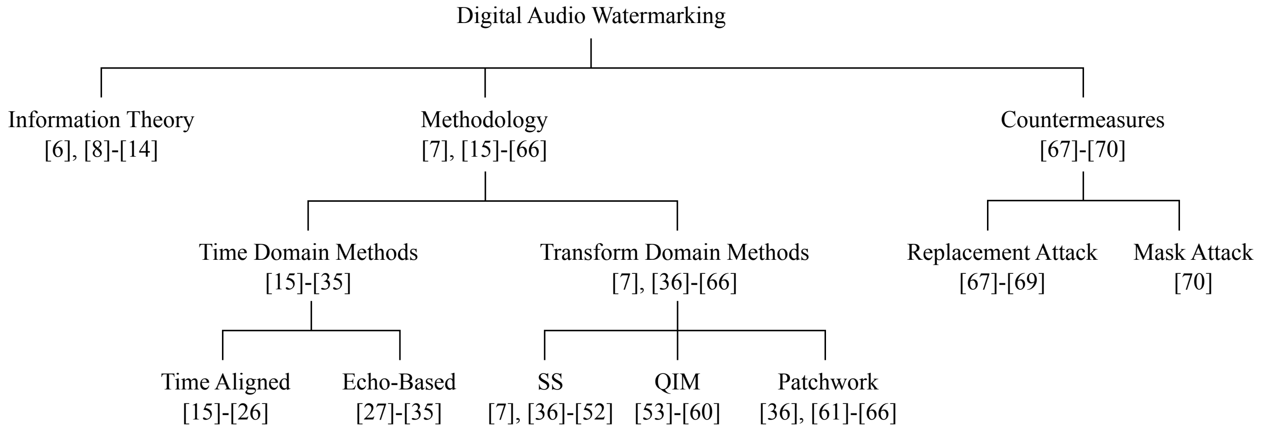

Figure 2: Categorization of existing digital audio watermarking works.

replacement attack [70]. Note that the replacement attack has been introduced for several years, but it has seldom been considered in system design works. iv) Audio watermarking applications are also intensively investigated. Specifically, we first review a series of US patents on audio watermarking and then describe the existing commercial solutions that utilize audio watermarking techniques.

The paper is organized as follows. Section 2 provides a systematic categorization of existing works on audio watermarking, followed by discussions on signal models, specific features, and a brief history of the development for each category of solutions. In Section 3, we focus on the imperceptibility issue, summarize heuristic and analytic approaches to control the imperceptibility, and discuss the choices of watermark embedding regions. Intensive robustness evaluations are provided in Section 4, in which a set of critical attacks are taken into consideration. Audio watermarking applications are discussed in Section 5. Conclusions are made in Section 6.

2. Categorization of Audio Watermarking Works

2.1. Categorization

The topic of watermarking for digital multimedia dates back to the time of digital revolution in early 1990's, after which, the major focus has been on digital content forms of still image [76] and audio [1]. For digital audio watermarking, several pioneering solutions were based on the incorporation and modifications of existing techniques from other research areas, e.g., spread spectrum (SS) from communication theory [7], and patchwork methods from image watermarking [62]. In early 2000's, a new class of embedding methods, termed as quantization index modulation (QIM), was introduced in [53], for general multimedia formats. Other than technique groups incorporated from other research areas, a series of time domain methods have been proposed as unique solutions to deal with audio format, among which the pioneering works are [15] and [27]. Noticeably, the work in [27] revealed a new paradigm termed as echo-based audio watermarking.

It can be seen from the above that many audio watermarking methods have been classified according to the specific signal processing techniques for watermark embedding. Nonetheless, other binary options have also been seen in the literature to characterize different solutions. These include but not limit to i) whether the watermarks are embedded in time domain or transform domain[1], ii) whether an additive or a multiplicative model is used for watermark embedding, iii) whether a psychoacoustic model is used in the design, and iv) whether the watermarking system is informed or non-informed. Being informed means that the properties of the host audio signal are explored to facilitate better watermark embedding and extraction procedures. Similar to the informed and non-informed option, we may call an audio watermarking system as a blind or non-blind system. A blind system does not require the original unmarked host signal during watermark extraction, but a non-blind system does. Based on the above characteristics, the SS system introduced in [36] is known as a transform domain, additive, without a psychoacoustic model, non-informed, and blind (detection) system. Although differentiating existing solutions using the above attributes is acceptable in many circumstances, some key features may still be missing, and this approach cannot be used efficiently with the consideration of methods defined by signal processing techniques. For example, SS and QIM methods are very likely to be categorized into the same group.

In this paper, we categorize the existing audio watermarking works using a tree structure, which is shown in Fig. 2. For a comprehensive treatment, we not only include the watermarking schemes, but also consider related works on information-theoretic analysis [6, 8–14] and investigations on uncommon attacks [67–70]. The approach to efficiently perform the categorization is to preserve the well established methods grouped by specific signal processing methods, and select the most important characteristics from the ones described in the above paragraph. Among the four binary options, the domain to perform watermarking is the most significant characteristic that can generally differentiate the existing solutions into two groups. In contrast, the remaining options are not truly critical, and the reasons are as follows. First, multiplicative models can

[1] Transform domain refers to various transforms that have been implemented in the literature, such as modulated complex lapped transform (MCLT), discrete Fourier transform (DFT), discrete cosine transform (DCT), discrete wavelet transform (DWT), etc.

be equivalently represented by additive models. In this way, the difference simplifies to whether watermark embedding is dependent on the host signal. Thus, this option becomes trivial since most of existing watermark embedding methods are informed. Second, the use of psychoacoustic models is more related to the improvement of the imperceptibility property, and this criterion is very likely to overlap with categories specified by signal processing techniques. For example, in the development works for echo-based watermarking systems, a psychoacoustic model is considered in a newer work [34], but not in the preceding work [32]. At last, being an informed (embedding) and blind (detection) system is currently a commonly preferable feature for audio watermarking systems.

Therefore, the existing watermarking methods can be generally categorized into time domain and transform domain ones, as shown in Fig. 2. After that, they are further divided into several subcategories, i.e., time domain ones contains time aligned [15–26] and echo-based (not aligned) [27–35] methods, while transform domain ones can be divided into spread spectrum (SS) [7, 36–52], quantization index modulation (QIM) [53–60], and patchwork [36, 61–66] methods. As a result, the technical groups defined by specific signal processing techniques are well incorporated in this respect. Let $x(n)$ and $X(k)$ be the time and transform domain representations of the host signal segment, where n and k are sample indices in the corresponding domains. Then the generic models for time and transform domain watermark embedding can be expressed as

$$y(n) = x(n) + \alpha w(n), \tag{1}$$

and

$$Y(k) = X(k) + \alpha m(k), \tag{2}$$

where α controls the watermark strength, $w(n)$ and $m(k)$ correspond to the watermarks in time and transform domain, respectively. Note that $w(n)$ and $m(k)$ can be non-modulated watermark bits or modulated watermark sequences. A simple realization of multiplicative model, using (1) as an example, can be obtained by replacing $w(n)$ with $x(n)w(n)$, yielding $y(n) = x(n)(1 + \alpha w(n))$. Next, we review the audio watermarking works summarized in Fig. 2 in detail.

2.2. Information-Theoretic Analysis

The information-theoretic analysis works have established the theoretical foundation for the design of digital audio watermarking systems. In 2003, several important works were published in the literature [9–12], which covered the general topic of information hiding [9], and digital watermarking for both image and audio forms [10–12]. A significant difference of these works as compared to practical system designs is that the attacks are usually modeled as a communication channel, whose output (signal after attacks) is characterized by conditional probability functions. Meanwhile, the *a priori* probability density function (PDF) of the host signal (usually in transform domain) is assumed to be available, e.g., generalized Gaussian distribution (GGD) in [11], and Weibull distribution [12], etc. The information-theoretic works dedicated to audio wateramrking are seen in [13] and [14]. In [13], a statistical analysis revealed that audio watermarking can be considered as a signal processing procedure to enhance the stationarity of the signal. In [14], an attack model with noise and desynchronization was proposed, and the system was designed based on game theory. In addition, the notion that desynchronization attacks are tougher to deal with than additive noise attacks was theoretically verified. Note that a comprehensive publication list and in-depth review of information-theoretical works on digital watermarking are not provided in this paper, since the main focus in this paper is on practical system designs. The brief review provided in this subsection is hence used for the completeness of the coverage for digital audio watermarking.

2.3. Countermeasures

In the application of audio watermarking for copyright protection, the system designer usually puts the robustness of the system against intentional and unintentional attacks in the first priority in the design. This is because robustness essentially determines whether the system is feasible before considering imperceptibility. There exist various types of attacks, and a comprehensive review and analysis of the robustness against attacks will be provided in Section 4. However, in this section, we introduce two types of uncommon attacks that have seldom been considered in existing system designs, i.e., the replacement attack [67–69], and the mask attack [70]. We classify these works into a single category in Fig. 2 to emphasis the challenges that the system designers are facing in the audio watermarking game against the adversaries. The works in [67–69] share a similar idea of the attack method, i.e., replacing signal blocks with other perceptually similar ones. This attack is especially effective because it is observed that audio content is highly repetitive. It has also been verified that the replacement attack can destroy SS- and QIM-based watermarking systems. Moreover, it can actually be used against all watermarking schemes, since the attack is generic, i.e., only the watermarked signal is required to perform the attack. Besides the replacement attack, the mask attack demonstrated in [70] is another type of intelligent attack, which takes advantage of the masking model used in SS schemes to improve the imperceptibility. Specifically, the watermarks embedded in the masked frequency regions are likely to be estimated by the adversaries by examining the masking model of the watermarked signal. More discussions will be provided in Section 4. In the following subsections, we look into the practical and ready-to-implement audio watermarking solutions.

2.4. Time Domain Methods

Digital audio watermarking systems that perform watermark embedding in the time domain generally provide straightforward solutions since they directly modify the audio samples. Therefore, except for the echo-based solutions [27–35] which have specific signal model and designing approaches, it is difficult to effectively categorize other existing time domain methods [15–26]. Usually, they are treated as ad-hoc solutions. In this paper, by noting the difference between echo-based solutions from other existing time domain ones, we classify the works in [15–26] into the "Time Aligned" category as shown in Fig. 2, since embedding the watermarks without imposing a time shift is one of the very few properties that they share in common. Note that all the time domain methods have informed watermarks, hence, in this subsection, we modify the generic single bit embedding equation (1) as

$$y(n) = x(n) + \alpha f[x(n), w(n)],$$

where $f(\cdot)$ represents a mechanism to generate the watermarks with the use of both host signal and the original watermark sequence. The works reviewed hereafter are hence simplified to the designs of $f[x(n), w(n)]$.

2.4.1. Time Aligned

The signal models for time aligned audio watermarking methods are summarized as follows. In [15],

$$f[x(n), w(n)] = \mathrm{LPF}\left[|x(n)|\, w(n)\right],$$

where $w(n) \in \{-1, 1\}$, and LPF stands for lowpass filtering process. In this work, a qualitative treatment of imperceptibility is achieved by the LPF, which suppresses the power spectral density (PSD) of the watermark signal below the PSD of the host signal. The corresponding watermark extraction scheme is based on exploring the test statistics of the mean value of the correlation function between the watermarked signal and $w(n)$. A similar approach is seen in [16], and the proposed embedding scheme is given by

$$f\left[x(n), w(n)\right] = w(n)\text{BPF}\left[x(n)\right],$$

where $w(n)$ is generated by passing the binary keys from $\{-1, 1\}$ into a conditioning circuit, and BPF stands for bandpass filtering process. For more efficient watermark detection, two watermark sequences, $w_1(n)$ and $w_2(n)$, generated from the same binary sequence by applying different circular shifts, are used with a delay of T, i.e.,

$$w(n) = w_1(n) + w_2(n - T).$$

Thus, two correlation peaks can be utilized in watermark detection. Note that the host audio is considered as a single frame for watermark embedding. The way that the authors in [16] achieve imperceptibility property of the system is by incorporating the 2-alternative forced-choice (2AFC) measurement paradigm [77] to determine the masking threshold, which is a manual setup. After that, α is selected is such a way that the PSD of the watermark is bounded by the masking threshold. In addition, the ITU-PEAQ stardard tool [78] is introduced to measure the perceptual quality of the watermarked signal. More details about imperceptibility measurement are provided in Section 3. In [17], A further design in the time domain for both imperceptibility and robustness is proposed. The embedding scheme is given by

$$y(n) = x(n) + \alpha h(n) * v(n),$$

where $v(n)$ is the modulated watermark signal mapped to a sequence of symbols, and $h(n)$ is a frequency shaping filter. The imperceptibility is controlled via an objective difference grade (ODG)-based mechanism that adaptively changes the scaling factor α, while the robustness is enhanced by using the host signal during the detection phase, and assigning synchronization bits. Note that using host signal in detection phase will limit the applicability of the system, while blind watermark detection is generally preferred to meet the requirements of more applications. Another time domain watermarking scheme is seen in [18], which proposes to modulate the binary pseudo noise (PN) sequence into sinusoidal patterns, i.e.,

$$y(n) = x(n) + \frac{\sqrt{2}}{\sqrt{N}} \sum_{i=0}^{I-1} \alpha(i) q_n(i) \sin\left[\frac{2\pi n\left(f_0 + i\right)}{N}\right],$$

where $n \in \{0, 1, \ldots, N-1\}$, $f_0 + i$ composes a series of frequency components corresponding to $q_n(i)$, and $q_n(i)$ corresponds to $w(n)$. Note that here, the length of the binary sequence $q_n(i)$ is modified to I, i.e., $i \in \{0, 1, \ldots, I-1\}$. The proposed sinusoidal patterns have very similar correlation properties as PN sequences, while they further enjoy some robustness against desynchronization attacks. In addition, the scaling factor $\alpha(i)$ becomes a function of frequency component, determined by the psychoacoustic model. This is one of the few early examples that emphasized the notion of changing α from a constant to a function based on the psychoacoustic model. The watermark detection mechanisms for [15–18], which are highly related to the embedding functions, are omitted here for conciseness, and the readers could refer to the references for more information.

Different from [15–18], the watermark embedding schemes proposed in [19] and [20] cannot be easily represented by mathematically equations. In [19], one binary watermark bit is embedded by modifying the energy conditions among three consecutive non-overlapping signal frames. With synchronization codes designed in a similar way as the one used in [17], watermark extraction is performed by reversely mapping the energy conditions to the watermark bits. In [20], we see a similar idea of embedding and extraction of watermarks based on examining consecutive signal frame energies, but such a mechanism is imposed on the histogram of the host signal, rather than the time domain raw data. Noticeably, the authors in [19] proposed a mechanism to control the imperceptibility using the absolute threshold of hearing (ATH) of human auditory system (HAS). Although the mechanism is heuristically approached, it is an automated solution. Similarly, the ODG-based heuristic tuning is incorporated in [20].

