SPRINGER BRIEFS IN ELECTRICAL AND COMPUTER ENGINEERING · SIGNAL PROCESSING

Yong Xiang
Guang Hua
Bin Yan

Digital Audio Watermarking

Fundamentals, Techniques and Challenges

Springer

SpringerBriefs in Electrical and Computer Engineering

Signal Processing

More information about this series at http://www.springer.com/series/11560

Yong Xiang · Guang Hua · Bin Yan

Digital Audio Watermarking

Fundamentals, Techniques and Challenges

Springer

Yong Xiang
School of Information Technology
Deakin University
Melbourne, VIC
Australia

Guang Hua
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan
China

Bin Yan
College of Electronics, Communication and Physics
Shandong University of Science and Technology
Qingdao, Shandong
China

ISSN 2191-8112 ISSN 2191-8120 (electronic)
SpringerBriefs in Electrical and Computer Engineering
ISSN 2196-4076 ISSN 2196-4084 (electronic)
SpringerBriefs in Signal Processing
ISBN 978-981-10-4288-1 ISBN 978-981-10-4289-8 (eBook)
DOI 10.1007/978-981-10-4289-8

Library of Congress Control Number: 2017934209

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

To my beloved Shan, Angie, and Daniel.

—Yong Xiang

To my beloved ones.

—Guang Hua

To my parents and my beloved family.

—Bin Yan

Preface

Digital watermarking has been an active research topic for over two decades. This book focuses on digital watermarking in the audio domain. It takes a general and comprehensive perspective and introduces audio watermarking techniques and system frameworks developed over the past two decades for various applications. It covers the fundamentals of audio watermarking, including performance criteria, basic system structure, as well as classical system designs, followed by the latest developments and state-of-the-art techniques in the literature. Furthermore, the emerging topics of reversible audio watermarking and audio watermarking with cryptography are also introduced in this book with illustrative design examples based on the state-of-the-art techniques. This book could serve as a tutorial material for readers with general background knowledge in signal processing, multimedia security, or a reference material for experienced researchers in watermarking area.

Melbourne, Australia — Yong Xiang
Wuhan, China — Guang Hua
Qingdao, China — Bin Yan

Acknowledgements

This work was partially supported by the Shandong Provincial Natural Science Foundation (No. ZR2014JL044) and the National Natural Science Foundation of China (NSFC) (No. 61272432).

Contents

Chapter 1
Introduction

Abstract Digital audio watermarking is the science and art of embedding a special type of data, such as a mark or a covert message, into digital audio content. The embedded information is then extracted at the receiving end for various purposes pertaining audio security. It has been an active research topic since the advent of digital era in mid 1990's Boney et al. (The Third IEEE international conference on multimedia computing and systems 473–480, 1996) [1], Hua et al. (128:222–242, 2016) [2]. In this chapter, we introduce the general background of digital audio watermarking, including its brief research and development history, applications, general system structure, as well as the performance criteria to evaluate an audio watermarking system.

1.1 Watermarking and Audio Watermarking

Data hiding, or information hiding, is a general term that describes the problem of embedding information in a cover content [3]. While the term "hiding" could refer to either making the embedded information imperceptible (watermarking) or covering the existence of the embedded information (steganography), this book is dedicated to the former scenario. While the information could be hidden in any form of a cover, e.g., computer programs, network protocols, multimedia content and so on, we mainly focus on data hiding in digital audio content, which defines the scope of this book, i.e., digital audio watermarking. Therefore, we also refer to the cover content as the host signal. Note that the watermarks could actually be either perceptible or imperceptible [4] especially for image watermarking, but in audio watermarking, the watermarks are generally imperceptible. Further, audio watermarks could be fragile or robust [4], where the former is usually associated with specific scenarios or with tamper-resisting purposes.

A brief clarification about the differences among similar terms are provided as follows. Data hiding and information hiding are considered to be interchangeable, and both terms represent the general concept of embedding a secret signal into a cover content. Watermarking, on the other hand, specifies the type of cover content as document, audio, image, video, etc., mainly multimedia content. It is also important to note that the watermarks to be embedded may or may not contain a message,

Y. Xiang et al., *Digital Audio Watermarking*, SpringerBriefs in Signal Processing, DOI 10.1007/978-981-10-4289-8_1

that is to say, the watermarks could simply be a mark, signature or logo, instead of a message. The designer of a watermarking system is concerned with how the watermarks could survive uncontrollable processing or attacks when the watermarked content is distributed. On the contrary, in the context of steganography, the designer is more concerned with concealing the existence of the hidden massage. More details about the similarities and differences could be found in [3, 4].

The very original concept of paper marks or paper watermarks emerged hundreds of years ago [3] but in this book, we will focus the modern concept of watermarking on digitized multimedia content. The original work on modern data hiding, i.e., dirty paper coding, was carried out from the theoretical perspective of communications, which appeared in 1980's [5]. Then, the first formally reported work on digital audio watermarking was seen in 1997 [1]. Over the last twenty years, a great amount of research and development work on digital audio watermarking has been conducted by researchers and practitioners with a major objective of securing digital audio content distributed among the end users in computer and communication networks [2]. Cryptography, although being the most important solution to ensure security and privacy of digital data, could not offer any help once the content has been decrypted. Therefore, as a complementary technique to cryptography, watermarking mainly protects multimedia data after decryption [6]. In this book, we will provide the readers with a comprehensive review of fundamental principles, latest techniques, and open problems of digital audio watermarking. Classical audio watermarking techniques, such as echo hiding, spread spectrum (SS), quantization index modulation (QIM), and patchwork, will be introduced. Further, the development history of each type of techniques will be reviewed, including the latest works. In addition, this book covers two special topics on audio watermarking, i.e., reversible audio watermarking and audio watermarking with cryptography.

1.2 Performance Criteria of Audio Watermarking Systems

Despite the performance criteria for an audio watermarking system depend on the specific application, common performance criteria could be summarized as follows.

- *Imperceptibility.* Except for the special type of perceptible image watermarking, imperceptibility is probably the most primary concern of a watermarking system. It characterizes how the watermarking system could manage the distortion of original cover content caused by watermark embedding. In audio watermarking, terms such as *fidelity, inaudibility, transparency* and *distortion*, refer to the similar concept. We will stick to the term imperceptibility in this book as it more precisely reflects the relationship between embedding distortion and human perception. For image and audio watermarking, imperceptibility is achieved via exploiting the characteristics of human visual system (HVS) and human auditory system (HAS) respectively.
- *Robustness.* The watermarked content, when released to end users, may undergo uncontrollable processing or even malicious attacks, before we attempt to extract

the watermarks from the content. Therefore, it is essentially important for the watermarks to be robust against unintentional or malicious attacks. Typical unintentional attacks include A/D and D/A conversion, equalization, re-encoding, etc., while malicious attacks could be nonlinear scaling, tampering, and so on. Besides, a certain attack could be caused either unintentionally or intentionaly. For example, desynchronization effects could be caused by either A/D and D/A conversion or malicious scaling. Also note that robustness is the most complicated performance criterion for a watermarking system.

- *Security.* In order not to confuse security with robustness, we confine the concept of watermarking security as the system's capability of preventing unauthorized party to access the watermarking channel, i.e., the system's capability of preventing unauthorized watermarking key detection and watermark extraction. This means that when security is concerned, the adversary's aim is assumed to be watermarking key detection and watermark extraction only, not including watermark removal. Unlike imperceptibility and robustness, the security of audio watermarking systems has been less studied, while the main component to ensure security is the pseudo-random noise (PN) which serves as the key.
- *Capacity.* The watermarks to be embedded in a given cover content is also called payload. The maximum size of payload, usually in terms of bits, is termed capacity. Since audio signal is a function of time and most audio watermarking schemes segment audio signal into frames before watermark embedding, the embedding capacity could also be characterized by embedding rate with the unit of bits per second or bits per frame.
- *Computational Complexity.* At last, an audio watermarking system is preferable to be more computationally efficient when it is designed for real-time applications such as broadcast monitoring and "second screen" systems.

Generally, there exists a trade-off among the above performance criteria, and improving one would probably result in compromising another (or a few others). A typical example is the most intensively studied trade-off between imperceptibility and robustness. Conceptually, if one would like the watermarks to be more imperceptible, then he or she tends to weaken the strength of the embedded watermarks, but this will result in the watermarks being more vulnerable to attacks, and vice versa. In the context of reversible audio watermarking, robustness is not considered, while the trade-off is established between imperceptibility and capacity. Specifically, if more watermark bits are embedded, then more distortion will be introduced to the cover signal, and vice versa.

1.3 Applications of Audio Watermarking

Watermarking systems were created with the original objective of copyright protection for multimedia content. Such a need has drawn more and more attentions since the era of digital technology in early 1990's. Since watermarking systems generally take effect only after copyright infringement having taken place, they

are rarely used for preventing it from happening. Several companies have developed their commercial audio watermarking products to suit the need of the market. Verance [7] has developed a series of robust audio watermarking solutions for copy protection purpose, where imperceptibility and robustness are key benefits of these solutions. Similarly, audio watermarking has been implemented to protect blue-ray and DVD movies by Cinavia [8]. In such a system, watermarks containing copy control information and identification bits are embedded in the DVD audio track repeatedly. During playback, if the detected watermarks do not match those of specific disc, then the playback will be halted. In addition, an emerging application that has recently drawn much attention from the community is the "second screen" application for enriched streaming services [9, 10]. In this application, while streaming subscribed multimedia content on a device such as a television, the subscriber could use another mobile device, e.g., a smartphone or a tablet, to receive enhanced viewing experience. In this way, immersive complementary content could be provided on the "second screen" of the mobile device. The communication between the second screen and the primary screen is established via audio watermarking. Specifically, audio watermarks are embedded into the audio channel of the multimedia content streamed on the primary device. Then, the second screen subscriber uses the alternative device to receive the audio signal from the primary device and extracts the imperceptible audio watermarks which trigger the service of enriched content on the second screen.

Other than commercial products, the development of audio watermarking technologies could also be reflected from existing patents. Many companies have patented their audio watermarking technologies in order to establish their solutions, including Microsoft Corporation [11–13], Digimarc Corporation [14, 15], Kent Ridge Digital Labs (Singapore) [16], Alcatel-Lucent USA Inc. [17], NEC (China) Co., Ltd. [18], The Nielsen Company [19], and Cisco Technology Inc. [20], etc. A brief patent review is provided in [2].

In some other situations, the users are specially concerned with the exact recovery of the host audio signal after watermark extraction, which means that it is desirable to not only obtain the watermark bits, but also restore the original audio content at the receiving end. Systems with such a property is called reversible audio watermarking. Reversible audio watermarking is applicable for the protect of sensitive data such as legal archives.

1.4 General System Structure

A watermarking system generally takes the structure depicted in Fig. 1.1. During watermark embedding, there exist several options to be considered. First, the host cover signal could be exploited to achieve some improved properties, which results in an informed watermark embedder. On the contrary, an embedder without considering the cover signal is called a non-informed embedder. For example, the basic SS method [21] has a non-informed embedder while the basic QIM method [22]

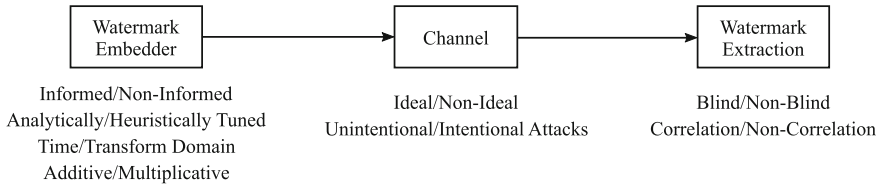


Fig. 1.1 Block diagram and associated attributes of a generic watermarking system

has an informed one. Second, imperceptibility control is carried out during watermark embedding. The system could either use the psychoacoustic model of HAS to analytically shape the watermarks or rely on heuristic tuning with a recursive "try-measure-tune" working flow. Third, we may pre-process the cover signal before watermark embedding, i.e., transform the signal into a "convenient" domain for better control of system performance. Alternatively, we may also carry out watermark embedding in the original time domain. Note that if a transform is applied, then after watermark embedding, the watermarked signal in transform domain will be inversely transformed back to time domain. At last, the designer could also decide between additive and multiplicative embedding, where the former simply adds the watermarks into the host signal while the latter does multiplication instead.

At the receiving end, if the copy of audio clip under test is identical to the watermarked copy, then we consider this as a closed-loop environment. In the context of reversible watermarking, a closed-loop environment is assumed. This is equivalent to assuming the channel is ideal, i.e., free of interference. However, in most situations and applications, the channel is non-ideal. For copyright protection, we have to deal with unintentional and intentional attacks to ensure the survival of watermarks. In the emerging second screen systems, the audio watermarks rendered from the loudspeakers of the primary device travel through an acoustic path to reach the device of the second screen. Therefore, the audio watermarking system designed for such an application needs to survive A/D conversion, acoustic propagation (multi-path effects), as well as possible environmental noise, which define the involved channel model. Usually, the embedding and extraction mechanisms will be designed with the consideration of specific channels.

While the watermark extraction mechanism may not necessarily be unique for a given embedding mechanism, it is closely related to the corresponding embedding mechanism, generally in a reverse manner. If the original cover signal is not required during watermark extraction, we call such a system as a blind system. Otherwise, it is a non-blind system. Further, it is also optional for the extraction mechanism to implement a correlator, depending on the corresponding embedding mechanism.

The remainder of the book is arranged as follows. In Chap. 2, we address the imperceptibility issues for audio watermarking systems by introducing the psychoacoustic models and imperceptibility control methods. After that, classical as well as recently developed audio watermark embedding and extraction techniques are introduced in Chap. 3. Two special topics about audio watermarking are discussed

in the chapters followed. In Chap. 4, techniques for reversible audio watermarking (RAW) are introduced, while in Chap. 5, some topics of combining cryptography and watermarking are addressed, including watermark embedding with encryption and replacing conventional normalized square transform matrix with over-complete dictionaries. Finally, concluding remarks and future works are provided in Chap. 6.

References

1. Boney L, Tewfik A, Hamdy K (1996) Digital watermarks for audio signals. In: The Third IEEE international conference on multimedia computing and systems, pp 473–480
2. Hua G, Huang J, Shi YQ, Goh J, Thing VLL (2016a) Twenty years of digital audio watermarking - a comprehensive review. Signal Process 128:222–242
3. Cox IJ, Miller ML, Bloom JA, Fridrich J, Kalker T (2008) Digital Watermarking and Steganography, 2nd edn. Morgan Kaufmann, Burlington
4. Swanson MD, Kobayashi M, Tewfik AH (1998a) Multimedia data-embedding and watermarking technologies. Proc IEEE 86(6):1064–1087
5. Costa MHM (1983) Writing on dirty paper. IEEE Trans Inf Theory 29(3):439–441
6. Cox IJ, Doërr G, Furon T (2006) Digital Watermarking: Watermarking Is Not Cryptography. In: Proceedings of 5th international workshop, IWDW 2006, Jeju Island, Korea, November 8–10, 2006. Springer, Heidelberg, pp 1–15
7. Music solutions. https://www.verance.com/products/music_embedders_dvd_audio_desktop.php. Accessed 21 July 2015
8. Cinavia technology. http://www.cinavia.com/languages/english/pages/technology.html. Accessed 21 July 2015
9. Syncnow^{TM} live sdi embedder. https://www.axon.tv/EN/second-screen-applications. Accessed 21 July 2015
10. Second screen synchronization. http://www.intrasonics.com/whatwedo.php. Accessed 21 July 2015
11. Kirovski D, Malvar H, Jakubowski MH (2007) Audio watermarking with dual watermarks
12. Kirovski D, Malvar H (2007) Stealthy audio watermarking
13. Kirovski D, Malvar H (2009) Audio watermarking with covert channel and permutations
14. Rhoads GB (2006) Methods for audio watermarking and decoding
15. Rhoads GB (2011) Methods for audio watermarking and decoding
16. Xu CS, Wu JK, Sun QB, Xin K, Li HZ (2004) Digital audio watermarking using content-adaptive, multiple echo hopping
17. Zhao JH, Wei YC, Hsueh MY (2011) Media program identification method and apparatus based on audio watermarking
18. Srinivasan V, Topchy A (2013) Methods and apparatus to perform audio watermarking and watermark detection and extraction
19. Geyzel Z (2014) Audio watermarking
20. Patfield KM (2010) Audio watermarking for call identification in a telecommunications network
21. Cox IJ, Kilian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Trans Image Process 6(12):1673–1687
22. Chen B, Wornell GW (2001) Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. IEEE Trans Inf Theory 47(4):1423–1443

Chapter 2
Human Auditory System and Perceptual Quality Measurement

Abstract This chapter introduces the human auditory system (HAS) and discusses the evaluation and control of imperceptibility. Since the HAS is very sensitive, the embedded watermarks may cause audible distortion if watermark embedding is not conducted properly. One central role in the perception of sound is the frequency-to-place transformation of the cochlear in inner ear. This transformation helps explain the masking effects, the perception of echoes and cochlear delay. The imperceptibility of audio watermarking is evaluated by subjective test, simple objective quality measures and objective perceptual quality measures. Finally, two imperceptibility control paradigms are reviewed, including heuristic control and analytic control. For heuristic control, the quality measures are utilized in a feedback framework to control imperceptibility. For analytic control, the watermark and/or the embedding strength are determined directly from the host audio.

2.1 HAS and Psychoacoustics

Instead of providing a comprehensive review of the HAS, we only discuss the key psychoacoustic functions of the HAS that are relevant to imperceptibility control in audio watermarking. A key to understanding many psychoacoustic facts is the frequency-to-place transformation in the cochlear. The absolute threshold of hearing and masking effects are utilized by many watermarking algorithms to find just noticeable distortion. Echo perception and cochlear delay are two psychoacoustic facts that are specific to audio perception.

2.1.1 Frequency-to-Place Transformation and Bark Scale

We briefly summarize the physiological basis of psychoacoustic models. The peripheral auditory system, i.e., the part of the HAS that is outside of the human brain, includes the outer ear, the middle ear and the inner ear. The *outer ear* consists of the pinna, ear canal, and ear drum. Its main function is to help focus the sound wave and

Y. Xiang et al., *Digital Audio Watermarking*, SpringerBriefs
in Signal Processing, DOI 10.1007/978-981-10-4289-8_2

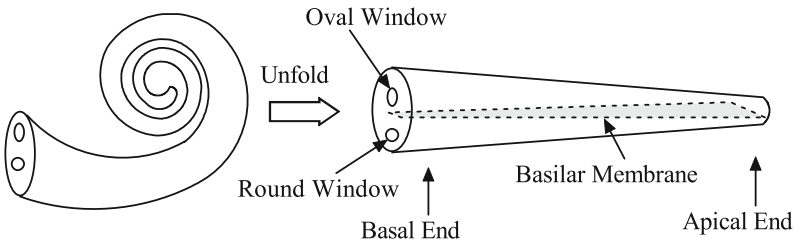


Fig. 2.1 Illustrative diagram of the cochlear in inner ear

guide it through the canal, which has resonant frequency around 4 kHz. In the *middle ear*, three bones, the hammer, the anvil and the stirrup, are connected to conduct the oscillation of the ear drum to the cochlear, transforming the air vibration in outer ear to fluid vibration in inner ear. The cochlear in the *inner ear* is crucial to the understanding of psychoacoustic facts, including perceptual masking, echo perception and cochlear delay.

As illustrated in Fig. 2.1, the cochlear has a spiral shape in order to save space in inner ear. It revolves around the auditory nerve, the output of inner ear to brain. To better illustrate the structure, the cochlear is also unfolded in Fig. 2.1. It can be thought as a bottle full of fluid with two openings, the oval window and the round window. Each window is covered with flexible membrane. The inside of the cochlear is separated into three cavities by two membranes, the basilar membrane and the Reissner's membrane (only the basilar membrane is shown). Auditory nerve is connected with the basilar membrane. Sound stimuli from the middle ear is applied to the oval window, causing the fluid to be compressed or expanded. Such vibration forces the basilar membrane to move accordingly and the vibration is converted into electrophysiology signal to the brain [1].

Each region along the basilar membrane has its own resonant frequency. The end near the oval and round window is called the *basal end* and the other end is the *apical end*. The basal end is stiffer, hence responds to higher frequency components; while the apical end is more flexible, hence responds to lower frequency components. Consequently, the cochlear acts as a spectrum analyzer, where different frequency components of the sound are resolved into responses at different places along the basilar membrane. This is usually referred to as the *frequency-to-place transformation*. This transformation is helpful in understanding the perceptual masking, perception of echo, and cochlear delay.

2.1.1.1 Bark Scale

The frequency-to-place transformation is nonlinear in Hz scale. In another scale, the Bark scale, this transform is linear. Using the Bark scale, it is also convenient to describe the spread of masking. The Bark scale is related to the masking effects, which refers to the phenomenon that a weaker sound is rendered imperceptible by a

stronger sound. A masking threshold is a sound pressure level (SPL) that the maskee is just noticeable in the presence of the masker. More details can be found in the Sects. 2.1.2 and 2.1.3.

In the tone-mask-noise experiment, the maskee is a narrow band noise centered at 2 kHz. The maskers are two tones centered symmetrically around the maskee, having SPL 50 dB and are Δf Hz apart from each other [2]. It is found that when Δf is below 300 Hz, the masking threshold is roughly a constant of 33 dB. As Δf increases further to be above 300 Hz, the masking threshold drops rapidly. Similar observation can be made for noise-mask-tone experiment. This observation suggests that masking is confined to a frequency band surrounding the frequency of the masker. The bandwidth of this band is called the *critical bandwidth*. The critical bandwidth can be fitted by the following nonlinear curve:

$$\Delta f = 25 + 75\left[1 + 1.4\left(\frac{f_c}{10^3}\right)^2\right]^{0.69} \text{(Hz)}, \tag{2.1}$$

where f_c is the center frequency of the band, which is also in Hz. The critical bandwidth Δf, although varying with the center frequency f_c, corresponds to a fixed distance along the basilar membrane. So a unit of frequency in terms of critical bandwidth may provide a linear mapping from frequency to location along the basilar membrane. Such a unit is called *Bark scale*. The conversion from frequency f in Hz to its Bark scale z can be obtained by treating (2.1) as $\frac{df}{dz}$. After inverting Δf followed by integration, it results in [2]

$$z(f) = 13 \arctan\left(\frac{7.6f}{10^4}\right) + 3.5 \arctan\left(\left(\frac{f}{7.5 \times 10^3}\right)^2\right). \tag{2.2}$$

This new measure of frequency is also called *critical band rate*. In psychoacoustic models such as Model I in MPEG-1 [3], the entire hearing range is approximated by 25 critical bands, with each band corresponding to a segment of 1.3 mm distance along the basilar membrane. Since the Bark scale is closely related to masking effect of the HAS, it is a natural scale to build a psychoacoustic model [3].