Amplitude modulation technique is applied in [21] to perform watermark embedding. The concept of subband modulation is introduced to audio watermarking herein. Specifically, the host signal is first divided into I subands where I is a positive even number,

$$x(n) = \sum_{i=1}^{I} \left[x_{2i-1}(n) + x_{2i}(n)\right] + x_{\text{Residual}}(n),$$

where $x_{\text{Residual}}(n)$ is the remaining high frequency component of $x(n)$ not used for watermark embedding. The embedding formula is then given by

$$\begin{aligned} y_i(n) =& x_{2i-1}(n) \left[1 + \alpha(i) \sin(2\pi f_0 + w_0(i) + w(i))\right] \\ & + x_{2i}(n) \left[1 - \alpha(i) \sin(2\pi f_0 + w_0(i) + w(i))\right], \end{aligned}$$

where f_0 is a low frequency envelope, $w_0(i)$ is a randomly generated initial phase for the amplitude modulation, and $w(i)$ is the hidden information determined by phase shift keying. For simplicity of notations, we omitted the grouping process in [21] before watermark embedding. Therefore, $w(i)$ corresponds to the phases of 2π evenly divided by I. Watermark extraction is performed by envelope extraction followed by phase comparison processes. To assess the imperceptibility, $\alpha(m)$ is determined by the subband envelope ratios, and the subjective difference grade (SDG) is used for measurement.

Further studies of HAS with the purpose of discovering available "spaces" for audio watermarking are presented in [22–25], where cochlear delay characteristics have been utilized to embed watermarks. Cochlear delay is the non-uniform delays of wave propagation in the basilar membrane, where lower frequency components require more time to be perceived. According to this fact, the binary watermark bits are represented by two first order infinite impulse response (IIR) all-pass filters, whose transfer functions (in z-transform) are given by

$$H_w(z) = \frac{-b_w + z^{-1}}{1 - b_w z^{-1}},$$

where b_w is the filter coefficient, $0 < b_w < 1$, and $w \in \{-1, +1\}$. Then, the embedding function is given be

$$y(n) = x(n) * h_w(n),$$

where $*$ denotes linear convolution. The major limitation of the above scheme in [22–24] is the need of the original host signal during watermark extraction. Recently, this limitation has been broken by the improved system proposed in [25]. Note that in [24], the authors proposed another subjective imperceptibility measurement method called post hoc test with analysis of variance (ANOVA).

A unique perspective for audio watermarking is seen in [26], where the author proposed a system to achieve audio watermarking by exploiting the properties of spatial masking and ambisonics.

Specifically, the embedded watermarks are rotated versions of the host signal, and the system can be efficiently realized by appropriate arrangement of loudspeakers. Therefore, the watermarks are embedded in time domain, while the effects are taken in spatial domain.

2.4.2. Echo-Based

The echo-based audio watermarking methods [27–35] accomplish audio watermark embedding by adding attenuated echoes to the host signal, and perform watermark extraction via cepstral analysis. Note that the theory of cepstrum, [79, 80], is the foundation based on which such a paradigm has been developed. Also note that the difference between echo-based methods and cochlear delay based ones [22–25] is that the former embed watermarks by adding extra echoes while the later cause delay effects in low frequency region on the host signal without adding in new signals. For echo-based methods, we have

$$f[x(n), w(n)] = w(n-d) * x(n),$$

and the embedding function can be represented by

$$y(n) = x(n) + \alpha w(n-d) * x(n), \tag{3}$$

where d is the sample delay parameter. Usually, (3) can be equivalently expressed as a convolution function, i.e.,

$$y(n) = [\delta(n) + \alpha w(n-d)] * x(n),$$

where the filter applied to $x(n)$ is termed echo kernel [30–34], and the echo portion, $w(n)$, is termed echo filter [34]. Therefore, the design of an echo kernel is in fact the problem to design the echo filter. Early designs [27–30] of echo kernels can be generalized as in [34]

$$\begin{aligned} h(n) = \delta(n) + \sum_{i=0}^{I-1} \Bigg[& \underbrace{\overbrace{\alpha_{1,i}\delta(n-d_{1,i})}^{\text{Positive}} \overbrace{-\alpha_{2,i}\delta(n-d_{2,i})}^{\text{Negative}}}_{\text{Forward}} \\ & + \underbrace{\alpha_{1,i}\delta(n+d_{1,i}) - \alpha_{2,i}\delta(n+d_{2,i})}_{\text{Backward}} \Bigg], \end{aligned} \tag{4}$$

where I counts for the number of echo sets. In the original proposal in [27], $I = 1$, and only the forward positive echo existed. In [28], the concept of using both positive and negative kernels were proposed. Backward kernels were introduced in [29]. An analysis-by-synthesis approach and an interlaced kernel, $I = 2$, were proposed in [30]. Noticeably, the analysis-by-synthesis approach considers several attacks during watermark embedding phase, whose results are used to heuristically tune the scaling factor. The above designs are in terms of the combinations of the signs "+" and "−" for the echo filter and delay value, respectively. However, they suffer from security issue that the echo kernels can be detected via cepstral analysis.

To secure the audio watermarking systems against unauthorized watermark detection, a secret key should be incorporated during watermark embedding and extraction phases. Time-spread echo-based methods are hence proposed to satisfy the security requirement [31]. Different from conventional single or multiple echo kernels, time-spread echoes are composed using a sequence, e.g., binary PN in [31], modified pseudo noise (MPN) in [32], colored PN in [33], and finite-impulse-response (FIR) filter coefficients in [34]. The designs of $w(n)$ using these sequences establish the security feature of the systems. As a result, in the corresponding watermark extraction phase, a correlator is involved to explore the auto-correlation properties of the above mentioned sequences.

Among the time-spread methods, the use of the binary PN sequence [31] is the pioneering work, based on which the MPN is proposed in [32] for improved imperceptibility property as well as robustness. In [33], a novel embedding signal model, i.e., a dual-channel scheme is proposed:

$$\begin{cases} y_{\text{odd}}(n) = [\delta(n) + 0.5\alpha w(n-d)] * x_{\text{odd}}(n), \\ y_{\text{even}}(n) = [\delta(n) - 0.5\alpha w(n-d)] * x_{\text{even}}(n), \end{cases}$$

where $x_{\text{odd}}(n)$ and $x_{\text{even}}(n)$ are composed by extracting odd and even samples of $x(n)$ respectively. The imperceptibility property of the system is further improved since the scaling factor is attenuated by 0.5 in both channels. A systematic approach that quantitatively controls the imperceptibility and robustness is presented in [34]. Specifically, this work addresses the problem from the perspective of digital filter design, and the advanced FIR filter design methods are incorporated. The imperceptibility is controlled by using the ATH and the proposed maximum power spectral margin (MPSM), while the robustness is quantified by constraining the peak and sidelobe values of the auto-correlation function of the echo filter coefficients. The MPSM consists of the maximum powers of each frequency bin that have existed in the host audio frames. The design thus becomes a feasibility (optimization) problem, whose general idea is summarized as

$$\begin{aligned} \text{find} \quad & w(n) \\ \text{s.t.} \quad & \text{constraints on power spectrum of } w(n), \\ & \text{constraints on auto-correlation of } w(n). \end{aligned}$$

In this formulation, the constraints on power spectrum ensures that the artifacts introduced by convolving the original signal with the echo kernel stay beneath the ATH of HAS. For the constraints on the auto-correlation function, the peak at the center is first set to a constant value. Then, an upper bound is imposed on absolute values of the sidelobes to ensure successful watermark detection. More discussions and variations can be found in [34]. To evaluate the imperceptibility of echo-based audio watermarking systems, a method called AXB listening test [81] has been commonly used in this category [31–34]. Note that the ODG score is used in [30].

Recently, the host signal interference rejecting property is first introduced to echo-based methods in [35] by using host signal characteristics during watermark embedding. The embedding function is a slight modification of (3), i.e.,

$$y(n) = x(n) + \eta\alpha w(n-d) * x(n),$$

where

$$\eta = \begin{cases} -1, & \sum_n c_x(n)w(n-d) < 0, \\ +1, & \sum_n c_x(n)w(n-d) \geqslant 0, \end{cases}$$

and $c_x(n)$ is the real cepstrum of $x(n)$. By priorly considering the correlation between the cepstrum of the host signal and the echo kernel, the host signal interference is turned to be a helpful parameter to enlarge the desired correlation peak during watermark extraction, while the imperceptibility is almost identical to the system based on (3) because $|\eta| = 1$.

2.5. Transform Domain Methods

While the time domain methods are more like ad-hoc solutions, the transform domain methods seem to be more favored by researchers and designers, yet there has not been rigorous evidences to support that transform domain methods are better than time domain ones. More discussions about watermark embedding domains are provided in Section 4.3. For transform domain methods, two

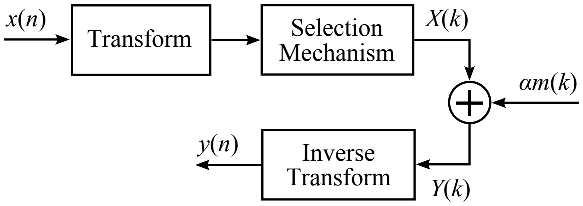

Figure 3: A generic block diagram of transform domain embedding.

more steps, i.e., forward and inverse transform before and after watermark embedding are needed, which is illustrated by Fig. 3. In this way, it is important to ensure that the modified transform domain samples, $Y(k)$, can take inverse transform with appropriate forms. For example, methods based on DFT should preserve the symmetric property of frequency domain samples within $[-\pi, \pi)$ in order to obtain real-valued samples of $y(n)$ after the inverse transform. Sometimes, not all of the transform coefficients are watermarked, and mechanisms to efficiently select the coefficients for watermark embedding are used for improved imperceptibility and robustness. In the following content, we discuss each subcategory of transform domain methods.

2.5.1. SS and Variations

The incorporation of SS technique to the application of digital audio watermarking was originally proposed in both [36] and [7], almost at the same time. In [36], a general introduction with SS signal processing model was given, while more details were provided in [7] in the context of image watermarking. Note that the watermark detection scheme proposed in [7] is signal dependent. A formal treatment of SS-based watermarking for audio signals was seen in [37], where the signal model is identical to (2), and $m(k) \in \{-1, 1\}$. In this work, the imperceptibility is improved by empirically avoiding embedding watermarks in blocks containing both silent and rich sound. Meanwhile, a series of strategic treatments were proposed to improve watermark detection, including embedding watermarks in audible regions, replacing correlation test by cepstrum filtering plus covariance test, block repetition embedding, etc.

Note that the systems in [7, 36] are non-informed, meaning the properties of $X(k)$ are not explored to facilitate watermark embedding. However, a novel SS-based method, termed improved spread spectrum (ISS), is proposed in [38], whose embedding scheme involves signal dependent watermarks. The embedding scheme in [38] is given by

$$Y(k) = X(k) + (\alpha - \lambda\Phi)\, m(k), \tag{5}$$

where $\Phi \triangleq \sum_k X(k)m(k)/\sum_k m^2(k)$, $m(k) \in \{-1,1\}$, and λ controls the removal of host signal interference during watermark detection. The major contribution of [38] lies in the introduction of $\lambda\Phi$, and it has also been discussed that the function of Φ can be arbitrary nonlinear functions. In addition, the conventional embedding model (2) can be obtained by setting $\lambda = 0$ in (5). The advantage of using (5) over (2) can be explicitly shown by the detection correlation result at zero-lag point, i.e.,

$$\frac{\sum_k Y(k)m(k)}{\sum_k m^2(k)} = \alpha + (1-\lambda)\,\Phi,$$

whereas the result using (2) is $\alpha+\Phi$. Statistical analysis further shows that $\lambda = 1$ is approximately the optimal solution for additive noise model. In this case, the interference term, Φ, caused by the host signal, is eliminated. In [39], the multiplicative spread spectrum (MSS) scheme is investigated. The equivalent additive model (informed) for conventional MSS given by

$$Y(k) = X(k) + \alpha X(k)m(k), \tag{6}$$

and the corresponding nominal detection test statistic is given by $\sum_k Y^2(k)m(k)$. Based on such a signal model, an improved embedding scheme is proposed as

$$Y(k) = X(k) + \alpha X(k)m(k) + g[\alpha, X(k)]X(k)m(k), \tag{7}$$

where $g[\alpha, X(k)]$ is a function designed to eliminate squared terms in the test statistic, so that the binary decision on watermark bit can be correctly made. While it seems more rigorous to call (7) a multiplicative model than (5), since the former has $X(k)$ explicitly shown in the watermark term, we note that the works in [38] and [39] share the same idea, i.e., both the proposed schemes aim at eliminating the interference terms existing in the detection test statistics. From this point of view, it then may be worth comparing the robustness of the two systems.

Apart from the above works dealing with the interference caused by the host signal in an additive noise framework, the works presented in [40–42] intensively focus on the robustness against desynchronization attacks. In [40], the key idea to track pitch scaling effects is to perform watermark embedding in a logarithm scale. The embedding equation is the same as (6), but the coefficients for watermark embedding follow a novel nonlinear mapping mechanism, which is illustrated in Fig. 4. In this figure, only the frequency region within $[0.12, 0.24)$ is watermarked, and in this region, multiple frequency indices are mapped to the same logarithmic index, indicating the same watermark is repeatedly embedded into the corresponding frequency bins. One of the drawbacks of the system proposed in [40] may be the processing of the host signal as a single frame, i.e., DFT is applied to the whole signal, and audio clips of 44100 Hz sampling rate and 15 seconds duration are used therein. This may compromise the imperceptibility property and computational efficiency especially when dealing with high quality long audio files. The imperceptibility test is performed using ODG in [40]. In [41], three embedding region selection methods and one watermark embedding method are presented. The selected regions are energy significant ones that are invariant under pitch-invariant time scaling attacks. The watermark embedding is performed by exchanging consecutive coefficient pairs within a selected low frequency portion of the magnitude spectrum according to $m(k)$ and energy comparison results, i.e.,

$$\begin{bmatrix} Y(2k) \\ Y(2k+1) \end{bmatrix} =$$

$$\begin{bmatrix} X\left(2k + \{0 \oplus \operatorname{sgn}\{[X(2k+1) - X(2k)]\, m(k)\}\}\right) \\ X\left(2k + \{1 \oplus \operatorname{sgn}\{[X(2k+1) - X(2k)]\, m(k)\}\}\right) \end{bmatrix},$$

where

$$\operatorname{sgn}(\theta) \triangleq \begin{cases} 1, & \theta > 0, \\ 0, & \theta \leq 0, \end{cases}$$

for an arbitrary real number θ, the symbol $\oplus$ is the logical "XOR" operator, and $m(k) \in \{-1, 1\}$. Although the embedding scheme is not strictly consistent with the expression (6), the embedding procedure still directly introduces some noise into the frequency spectrum, and hence we have categorized this work into the SS-based category. Note that in this scheme, the length of $m(k)$ is half of the length of $X(k)$ since two samples of $X(k)$ are needed to embed one bit of watermark. The SDG, which is corresponding to the ODG, is used in [41] to evaluate the imperceptibility property of the system. The concept of embedding watermarks in localized regions is further explored in [42], where the authors have incorporated feature points detector techniques in image processing, and proposed a robust audio segments extractor (RASE) algorithm. The selected feature samples are energy significant ones, and each segment is centered at the local peak. The embedding process is performed on the level three stationary wavelet transform (SWT) coefficients of the feature segments, with a model identical to (2), but the watermark bits, $m(k)$, are Gaussian

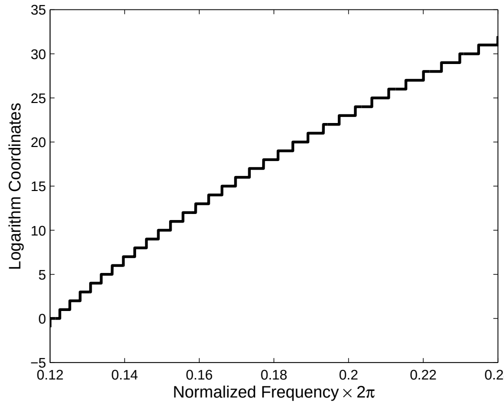

Figure 4: An example of the nonlinear mapping proposed in [40].

distributed, rather than a binary PN sequence. The SDG is also used in [42] to evaluate the imperceptibility.