2.1.2 Absolute Threshold of Hearing

We start reviewing the psychoacoustic aspect from the perception of a single sound, the absolute threshold of hearing (ATH). The intensity measure of sound will be introduced first.

2.1.2.1 Intensity Measure of Sound

The change of pressure due to the existence of sound is *sound pressure*, with its unit Pascal defined as 1 N per square meter. The effect of instantaneous sound pressure $p(t)$ can be captured by root mean square (RMS) sound pressure p_{RMS}:

$$p_{\text{RMS}} = \sqrt{\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} p^2(t) dt}, \tag{2.3}$$

where $[t_1, t_2]$ is the time interval of interest. For periodic sound, such as pure tone, one may choose $t_1 = 0$ and $t_2 = T$, where T is the period of the signal [4]. The *sound pressure level, or SPL,* is a relative quantity in logarithmic scale defined as

$$L(\text{dB}) = 20 \log_{10} \frac{p_{\text{RMS}}}{p_0}, \tag{2.4}$$

where $p_0 = 20\,\mu\text{Pa}$ is a reference RMS value. This p_0 is chosen such that, for a person with normal hearing, the chances of perceiving the existence of a 2 kHz tone with sound pressure p_0 is around a half [2].

The SPL is an objective measure of the sound intensity. The subjective perception of sound intensity is called *loudness*. In general, for a given frequency, a tone with higher SPL is perceived as having greater loudness. However, for a given SPL, the perceived loudness varies with the frequency of the tone. The equal loudness curve describes this variation [2].

2.1.2.2 Definition of ATH

The ATH is the minimum SPL that a pure tone can be perceived by HAS in a noiseless environment. This threshold is frequency-dependent, being smaller for the frequencies that the HAS is sensitive to. The ATH can be well approximated as follows [5]:

$$T_{\text{ATH}}(f) = 3.64 \left(\frac{f}{10^3}\right)^{-0.8} - 6.5 \exp\left(-0.6 \left(\frac{f}{10^3} - 3.3\right)^2\right) + 10^{-3} \left(\frac{f}{10^3}\right)^4, \tag{2.5}$$

where f is frequency in Hz and $T_{\text{ATH}}(f)$ is SPL in dB.

The ATH gives the lower limit of hearing. For the upper limit, typical speech is below 80 dB SPL and music is below 95 dB SPL. When the SPL is above 100 dB, the HAS may be damaged in the presence of such sound [2]. So, in audio signal processing, the range of SPL from around 0 dB to as high as 100 dB is usually assumed.

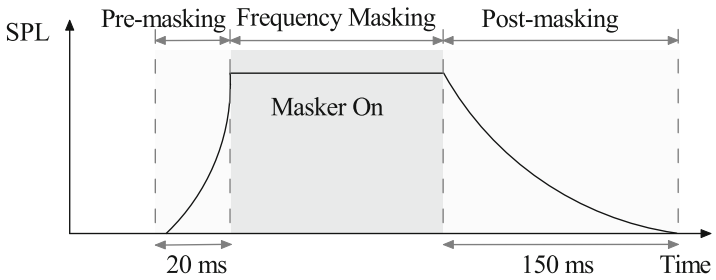


Fig. 2.2 Time masking and frequency masking

2.1.3 Masking Effects

As illustrated in Fig. 2.2, masking may occur in frequency domain or time domain. If two or more sounds that are either close in frequency domain or in time domain are presented to the HAS, the perception of the weaker sound (the maskee) may be masked by the stronger one (the masker). This masking effect helps increase the perception threshold, hence helps enhance the strength and robustness of the embedded watermarks [6, 7]. If the masker and the maskee are presented simultaneously in time domain and are close in frequency domain, then the masking is called *frequency masking* or *simultaneous masking*. In time domain, before the masker is turned on and after the masker is turned off, it may also influence the perception of the maskee. This is referred to as the *time masking* or *non-simultaneous masking*.

Frequency masking occurs due to the fact that when the basilar membrane is excited by a stronger sound, its response to weaker sound is weakened within the same critical band. The frequency masking model for real audio signals is based on psychoacoustic experiments on masking between the testing signals. For the testing signals, each masker and maskee can be either a pure tone or a narrow-band noise. So, there are four possible masking effects to consider: noise-masking-tone, tone-masking-tone, noise-masking-noise and tone-masking-noise. Among them, the noise-masking-tone (NMT) and tone-masking-tone (TMT) masking effects have been well-studied [1, 2]. For each masking effect, the masking curve depends on both the frequency and the SPL of the masker, and a typical curve for NMT is shown in Fig. 2.3. More detailed properties about these two types of masking can be found in [1].

Time masking is the masking phenomenon shortly before the masker is turned on (pre-masking) or after the masker is turned off (post-masking). When the excitation of the sound applied to the HAS is turned off, the HAS requires a certain time to build the perception of another sound. The pre-masking is shorter (roughly 20 ms) than the post-masking (up to 150 ms). In audio watermarking, pre-masking is usually ignored and only post-masking is considered. For example, in [6], a damping exponential curve is used to approximate the envelope of the audio, in order to provide approximate post-masking for the embedded watermark signal.

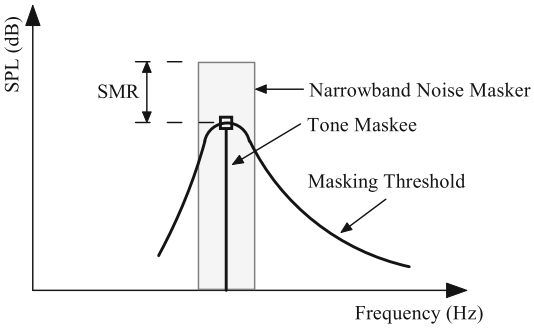


Fig. 2.3 Masking threshold for NMT, where SMR is the signal to mask ratio

2.1.4 Human Sensitivity to Echoes

As sound wave travels from its source (e.g., musical instrument, speaker, etc.) to receiver (such as HAS, microphone), it may be reflected by walls, furniture, or even buildings. So, the receiver may receive multiple delayed and attenuated versions of the emitted sound wave. The reflected versions of the sound stimulus are called echoes. Under certain conditions, such echoes are not perceivable or not annoying to the HAS. Hence, one can identify these conditions in order to embed secret bits by introducing additional echoes into audio signals. The conditions under which the introduced echoes are not perceptible may depend on the quality of the audio, the type of the music or even the listener. The delay time and attenuation are two key factors to the imperceptibility of echoes. For a single echo, if the delay time is greater than 50 ms, then a clear echo can be heard. If the introduced delay is smaller than 2 ms, then the HAS cannot perceive a clear echo, but only feel that the timbre of the sound is changed, usually more pleasing to the HAS. The change of timbre is called *coloration*.

An additional echo can be introduced into an audio signal by adding an attenuated and delayed replica of the signal to itself:

$$s_{\mathrm{w}}(n) = s(n) + \alpha s(n - d), \tag{2.6}$$

where the attenuation factor α is referred to as initial amplitude and the delay d between the original sound and the replica is called offset. The parameters α and d can be chosen based on psychoacoustic experiments to achieve imperceptibility. The imperceptible region of echoes is a region in a 2D plane of initial amplitude α and offset d. One such region as found in [8] is shown in Fig. 2.4.

Echo adding can be realized by convolving the original host audio signal with a linear filter, the echo kernel [7]. The frequency response of the echo kernel affects the perceptual quality. Oh *et al.* found that there is a close connection between the frequency characteristic of the echo kernel in Bark scale and the perceived distortion [9].

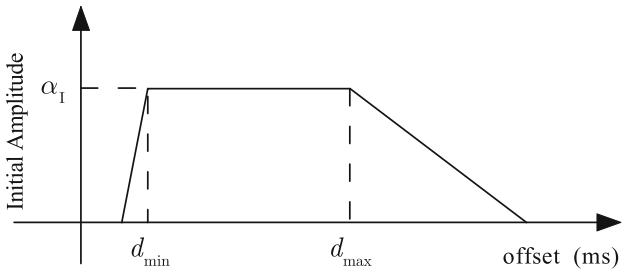


Fig. 2.4 The imperceptible region for a single echo, where $\alpha_{\mathrm{I}} = 0.31$, $d_{\min} = 0.9$ ms, and $d_{\max} = 3.4$ ms

- In Bark scale, the spectral envelope affects the perceptual quality more than the fine details of the spectrum.
- The bands between 0 to 10 Barks are crucial to the perceived quality. Furthermore, the lowest serval critical bands are especially important.

These experimental results suggest that, for an echo kernel to be imperceptible, the frequency response should be as flat as possible in the lower Bark bands. Besides, large fluctuations should be pushed to higher Bark bands. In this respect, the echo kernel with only one single delay is not optimum. In [10], a bipolar echo kernel with closely located positive and negative echoes are designed, which provides much flatter frequency response in low critical bands. Further improvements will be discussed in Chap. 3.

2.1.5 Characteristics of Cochlear Delay

Cochlear delay is the inherent delay in HAS as audio stimuli propagates along the cochlear. Psychoacoustic study reveals that the HAS cannot perceive the difference between the original sound and a processed sound with enhanced delay [11, 12]. This fact was utilized to embed watermark by adding group delays similar to the characteristics of the cochlear delay [13–17].

The physiology basis of cochlear delay relies on the structure of the cochlear and the sound propagation within the cochlear. As we know, the cochlear in the inner ear performs a 'frequency-to-place' conversion, where different frequency components of the stimuli excite different locations along the basilar membrane within the cochlear. At the basal side of the cochlear, the BM is stiffer hence responses to higher frequency components. At the apical side, the basilar membrane is more flexible and hence responses to lower frequency components. As the sound stimulus travels from the basal side to the apical side, different frequency components are 'picked up' by the basilar membrane sequentially. So, higher frequency components need less time to travel from the base to its location, while lower frequency components need

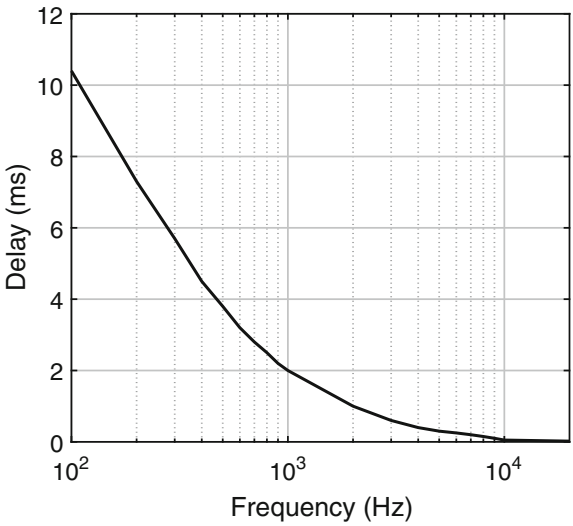


Fig. 2.5 Typical frequency-dependent cochlear delay [13]

more time. Consequently, lower frequency components are delayed more than higher frequency components. This frequency-dependent delay affects the phase spectrum of the perceived sound. Instead of using a filterbank model to model the cochlear, the 'transmission line filterbank' can better describe this sequential 'frequency-to-place' conversion. The frequency-dependent cochlear delay is shown in Fig. 2.5.

The psychoacoustic study by Aiba *et al.* reveals that [12], if an intentional delay with similar characteristic of the cochlear delay of HAS is introduced into the audio signal, then the HAS cannot perceive the difference between the sounds before and after processing. To study the effect of cochlear delay on the perception of audio stimulus, three signals are used. The first one is an impulse signal that contains all the frequency components and is delayed by the cochlear delay. The other two signals are chirp signals, a down-chirp and an up-chirp. Using chirp signals, both the frequency content and the time domain delay can be adjusted. The up-chirp signal starts with low frequency sinusoidal and sweeps up to high frequency sinusoidal. So in up-chirp, low frequency components lead ahead the high frequency components. By adjusting the frequency sweeping speed, one can compensate for the intrinsic cochlear delay in HAS. On the other hand, the down-chirp signal starts with high frequency sinusoidal and sweeps down to low frequency sinusoidal signal. So using a down-chirp, the low frequency components are further delayed than intrinsic cochlear delay. The sweeping speed is designed to provide similar delay characteristics as the cochlear delay.

Aiba *et al.* used the three signals to study the delay threshold needed for a subject to detect the onset asynchrony between two identical signals [12]. They found that using the up-chirp with compensated cochlear delay does not increase the ability of the subject in detecting the onset asynchrony. In addition, the down-chirp with enhanced cochlear delay sounds similar to the impulse signal. Similar result was also found

by Uppenkamp *et al.* [18]. These findings suggest that if additional delays that are similar to the cochlear delay are introduced into the audio signal, then the HAS may not be able to perceive the change. Furthermore, if two delay patterns are used, then one bit of watermark can be embedded into the audio signal. To this end, appropriate filters must be designed to introduce the additional delays.

Unoki *et al.* investigated the use of all-pass filters to approximate the cochlear delay characteristic [19]. A first order all-pass filter with z-transform

$$H(z) = \frac{-b + z^{-1}}{1 - bz^{-1}} \tag{2.7}$$

is used to introduce delay. Let the group delay introduced by $H(z)$ be

$$\tau(f) = -\frac{1}{2\pi}\frac{d}{df}\arg\left(H(f)\right), \tag{2.8}$$

where $H(f) = H(z)|_{z=e^{j2\pi f}}$. Then, the parameter of this filter, i.e., b, can be determined by minimizing

$$E = \int_{-1/2}^{1/2} \left[\alpha\tau^*(f) - \tau(f)\right]^2 df, \tag{2.9}$$

where $\alpha < 1$ is used to ensure that only slight cochlear delay is introduced. Using the least mean square (LMS) algorithm, the optimum b was found to be 0.795 if α is set as 1/10 [19].

2.2 Evaluation of Imperceptibility

For the purpose of designing imperceptible watermarks, the imperceptibility must be quantified and measured, which can then be fed back to off-line or online tuning stage. Current evaluation measures of imperceptibility can be classified into three categories: subjective measures, objective measures, and objective perceptual measures.

2.2.1 Subjective Measures

Since the receiver of a watermarked audio signal is the HAS, the subjective judgement of the quality or distortion of the watermarked audio signal is the ultimate way of evaluating perceptual quality. Furthermore, the result from subjective test can be used to calibrate objective measures. For example, correlation analysis between subjective measures and frequency weighted signal-to-noise ratio (SNR) can be used to determine optimum parameters, as will be discussed in Sect. 2.2.2. This section

reviews several popular subjective measures in audio watermarking, such as ABX test, mean opinion score (MOS) and subjective difference grade (SDG).

2.2.1.1 Evaluating Transparency: ABX

For high quality audio, such as music CD, the watermarked audio signal is usually required to attain 'transparent' quality. 'Transparent' means that the listener cannot perceive any difference between the original audio signal and the watermarked audio signal. ABX test can be used in such a context.

In ABX test, the listener is presented with three signals: the original audio signal (marked as A), the watermarked audio signal (marked as B), and a signal X that is randomly chosen from A or B. Then the listener is asked to judge if X is A or B. The ratio of correct answers r can be used to decide if the watermarked audio is of 'transparent' quality. If r is above a threshold, say τ, then we may state with high confidence that the watermarked audio signal is of transparent quality. If the listener's response is based purely on random guess, then r is around 50%. So a popular choice is $\tau = 0.75$ [20, 21]. In general, the ABX test can be put into a hypothesis testing framework. The number of experiments and the threshold can then be determined from the significant level [22, 23].

2.2.1.2 Evaluating Absolute Quality: MOS

In applications where certain degradation to the audio is acceptable, rating the absolute quality of the watermarked audio is needed. The MOS provides an absolute measure for perceptual degradation [24], where only the watermarked audio is evaluated. The testing procedure, including the room space, noise level, etc., are specified in ITU-T recommendation P.800 [25]. After listening to the testing audio (i.e., the watermarked audio), the listener choose a level from a five level scale to evaluate the quality of the audio. The criterion for choosing the levels is listed in Table 2.1.

2.2.1.3 Evaluating Small Impairments: SDG

While MOS gives the absolute measure of the watermarked audio signal, it is often desirable to measure the small distortions in the watermarked audio signal. The ITU

Table 2.1 Mean opinion score

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

standard ITU-R BS. 1116 provides a procedure to evaluate subjective quality when the impairment is expected to be small.[1]

The procedure follows the 'double-blind, triple-stimuli with hidden reference' approach. During each round, the listener is presented with three audio files (triple-stimuli): a reference signal A, and two testing signals B and C. The original one is always marked as A. One of B and C is the original audio signal (hidden reference) and the other is the watermarked audio signal. The two testing signals are permutated before presenting to the listener. 'Double blind' means that neither the administrator nor the listener knows which one among B and C is the reference signal (the original signal). After listening to the three signals, the listener is asked to grade the quality of B and C, when compared to the known reference signal A, respectively. As shown in Table 2.2, the grading is on a continuous scale from 1.0 to 5.0 with recommended precision of 0.1. Since one of B and C is the reference signal, so at least one of B and C should be graded as 5.0. At least 20 subjects are required, and each grading session includes 10 to 15 trails, with each testing signal having a length between 10 to 20 s. After the experiments, the score of the hidden reference signal S_{HR} and the score of the watermarked signal S_{W} are gathered. After gathering the raw data, the SDG is calculated as $\mathrm{SDG} = S_{\mathrm{W}} - S_{\mathrm{HR}}$. The last column of Table 2.2 shows the SDG values versus the impairment scale. So, $\mathrm{SDG} = 0$ means the watermark is imperceptible, while $\mathrm{SDG} = -4.0$ corresponds to very annoying distortion. The testing result is usually presented using the mean SDG along with the 95% confident interval for each type of testing tracks. Further statistical analysis such as ANOVA can be performed to test if the means of the different tracks are equal.

Although subjective listening test is an effective approach to evaluating the perceptual quality of watermarked audio signals, it is often time consuming and costly. Therefore, it is often used at the final stage of audio watermarking algorithm development. At the early and middle stages of algorithm development, non-subjective metrics for perceptual quality evaluation are desired. There are two types of such metrics, the simple SNR-based objective measures and the objective perceptual measures that can mimic the function of HAS, such as the perceptual evaluation of audio quality (PEAQ) measure [26].

2.2.2 Objective Measures

For objective measures, the psychoacoustic model or auditory model is not explicitly incorporated. Instead, they exploit the concept of SNR.

[1] The ITU-R BS.1116 standard is recommended only when the expected impairment of the watermarked audio signal is small. For a watermarked audio signal with intermediate quality, the ITU-R BS. 1534 standard is suitable [2].

Table 2.2 The five grade continuous scale in BS. 1116

Score	Quality	Impairment	SDG
5.0	Excellent	Imperceptible	0.0
4.9–4.0	Good	Perceptible but not annoying	−1.0
3.9–3.0	Fair	Slightly annoying	−2.0
2.9–2.0	Poor	Annoying	−3.0
1.9–1.0	Bad	Very annoying	−4.0

2.2.2.1 SNR

Let $s(n)$ be the host audio signal and $s_{\mathrm{w}}(n)$ be the watermarked audio signal. The average power of the watermarks can be an indicator of distortion introduced by watermarking. For this purpose, the SNR defined below can be utilized as an objective measure:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1}\left[s(n) - s_{\mathrm{w}}(n)\right]^2}, \tag{2.10}$$

where N is the sample size of the audio signals. Due to the averaging effect, the global SNR could be small even though there exist some large local differences between $s(n)$ and $s_{\mathrm{w}}(n)$. It is shown that SNR correlates badly with the subjective evaluation scores [2].

2.2.2.2 Segmental SNR

To better capture the local variation of SNR, the long duration signal is split into small segments. SNR is calculated for each segment and then the obtained SNRs from all segments are averaged. There are two segmental SNR: time domain segmental SNR and frequency-weighted segmental SNR.

For time domain segmental SNR [2], the local SNR is calculated in time domain. No frequency features are incorporated. Let M be the total number of segments and $\mathrm{SNR}(m)$ be the SNR for the mth segment. Then, the time-domain segmental SNR is calculated as:

$$\mathrm{segSNR} = \frac{1}{M}\sum_{m=0}^{M-1}\max\left\{\min\left\{35, \mathrm{SNR}(m)\right\}, -10\right\}, \tag{2.11}$$

where the frame SNR is limited to be between $[-10, 35]$ dB. For the frames with SNRs greater than 35 dB, the large SNRs will not make a difference to perceptibility but may only bias the final average toward larger values. So, the maximum frame SNR is clipped at 35 dB. For silent frames with only background noise, the signal

and watermark components may have competing amplitudes, making the SNR very negative. This may bias the final average towards smaller values. So the minimum frame SNR is clipped at -10 dB.

For frequency-weighted segmental SNR [27], the SNR is calculated in each critical band and weighted according to the signal strength in the same critical band. Let K be the number of critical bands. The quantity $X_m(k)$ is the spectrum for band k and frame m, obtained by summing up all spectrum components in band k. The corresponding spectrum component for the watermarked signal is denoted as $X_{m,\mathrm{w}}(k)$. A normalized weight $W_m(k)$ is chosen for each band, where $W_m(k) > 0$ and $\sum_{k=1}^{K} W_m(k) = 1$. Then, the frequency-weighted segmental SNR is calculated as:

$$\mathrm{fwSNR} = \frac{1}{M}\sum_{m=0}^{M-1}\left(\sum_{k=1}^{K} W_m(k)\cdot 10\log_{10}\frac{|X_m(k)|^2}{\left(|X_m(k)| - \left|X_{m,\mathrm{w}}(k)\right|\right)^2}\right). \tag{2.12}$$

The weight $W_m(k)$ can be designed to be proportional to $|X_m(k)|^\gamma$, with $\gamma \in [0.1, 2]$ optimized to get maximum correlation between the fwSNR measure and the subjective evaluation scores.

There are also other objective quality measures that are designed for speech signals, e.g., the Itakura-Saito distortion measure, Itakura distance, and log-area ration measure. These measures are closely connected to the speech production models, such as the all-pole filter model. They are suitable for measuring the quality of speech signals after processing, such as speech enhancement and speech synthesis [27].

The range of segmental SNRs for transparent audio codecs varies from 13 to 90 dB [3], meaning that they also correlate badly with the subjective quality. So, it is necessary to explore the auditory models in designing objective quality measures.