In transform domain watermarking schemes, if the transform is complex valued, then usually the watermarks are embedded into the magnitudes of the transformed coefficients. This can be seen from [40] and [41], where DFT is used. However, the work reported in [43] has revealed that the watermarks can also be embedded in the phase terms of the DFT coefficients, i.e.,

$$Y(k) = X(k)e^{j\alpha(k)m(k)},$$

or more explicitly,

$$\angle Y(k) = \angle X(k) + \alpha(k)m(k),$$

where $\angle$ denotes the angle of a complex number, and the watermarks are complex values on the unit circle. Similar to the concept of manipulating α in [18], intensive discussions are provided for the implementation of the psychoacoustic model to obtain a global masking curve, based on which $\alpha(k)$ is determined. Both ODG and SDG are used to for the imperceptibility test. To deal with resampling attacks, a heuristic search mode is proposed to restore synchronization. In the detection phases of the systems proposed in [40–43], approaches similar to the corresponding embedding schemes are used to first extract the watermark bits. After that, correlation or bit error rate (BER) is calculated to quantify the robustness of the systems. It is important to note that amplitude and phase modulation (modification) are a very common means to perform watermark embedding, either in time [19–21] or in transform domain [43]. More of such examples in frequency domain are reported in [44–46].

The remaining SS-based works [47–52] can reflect the strong vitality in this research subdomain, in which various aspects regarding system designs and applications have been proposed. For brevity, detailed signal models are not discussed in this paper. Instead, we provide a few comments to these works in the reminder of this subsection, and interested readers could refer to the original works for more details.

Recently, an solution is proposed in [47] to substantially increase the embedding capacity of SS-based methods without compromising the robustness. The major contribution of this work is the creation of a set of, instead of two (corresponding to "-1" and "+1" respectively), PN sequences for watermark embedding. This is validated by discovering the preserved correlation properties between a PN sequence and its circularly shifted versions. In this way, multiple information bits can be embedded in a single frame instead of the conventional one-bit-per-frame situation. Specifically, M bits of information could be embedded in a single frame by generating 2^M circularly shifted PN sequences of a length greater than M, and selecting the PN sequence corresponding to the M bits information for watermark embedding. In [48], another solutions to increasing capacity while preserving the robustness is proposed. Interestingly, it is based on the use of Fibonacci numbers. Due to the special structure of Fibonacci numbers, watermark embedding only modifies a few FFT samples but can achieve high embedding capacity. The embedding distortion is also effectively controlled by the properties of Fibonacci numbers. A special solution to speech watermarking is proposed in [49] and [50], based on formant enhancement. Formant is the concentrated frequencies close to the resonance frequency of the vocal tract. By using the estimated formants, watermark bit "-1" is embedded by enhancing the sharpest formant while watermark "+1" is embedded by enhancing the second sharpest one. The applicability of this method to speech tampering detection is discussed in [50]. Another interesting work is presented in [51], where the SS technique is combined with the amplitude expansion technique commonly used in reversible data hiding literature, to simultaneously achieve robustness and reversibility of an audio watermarking system. Reversible (or lossless) data hiding refers to exactly restoring the original host signal (zero recovery error) after watermark extraction, which is a very important requirement in several applications [82]. Finally, compressed domain embedding is proposed in [52], where the watermarks are directly embedded in copmressed MPEG data in 32 frequency subbands, either by tuning the scaling factor during quantization and bit allocation or by modifying encoded audio sample. Te general objective of this propose is to achieve real-time stream embedding without extra processes required for re-compression.

Generally, it can be seen, from the literature, that SS-based methods are favored by researchers and practitioners due to the long history and maturity of the SS theory established in digital communications.

2.5.2. QIM and Variations

The QIM technique refers to modulating the watermarks in the indices of a series of quantizers which are then applied to the host signal. The original idea is well elaborated in [53], where the authors sufficiently investigated this technique from information-theoretic point of view to practical realization examples. Although it was designed for general multimedia contents, more emphasis on the use cases had actually been put on image contents. One of the motivations of creating such a paradigm is from noting the inability of dealing with host interferences in conventional SS-based systems. Recall that [38] and [39] are dedicated to dealing with this drawback in the context of SS technique.

Original implementations of the QIM technique [53] are in terms of i) dither modulation, which is later commonly referred to equivalently as QIM, ii) distortion-compensated QIM, and iii) spread-transform dither modulation (STDM). The simplest realization of dither modulation, which is also usually used in subsequent works [54–60], is illustrated in Fig. 5. The embedding scheme presented in Fig. 5 can be mathematically expressed as

$$Y(k) = \begin{cases} \left[\lfloor X(k)/\Delta \rfloor + 3/4\right]\Delta, & m(k) = 1, \\ \left[\lfloor X(k)/\Delta \rfloor + 1/4\right]\Delta, & m(k) = -1, \end{cases}, \tag{8}$$

where $\lfloor \theta \rfloor$ denotes the maximum integer less than θ. The corresponding extraction scheme is somewhat straightforward, in which (under no attacks for example) one only needs to calculate the quantization residual of $Y(k)$ according to Δ and decide $m(k)$ according to a threshold of $\Delta/2$. To comply with the generic scheme (2), we can rewrite (8) as

$$Y(k) = X(k) + \frac{[m(k) + 2]\,\Delta}{4} - r\left[X(k), \Delta\right], \tag{9}$$

where $r(\theta, \Delta)$ denotes the quantization residual, e.g., $r(2.1, 1) = 0.1$. Note that the larger the Δ, the more robust the system. However, it is also more likely to cause audible artifact. To improve such a trade-off, distortion-compensated QIM comes into the scene, whose embedding function is modified from (9) to

$$Y(k) = X(k) + \frac{[m(k)+2]\,\Delta}{4\beta} - \beta r\left[X(k), \frac{\Delta}{\beta}\right], \tag{10}$$

where $0 < \beta \leq 1$. First, the quantization step-size is scaled up by β, which causes more distortion during watermark embedding process. However, the distortion is compensated by the last term of (10), which is not hard to verify. if β approaches 1, then (10) approaches (9), indicating a full quantization but with the smallest step-size Δ/β. On the other hand, if β is small, then the portion $\{X(k) - \beta r\,[X(k), \Delta/\beta]\} \longrightarrow X(k)$, meaning a $X(k)$ is weakly quantized, although the watermark values become larger. Another effective treatment is the STDM, which embeds a watermark bit into the projection of a whole frame of the host signal on a random vector. Let $\mathbf{x} = [X(0), X(1), \ldots, X(K-1)]^T$, where $\{\cdot\}^T$ is the transpose operator, and $\mathbf{v} \in \mathbb{R}^{K\times 1}$ be the unit-energy random vector, then the STDM embedding scheme is given by

$$\mathbf{y} = \mathbf{x} + \left[\frac{(m+2)\,\Delta}{4} - r\left(\mathbf{x}^T\mathbf{v}, \Delta\right)\right]\mathbf{v}, \tag{11}$$

where $\mathbf{y}$ is formed in the same way of $\mathbf{x}$, and $m \in \{-1, 1\}$ is a single watermark bit for the specific frame $\mathbf{x}$. For simplicity, we omitted the subscript representing frame indices. The corresponding watermark extraction scheme is a slightly different from the straightforward ones in dither modulation or distortion-compensated QIM, because it involves a correlation process due to the introduction of $\mathbf{v}$. Assuming a closed-loop environment, then the test statistic is given by the quantization residual of $\mathbf{y}^T\mathbf{v}$, i.e.,

$$\begin{aligned} r\left(\mathbf{y}^T\mathbf{v}, \Delta\right) &= \mathbf{y}^T\mathbf{v} - \left\lfloor \frac{\mathbf{y}^T\mathbf{v}}{\Delta} \right\rfloor \\ &= \mathbf{x}^T\mathbf{v} + \frac{(m+2)\,\Delta}{4} - r(\mathbf{x}^T\mathbf{v}, \Delta) - \left\lfloor \frac{\mathbf{x}^T\mathbf{v}}{\Delta} \right\rfloor \\ &= \frac{(m+2)\,\Delta}{4}, \end{aligned}$$

where $\left\lfloor \mathbf{x}^T\mathbf{v}/\Delta \right\rfloor = \mathbf{x}^T\mathbf{v} - r(\mathbf{x}^T\mathbf{v}, \Delta)$. Therefore, the same watermark extraction threshold as used in dither modulation, $\Delta/2$, is used for the decision of the embedded bit. It is then verified in [53] that STDM can lead to improved capacity and better signal-to-noise ratio (SNR) in watermark extraction phase as compared to the SS counterpart.

It should be noted that, in the examples provided above, the dither modulation (9) has K times

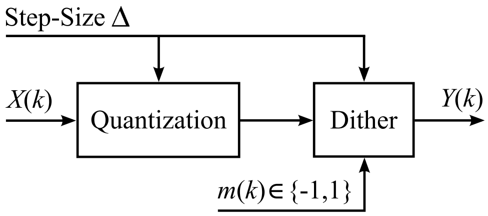

Figure 5: A dither modulation audio watermark embedding scheme.

larger embedding capacity as compared to STDM (11), because in the former, the watermark bits $m(k)$ are embedded in each sample of $X(k)$, while in the latter, one watermark bit m is embedded in one frame which is the ensemble of $X(k)$. However, this is not necessarily true, since we have omitted a modulation step of the watermark bits in the dither modulation. In fact, in the original descriptions of dither modulation [53], each quantizer is a function of both k and m, indicating that one watermark bit is embedded in one frame rather than one sample. Here, the simplification is in terms of a constant Δ among different frames. Such a simplified treatment is adopted in all subsequent works [54–60].

It can be seen from the above that the theory of QIM for audio watermarking has been well developed by its inventor in [53]. Therefore, instead of the investigations on modified or improved QIM algorithms, the subsequent works are more towards incorporating QIM into their ad-hoc designs, in which various transforms have been incorporated in the systems, including DWT [54, 58], empirical mode decomposition (EMD) [55], and combined DWT and DCT [56, 57, 59]. Among these works, synchronization codes are commonly used to cope with desynchronization attacks [54–59]. A series of heuristic algorithms are incorporated in the designed systems, such as the differential evolution algorithm [56] used to control the embedding strength in an analysis-by-synthesis approach similar to [30], support vector machine (SVM) used to determine optimal watermark embedding regions [58], and better watermark extraction [59]. A special treatment is presented in [60] to deal with the vulnerability of QIM methods (see Introduction of [38]) against amplitude scaling attacks. This is achieved by designing a maximum likelihood estimator of the amplitude scaling factor under a scaling-plus-noise attack signal model. The distributions of the host signal, dithering sequence, and additive noise (all statistically modeled) are assumed known during watermark extraction phase. The solution to the maximum likelihood estimation is obtained by brute force peak searching which is very computationally expensive, as mentioned by the authors. The estimated amplitude scaling factor is then compensated at the detector before QIM watermark extraction. Performance degradations using practical audio samples are discussed, where the reasons mainly lie in the distribution mismatch between the host and watermarked signal, and the fact that practical audio signals are non-stationary and correlated. It should be noted that in the category of QIM methods, apart from ODG [55] and SDG [56] measurements, no active control of imperceptibility has been deployed.

2.5.3. Patchwork and Variations

Literally, patchwork refers to small elements to be added into a host subject. As such, it was originally introduced for image watermarking in [36]. Then, the work in [61] has brought this technique into audio watermarking system designs. The patchwork technique is by nature a dual-channel scheme, indicating that the samples of transformed host signal are separated into two sets, in which the watermarks are embedded in slightly different ways. Note that the dual-channel concept has also been used in [33] and [41]. However, instead of using deterministic channels as in [33] and [41], patchwork methods randomly choose from transform domain samples to form the two channels. In this way, the generic equation (2) should be modified to a dual form as well. Let the randomly selected non-overlapping channels of $X(k)$ be $X_i(l)$, where $i \in \{0, 1\}$, and $l \in \{0, 1, \ldots, L-1\}$, $L < K$, is the dimension of the subsets, then the watermark embedding function is given by

$$Y_i(l) = X_i(l) + (-1)^i m, \tag{12}$$

where $m > 0$ is an arbitrary scalar that represents the strength of the patchwork. The corresponding watermark detection scheme is straightforward. Specifically, according to the indices of the selected samples of $X_i(l)$, $Y_i(l)$ is formed from $Y(k)$ in the same way. Then, m can be detected according to the calculation of

$$\frac{\sum_l [Y_0(l) - Y_1(l)]}{L} = \frac{\sum_l [X_0(l) - X_1(l)]}{L} + 2m, \tag{13}$$

where the first term in the right hand side is the host signal interference. Ideally, the interference term equals to 0 if $X_i(l)$ has a zero mean value. Based on (13), an improved design is proposed in [62], in which the binary watermark bits are modulated to the patches before watermark embedding in DCT domain. Suppose the subsets of DCT coefficients for embedding a watermark bit $m \in \{-1, 1\}$ are $X_{0,m}(l)$ and $X_{1,m}(l)$, respectively, then the embedding function is given by

$$Y_{i,m}(l) = X_{i,m}(l) + (-1)^i \alpha A(m) S(m), \tag{14}$$

where

$$A(m) = \text{sgn}\left\{\sum\nolimits_l \left[X_{0,m}(l) - X_{1,m}(l)\right]\right\},$$

and

$$S(m) = \sqrt{\frac{\sum_i \sum_l \left\{X_{i,m}(l) - \sum_l X_{i,m}(l)/L\right\}^2}{L(L-1)}}.$$

The advantage of (14) over (12) is that whenever $A(m) = 1$ or -1, the larger quantities between $X_{0,m}(l)$ and $X_{1,m}(l)$ will always become even larger, while the smaller ones will always become even smaller. In this way, the test statistic will be more distinguishable, and the watermark extraction could be more robust against common attacks. Some modifications of (14) are presented in [63], in which psychoacoustic model is utilized for better imperceptibility.

The multiplicative patchwork method is intensively discussed in [64], in which only one of the channels is embedded with watermarks, and the other is unchanged. Suppose Channel 0, i.e., $X_0(l)$ is selected for embedding, then original embedding function is given by

$$Y_0(l) = \alpha^m X_0(l),$$

whose equivalent additive model is given by

$$Y_0(l) = X_0(l) + (\alpha^m - 1)\, X_0(l).$$

where $0 < \alpha < 1$. During the watermark extraction phase, where we also suppose a close-loop environment, the energies of $Y_0(l)$ and $Y_1(l)$ are compared to decide which watermark bit is embedded. Therefore, it is preferred that the selected two subsets, $X_0(l)$ and $X_1(l)$, have similar energies. A heuristic approach based on ODG is used to modify α in order to achieve the desired imperceptibility.

The latest proposals for patchwork-based systems are proposed in [65] and [66], where very sophisticated embedding and extraction schemes are investigated. The procedures to perform watermark embedding [65] consist of i) equally divide a time domain frame $x(n)$ into two parts, termed as front and rear part respectively; ii) perform DCT to each part and select low-to-middle frequency components for embedding; iii) further divide the selected components into sub-frames of equal size; iv) for each sub-frame, selecte two channels of coefficients according to random indices; v) perform watermark embedding in a way similar to [19], i.e., embedding watermarks via manipulating the energy levels. If we denote the two channels of samples in the front and rear parts by $X_{i,\mathrm{F}}(l)$, and $X_{i,\mathrm{R}}(l)$, respectively, $i \in \{0,1\}$, then the function to embed one watermark bit, $m \in \{-1,1\}$, is given by

$$\begin{cases} Y_{i,\mathrm{F}}(l) = X_{i,\mathrm{F}}(l) + (0.5 - i)F(\alpha, m)X_{i,\mathrm{F}}(l), \\ Y_{i,\mathrm{R}}(l) = X_{i,\mathrm{R}}(l) - (0.5 - i)R(\alpha, m)X_{i,\mathrm{R}}(l), \end{cases}$$

where

$$F(\alpha, m) = \\ \alpha \operatorname{sgn}\left\{\sum\nolimits_l \left[(2m + \alpha)\left|X_{0,\mathrm{F}}(l)\right| - (2m - \alpha)\left|X_{1,\mathrm{F}}(l)\right|\right]\right\},$$

and

$$R(\alpha, m) = \\ \alpha \operatorname{sgn}\left\{\sum\nolimits_l \left[(2m + \alpha)\left|X_{1,\mathrm{R}}(l)\right| - (2m - \alpha)\left|X_{0,\mathrm{R}}(l)\right|\right]\right\}.$$

Note that $F(\alpha, m)$ and $R(\alpha, m)$ are binary quantities resulting from comparison of energy related terms, and they only take values from 0 and α. If 0 is the result, then the corresponding coefficients are unchanged during the watermark embedding process. Also note that, here, the scaling factor α determines the values of $F(\alpha, m)$ and $R(\alpha, m)$, which indirectly controls the watermark strength. The corresponding watermark extraction scheme follows the same mechanism as the embedding process. The work presented in [66] modifies the system in [65] by, first, processing the host signal as a single frame, which is similar to [40], and then, adding a re-scaling mechanism with synchronization bits, to achieve the robustness against desynchronization attacks.

Remark 1: Some clarifications about the methodology for the categorization of existing audio watermarking works are made here. i) Time domain and transform domain methods are strictly distinguished by the explicit domain in which the watermarks are embedded. In other words, an audio watermarking system without the forward and inverse transform procedures shown in Fig. 3 can be categorized as a time domain method. We emphasize on this criterion because a transform domain method can also be loosely considered as a time domain method whose time domain watermarks are obtained by finding the difference between the original and watermarked signals in time domain, and vice versa. ii) In fact, all transform domain methods could be loosely considered as an SS-based method since the watermark signal is added (spread in some sense) into the audio samples represented in a specific spectral domain (e.g., FFT, DCT, DWT, etc.). However, such a loose categorization could result in no benefit but losing the distinctions among different transform domain subcategories. Therefore, we stick to three subcategories in Fig. 2, in which we emphasize the classical technique and further extensions in SS-based works, the specific methodology of embedding watermarks through controlled quantizations in QIM-based works, and the dual-channel and random indexing nature in patchwork-based works.

2.6. Summary

The digital audio watermarking works reviewed in this section can effectively reflect the situations of the research and development for this topic in the last two decades. It can be seen that the primary concerns of the designers are mainly on imperceptibility and robustness, which is largely determined by the watermark embedding schemes. Meanwhile, the works reviewed in Section 2.4 are specifically dedicated to audio media, as compared to the ones discussed in Section 2.5 that are applicable to multiple multimedia formats. The variety of solutions also indicates the complexity of the problems in audio watermarking system design, hence we observe that many works are dedicated to specific problems or under constrained situations, e.g., some aiming for host interference rejection, while others dealing with desynchronization attacks, or improving imperceptibility without compromising the established robustness, etc. In the sequel, we provide comprehensive reviews on imperceptibility and robustness properties of existing audio watermarking systems, which summarizes current state-of-the-art solutions and reveals several open challenges.

3. Imperceptibility – Preserving Audio Quality

As one of the most important characteristics of a digital audio watermarking system, the imperceptibility property has become a must that should be considered during system designs. However, this is not an easy problem since it is difficult to claim that the watermarks are strictly imperceptible. First, the hearing abilities among individuals can be very different, and the hearing abilities of an individual also change over time. Second, the rapid development of data storage devices and high quality audio systems have enabled high resolution playbacks, which raises the risk of revealing the hidden watermarks. In this section, we first comprehensively review existing proposals to deal with imperceptibility, commenting on the efficiency and accuracy. Then, we discuss the choices of watermark embedding regions and the trade-off between the use of the psychoacoustic model and the robustness against lossy compression.

3.1. Imperceptibility Measurement and Control

An intuitive measurement of imperceptibility of the watermarks would be the embedding SNR, i.e.,

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_n x^2(n)}{\sum_n \left[y(n) - x(n)\right]^2}.$$

However, it can only provide limited information about the degradation of sound quality. Therefore, some modified measurements, which weight different frequency bands differently, are utilized to more accurately measure the imperceptibility. One of the examples is the frequency-weighted segmental signal-to-noise ratio (fwsSNR) [83], which is given by

$$\mathrm{fwsSNR} = \frac{10}{N_{\mathrm{Seg}}}\sum_{i=0}^{N_{\mathrm{Seg}}-1}\frac{\sum_k \left|X_i(k)\right|^2 \log_{10}\frac{|X_i(k)|^2}{\left[|Y_i(k)|-|X_i(k)|\right]^2}}{\sum_k \left|X_i(k)\right|^2},$$

where N_{Seg} is the number of non-overlapped frames of the original and watermarked signal, and i is the index of the frames. However, the above measurements are still not adequate in revealing the impacts of the added artifacts to human perception. More popular performance metrics to study the imperceptibility rely on two types of measurements more related to human hearing abilities, i.e., subjective listening tests, and objective measurements with well designed human auditory models. The former usually includes the 2AFC paradigm [77], the AXB paradigm [81], post hoc test with ANOVA [24], and the SDG score [78], whereas the later refers to the ODG score [78]. The 2AFC test is used to determine the masking curve according to the listeners responses between the original audio and the watermarked versions with different embedding levels. The AXB test involves three versions of testing audio clips, marked as A, B, and X, in which A and B (cannot be the same) are randomly chosen from original and watermarked signals, and X is randomly chosen from A and B. The listener is asked to determine which one from A and B is the same as X. In the post hoc test, the listener is asked to grade two audio clips using a number from $\{0, 1, 2, 3\}$, where 0 corresponds to exactly the same and 3 corresponds to completely different. The two clips are an original clip plus another one randomly chosen from the original and watermarked clips. Then ANOVA is applied to the scores given by all the listeners. Both SDG and ODG are from ITU-R BS.1837-1 recommendation [78]. In the SDG test, three audio clips marked as A, B, and

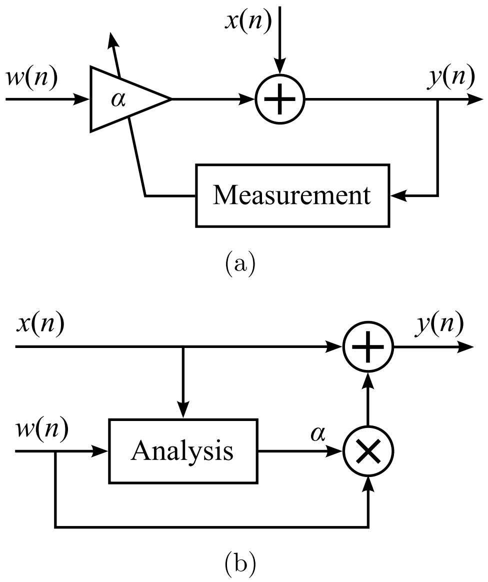

Figure 6: Typical block diagrams of (a) heuristic and (b) analytical schemes to determine α in imperceptibility control.

C are presented. The listeners are asked to tell between B and C the one indiscernible from the original audio A. Then a grade within $\{0, -1, -2, -3, -4\}$ is assigned to the selected piece, with 0 corresponding to imperceptible and -4 corresponding to very annoying. The ODG has the same output score values as SDG, but it is an automated test without peer listeners. In addition, researchers have also looked into objective sound quality measurement with the consideration of watermark robustness, which is presented in [84].

Furthermore, the above methods can be used not only to evaluate the imperceptibility property of the system, but also to automatically control the strength of the watermarks[2], i.e., α. This is usually realized by establishing a heuristic tuning mechanism, and a typical example is shown in Fig. 6 (a). Specifically, a tentative value of α is first used to generated the watermarked signal. Then, the measurement procedure decides whether the watermarks exceed some predefined tolerance. If it turns to be audible, then α is reduced accordingly. Otherwise, α is increased. This process is repeated until the distortion level is within the desired range. Examples can be seen in [16, 17, 19, 20, 30], and [37]. Although being effective, such a mechanism is less efficient since many design cycles will be performed. It should also be noted that there lacks sufficient considerations of imperceptibility in categories of QIM and patchwork methods. We observe that for QIM, SNR is a common measurement [54, 56–59], while for patchwork, the ODG values are set to around -1,

[2] Note that the subjective (manual) approaches which require feedbacks of human participants are generally excluded in automated designs. Instead, they are more often used for evaluations.

e.g., [64], [66], meaning the watermarks are "perceptible but not annoying" [78]. In fact, the above mentioned heuristic tuning mechanism can be effectively utilized in these methods to discover the imperceptibility-robustness trade-off. However, proactive or systematic control of imperceptibility, which will be discussed in the following content, seems not easy to be incorporated in QIM and patchwork methods, due to the nature of the dual-channel watermarking embedding mechanisms.

A more proactive approach to control the imperceptibility is to analytically determine the value of α (or some times together with $w(n)$), before watermark embedding, according to some imperceptibility rules, which is illustrated in Fig. 6 (b). The key component of the analysis module in Fig. 6 (b) is the psychoacoustic model of HAS, including the masking curve and ATH. The imperceptibility is hence achieved by embedding the watermarks under the masking threshold or ATH so that they cannot be perceived by human ears. In [15], a lowpass filter is applied to the watermarks, making the resultant watermark power spectrum lying below that of the host signal. Although this treatment is not strictly based on a psychoacoustic model, the basic idea of masking is shared. A special case is seen in [17], where α is heuristically controlled as described in the previous paragraph, while $w(n)$ is filtered below the local masking threshold. A perceptual shaping procedure to obtain the global masking curve for phase shifts is proposed in [43]. Examples of using the ATH are seen in [19] (heuristic) and [34] (analytical). In [34], the ATH is combined with the proposed MPSM to obtain a desired filter response that can suppress the power spectra of all locally embedded watermarks under the ATH.

3.2. Watermark Embedding Regions

A primary question for audio watermarking system designers, although not explicitly mentioned in many existing works, is where the best locations to embed the watermarks are. The widely accepted answer to this question is raised in [7], which states that the watermark should be embedded in perceptually significant regions, i.e., low to middle frequency bands, so that the attackers can only remove the watermarks at the price of destroying the host signal. However, we have also seen solutions with watermarks embedded in high frequency bands (e.g., echo-based solutions [31–34]). In addition, the use of a psychoacoustic model is also optional in the existing designs. Therefore, we classify watermark embedding regions into i) perceptually significant regions with psychoacoustic constraints, e.g., [16, 19], ii) perceptually significant regions without psychoacoustic constraints, e.g., [37, 40–43, 57–59, 64–66], and iii) perceptual insignificant regions, e.g., [31–34]. It is worth noting that embedding watermarks in frequency insignificant regions is in fact very similar to those based on a psychoacoustic model, since the masking curve and ATH in very low and high frequency regions contain very strong energies which ensure that the watermarks are perceptually bounded. In general, the preferences on watermark embedding locations would be finally determined according to specific applications with different requirements on the imperceptibility and robustness properties.

The degradation of audio quality after watermark embedding can be generically characterized from the perspectives of digital filter or psychoacoustic models. If we model the watermark embedding as a filtering process with transfer function denoted as $H_{\mathrm{W}}(k)$, then we have

$$|H_{\mathrm{W}}(k)|^2 = \frac{|Y(k)|^2}{|X(k)|^2}, \tag{15}$$

for a single frame. This is usually used in echo-based designs, since the echo kernels are usually fixed for each frame. Examples are provided in Fig. 7. However, other methods can also use (15) to quantify the added artifacts. Note that $H_{\mathrm{W}}(k)$ serves as an equalizer which modifies different frequency components. In addition, for long audio signals which requires time-frequency analysis, efficient consolidation of $H_{\mathrm{W}}(k)$ for each frame may not be an easy task. In this respect, the use of $H_{\mathrm{W}}(k)$ seems less informative in telling the degradation of audio quality. Therefore, psychoacoustic models have been used to more straightforwardly control the audibility of the artifacts. For example, in a realization of the proposed design in [34], the MPSM of the echo portion in (3), i.e., $y(n) - x(n) = \alpha w(n-d) * x(n)$, is almost upper bounded by the ATH, which is shown in Fig. 8. In contrast, due to lack of consideration of the imperceptibility, the patchwork-based watermarks designed in [62], i.e., (14), leak out of ATH within Bark scale 13 to 18 Barks. Although the masking effects of host signal can cover some part of this leakage, artifacts can be perceived in the resultant watermarked signal, according to our experiments.