2.2.3 Objective Perceptual Measures

Objective perceptual measures are objective (computational) measures which utilize psychoacoustic or/and higher-level cognition models. Such measures output an objective difference grade (ODG) by comparing the watermarked audio signal with the original audio signal. Examples of such objective perceptual measures include the PEAQ measure as standardized in the ITU-R BS. 1387 [26], and the more recent PEMO-Q measure [28]. As reported in [29], the ODG from PEAQ correlates well with the SDG score from subjective listening test. As a result, the PEAQ measure has been widely used in audio watermarking systems to evaluate perceptual quality [30, 31].

The block diagram of PEAQ is shown in Fig. 2.6. For every pair of original and watermarked audio signals, the output of PEAQ is a single number, the ODG value, which has the same specification as the SDG scale. Firstly, the audio signals are mapped to frequency domain using DFT and/or filter banks, followed by

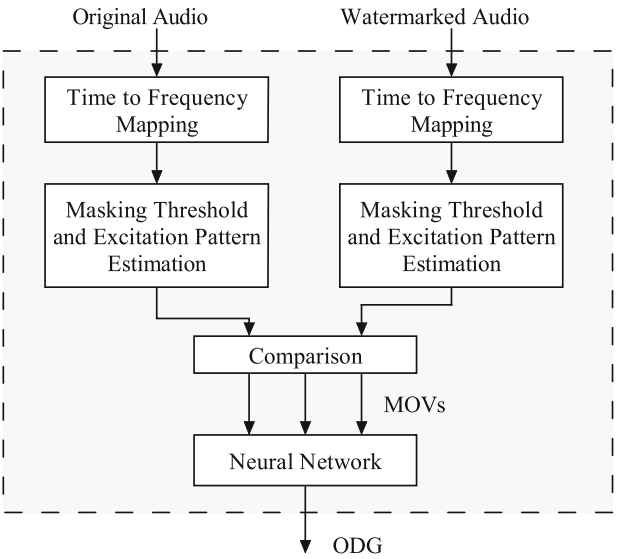


Fig. 2.6 Block diagram of PEAQ [26]

performing psychoacoustic analysis of the frequency components using masking threshold method and excitation pattern method. For the masking threshold method, the masking threshold of the original audio signal is analyzed and compared to the spectrum of the watermarked audio signal. The inaudible components, i.e., those below the masking threshold, are identified. In the excitation pattern method, the excitation pattern of each audio signal on the cochlear is estimated. Secondly, the internal representations are compared to obtain a set of model output values (MOVs), such as noise loudness, noise-to-mask ratio, average distorted block, and so on. Finally, a neural network is utilized to predict the SDG score, or equivalently the ODG value, from the MOVs. Such a neural network is trained with adequate testing signals and their corresponding SDG scores.

2.3 Control of Imperceptibility

The perceptual quality measures are mostly used to evaluate the performance of the designed watermarking algorithms. However, the evaluation result can also be fed back to the design phase or encoder to achieve better imperceptibility [32]. Moreover, one can systematically explore the psychoacoustic model to help design audio watermarking algorithms, both in frequency domain and time domain [6]. So in general, there are two approaches to control imperceptibility: the heuristic approach (feed-back approach) and the analytic approach (feed-forward approach), which will be reviewed separately below.

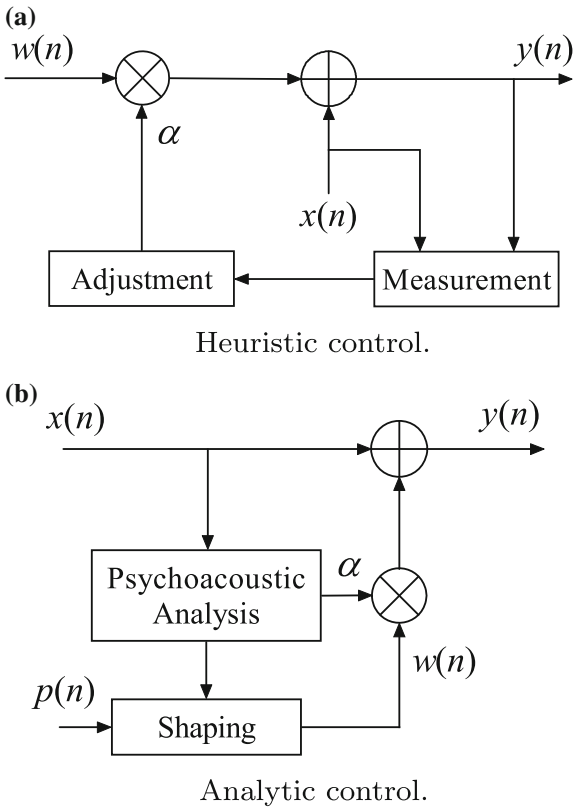


Fig. 2.7 Approaches to imperceptibility control

2.3.1 Heuristic Control of Imperceptibility

Figure 2.7a illustrates the heuristic control of spread spectrum audio watermarking. The embedding starts with an initial amplitude α and a unit power watermark $w(n)$: $y(n) = x(n) + \alpha \cdot w(n)$, where $x(n)$ is the original audio signal and $y(n)$ is the watermarked audio signal. Right after embedding, the perceptual quality of the watermarked audio signal $y(n)$ is measured with either a subjective or an objective quality measure as introduced in Sect. 2.2. The output of this measurement, in the form of SNR, SDG or ODG, is fed back to the embedding stage to adjust the embedding strength α. If the perceptual quality is bad, then α is reduced, until the desired perceptual quality is attained. The heuristic control of imperceptibility usually involves several rounds of tuning. The watermarking algorithms adopting this approach can be found in [33–38]. For the quantization index modulation (QIM) based algorithm, one can utilize similar approach to determine its parameters, such as the quantization step size Δ.

In general, the heuristic control approach is mostly suitable for off-line tuning of global embedding parameters, such as the global embedding strength. By adopting

ODG, online tuning is also possible. However, the online watermark design, i.e., designing the time and frequency characteristics of the watermarks according to the original audio signal, benefits less from heuristic control. The reason is that the perceptual quality measure acts as a black box in heuristic control, making it difficult to explore the perceptual redundancy.

2.3.2 Analytic Control of Imperceptibility

The general structure of analytic control is illustrated in Fig. 2.7b for spread spectrum watermarking. Using this approach, the embedding strength α and/or the watermark $w(n)$ is designed directly from analyzing the original host audio signal $x(n)$. An explicit psychoacoustic model is utilized here to analyze the perceptual redundancy within $x(n)$, which may be the masking threshold in time or frequency domain. This perceptual information is then used to determine the embedding strength α before actual embedding. The watermark $w(n)$ can also be designed by shaping the pseudo-random sequence $p(n)$. Depending on the shaping algorithm, the embedding strength may also be determined from the shaping process.

Compared to heuristic control, no measurement of perceptual quality and feedback is needed. In addition, the watermark $w(n)$ can also be designed from this approach, and can be made adaptive to local characteristic of the host audio signal. To present the specific analytic control methods, we start with reviewing a popular psychoacoustic model in MPEG-1 audio.

2.3.2.1 A Typical Psychoacoustic Model

As a typical example of computational psychoacoustic models, the psychoacoustic model I in MPEG-1 is widely used in audio watermarking. In the following, the computational steps for sampling rate 44.1 kHz and 16 bits quantization are briefly outlined. More details can be found from audio coding monographs and standards [2, 3, 5].

The input to this model is one frame of audio signal consisting of 512 samples. The output is a masking threshold $M(f)$ in frequency domain, which specifies the just noticeable distortion (JND) that the watermark can introduce into the host audio signal. First, the individual masking thresholds of tone-like and noise-like components are estimated. Then, by accounting for the spread of masking among different critical bands, the global masking threshold is obtained.

Step 1. Spectrum calculation and normalization: For each frame of input signal $s(n)$, the power spectrum $P(k)$ is calculated and normalized.

Step 2. Determination of tone-like and noise-like maskers: A spectrum component is classified as tone-like if it is greater than its surrounding components by at least 7 dB. For each of the tone maskers, its strength is found by adding itself with two closest neighbors:

$$P_{\mathrm{TM}}(k) = 10\log_{10}\left[10^{0.1P(k-1)} + 10^{0.1P(k)} + 10^{0.1P(k+1)}\right]. \tag{2.13}$$

After excluding the components that are close to a tone, the remaining components are treated as noise. Here, 'close' is quantified by Δ_k:

$$\Delta_k \in \begin{cases} 2, & 2 < k < 63; \\ \{2, 3\}, & 63 \le k < 127; \\ \{2, \cdots, 6\}, & 127 \le k \le 256. \end{cases}$$

Therefore, the strength of the noise-like component in each critical band is

$$P_{\mathrm{NM}}(\bar{k}) = 10\log_{10}\left[\sum_j 10^{0.1P(j)}\right], \tag{2.14}$$

where the summation is over all components that are more than $\pm\Delta_k$ from the tonal components in the current critical band, and $\bar{k}$ is the geometric mean of the frequencies within the current critical band.

Step 3. Decimation of maskers: First, the maskers below the ATH are removed. Then, for any maskers that are less than 0.5 Barks from each other, the weaker one is removed. Finally, the maskers from 18 to 22 Barks are decimated by 2:1, and those from 22 to 25 Barks are decimated by 4:1.

Step 4. Calculation of individual masking thresholds: The masking effects also spread across critical bands. The influence of a masker at band j on the threshold at band i can be approximated by the following piecewise linear function:

$$SF(i, j) = \begin{cases} 17\Delta_z - 0.4P(j) + 11, & -3 \le \Delta_z < -1 \\ (0.4P(j) + 6)\,\Delta_z, & -1 \le \Delta_z < 0 \\ -17\Delta_z, & 0 \le \Delta_z < 1 \\ (0.16P(j) - 17)\,\Delta_z - 0.15P(j), & 1 \le \Delta_z < 8 \end{cases} \tag{2.15}$$

where $P(j)$ is the SPL of either the tone-like masker or the noise-like masker, and $\Delta_z = z(i) - z(j)$ is the Bark scale masker-maskee separation, with $z(i)$ and $z(j)$ being the Bark scales of band i and j, respectively. The masking thresholds for tone masker and noise masker are given by

$$T_{\mathrm{TM}}(i, j) = P_{\mathrm{TM}}(j) - 0.275z(j) + SF(i, j) - 6.025, \tag{2.16}$$

$$T_{\mathrm{NM}}(i, j) = P_{\mathrm{NM}}(j) - 0.175z(j) + SF(i, j) - 2.025. \tag{2.17}$$

Step 5. Calculate global masking threshold: The global masking threshold accumulates the ATH and contributions from other critical bands:

$$M(i) = 10\log_{10}\left(10^{0.1T_{\mathrm{ATH}}(i)} + \sum_{\ell=1}^{N_{\mathrm{TM}}} 10^{0.1T_{\mathrm{TM}}(i,\ell)} + \sum_{\ell=1}^{N_{\mathrm{NM}}} 10^{0.1T_{\mathrm{NM}}(i,\ell)}\right), \tag{2.18}$$

where N_{TM} and N_{NM} are the numbers of tonal maskers and noise maskers, respectively, and $T_{\mathrm{ATH}}(i)$ is the value of ATH at band i.

2.3.2.2 Frequency Domain Shaping

The output of the psychoacoustic model is the masking threshold. The masking threshold in frequency domain is a function of frequency, which provides the watermark embedder with the following information: (1) any frequency component that is below the masking threshold is inaudible, and (2) how to distribute the power of the watermark such that the watermarked signal has less perceptual distortion. These information can be used to design the time and frequency characteristics of the watermarks. Depending on how these information is used, frequency domain shaping can be done by either frequency domain multiplication, direct substitution, or perceptual filtering.