Unfortunately, psychoacoutsic models are also the key components in lossy audio compressions, in which the most widely commercialized compression formats are MP3 and AAC. While audio watermarking methods explore the imperfection of HAS and embed the watermarks in perceptual insignificant regions, lossy audio compression methods just discover the “redundant” data in these regions and achieve lower bit rates without much compromising audio quality. In other words, if a watermarking method can identify insensitive regions for watermark embedding, then the lossy compression algorithms can efficiently remove the embedded watermarks. For example, the watermarks with energy focused on high frequency regions in echo-based methods are likely to

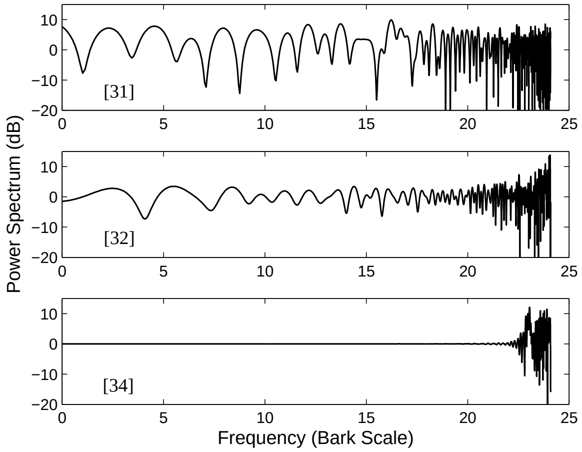

Figure 7: Examples of power spectra of echo kernels designed in [31, 32], and [34], respectively. The imperceptibility is improved from [31] to [32] by reducing the modifications in low to middle frequency bands, while the design in [34] ensures no artifacts within [0, 21] Bark scale, approx. [0, 10] kHz.

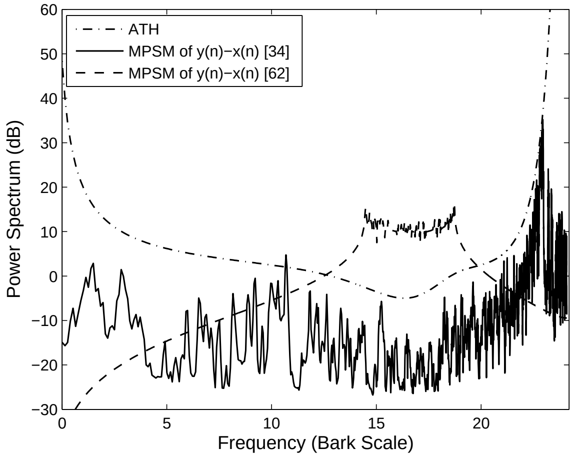

Figure 8: Evaluation of the imperceptibility property using ATH, where the echo-based design [34] and the patchwork design [62] are implemented for example. Due to a systematic control, the former has much better imperceptibility than the later.

be removed by the common lowpass filtering processing with cut-off frequency at e.g., 16 kHz. Therefore, we can observe from [30–32] that the watermark detection rates drop drastically under lossy compression attacks at $64-96$ kbps, and satisfactory results can only be obtained when the compression bit rate increases to 128 kbps and above. Note that this trade-off also exits in heuristic imperceptibility control systems, because the tuned embedding strength will finally determine that the watermarks are embedded with less perceivable magnitudes.

Based on the above analysis, we make the following conclusions in this section. i) As long as lossy compressions are the predominant tool to process audio data, imperceptibility is generally

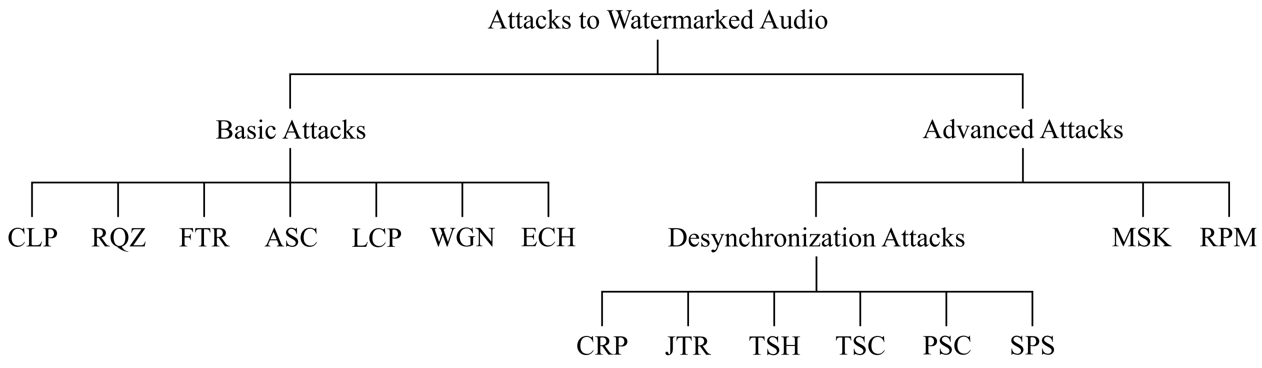

Figure 9: Categorization of existing attacks to watermarked audio.

hard to achieve. Hence, by embedding watermarks into an audio file, the sound quality is inevitably compromised. ii) Instead of systematically implement psychoacoustic model, heuristic tuning based on [77], [78], or [81], may be a better means to control the imperceptibility. Although heuristic tuning is more computationally intensive, it is achieved via direct perception measurements. iii) If we stick to systematic control, then some tolerance parameters should be introduced to quantify the power spectral leakage of the watermarks against the masking curve or ATH. In addition, there may be alternative means to effectively measure and control the imperceptibility property and also facilitate watermark embedding region selection. For example, since an audio signal is a function of time, we can efficiently make use of this property and calculate the SNR (or psychoacoustic modified ratios) evolution versus time. In this way, we may create some post-processing mechanisms that fine tunes the embedded watermarks as a function of time so that the imperceptibility can be further enhanced. More generally, discovering audio watermarking regions in time-frequency domain is a relevant problem.

4. Robustness – Facing Attacks

The robustness of an audio watermarking system is usually the first priority during system designs. While, generally, there exists a trade-off between imperceptibility and robustness, sometime, to ensure the validity of the system, imperceptibility can be compromised to achieve enhanced robustness, especially when dealing with advanced attacks. For example, to effectively deal with desynchronization attacks, the ODG values are set to around -1 in [40] and [66], indicating that the watermarks are actually perceptible. In this section, we provide a comprehensive review of the robustness of existing audio watermarking systems, with the consideration of all important attacks. Due to the sophistication and variety of existing attacks, it is much more complicated to

Table 1: Summary and abbreviations of existing attacks

Attacks	Abbrev.
Closed-Loop	CLP
Requantization	RQZ
Filtering	FTR
Amplitude Scaling	ASC
Lossy Compression (MP3/AAC)	LCP
Adding white Gaussian noise	WGN
Adding Echoes	ECH
Cropping	CRP
Jittering	JTR
Time Shifting	TSH
Pitch-Invariant Time Scaling	TSC
Time-Invariant Pitch Scaling	PSC
Speed Scaling	SPS
Mask Attack	MSK
Replacement	RPM

deal with system robustness than imperceptibility. Nonetheless, certain imperceptibility conditions still need to be imposed during the optimization of robustness.

4.1. Categorization of Attacks

There are several ways to categorize the attacks towards audio watermarking systems. Intuitively, the watermarked audio may be attacked intentionally or unintentionally, which is determined by the motivations of the corresponding users. For example, lossy compression is likely to be an unintentional attack while adding noise could be considered as an intentional attack. An important contribution to the evaluations of an audio watermarking system against attacks has been made in [85], in which the Stirmark benchmark system is proposed. In this work, the authors differentiate the attacks into musical effects and algorithmic attacks. In addition, potential attacks during audio transmission and playback of audio stored in CDs are also discussed. The Stirmark benchmark system has been commonly used to evaluate the robustness of audio watermarking systems, which can be seen in [17, 20, 37, 40–42], and [56]. However, generally, the ways to evaluate robustness are not as standardized as the ways to evaluate imperceptibility as discussed in Section 3. As a result, the attacks considered in individual robust system designs are usually limited, which raises the risk of watermarks being destroyed by experienced adversaries. Therefore, a robust copyright watermarking system is preferred to be evaluated against as many attacks as possible to ensure its robustness and applicability to different situations.

In the scope of this paper, the attacks to be discussed are summarized in Table 1 and categorized in Fig. 9. The selection of attacks shown in Table 1 is based on the following considerations. i) We emphasize more on the impact of signal processing to the audio content, which is more likely to take place at the attackers' side. ii) The considered attacks are a combination of musical effects and algorithmic attacks, if judged from the perspective of [85]. iii) The considered attacks are generally a union set of all the attacks considered in [7, 15–70]. iv) The adversaries are also constrained in such a way that the attacks should not alter the original content to be delivered in the audio data.

In Fig. 9, we categorize the selected attacks into basic and advanced attacks. The basic attacks are generally easier to deal with than advanced attacks. The CLP attack is a closed-loop environment where the watermarked signal is in fact not attacked. This attack is considered because it is related to host interference non-rejecting methods [14] in which during watermark extraction, the host signal plays as the interference. In contrast, host interference rejecting methods are interference free under the CLP attack since the knowledge of the host signal is used during watermark extraction. Generally, Echo-based, SS, and patchwork methods are host interference non-rejecting, while QIM method is host interference rejecting. The RQZ attack refers to applying a different quantization level to the watermarked signal, e.g., from 16-bit to 8- or 24-bit. The FTR attacks include but not limit to lowpass, bandpass, highpass filtering attacks. However, the FTR attack should not remove the original information to be delivered in the watermarked signal. Note that the musical effects introduced by an equalizer can also be considered as FTR. The ASC attack scales the amplitude to a different level. The LCP attacks can be MP3 or AAC compressions with different but reasonable bit rates (e.g., bit rates less than 128 kbps may be excluded since today's digital technology can easily support higher fidelity). The WGN attack considers reasonable levels of injected noise, and usually, we constrain the resultant SNR above 20 dB. The ECH attack embeds delayed and attenuated copies into the watermarked signal.

The above basic attacks have been considered in most of the existing works, while few efforts have been devoted into a comprehensive study of advanced attacks. Although many works are dedicated to coping with desynchronization attacks, the definitions of desynchronization attacks have been very different among the existing works, which can be seen in [14, 20, 40–42], and [66]. A rigorous definition of decynchronization attacks should characterize all the attacks that introduce misalignments of the watermark positions between attacked and original copies, or, at least, it should include the ones summarized in Fig. 9. The CRP attack refers to random

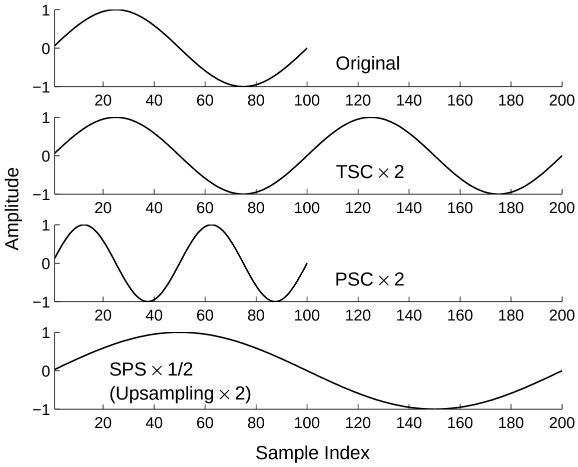

Figure 10: Examples of TSC, PSC, and SPS attacks on a sine wave.

cropping. However, it should be noted that this attack should not remove original information in the host signal. The JTR attack removes samples in each of the non-overlapping frames with a fixed duration. The TSH attack applies a time shift to the audio data. The TSC attack scales the time of the signal while preserving the pitch. It is also known as time-scale modification (TSM) attack. The PSC attack, on the other hand, scales the pitch while preserving the time of the signal. At last, The SPS attack modifies time and pitch at the same time, which is also called resampling attack. Examples of TSC, PSC, and SPS attacks are provided in Fig. 10, where the four signals are played with the same sampling frequency. Although all these desynchronization attacks can be easily implemented using pervasive audio players (e.g., Foobar2000, Audacity, and Adobe Audition, etc.), they have been less investigated in the existing works. It is therefore implied that the adversaries can easily use a combination of these attacks to counter the watermark systems.

The MSK attack is performed by taking advantages of masking curves used to control the imperceptibility during the watermark embedding phase. The rationale is as follows: if the psychoacoustic model is efficiently implemented, then the masking curve of the watermarked signal would be a good approximation of the masking curve of the original host signal. In this way, the watermarks can be efficiently estimated below the masking curve of the watermarked signal, and hence subtracted from the watermarked signal with a parameter to trade off attack distortion and removal strength. In fact, this attack shares a similar idea with lossy compressions, because both utilize the psychoacoustic model and remove signals in perceptually insensitive regions. The original idea of MSK attack proposed in [70] intensively rely on the SS signal model and statistical estimations. Nonetheless, the same idea could be applied to other embedding models. Next, we

Table 2: Comprehensive Evaluations of Robustness (Superscript * indicates single frame watermark embedding)

		CLP	RQZ	FTR	ASC	LCP	WGN	ECH	CRP	JTR	TSH	TSC	PSC	SPS	MSK	RPM	ODG
Aligned	[16]*	−	+	+	+	+	+	+	+	+	+	−	−	−	+	−	$[-0.1, 0]$
	[17]	−	+	+	+	+	+	+	+	+	+	−	−	−	+	−	$[-1.1, -0.9]$
	[19]	+	+	+	+	+	−	−	+	+	+	−	−	−	+	+	-0.17
	[20]	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	$[-1.3, -0.7]$
Echo	[32]	−	+	+	+	+	+	+	−	−	+	−	−	−	+	−	Moderate
	[33]	−	+	−	+	+	+	+	−	−	+	−	−	−	+	−	Better
	[34]	−	+	−	+	+	+	+	−	−	+	−	−	−	+	−	Best
SS	[38]	+	+	+	+	+	+	+	−	−	+	−	−	−	+	−	N.A.
	[40]*	−	+	+	+	+	+	+	+	+	+	+	+	+	+	−	$[-1, -0.4]$
	[41]	+	+	−	+	+	+	−	−	+	+	+	−	−	+	+	-0.1
	[42]	−	+	+	+	+	+	+	+	+	+	+	+	+	−	−	-0.1
	[43]	−	+	+	+	+	+	−	−	−	+	+	−	−	−	−	$[-2, -0.4]$
QIM	[54]	+	+	+	−	+	+	+	+	−	+	−	−	−	+	−	N.A.
	[56]	+	+	+	−	+	+	+	+	+	+	+	−	−	+	−	$[-0.2, -0.1]$
	[58]	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	< -1
Patch	[62]	−	+	−	+	+	+	−	−	−	−	−	−	−	+	−	N.A.
	[64]	−	+	+	+	+	+	+	−	−	−	−	−	−	+	−	$[-1.7, -0.6]$
	[66]*	−	+	+	+	+	+	+	+	+	+	+	+	+	+	−	-1

discuss a more challenging attack, i.e., the RPM attack.