Let the masking threshold from psychoacoustic analysis be $M(f)$. For the *frequency domain multiplication* approach [6], the spectrum of the pseudo-random sequence is multiplied with $M(f)$, and then is transformed back to time domain. This can be regarded as a frequency domain filtering operation.

For the *direct substitution* approach [39], the signal components that fall below the masking threshold $M(f)$ are replaced with frequency components from a pseudo-random sequence. Let $P_x(f)$ be the power spectrum of the audio signal, $X(f)$ and $W(f)$ be the spectrum of the audio signal and the pseudo-random sequence, respectively. Then the spectrum of the watermarked audio is determined by

$$X_{\mathrm{w}}(f) = \begin{cases} X(f), & \text{if } P_x(f) \geq M(f) \\ \alpha \cdot W(f), & \text{if } P_x(f) < M(f), \end{cases} \tag{2.19}$$

where the parameter α controls the embedding strength. This approach embeds the watermark during the shaping process.

For the *perceptual filtering* approach [40], an all-pole filter $H(f)$ is designed to approximate the masking threshold $M(f)$ within each frame:

$$\min_{H(f)} \int_{-1/2}^{1/2} \left[\sigma^2 |H(f)|^2 - M(f)\right]^2 \mathrm{d}f, \tag{2.20}$$

where the power spectral density (PSD) of the pseudo-random sequence $p(n)$ is assumed to be a constant σ^2. Then, $p(n)$ is filtered with $H(f)$. This ensures that the spectrum of the watermark has similar shape with the masking curve. After appropriate tuning of the embedding strength α, the PSD of the watermark can be controlled to be below the masking threshold.

The computation of the masking threshold $M(f)$ is usually time consuming. Using the ATH as the masking curve may drastically reduce the computational load, and has been utilized in some recent works, such as [7, 35].

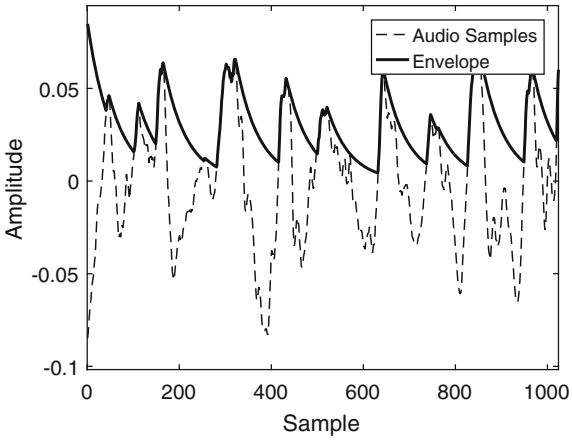


Fig. 2.8 Time domain envelope extraction

2.3.2.3 Time Domain Shaping

The effects of spectrum shaping or substitution in frequency domain will be distributed in the whole time frame in time domain, due to the time-frequency location property. So approximate time domain shaping usually follows the frequency domain shaping.

One way to shape the watermark in time domain is to use signal envelope $t(n)$ to approximate the post masking. In [6], the envelope is extracted by a damping exponential signal. Let $s_r(n)$ be a half-rectified signal: $s_r(n) = s(n)$, if $s(n) > 0$ and $s_r(n) = 0$, otherwise. Then the envelop can be extracted as

$$t(n) = \begin{cases} s_r(n), & \text{If } s_r(n) > t(n-1)e^{-\beta} \\ t(n-1)e^{-\beta}, & \text{Otherwise} \end{cases}$$

where $\beta > 0$ is the damping ratio. An example of the extracted envelop is shown in Fig. 2.8. Finally, the watermark $w(n)$ in time domain (or transformed from frequency domain) are multiplied with sqaured and normalized $t(n)$ to achieve time domain shaping [40]:

$$\hat{w}(n) = w(n)\frac{t^2(n)}{\sum_{k=0}^{N-1} t^2(k)}. \tag{2.21}$$

2.4 Summary and Remark

The various aspects pertaining to imperceptibility of audio watermarking is reviewed in this chapter. By exploring the basic frequency-to-place transformation in cochlear, the psychoacoustic facts such as ATH, masking effects, perception of echoes and cochlear delay are outlined. These psychoacoustic effects are exploited to evaluate

and control the imperceptibility of audio watermarking. These backgrounds can help understand the design of audio watermarking algorithms.

The masking effects are exploited in frequency domain and time domain separately. The joint masking effect in time-frequency plane can be further utilized to make audio watermarking more robust. Since the heuristic control is related to direct perception and imperceptibility evaluation, it may provide better robustness and perceptibility tradeoff.

References

1. Fastl H, Zwicker E (2007) Psychoacoustics: facts and models, 3rd edn. Springer, Berlin
2. Bosi M, Goldberg RE (2002) Introduction to digital audio coding and standards. Kluwer Academic Publishers, Norwell
3. Spanias A, Painter T, Atti V (2007) Audio signal processing and coding. Wiley, New Jersey. Chapter 5
4. Speaks CE (1992) Introduction to sound: acoustics for the hearing and speech sciences. Springer Science + Business Media Dordrecht, San Diego
5. Painter T, Spanias A (2000) Perceptual coding of digital audio. Proc IEEE 88(5):451–515
6. Swanson MD, Zhu B, Tewfik AH, Boney L (1998b) Robust audio watermarking using perceptual masking. Sig Process 66(3):337–355
7. Hua G, Goh J, Thing VLL (2015a) Time-spread echo-based audio watermarking with optimized imperceptibility and robustness. IEEE/ACM Trans Audio Speech Lang Process 23(2):227–239
8. Oh HO, Kim HW, Seok JW, Hong JW, Youn DH (2001a) Transparent and robust audio watermarking with a new echo embedding technique. In: IEEE international conference on multimedia and expo (ICME 2001), pp 317–320
9. Oh HO, Seok JW, Hong JW, Youn DH (2001b) New echo embedding technique for robust and imperceptible audio watermarking. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)
10. Oh HO, Youn DH, Hong JW, Seok JW (2002) Imperceptible echo for robust audio watermarking. In: Audio engineering society convention 113, Los Angeles, USA
11. Dau T, Wegner O, Mellert V, Kollmeier B (2000) Auditory brainstem responses with optimized chirp signals compensating basilar-membrane dispersion. J Acoust Soc Am 107(3):1530–1540
12. Tanaka S, Unoki M, Aiba E, Tsuzaki M (2008) Judgment of perceptual synchrony between two pulses and verification of its relation to cochlear delay by an auditory model. Japan Psychol Res 50(4):204–213
13. Unoki M, Imabeppu K, Hamada D, Haniu A, Miyauchi R (2011a) Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics. J Inf Hiding Multimedia Signal Process 2(1):1–23
14. Unoki M, Miyauchi R (2008) Reversible watermarking for digital audio based on cochlear delay characteristics. In: International conference on intelligent information hiding and multimedia signal processing, 2008. IIHMSP'08, pp 314–317
15. Unoki M, Miyauchi R (2013) Method of digital-audio watermarking based on cochlear delay characteristics. In: Multimedia information hiding technologies and methodologies for controlling data. IGI Global, pp 42–70
16. Unoki M, Miyauchi R (2015) Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay. IEICE Trans Inf Syst E98–D(1):38–48
17. Unoki M, Hamada D (2010) Method of digital-audio watermarking based on cochlear delay characteristics. Int J Innovative Comput Inf Control 6(3(B)):1325–1346
18. Uppenkamp S, Fobel S, Patterson RD (2001) The effects of temporal asymmetry on the detection and perception of short chirps. Hear Res 158(12):71–83

19. Unoki M, Hamada D (2008) Audio watermarking method based on the cochlear delay characteristics. In: International conference on intelligent information hiding and multimedia signal processing, pp 616–619
20. Xiang Y, Peng D, Natgunanathan I, Zhou W (2011) Effective pseudonoise sequence and decoding function for imperceptibility and robustness enhancement in time-spread echo-based audio watermarking. IEEE Trans Multimedia 13(1):2–13
21. Ko BS, Nishimura R, Suzuki Y (2005) Time-spread echo method for digital audio watermarking. IEEE Trans Multimedia 7(2):212–221
22. Boley J, Lester M (2009) Statistical analysis of ABX results using signal detection theory. In: Audio engineering society convention 127
23. Lu ZM, Yan B, Sun SH (2005) Watermarking combined with CELP speech coding for authentication. IEICE Trans 88–D(2):330–334
24. Kleijn WB, Paliwal KK (eds) (1995) Speech coding and synthesis. Elsevier Science Inc., New York
25. ITU-T (1996) ITU-T recommendation p.800: Methods for objective and subjective assessment of quality. http://www.itu.int/
26. ITU-T (1998–2001) Recommendation ITU-R BS. 1387-1: Method for objective measurements of perceived audio quality
27. Hu Y, Loizou PC (2008) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 16(1):229–238
28. Huber R, Kollmeier B (2006) PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception. IEEE Trans Audio Speech Lang Process 14(6):1902–1911
29. Treurniet WC, Soulodre GA (2000) Evaluation of the ITU-R objective audio quality measurement method. J Audio Eng Soc 48(3):164–173
30. Xiang Y, Natgunanathan I, Guo S, Zhou W, Nahavandi S (2014) Patchwork-based audio watermarking method robust to de-synchronization attacks. IEEE/ACM Trans Audio Speech Lang Process 22(9):1413–1423
31. Kalantari NK, Akhaee MA, Ahadi SM, Amindavar H (2009) Robust multiplicative patchwork method for audio watermarking. IEEE Trans Audio Speech Lang Process 17(6):1133–1141
32. Hua G, Huang J, Shi YQ, Goh J, Thing VLL (2016a) Twenty years of digital audio watermarking - a comprehensive review. Signal Process 128:222–242
33. Lemma AN, Aprea J, Oomen W, Kerkhof LVD (2003) A temporal domain audio watermarking technique. IEEE Trans Signal Process 51(4):1088–1097
34. Baras C, Moreau N, Dymarski P (2006) Controlling the inaudibility and maximizing the robustness in an audio annotation watermarking. IEEE Trans Audio Speech Lang Process 14(5):1772–1782
35. Lie WN, Chang LC (2006) Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification. IEEE Trans Multimedia 8(1):46–59
36. Xiang S, Huang J (2007) Histogram-based audio watermarking against time-scale modification and cropping attacks. IEEE Trans Multimedia 9(7):1357–1372
37. Chen OTC, Wu WC (2008) Highly robust, secure, and perceptual-quality echo hiding scheme. IEEE Trans Audio Speech Lang Process 16(3):629–638
38. Kirovski D, Malvar HS (2003) Spread-spectrum watermarking of audio signals. IEEE Trans Signal Process 51(4):1020–1033
39. Garcia RA (Sep 1999) Digital watermarking of audio signals using a psychoacoustic auditory model and spread spectrum theory. In: Audio engineering society convention 107
40. Boney L, Tewfik A, Hamdy K (1996) Digital watermarks for audio signals. In: The third IEEE international conference on multimedia computing and systems, pp 473–480

Chapter 3
Classical Techniques and Recent Developments

Abstract In this chapter, we introduce the research and development works for robust audio watermarking over the past decades. An audio watermarking system could be categorized into a time or transform domain system, simply by examining whether the watermarks are embedded in the original or transformed audio samples Hua et al. (Signal Process 128:222–242, 2016) [1]. Here, we take a different perspective to review audio watermarking techniques. We start from introducing the three classical techniques, i.e., echo hiding, spread spectrum, and quantization index modulation, followed by their further improvements and advanced designs. For the advanced designs, we categorized them according to what the watermarking systems are aimed for, instead of in which domain the watermarks are embedded. We also introduce several novel perspectives on audio watermarking at the end of this chapter.