The rationale behind the RPM attack, i.e., multimedia content is highly repetitive, is indeed simple but highly effective. For audio formats, voiced segments that correspond to the same word or vowel from a single speaker, or the musical segments in which the same instrument plays the same tone, etc., can all be very similar. Therefore, replacing audio segments with the corresponding perceptual similar counterparts within the same audio signal would result in marginal attack distortion but totally disordered watermarks. Note that the RPM attack can be applied to any watermarking methods, since the only requirement to perform this attack is the watermarked signal. Using the squared Euclidean distance as the similarity metric, steps to perform RPM attack are summarized as

1. Partition $y(n)$ into overlapping blocks[3].
2. For each block, find a set of most similar blocks.
3. Linearly combine these blocks to optimize similarity.
4. Replace the resultant block with the original one.

The above procedures reflect the main idea of [68] and [69]. Note that the effectiveness of RPM attack is also in terms of the availability of modifications on the attack steps. For example, the least-squares problem associated with Step 3) can be replaced by random permutation of the discovered similar blocks [67], which is called blind pattern matching attack. It has been verified in [67–69] that the RPM attack is highly effective against SS and QIM methods.

In the existing literature, although the robustness has been considered more intensively than other criteria (e.g., imperceptibility, capacity, etc.) for an audio watermarking system, it is indeed a very challenging task to consider as many attacks as possible in a single system design. However, this does not mean that the adversaries will not make the best efforts to implement multiple and advanced attacks. In view of this, the attacking scenarios considered in existing works seem to be somewhat optimistic, especially in those applications that require the watermarks to be robust. From the perspective of an adversary, am intuitive setup of attacks could be more than enough to counter the watermarks. For example, in accordance with the arguments about the good regions to embed watermarks as described in Section 3.2, a simple combination of lowpass FTR (say

[3] In [67–69], the recommended representation of the watermarked signal is in transform domain, i.e., real MCLT coefficients in logarithm scale, and only the $2-7$ kHz sub-band is considered. However, time domain attacks may also be effective if the Euclidean distance is the similarity metric.

cutoff frequency at 10 kHz) and the MSK attack will probably destroy the watermarks, where the former wipes out watermarks in high frequency bands, and the latter removes watermarks in masked low to middle frequency bands. For watermarks embedded in perceptually significant regions without considering any psychoacoustic model, the RPM attack may become more effective as a countermeasure - instead of calculating the replacement blocks for each block, one only needs to replace the blocks containing more energy. Therefore, the key research challenge for advanced audio watermarking solutions should lie in the improvement of system robustness against advanced attacks. Next, we comprehensively evaluate the robustness of the existing solutions against all the attacks summarized in Fig. 9.

Remark 2: Among the processing stages in audio watermarking systems as depicted in Fig. 1, audio quality degradation actually comes from two phases, i.e., watermark embedding and attacks, respectively. The degradation of audio quality during the watermark embedding phase is closely related to imperceptibility property and has been intensively discussed in Section 3. Here, it is important to note that both basic and advanced attacks will also cause audio quality loss, and the adversaries are facing a trade-off between successful watermark removal and audio quality degradation. Therefore, a good watermark embedding scheme would be characterized as a scheme such that the embedded watermarks cause minimum perceptible artifacts while an adversary is not able to remove the watermarks without severe degradations. However, in some extreme cases, the adversary would still consider an attack to be successful even the audio quality is severely compromised, e.g., adding 0 dB noise. Among the attacks considered in this paper, the basic attacks are more likely to cause perceptible quality loss if the strengths of the attacks are out of control. But if the strengths are carefully bounded, then the attacks may not be successful. In contrast, advanced attacks are more likely to be successful without noticeable quality degradation. For example, the desynchronization attacks, especially the TSC, PSC, and SPS, can successfully remove the watermarks even if the scaling ratio is less than 5%. This is highly because of these attacks being highly nonlinear operations. the watermarked signal after the TSC or PSC attack could be considered as being processed through a nonlinear system. It is hence very hard to recover a signal gone through a nonlinear process with linear operations. Lastly, the MSK and RPM attacks could be more effective since quality degradation is considered during these attacks, in which watermarks may be removed without any noticeable quality loss.

4.2. Comprehensive Evaluations of Robustness

The main objective of this subsection is to provide the readers with comprehensive evaluation results on the robustness properties of all important existing watermarking systems in Fig. 2, against all important attacks in Fig. 9. The results serve as building blocks that could be potentially considered to synergize efficient strategies against the variety of attacks. To characterize the robustness of the each considered solution, we use "+" to represent positive results, i.e., the specific method is robust against the corresponding attack, while the sign "–" indicates the opposite. The evaluations are carried out in a qualitative way, in which the robustness is determined according to the results reported in the existing works as well as the analysis of the corresponding watermark signals and embedding/extraction schemes. Note that a quantitative evaluation setup is possible, in which the testing results can be obtained from implementations of all considered solutions. However, such a complicated setup may not guarantee more accurate results. The reasons are i) audio signals are nonstationary, and the audio samples used in existing works have been very different. It is then highly possible that the same method yields different results for different testing databases. ii) The imperceptibility properties among different designs (correlated with robustness) are also quite different, indicating that it is difficult to realize a fair comparison of robustness. Some designs could be very robust against various attacks, but they are achieved via largely compromising the imperceptibility. iii) Experimental results also vary when random signals are realized from time to time, e.g., random keys, PN sequences, and noise. Therefore, in current situations, it suffices to provide qualitative results while noting that a comprehensive analysis has not yet been seen in the literature. In the qualitative evaluations, we aim at identifying whether a specific method has the capability to resist certain types of attacks.

The evaluation results are provided in Table 2, where for each category of audio watermarking methods, several representative solutions are chosen for comparison. The first column only indicates whether the method is host interference rejecting (+) or non-rejecting (–). Usually, in a closed-loop environment, the watermark detection accuracy approaches 100% for host interference non-rejecting methods, while for host interference rejecting ones, the accuracy is guaranteed to be 100%. The results in the second column indicate that, requantization, although commonly considered in the existing works, is in fact a trivial attack. The corresponding imperceptibility situations of the considered works in terms of ODG (according to what have been reported in these works) are shown in the last column of Table 2.

Furthermore, several important implications from the results in Table 2 are noted here. i) All the solutions in Table 2 are robust against LCP attacks. This is closely related to the control of imperceptibility. If the artifacts of watermarks are audible, then the watermarks cannot be removed by lossy compression. Note that even for the systematical design in [34], there exists a small spectral region in which the watermark energy leaks out from the ATH. ii) It is very difficult to achieve the robustness against all the considered attacks, and it would become more difficult to simultaneously maintain a satisfactory imperceptibility level. Among the robustness analysis of the existing works (not limited to the ones cited in this paper), the ones robust against the a majority of attacks are only [20] in time domain, and [40, 42, 58], and [66] in transform domain. However, no active and analytical control of imperceptibility has been seen in these works. The ODG scores in these works are all near or worse -1, except in [42], indicating the artifacts are perceptible. Note that the imperceptibility property of the system proposed in [58], as mentioned by the authors, has been largely compromised. iii) In echo based methods, ODG is not used to measure the imperceptibility, but the power spectra of the echo kernels can well reflect the imperceptibility situations. In [34], embedding distortions are almost imperceptible, indicating that the ODG scores approach zero value. Although generally, echo-based methods have better imperceptibility properties than other methods, a common drawback of these methods is the vulnerability against desynchronization attacks. It is hence worth carrying out further research to tackle this issue for this category of solutions. iv) Among the five works mentioned in ii), [40] and [66] are based on single frame watermark embedding schemes. The underlying rationale is somewhat straightforward: after desynchronization attacks, a single frame still remains a single frame, hence it naturally becomes much easier to restore synchronization than the methods with multiple frames. However, on the one hand, the imperceptibility would be problematic due to the ultra-high frequency resolution from long-sample transforms; on the other hand, a heavy computational load is introduced (e.g., an audio of 5 minutes and 44100 Hz sample rate requires an 2^{24}-point fast Fourier transform (FFT)), which may not be suitable for applications with efficiency requirements (secured data transmission, broadcasting, etc.). v) Because the MSK attack [70] is derived based on SS signal model, it would be more effective in countering SS-based systems with the consideration of a psychoacoustic model, i.e., [42] and [43]. vi) It is implied, from Table 2, that among the desynchronization attacks, the CRP, JTR, and TSH attacks are easier to deal with than the TSC, PSC and SPS. Furthermore, the TSC attack (equivalent to the TSM), has

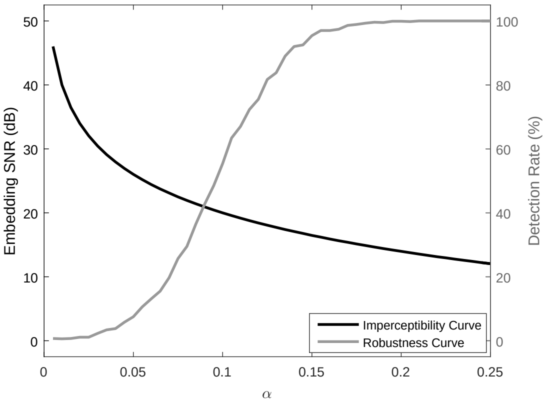

Figure 11: An illustration of the trade-off between imperceptibility and robustness, where a generic SS system in (2) is used as an example.

attracted more research attention than the PSC and SPS attacks, although the latter can also be easily implemented. vii) Lastly, the RPM attack is the most challenging attack to counter. It is observed that only three solutions, i.e., [19, 20], and [41], can potentially be robust against this attack. These solutions share the same idea of embedding watermarks according to the comparison of energies among consecutive frames (time frames in [19], consecutive bins according to histogram in [20], and FFT magnitudes in [41]). Since the replaced blocks are similar in terms of Euclidean distance, the resultant energy difference before and after attack would also be minimal.

4.3. Imperceptibility Robustness Trade-Off

Before we proceed to the discussions of research potentials to improve system robustness, the trade-off between imperceptibility and robustness is addressed in this subsection. Conceptually, if one wants to make a signal (e.g., the watermarks) imperceptible, then he or she would intuitively reduce the energy of this signal (decreasing α). On the other hand, the signal's general robustness property is closely related to its energy, because a stronger signal (increasing α) is less likely to be destroyed by noise or other manipulations. Therefore, the trade-off between imperceptibility and robustness can be described as the problem that the energy of the watermarks should be neither too big nor too small, which calls for an appropriate balance. This trade-off is visualized in Fig. 11, where a simplest SS system based on the model in (2) is used as an example. In this example, a single frame synthetic host signal of 1024 samples is generated as a white noise with normal distribution, and $m(k)$ is a binary PN sequence with the same length. Watermark detection is performed by correlating $m(k)$ with $y(k)$, and examining whether the peak is located at the 1024th sample in the correlation function. Then the imperceptibility (black) and robustness (gray) curves are obtained by averaging 1024 independent realizations of the host signal and PN sequence. The trade-off relationship between imperceptibility and robustness is clearly shown in this figure, where the two curves are inversely related to each other. Note that the above example is corresponding to host signal interference non-rejecting methods under the CLP attack. For host signal interference rejecting methods, the benchmark robustness needs to be tested against the WGN attacks instead of the CLP attack, since under the CLP attack, the detection rate will constantly stay at 100%. The above example and analysis indicate the importance of host interference rejecting methods for audio watermarking, which have to be achieved via informed embedding mechanisms.

4.4. Discussions

The results shown in Table 2 reveal the challenges in the designs of audio watermarking systems against intelligent adversaries. Nonetheless, Table 2 has also provided us with building blocks to synergize better solutions in terms of both imperceptibility and robustness. We briefly discuss the following problems and potentials that have not been reported in the literature.

4.4.1. Time Domain or Transform Domain?

Despite the variety of audio watermarking system designs, a fundamental and important question has not yet been answered, that is, should we performed a transform before embedding the watermarks? In other words, what is the basic differences between time domain methods (1) and transform domain methods (2)? Intuitively, one would consider frequency domain methods a better choices to achieve improved imperceptibility because of the availability of psychoacoustic models in frequency domain. However, a time domain system with a carefully designed heuristic tuning mechanism could also achieve arbitrarily high levels of imperceptibility. For robustness, security, and capacity, even less clues have been revealed to indicate which domain is a better choice. Furthermore, among transform domain methods, many transforms have been proposed, but the features or advantages of a specific transform over other transforms are missing in the literature. To reveal the answer to these questions, a systematic and fair comparison setup using appropriate imperceptibility and robustness evaluation tools need to be designed.

4.4.2. Random Sequence vs. Pattern Watermarks

Although the majority of the existing watermarks are modulated or non-modulated binary random sequences, there exist several successful designs where the watermarks are two dimensional patterns, e.g., [55, 56], [58], and [59]. Note that the detection schemes for systems with random watermarks first extract the watermark bits which are then used to calculate the BER, or generate cross-correlation results to compare with a predefined threshold, for final decision. However, for systems using two dimensional patterns as watermarks, the extraction phase restores the pattern after extracting each pixel (binary quantity), and then manually determines the existence of the watermarks. Although systems with random watermarks perform automated detection, it is relatively more difficult to control false positive or false negative decisions. For example, a BER of 20% may not be strong enough to prove the existence of watermarks for such systems. However, for the case of using two dimensional patterns, the decision process tends to become more accurate and convincing, since the human visual system (HVS) is very robust in recognizing noisy images. For example, the watermark pattern in [56] can still be identified when the BER is larger than 20%. Therefore, it is worth of investigations on the effectiveness of using image patterns as watermarks in various categories of audios watermarking systems.

4.4.3. Synchronization Issue

Desynchronization attacks are the most difficult attacks in audio watermarking systems. This is intrinsically because the TSC, PSC, and SPS attacks apply a nonlinear processing system to the host signal which could not be sufficiently restored by linear operations. Also note that inserting synchronization bits [19, 46, 56–59, 61] (e.g., a series of symbol “1”) is in fact very vulnerable to desyncrhonization attacks. Therefore, effective means to deal with such attacks should be thoroughly re-considered starting from the embedding process with the exploration of desynchronization-invariant features. This can be reflected by evaluating the systems proposed in [20] and [42]. The essential contributions of the two works lie in the proposals of using the histogram and RASE, respectively. The watermark embedding regions associated with these features can then be used for many possible embedding schemes. For example, the system proposed in [42] is highly effective in dealing with desynchronization attacks, although only the simplest realization of SS method is used therein. Combining histogram or RASE with other embedding and extraction schemes would hence be interesting for further research attentions. Generally, future system designers should continue to discover desynchronization-invariant features.

4.4.4. Framing

Although single frame processing has been proposed in several works to enhance the robustness against desynchronization attacks, e.g., [16, 40, 66], it may not be recommended for practitioners due to the considerations of imperceptibility and computation issues. However, there may exist alternative ways to frame the host signal during watermark embedding process, which not only preserve the efficiency of frame based processing, but also gain robustness against desynchronization attacks. For example, one may set constant frame numbers instead of constant frame length that has been used in most existing works. The immediate advantage of using a constant frame number is in terms of the special invariance property against JTR, TSC, PSC, and SPS attacks. Furthermore, if constant frame number is combined with localized watermark embedding, then the resultant system could also be robust against CRP, TSH, and RPM attacks, making a system simultaneously robust against RPM and all desynchronization attacks. It is hence worth comprehensively evaluating the effectiveness of constant frame number for all existing frame based audio watermarking systems.