3.1 Classical Techniques

The three classical digital audio watermarking techniques are echo hiding [2], spread spectrum (SS) [3], and quantization index modulation (QIM) [4], respectively. While the echo hiding technique is unique for audio watermarking, SS and QIM techniques have been widely applied in image and video watermarking as well. The echo hiding technique performs watermark embedding by adding attenuated and delayed replica of the original cover signal, while watermark extraction relies on cepstral analysis techniques. Such a paradigm exploits the insensitivity of human auditory system (HAS) towards echoes in audio signals, which allows the watermarks in the form of echoes to exist in the cover audio signal imperceptibly. The SS technique originates from communications theory [3]. In SS communications, a signal to be transmitted is deliberately spread in frequency domain with a wider bandwidth for better detection performance. The spreading sequence in SS based watermarking is analogous to the spreading signal in SS communications. The QIM technique is a special technique dedicated for watermarking. It modulates the watermarks into the indices of a series of quantizers of the cover signal. We assume the binary watermark bit b is drawn from either $\{-1, +1\}$ or $\{0, 1\}$, where appropriate.

Y. Xiang et al., *Digital Audio Watermarking*, SpringerBriefs in Signal Processing, DOI 10.1007/978-981-10-4289-8_3

3.1.1 Echo Hiding

Let a frame of the cover signal be $s(n)$, where the frame index (subscript) is dropped for simplicity. The watermark, in the form of an echo, is given by $\alpha s(n-d)$, where α controls the watermark strength, and $d \in \mathsf{Z}_+$ is the delay parameter. The watermarked signal, $s_{\mathrm{w}}(n)$, is then given by

$$s_{\mathrm{w}}(n) = s(n) + \alpha s(n-d), \tag{3.1}$$

which could be rewritten in convolution form

$$s_{\mathrm{w}}(n) = s(n) \otimes [\delta(n) + \alpha\delta(n-d)], \tag{3.2}$$

where $\delta(n)$ is the Dirac delta function. The watermark bit to be embedded into the host signal is modulated by the delay time d, i.e., there are two values for d to represent, which are "0" and "1", respectively. In general, the delay time is preferable to be at around one thousand of a second for the best imperceptibility [2]. Based on the introductory work in [2], further modifications have been reported in subsequent works. In [5], the concept of both positive and negative echoes is introduced, while in [6], the concept of both forward and backward echoes is presented. For clarity, we call the term convolved with $s(n)$ as *echo kernel*, denoted by $h(n)$, while we refer to the term in echo kernel excluding $\delta(n)$ as *echo filter*. Summarizing the works in [2, 5, 6], the general echo kernel for early echo hiding watermarking is given by

$$h(n) = \delta(n) + \sum_i \left[\overbrace{\underbrace{\alpha_{1,i}\delta(n-d_{1,i})}_{\text{Positive}} \underbrace{-\alpha_{2,i}\delta(n-d_{2,i})}_{\text{Negative}}}^{\text{Forward}} + \underbrace{\alpha_{1,i}\delta(n+d_{1,i}) - \alpha_{2,i}\delta(n+d_{2,i})}_{\text{Backward}} \right], \tag{3.3}$$

which consists of positive, negative, forward, and backward echo filters, where i indexes the number of echos. Note that the echo kernel in (3.2) is a special case of (3.3). Let us denote the echo filter by $w(n)$, then the general echo-based audio watermark embedding function is given by

$$s_{\mathrm{w}}(n) = s(n) \otimes [\delta(n) + \alpha w(n-d)], \tag{3.4}$$

and we will use this notation to derive watermark extraction method based on cepstral analysis [7, 8].

First, let the uppercase letters $X(\omega)$, $H(\omega)$, and $W(\omega)$ denote the discrete-time Fourier transform (DTFT)[1] of the time domain signals $s(n)$, $h(n)$, and $w(n)$ respectively. Then, (3.4) can be written in frequency domain as

$$\ln |X_{\mathrm{w}}(\omega)|^2 = \ln |X(\omega)|^2 + \ln |H(\omega)|^2, \tag{3.5}$$

where we have

$$\begin{aligned} \ln |H(k)|^2 &= \ln \left\{ \left[1 + \alpha W(\omega) e^{-j\omega d}\right] \left[1 + \alpha W^*(\omega) e^{j\omega d}\right] \right\} \\ &= \ln \left\{ 1 + 2\alpha \Re \left[W(\omega) e^{-j\omega d}\right] + \alpha^2 |W(\omega)|^2 \right\} \\ &\approx \alpha \Re \left[W(\omega) e^{-j\omega d}\right], \end{aligned} \tag{3.6}$$

where $\{\cdot\}^*$ denotes complex conjugation, and the approximation is obtained via Taylor series expansion by noting α is small [9]. The inverse DTFT of (3.6) yields the cepstrum of $h(n)$, i.e.,

$$\mathcal{C}_H(n) \approx \frac{\alpha}{2} \left[w(n-d) + w(-n-d) \right], \tag{3.7}$$

and the cepstrum of $s_{\mathrm{w}}(n)$ is then given by

$$\mathcal{C}_{X_{\mathrm{w}}}(n) \approx \frac{\alpha}{2} \left[w(n-d) + w(-n-d) \right] + \mathcal{C}_X(n). \tag{3.8}$$

It can be seen from (3.8) that the peak value located at delay d indicates the echo location, while the cepstrum of the host signal is an interference term during watermark extraction. It is also indicated that such an extraction scheme is insecure, since the echo location is open to standard cepstral analysis. Therefore, systems using echoes from (3.3) do not have security property.

To overcome this disadvantage, time-spread echoes have been proposed in [9], in which the echo filter coefficients are drawn from binary pseudo-noise (PN) sequence, $p(n)$. In this way, the result obtained in (3.8) is noise-like, and a correlation process is needed to detect the existence of $w(n)$. The block diagram of a time-spread echo-based audio watermarking system, including embedding and extraction mechanisms, is shown in Fig. 3.1. More discussions on cepstral analysis process for echo-based audio watermark detection are provided in [10]. In [11], the PN sequence is further modified for improved imperceptibility by using the following processing formula. Let the modified pseudo-noise (MPN) sequence be $\tilde{p}(n)$, then we have

$$\tilde{p}(n) = \begin{cases} p(n), \ n = 0, \ \text{or} \ n = L-1, \\ (-1)^{y(n)} \cdot p(n), \ 1 < n < L-1, \end{cases} \tag{3.9}$$

[1] Here DTFT is used instead of DFT for the ease of analysis and notation. Note that in practical implementations, the fast algorithm of DFT, i.e., fast Fourier transform (FFT) is used.

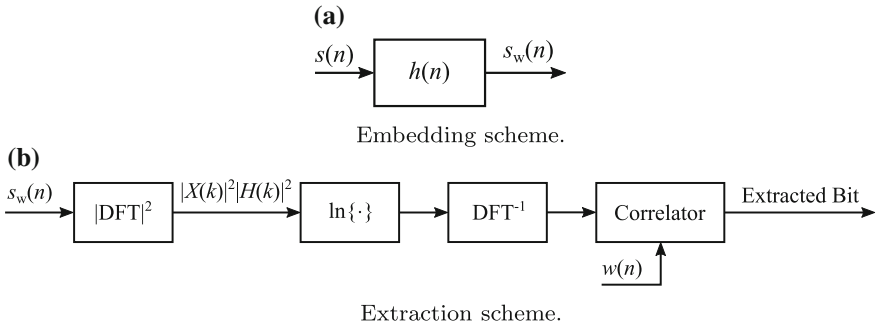


Fig. 3.1 Block diagram of an echo-based audio watermarking system using time-spread echoes

where L is the length of the PN sequence, and

$$y(n) = \text{round}\left(\frac{\tilde{p}(n-1) + p(n-1) + p(n) + p(n+1)}{4}\right), \tag{3.10}$$

where round(x) is a function that rounds x to the nearest integer. The corresponding watermark extraction scheme is also modified accordingly by considering two adjacent samples at the two sides of the correlations peak, in order to improve watermark extraction accuracy [11]. More advanced echo-based solutions, i.e., [12–14] will be introduced in Sect. 3.2.

3.1.2 Spread Spectrum

While the echo-based audio watermarking is a typical time domain audio watermarking technique, many techniques perform watermark embedding and decoding in transform domain. Note that Fig. 3.1a shows a special case of time domain watermark embedding, and a general scheme could be obtained by replacing the filter block with a generic watermark embedding block. Different from time domain methods, transform domain watermarking involves two more steps, i.e., the forward and the inverse transforms before and after watermark embedding. A generic block diagram of transform domain watermark embedding is shown in Fig. 3.2. Normally, both the SS and QIM techniques follow such a framework.

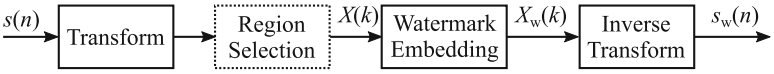


Fig. 3.2 A generic block diagram of transform domain watermark embedding

3.1.2.1 Direct Sequence Spread Spectrum

The SS technique that has been widely used in watermarking systems is more precisely the direct sequence spread spectrum (DSSS), in which the binary watermark bits are modulated by a spreading sequence, usually in the form of a PN sequence. Considering a single frame $X(k)$ and the watermark b, together with the spreading PN sequence $p(k)$, the watermark embedding function is given by

$$X_{\mathrm{w}}(k) = X(k) + \alpha b p(k). \tag{3.11}$$

In the above formula, the frame index i is omitted for simplicity, and $X(k)$ usually takes an appropriately selected portion of the transformed host signal, in order to achieve improved imperceptibility and robustness properties [3, 15]. It can be seen from (3.11) that the SS method is non-informed, i.e., watermark embedding is independent of the host signal. At the receiving end, the user is assumed to have access to the spreading sequence $p(n)$, and watermark extraction is achieved via the following function

$$\begin{aligned}\hat{b} &= \mathrm{sgn} \sum_k \frac{X_{\mathrm{w}}(k)p(k)}{p^2(k)} \\ &= \mathrm{sgn}\left(\sum_k \frac{X(k)p(k)}{p^2(k)} + \alpha b\right) \\ &\triangleq \mathrm{sgn}\left(\Psi + \alpha b\right),\end{aligned} \tag{3.12}$$

where Ψ is the host signal interference term defined by the normalized inner product between $X(k)$ and $p(k)$.

To deal with the host signal interference, an improved spread spectrum (ISS) method is proposed in [16], whose watermark embedding function is modified to

$$X_{\mathrm{w}}(k) = X(k) + (\alpha b - \lambda \Psi) p(k), \tag{3.13}$$

where λ is the host interference removal factor. (3.13) is an informed watermark embedding function as it incorporates the host signal interference Ψ during watermark embedding. The corresponding watermark extraction function is given by

$$\hat{b} = \mathrm{sgn}\left((1-\lambda)\Psi + \alpha b\right). \tag{3.14}$$

If $\lambda = 1$, the system fully removes the host signal interference. However, the interference term Ψ is not controllable because it is dependent on the host signal. As a result, the system loses certain imperceptibility control capability. Further analysis in [16] indicates that the system tends to be asymptotically optimal when $\lambda = 1$.