4.4.5. Enhanced Echo-Based Methods

Being vulnerable to desynchronization attack is the major drawback of echo-based audio watermarking systems. However, based on the above analysis, it is possible to endow echo-based methods with the robustness against desynchronization attacks. For example, combining echo-based systems with the concepts of constant frame number and a localized embedding scheme would achieve this goal. A constant frame number will ensure that each frame after desynchronization attack is aligned with the original frame but with a different frame length, while localized embedding can identify those frames with stronger capability to deal with desynchronization attacks. In addition, time domain desynchronization-invariant features which are compatible with echo kernel and linear filtering process are worth of investigations.

4.4.6. Preserving Imperceptibility

Generally, imperceptibility properties are very likely to be compromised when robustness issues are intensively considered. If a global solution to robustness is approachable, then the immediate problem to be tackled is to optimize the imperceptibility properties, given the obtained robustness. By noting from Section 3 that systematically control imperceptibility via a psychoacoustic model can be largely destroyed by pervasive lossy compressions, heuristic methods may be a better choice as a remedy. In this way, the ODG based heuristic tuning, or generally the scheme of Fig. 6 (a), would be the optimal under practical considerations, because it enables the system being both automated and effective.

4.4.7. Time-Frequency Domain Approach

It has been indicated from Section 3 that psychoacoustic modeling seems to be the only effective means to systematically control the imperceptibility property of the system, but watermarks tuned by a psychoacoustic model are likely to be removed by lossy compression. However, if we make use of the fact that audio signals are a function of time, then a systematic study of the host signal in time-frequency domain may become effective. Specifically, one may use time-frequency domain measurements to capture the uniqueness of audio data. Therefore, a new problem can be identified as how we can appropriately design watermark embedding regions and mechanisms in time-frequency domain, to minimize the embedding perceptual distortion while preserving good robustness properties.

The above discussions have provided a few concepts and general examples to reveal possible strategies to deal with current challenges in developing practically robust audio watermarking systems. The great potentials for further improvements have been revealed, which call for future research attentions. Next, we briefly discuss audio watermarking from industry point of view.

5. Audio Watermarking Applications

Digital audio watermarking technologies have been applied to industry applications since the early development stage in late 1990's. This section consists of two parts. In the first part, we review ten US patents [86–95], and generally describe the techniques used therein. It can be seen that in industry, the solutions are more application-oriented, and the incorporated techniques are usually easy-to-implement and cost efficient ones. The second part describes the existing commercial applications of audio watermarking [96–101].

5.1. Patent Review

The selected US patents are owned by Microsoft Corporation [86–88], Digimarc Corporation [89, 90], Kent Ridge Digital Labs (Singapore) [91], Alcatel-Lucent USA Inc. [92], NEC (China) Co., Ltd. [93], The Nielsen Company [94], and Cisco Technology Inc. [95], respectively. The inventors of [86–88] are also the authors of [37], hence the techniques involved for these patents are originated from [37], but with more details on system implementations and parameter settings. In addition, the invention in [86] proposes a scheme of dual watermarks with a mechanism to select a strong or a weak watermark according to the psychoacoustic model. Such a scheme improves the system's imperceptibility property. A covert communication system is proposed in [88], where each watermark bit is corresponding to a covert message chip. A secret key is used to permute the watermark bits, and it is also used for watermark detection.

In [89] and [90], a time domain aligned watermarking method is proposed, which requires the original signal during watermark detection phase. Let the watermark bits be $w(i) \in \{-1, 1\}$, and of length N_w, then the embedding scheme first generates N_w random sequences with the same length as $x(n)$, denoted as $c_i(n)$, $i \in 0, 1, \cdots, N_w - 1$. These sequences are then used to create a composite watermerk signal and added to the host signal, and the embedding scheme is given by

$$y(n) = x(n) + \alpha \sum_{i=0}^{N_w-1} w(i)c_i(n),$$

where α is selected based on SNR measurements. The corresponding watermark extraction scheme (in a closed-loop environment) first subtracts $x(n)$ from $y(n)$, and then calculates the cross-correlations between the subtracted signal and each of the N_w random sequences. Typically, the watermark bits consist of 32 samples, including first four bits $\{-1, 1, -1, 1\}$ as calibration codes, 16 bits of watermark sequence, 8 bits of version indicator, as well as the last 4 error checking bits.

In [91], the essential idea of echo-based watermarking is utilized to develop a more sophisticated audio watermarking system, involving techniques such as hashing, Mel scale frequency analysis, feature extraction and clustering, random hopping, and dynamic time warping (DTW), among others. Specifically, the original watermark information is encrypted by the audio hash function, which produces a content dependent watermark sequence. Via Mel scale frequency analysis and clustering, audio frames are classified into four classes, and each class corresponds to a particular set of embedding parameters. According to the comparisons between the energy values of frequency bands below and above 1 kHz, each frame is embedded with one up to four random hopping echoes within its class. During watermark extraction phase, the host signal is required to determine the synchronization of the watermarked signal. As a result, the system becomes robust against most of the desynchronization attacks. Another implementation of echo-based method is presented in [92], in which the embedding scheme can be efficiently expressed as (4), without the backward echo potion. Noticeably, each echo is repeated for several times, and for each positive echo, there is a corresponding negative one. Since a random sequence is not used in this design, the echoes are prone to malicious detection and removal.

An interesting transform domain (more specifically, FFT domain) watermarking system is described in [93]. Noticeably, it does not belong to any of the subcategories of transform domain methods in Fig. 2. The basic idea is to modify randomly selected frequency bins in the low frequency region of host audio frames according to local masking curves. Let the selected frequency bins for a specific frame be κ, then the watermark embedding process is given by

$$Y(k) = X(k) + \alpha(k)m(k),$$

where

$$m(k) = \begin{cases} 1, & k \in \kappa, \\ 0, & k \notin \kappa. \end{cases}$$

Note that the above embedding scheme looks quite similar to the original form of SS method, it is intrinsically a different method, because the watermarks are actually not embedded in a spread manner. Instead, only the magnitudes of a few (e.g., 7 in [93]) frequency bins are modified while the rest are kept unchanged. In addition, κ is determined randomly according to a look-up table which can also be specified randomly. The look-up table used for watermark embedding is also needed during the watermark exaction phase. Since the psychoacoustic model used in [93] is identical to the one for MPEG AAC compression, the system can have very good imperceptibility properties, but it becomes highly vulnerable to lossy compressions as discussed in Section 3.2. Similar to [93], another transform domain (also in FFT domain) method is proposed in [94]. The watermarks are embedded by exploring the insensitivity of HAS to harmonic components. Therefore, a psychoacoustic model is not considered therein. Here, we introduce the simplest implementation of this method. Let the fundamental frequency (which is determined by the peak value of the FFT magnitudes of a frame) be k_{F}, then the watermarks are embedded in terms of modifying the value of $X(2k_{\mathrm{F}})$. Specifically, the condition for a valid watermark embedding frame is

$$X(k) < 0.25X(k_{\mathrm{F}}), \quad k \in [2k_{\mathrm{F}} - 12\%k_{\mathrm{F}}, 2k_{\mathrm{F}} + 12\%k_{\mathrm{F}}].$$

Then, the watermark embedding process is given by

$$\begin{cases} Y(k) = X(k), & k \neq k_{\mathrm{F}}, \forall m, \\ Y(2k_{\mathrm{F}}) = (1 - 0.75m)\, X\left[(2-m)\, k_{\mathrm{F}}\right], & m = \{0, 1\}, \end{cases}$$

which indicates that when embedding “0”, the first order harmonic is not changed, while when embedding “1”, the first order harmonic is scaled to 1/4 of the fundamental frequency.

In [95], an audio watermarking technology is incorporated in a telecommunication network, in which the originating half-call is watermarked in the Internet protocol (IP) gateway. Then, when the terminating half-call reaches the gateway, the watermarks are detected to identify that the two half-calls are associated ones. In this invention, a general audio watermarking technology is applied, indicating the availability of all the methods reviewed in this paper.

5.2. Commercial Products

In the specification for enhanced content protection from MovieLabs [96], the applications of forensic watermarking and playback control watermarking are discussed to mitigate piracy problems, where robustness against corruption is the key performance criterion [97]. Verance [98] has developed a series of solutions to embed robust watermarks within digital audio waveform to enable copy protection, and imperceptibility and robustness are the “key benefits” of the solutions. Similarly, Cinavia’s solutions [99] are designed to protect blue-ray and DVD movies whose audio tracks are watermarked by “inaudible codes” provided by the company. In [100] and [101], the “second screen” applications have been developed to enable enriched streaming broadcasting services, where the embedded audio watermarks in the “second screen” are used for content identification and synchronization.

The general differences between academic and industrial audio watermarking solutions are noted as follows. i) Industry solutions consider more on imperceptibility than robustness. The reason is that each industry solution defines a specific application for the audio watermarking systems, in which the attacks may not need to be exhausted. ii) Industry solutions emphasize more on efficient implementations and ad-hoc designs for specific purposes. All the watermark embedding methods reviewed in this section are of the simplest form of the methods described in Section 2, while the ones for [93] and [94] are even simpler than the original form of the SS method. An example of ad-hoc treatment can be seen in [93], where the watermark bits are designed in such a way that one portion is changed in every 6 seconds while another portion is changed in every 2 seconds.

6. Conclusion

A comprehensive review of the last twenty years of digital audio watermarking is provided in this paper. We first systematically categorized the existing audio watermarking works with detailed analysis of the embedding/extraction schemes and key features of each considered work using generic signal models. Then, we intensively investigated the most important performance criteria for audio watermarking systems, i.e., imperceptibility and robustness. In particular, all the existing treatments to ensure acceptable imperceptibility properties are summarized and analyzed. Meanwhile, the watermark embedding regions and the trade-off between psychoacoustic model based imperceptibility control and the robustness against lossy compression are discussed. To study the robustness properties, we considered a comprehensive list of attacks which serve as a union set of all important attacks that have been considered in the existing works. The attacks are further categorized into basic and advanced attacks in Fig. 9. Furthermore, a rigorous definition of desynchronization attacks is given in this paper, which consists of 6 different attacks. Comprehensive evaluations of robustness against all the considered attacks are provided in Table 2. Current challenges for audio watermarking system design are revealed by the results shown in Table 2. In addition, the remaining problems and possible strategies for further improvements of audio watermarking systems are discussed. Finally, we briefly described the situations of the audio watermarking technology in industry, with reviews of several US patents and commercial products. The differences between academic and industrial solutions are also discussed.

Generally, digital audio watermarking is an important branch of the topic of data hiding in multimedia. Although numerous solutions have been developed within last few decades, it is still relatively easier for the adversaries to counter the watermarking system than for designers to protect it. This is due to the variety and flexibility of the attacks. Therefore, it is worth investigating on global solutions to enhance the robustness properties. Meanwhile, alternative solutions that discover available embedding spaces other than using psychoacoustic models (e.g., harmonics [94]) are also a potential research focus that could lead to substantial performance improvement.

7. References

[1] L. Boney, A. Tewfik, K. Hamdy, Digital watermarks for audio signals, in: The Third IEEE International Conference on Multimedia Computing and Systems, 1996, pp. 473–480.

[2] A. Spanias, T. Painter, V. Atti, Audio Signal Processing and Coding, John Wiley & Sons, 2007.

[3] M. Swanson, M. Kobayashi, A. Tewfik, Multimedia data-embedding and watermarking technologies, Proc. IEEE 86 (6) (1998) 1064–1087.

[4] F. A. P. Petitcolas, R. Anderson, M. G. Kuhn, Information hiding-a survey, Proc. IEEE 87 (7) (1999) 1062–1078.

[5] F. Hartung, M. Kutter, Multimedia watermarking techniques, Proc. IEEE 87 (7) (1999) 1079–1107.

[6] M. H. M. Costa, Writing on dirty paper, IEEE Trans. Inf. Theory 29 (3) (1983) 439–441.

[7] I. J. Cox, J. Kilian, F. T. Leighton, T. Shamoon, Secure spread spectrum watermarking for multimedia, IEEE Trans. Image Process. 6 (12) (1997) 1673–1687.

[8] J. Eggers, J. Su, B. Girod, A blind watermarking scheme based on structured codebooks, in: Inst. Elec. Eng., Secure Images and Image Authentication, Vol. 4, 2000, pp. 1–6.

[9] P. Moulin, J. A. O'Sullivan, Information-theoretic analysis of information hiding, IEEE Trans. Inf. Theory 49 (3) (2003) 563–593.

[10] P. Moulin, A. Ivanovic, The zero-rate spread-spectrum watermarking game., IEEE Trans. Signal Process. 51 (4) (2003) 1098–1117.

[11] Q. Cheng, T. S. Huang, Robust optimum detection of transform domain multiplicative watermarks., IEEE Trans. Signal Process. 51 (4) (2003) 906–924.

[12] M. Barni, F. Bartolini, A. D. Rosa, A. Piva, Optimum decoding and detection of multiplicative watermarks, IEEE Trans. Signal Process. 51 (4) (2003) 1118–1123.

[13] S. Larbi, M. J. Saidane, Audio watermarking: A way to stationnarize audio signals, IEEE Trans. Signal Process. 53 (2) (2005) 816–823.

[14] A. Zaidi, R. Boyer, P. Duhamel, Audio watermarking under desynchronization and additive noise attacks, IEEE Trans. Signal Process. 54 (2) (2006) 570–584.

[15] P. Bassia, I. Pitas, N. Nikolaidis, Robust audio watermarking in the time domain, IEEE Trans. Multimedia 3 (2) (2001) 232–241.

[16] A. N. Lemma, J. Aprea, W. Oomen, L. V. D. Kerkhof, A temporal domain audio watermarking technique, IEEE Trans. Signal Process. 51 (4) (2003) 1088–1097.

[17] C. Baras, N. Moreau, P. Dymarski, Controlling the inaudibility and maximizing the robustness in an audio annotation watermarking system, IEEE Trans. Audio, Speech, Language Process. 14 (5) (2006) 1772–1782.

[18] Z. Liu, A. Inoue, Audio watermarking techniques using sinusoidal patterns based on pseudorandom sequences, IEEE Trans. Circuits Syst. Video Technol. 13 (8) (2003) 801–812.

[19] W. N. Lie, L. C. Chang, Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification, IEEE Trans. Multimedia 8 (1) (2006) 46–59.

[20] S. Xiang, J. Huang, Histogram-based audio watermarking against time-scale modification and cropping attacks., IEEE Trans. Multimedia 9 (7) (2007) 1357–1372.

[21] A. Nishimura, Audio watermarking based on subband amplitude modulation, Acoust. Sci. & Tech. 31 (5) (2010) 328–336.

[22] M. Unoki, D. Hamada, Method of digital-audio watermarking based on cochlear delay characteristics, International Journal of Innovative Computing, Information and Control 6 (3(B)) (2010) 1325–1346.

[23] M. Unoki, R. Miyauchi, Multimedia Information Hiding Technologies and Methodologies for Controlling Data, IGI Global, 2013, Ch. Method of Digital-Audio Watermarking Based on Cochlear Delay Characteristics, pp. 42–70.

[24] M. Unoki, K. Imabeppu, D. Hamada, A. Haniu, R. Miyauchi, Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics, Journal of Information Hiding and Multimedia Signal Processing 2 (1) (2011) 1–23.

[25] M. Unoki, R. Miyauchi, Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay, IEICE Trans. Inf. & Syst. E98-D (1) (2015) 38–48.

[26] R. Nishimura, Audio watermarking using spatial masking and ambisonics, IEEE Trans. Audio, Speech, Language Process. 20 (9) (2012) 2461–2469.

[27] D. Gruhl, W. Bender, Echo hiding, in: Proc. Information Hiding Workshop, Cambridge, U.K., 1996, pp. 295–315.

[28] H. O. Oh, J. W. Seok, J. W. Hong, D. H. Youn, New echo embedding technique for robust and imperceptible audio watermarking, in: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2001, pp. 1341–1344.

[29] H. J. Kim, Y. H. Choi, A novel echo-hiding scheme with backward and forward kernels, IEEE Trans. Circuits Syst. Video Technol. 13 (8) (2003) 885–889.

[30] Oscal, T. C. Chen, W. C. Wu, Highly robust, secure, and perceptual-quality echo hiding scheme, IEEE Trans. Audio, Speech, Language Process. 16 (3) (2008) 629–638.

[31] B. S. Ko, R. Nishimura, Y. Suzuki, Time-spread echo method for digital audio watermarking, IEEE Trans. Multimedia 7 (2) (2005) 212–221.

[32] Y. Xiang, D. Peng, I. Natgunanathan, W. Zhou, Effective pseudonoise sequence and decoding function for imperceptibility and robustness enhancement in time-spread echo-based audio watermarking, IEEE Trans. Multimedia 13 (1) (2011) 2–13.

[33] Y. Xiang, I. Natgunanathan, D. Peng, W. Zhou, S. Yu, A dual-channel time-spread echo method for audio watermarking, IEEE Trans. Inf. Forensics Security 7 (2) (2012) 383–392.

[34] G. Hua, J. Goh, V. L. L. Thing, Time-spread echo-based audio watermarking with optimized imperceptibility and robustness, IEEE/ACM Trans. Audio, Speech, Language Process. 23 (2) (2015) 227–239.

[35] P. Hu, D. Peng, Z. Yi, Y. Xiang, Robust time-spread echo watermarking using characteristics of host signals, Electronics Letters 52 (1) (2016) 5–6.

[36] W. Bender, D. Gruhl, N. Morimoto, A. Lu, Techniques for data hiding, IBM Syst. J. 35 (3.4) (1996) 313–336.

[37] D. Kirovski, H. S. Malvar, Spread-spectrum watermarking of audio signals, IEEE Trans. Signal Process. 51 (4) (2003) 1020–1033.

[38] H. Malvar, D. Florencio, Improved spread spectrum: A new modulation technique for robust watermarking, IEEE Trans. Signal Process. 51 (4) (2003) 898–905.

[39] A. Valizadeh, Z. J. Wang, An improved multiplicative spread spectrum embedding scheme for data hidingg, IEEE Trans. Inf. Forensics Security 7 (4) (2012) 1127–1143.

[40] X. Kang, R. Yang, J. Huang, Geometric invariant audio watermarking based on an lcm feature, IEEE Trans. Multimedia 13 (2) (2011) 181–190.

[41] W. Li, X. Xue, P. Lu, Localized audio watermarking technique robust against time-scalecale modification., IEEE Trans. Multimedia 8 (1) (2006) 60–69.

[42] C. M. Pun, X. C. Yuan, Robust segments detector for de-synchronization resilient audio watermarking, IEEE Trans. Audio, Speech, Language Process. 21 (11) (2013) 2412–2424.

[43] M. Arnold, X. Chen, P. Baum, U. Gries, G. Doërr, A phase-based audio watermarking system robust to acoustic path propagation, IEEE Trans. Inf. Forensics Security 9 (3) (2014) 411–425.

[44] N. M. Ngo, M. Unoki, Digital-Forensics and Watermarking: 13th International Workshop, IWDW 2014, Taipei, Taiwan, October 1-4, 2014. Revised Selected Papers, Springer International Publishing, 2015, Ch. Watermarking for Digital Audio Based on Adaptive Phase Modulation, pp. 105–119.

[45] M. N. Ngo, M. Unoki, Method of audio watermarking based on adaptive phase modulation, IEICE Trans. Inf. & Syst. E99-D (1) (2016) 92–101.

[46] D. Megas, J. Serra-Ruiz, M. Fallahpour, Efficient self-synchronised blind audio watermarking system based on time domain and fft amplitude modification, Signal Processing 90 (12) (2010) 3078 – 3092.

[47] Y. Xiang, I. Natgunanathan, Y. Rong, S. Guo, Spread spectrum-based high embedding capacity watermarking method for audio signals, IEEE/ACM Trans Audio, Speech, Language Process. 23 (12) (2015) 2228–2237.

[48] M. Fallahpour, D. Megas, Audio watermarking based on fibonacci numbers, IEEE/ACM Trans Audio, Speech, Language Process. 23 (8) (2015) 1273–1282.

[49] S. Wang, M. Unoki, Speech watermarking method based on formant tuning, IEICE Trans. Inf. & Syst. E98-D (1) (2015) 29–37.

[50] S. Wang, R. Miyauchi, M. Unoki, N. S. Kim, Tampering detection scheme for speech signals using formant enhancement based watermarking, Journal of Information Hiding and Multimedia Signal Processing 6 (6) (2015) 1264–1283.

[51] A. Nishimura, Digital-Forensics and Watermarking: 13th International Workshop, IWDW 2014, Taipei, Taiwan, October 1-4, 2014. Revised Selected Papers, Springer International Publishing, 2015, Ch. Reversible and Robust Audio Watermarking Based on Spread Spectrum and Amplitude Expansion, pp. 215–229.

[52] L. Qiao, K. Nahrstedt, Noninvertible watermarking methods for mpeg-encoded audio, Proc. SPIE 3657 (1999) 194–202.

[53] B. Chen, G. W. Wornell, Quantization index modulation: A class of provably good methods for digital watermarking and information embedding, IEEE Trans. Inf. Theory 47 (4) (2001) 1423–1443.

[54] S. Wu, J. Huang, D. Huang, Y. Q. Shi, Efficiently self-synchronized audio watermarking for assured audio data transmission, IEEE Trans. Broadcasting 51 (1) (2005) 69–76.

[55] K. Khaldi, A. O. Boudraa, Audio watermarking via EMD, IEEE Trans. Audio, Speech, Language Process. 21 (3) (2013) 675–680.

[56] B. Lei, I. Y. Soon, E. L. Tan, Robust svd-based audio watermarking scheme with differential evolution optimization, IEEE Trans. Audio, Speech, Language Process. 21 (11) (2013) 2368–2377.

[57] X. Y. Wang, H. Zhao, A novel synchronization invariant audio watermarking scheme based on DWT and DCT, IEEE Trans. Signal Process. 54 (12) (2006) 4835–4840.

[58] X. Y. Wang, P. P. Niu, H. Y. Yang, A robust, digital-audio watermarking method, IEEE Multimedia 16 (3) (2009) 60–69.

[59] X. Y. Wang, W. Qi, P. P. Niu, A new adaptive digital audio watermarking based on support vector regression, IEEE Trans. Audio, Speech, Language Process. 15 (8) (2007) 2270–2277.

[60] I. Shterev, R. Lagendijk, Amplitude scale estimation for quantization-based watermarking, IEEE Trans. Signal Process. 54 (11) (2006) 4146–4155.

[61] M. Arnold, Audio watermarking: Features, applications and algorithms, in: IEEE International Conference on Multimedia and Expo, 2000, (ICME 2000), Vol. 2, IEEE, 2000, pp. 1013–1016.

[62] I. K. Yeo, H. J. Kim, Modified patchwork algorithm: A novel audio watermarking scheme, IEEE Speech Audio Process. 11 (4) (2003) 381–386.

[63] H. Kang, K. Yamaguchi, B. M. Kurkoski, K. Yamaguchi, K. Kobayashi, Full-index-embedding patchwork algorithm for audio watermarking, IEICE Transactions E91-D (11) (2008) 2731–2734.

[64] N. K. Kalantari, M. A. Akhaee, S. M. Ahadi, H. Amindavar, Robust multiplicative patchwork method for audio watermarking, IEEE Trans. Audio, Speech, Language Process. 17 (6) (2009) 1133–1141.

[65] I. Natgunanathan, Y. Xiang, Y. Rong, W. Zhou, S. Guo, Robust patchwork-based embedding and decoding scheme for digital audio watermarking, IEEE Trans. Audio, Speech, Language Process. 20 (8) (2012) 2232–2239.

[66] Y. Xiang, I. Natgunanathan, S. Guo, W. Zhou, S. Nahavandi, Patchwork-based audio watermarking method robust to de-synchronization attacks, IEEE/ACM Trans. Audio, Speech, Language Process. 22 (9) (2014) 1413–1423.

[67] D. Kirovski, F. A. P. Petitcolas, Blind pattern matching attack on watermarking systems, IEEE Trans. Signal Process. 51 (4) (2003) 1045–1053.

[68] G. Doerr, J. Dugelay, D. Kirovski, On the need for signal-coherent watermarking, IEEE Trans. Multimedia 8 (5) (2006) 896–904.

[69] D. Kirovski, F. A. P. Petitcolas, Z. Landau, The replacement attack, IEEE Trans. Audio, Speech, Language Process. 15 (6) (2007) 1922–1931.

[70] A. Robert, J. Picard, On the use of masking models for image and audio watermarking, IEEE Trans. Multimedia 7 (4) (2005) 727–739.

[71] I. J. Cox, G. Doërr, T. Furon, Digital Watermarking: 5th International Workshop, IWDW 2006, Jeju Island, Korea, November 8-10, 2006. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, Ch. Watermarking Is Not Cryptography, pp. 1–15.

[72] P. Bas, T. Furon, A new measure of watermarking security: The effective key length, IEEE Trans. Inf. Forensics Security 8 (8) (2013) 1306–1317.

[73] Y. F. Huang, S. Tang, J. Yuan, Steganography in inactive frames of VoIP streams encoded by source codec, IEEE Trans. Inf. Forensics Security 6 (2) (2011) 296–306.

[74] D. Jang, C. Yoo, S. Lee, S. Kim, T. Kalker, Pairwise boosted audio fingerprint, Information Forensics and Security, IEEE Transactions on 4 (4) (2009) 995–1004.

[75] M. Chen, Y. He, R. L. Lagendijk, A fragile watermark error detection scheme for wireless video communications, IEEE Trans. Multimedia 7 (2) (2005) 201–211.

[76] F. Boland, J. Ruanaidh, C. Dautzenberg, Watermarking digital images for copyright protection, in: Fifth International Conference on Image Processing and its Applications, 1995, pp. 326–330.

[77] S. A. Gelfand, Hearing: An Introduction to Psychological and Physiological Acoustics, 3rd Edition, Marcel Dekker, Basel, Switzerland, 1998.

[78] Recommendation itu-r bs.1387-1: Method for objective measurements of perceived audio quality (1998-2001).

[79] D. G. Childers, D. Skinner, R. Kemerait, The cepstrum: A guide to processing, Proc. IEEE 65 (10) (1977) 1428–1443.

[80] A. Oppenheim, R. W. Schafer, From frequency to quefrency: a history of the cepstrum, IEEE Signal Process. Mag. 21 (5) (2004) 95–106.

[81] B. C. J. Moore, An Introduction to the Psychology of Hearing, 4th Edition, New York: Acedamic, 1997.

[82] G. Gao, Y. Q. Shi, Reversible data hiding using controlled contrast enhancement and integer wavelet transform, IEEE Signal Processing Letters 22 (11) (2015) 2078–2082.

[83] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Trans. Audio, Speech, Language Process. 16 (1) (2008) 229–238.

[84] A. Nishimura, M. Unoki, K. Kondo, A. Ogihara, Objective evaluation of sound quality for attacks on robust audio watermarking, Proceedings of Meetings on Acoustics 19 (1) (2013) 1–9.

[85] M. Steinebach, F. A. P. Petitcolas, F. Raynal, J. Dittmann, C. Fontaine, C. Seibel, N. Fats, L. C. Ferri, Stirmark benchmark: Audio watermarking attacks, in: Proc. IEEE Int. Conf. Inf. Technol.: Coding Comput., Las Vegas, NV, 2001, pp. 49–54.

[86] D. Kirovski, H. Malvar, M. H. Jakubowski, Audio watermarking with dual watermarks (Apr. 2007).

[87] D. Kirovski, H. Malvar, Stealthy audio watermarking (Sep. 2007).

[88] D. Kirovski, H. Malvar, Audio watermarking with covert channel and permutations (Jun. 2009).

[89] G. B. Rhoads, Methods for audio watermarking and decoding (Jan. 2006).

[90] G. B. Rhoads, Methods for audio watermarking and decoding (Nov. 2011).

[91] C. S. Xu, J. K. Wu, Q. B. Sun, K. Xin, H. Z. Li, Digital audio watermarking using content-adaptive, multiple echo hopping (Jan. 2004).

[92] J. H. Zhao, Y. C. Wei, M. Y. Hsueh, Media program identification method and apparatus based on audio watermarking (Jun. 2011).

[93] V. Srinivasan, A. Topchy, Methods and apparatus to perform audio watermarking and watermark detection and extraction (Jan 2013).

[94] Z. Geyzel, Audio watermarking (Jun. 2014).

[95] K. M. Patfield, Audio watermarking for call identification in a telecommunications network (Jun. 2010).

[96] Movielabs specification for enhanced content protection – version 1.1, [Accessed: Jul. 21, 2015].
URL http://www.movielabs.com/ngvideo

[97] Nexguard forensic watermarking 101, [Accessed: Jul. 21, 2015].
URL http://www.nexguard.com/forensic-watermarking-introduction/

[98] Music solutions, [Accessed: Jul. 21, 2015].
URL https://www.verance.com/products/music_embedders_dvd_audio_desktop.php

[99] Cinavia technology, [Accessed: Jul. 21, 2015].
URL http://www.cinavia.com/languages/english/pages/technology.html

[100] Syncnow^{TM} live sdi embedder, [Accessed: Jul. 21, 2015].
URL https://www.axon.tv/EN/second-screen-applications

[101] Second screen synchronization, [Accessed: Jul. 21, 2015].
URL http://www.intrasonics.com/whatwedo.php