A further modification by considering the sign of Ψ during watermark embedding is proposed in [17], which is termed correlation-and-bit aware improved spread

spectrum (CAISS) method. According the the sign of Ψ and the watermark bit b, the embedding function is given by

$$X_{\mathrm{w}}(k) = \begin{cases} X(k) + \alpha_1 p(k), & \Psi \geq 0, \quad b = +1, \\ X(k) - (\alpha_2 + \lambda\Psi) p(k), & \Psi \geq 0, \quad b = -1, \\ X(k) - \alpha_1 p(k), & \Psi < 0, \quad b = -1, \\ X(k) + (\alpha_2 - \lambda\Psi) p(k), & \Psi < 0, \quad b = +1, \end{cases} \tag{3.15}$$

where the strength factor α is modified to α_1 and α_2 for different cases. The idea behind (3.15) is simple: if b and Ψ have the same sign (the first and third case in (3.15)), then the host interference actually enhanced the watermark, and for the concern of imperceptibility, there is no need to include Ψ anymore. On the other hand, if b and Ψ have different signs, then the embedding function flips the sign of Ψ during watermark embedding, and hence preserves the sign of b by using λ. The resultant watermark extraction function, taking the same form as (3.12), is given by

$$\hat{b} = \begin{cases} \alpha_1 + \Psi, & \Psi \geq 0, \quad b = +1, \\ -\alpha_2 + (1 - \lambda)\Psi, & \Psi \geq 0, \quad b = -1, \\ -\alpha_1 - \Psi, & \Psi < 0, \quad b = -1, \\ \alpha_2 + (1 - \lambda)\Psi, & \Psi < 0, \quad b = +1. \end{cases} \tag{3.16}$$

It can be seen from (3.16) that when Ψ and b have different signs, the estimation result $\hat{b}$ has the same expression as the one given in (3.14).

3.1.2.2 Patchwork

The essential idea behind SS based techniques is to spread small-valued noise-like sequences into the transform domain representation of the host signal. Similarly, an alternative approach, termed patchwork, has been proposed to perform the task [18–22]. The term patchwork refers to small elements to be added into a host subject. It was originally introduced for image watermarking [23] and then was incorporated for audio watermarking in [24]. Although patchwork based methods could be considered as a special case of SS, it differs from the SS based techniques in the following ways.

- Patchwork is based on dual-channel watermark embedding. While a patch is added to one channel, it is subtracted from the other channel.
- The transform domain samples for watermark embedding are randomly selected within a predefined frequency band. This makes the system more secure against unauthorized watermark extraction.
- Instead of modulating the watermark bits into the spreading sequence, bits 0 and 1 are distinguished by choosing different set of samples for embedding.
- Instead of using correlation based watermark bit extraction, patchwork method compares the test statistics obtained from the two sets of samples corresponding to bits 0 and 1 respectively, along with a predefined threshold, to make a decision.

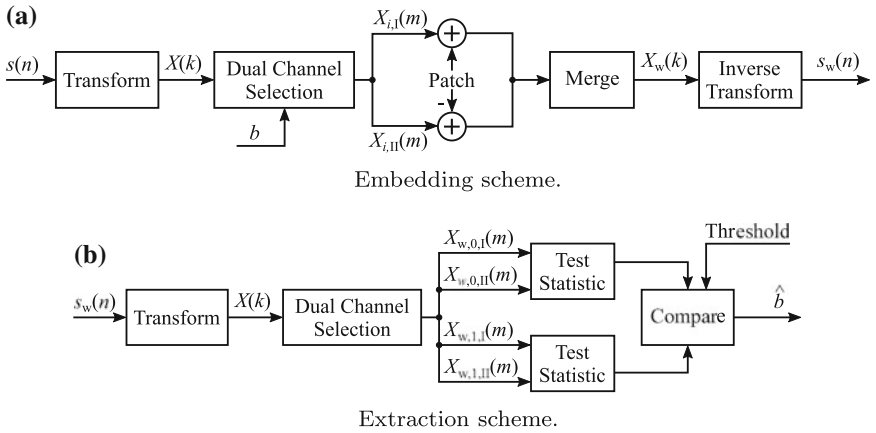


Fig. 3.3 Block diagram of patchwork based audio watermarking

The block diagram of a patchwork based audio watermarking system is shown in Fig. 3.3. During watermark embedding, the host signal frame is first transformed into, e.g., frequency domain via discrete cosine transform (DCT). To ensure the robustness against compression, only mid-frequency bands are reserved for embedding. The mid-frequency bands are then randomly grouped into two subsets, which will be used to embed "0" and "1" respectively. Specifically, if "0" is supposed to be embedded, then the first subset of DCT samples is used. Otherwise, the second subset is used. The selected subset is further divided into two halves, which form the dual-channel for watermark embedding. The patch, which could be designed in different ways e.g., [18, 19, 22], etc., is then added to the first channel and subtracted from the second channel. The modified samples in both channels are then merged, along with the unchanged samples, to form the watermarked signal in transform domain, followed by inverse transform to obtain the watermarked signal in time domain. During watermark extraction, the same mid-frequency bands are obtained and grouped in the same way as during watermark embedding. Then, the dual-channels of both subsets for "0" and "1" are extracted to calculate the test statistics for "0" and "1" respectively. A threshold value serving as a lower bound on the test statistics is also incorporated to make the decision on the estimation of embedded bit.

Let the dual-channel host signal in the selected subset be $X_{i,\mathrm{I}}(m)$ and $X_{i,\mathrm{II}}(m)$, where i denotes the watermark bit which is either "0" or "1", and let the patch be a simple constant d. The embedding function is given by

$$\begin{cases} X_{\mathrm{w},i,\mathrm{I}}(m) = X_{i,\mathrm{I}}(m) + d, \\ X_{\mathrm{w},i,\mathrm{II}}(m) = X_{i,\mathrm{II}}(m) - d. \end{cases} \tag{3.17}$$

Suppose the embedded watermark bit is "0", i.e., $i = 0$, then, during watermark extraction, the test statistics are given by

$$T_0 = \frac{1}{M}\sum_m \left(X_{\mathrm{w},0,\mathrm{I}}(m) - X_{\mathrm{w},0,\mathrm{II}}(m)\right) = \frac{1}{M}\sum_m \left(X_{0,\mathrm{I}}(m) - X_{0,\mathrm{II}}(m) + 2d\right), \tag{3.18}$$

$$T_1 = \frac{1}{M}\sum_m \left(X_{\mathrm{w},1,\mathrm{I}}(m) - X_{\mathrm{w},1,\mathrm{II}}(m)\right) = \frac{1}{M}\sum_m \left(X_{1,\mathrm{I}}(m) - X_{1,\mathrm{II}}(m)\right), \tag{3.19}$$

where M is the signal length. Since the subset for embedding "1" is not used, we have

$$E\{T_0\} \approx E\left\{X_{0,\mathrm{I}}(m) - X_{0,\mathrm{II}}(m)\right\} + 2d \approx 2d, \tag{3.20}$$

and

$$E\{T_1\} \approx E\left\{X_{1,\mathrm{I}}(m) - X_{1,\mathrm{II}}(m)\right\} \approx 0. \tag{3.21}$$

This is based on assuming the host signal samples in each channel have zero mean. If "1" is embedded, then we will have $E\{T_0\} \approx 0$ and $E\{T_1\} \approx 2d$ instead. The threshold T is chosen to be smaller than $2d$, and only if $\max\{E\{T_0\}, E\{T_1\}\} > T$ should the extraction be successful.

In [18], the patch, d is modified to

$$d = \alpha\,\mathrm{sgn}\left(\frac{1}{M}\sum_m \left(X_{i,\mathrm{I}}(m) - X_{i,\mathrm{II}}(m)\right)\right)\sqrt{\frac{\mathrm{var}\left\{X_{i,\mathrm{I}}(m)\right\} + \mathrm{var}\left\{X_{i,\mathrm{II}}(m)\right\}}{4(M-1)}}, \tag{3.22}$$

where

$$\mathrm{var}\left\{X_{i,\mathrm{I\,(or\,II)}}(m)\right\} = \frac{1}{M}\sum_m \left(X_{i,\mathrm{I\,(or\,II)}}(m) - \frac{1}{M}\sum_m X_{i,\mathrm{I\,(or\,II)}}(m)\right)^2. \tag{3.23}$$

The advantage of using (3.22) is that the sign of the average difference between the two channels is taken into consideration during watermark embedding, making the test statistics at the receiver separate further. The essential idea is in fact quite similar to the development from ISS to CAISS method, whose embedding function (3.15) also consider the signs therein. Further development of patchwork based method could be seen in [19–21].

3.1.3 Quantization Index Modulation

The QIM technique modulates the watermarks in the indices of a series of quantizers that are used to quantize the host signal. It can be used to watermark any cover work as long as the work is digitally stored. The proposal of the QIM technique is motivated by noting that conventional SS based methods are not able to deal with host signal interference. Original implementations of the QIM technique are in terms of dither

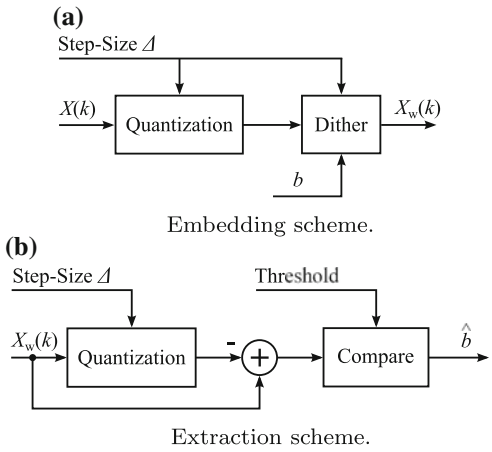


Fig. 3.4 Block diagram of dither modulation audio watermarking

modulation, distortion-compensated QIM, and spread-transform dither modulation (STDM), respectively [4]. In fact, the choices of quantizers are quite flexible in QIM based watermarking systems, and for each host signal sample, a different quantizer could be used [4]. For simplicity and to capture the essence of QIM, we only consider a single quantizer across the embedding process.

The block diagram of dither modulation audio watermarking is shown in Fig. 3.4. The transform domain host signal samples are first quantized according to the quantization step-size Δ, and then watermarked by adding different scales of quantization residuals. Specifically, the watermark embedding function is given by

$$X_{\mathrm{w}}(k) = \begin{cases} \left(\left\lfloor \frac{X(k)}{\Delta} \right\rfloor + \frac{1}{4}\right)\Delta, & b = 0, \\ \left(\left\lfloor \frac{X(k)}{\Delta} \right\rfloor + \frac{3}{4}\right)\Delta, & b = 1, \end{cases} \tag{3.24}$$

where $\lfloor \cdot \rfloor$ is the ceiling operator. In the above scheme, the threshold used for watermark extraction is $\Delta/2$. During watermark extraction assuming a closed-loop environment, the system simply re-quantizes $X_{\mathrm{w}}(k)$ using the same parameter Δ and obtains the quantization residual, either $\Delta/4$ or $3\Delta/4$. If the residual is not smaller than $\Delta/2$, the estimated watermark bit is 0. Otherwise, the system returns "1". Note that a large Δ could improve system robustness while the watermarks would tend to be more perceptible, and vice versa.

To improve the trade-off between imperceptibility and robustness, the distortion-compensated QIM adds an extra controlling parameter, $0 \leq \beta \leq 1$, to the embedding function, which yields

$$X_{\mathrm{w}}(k) = \begin{cases} X(k) - \beta\left(X(k) - \left\lfloor \frac{\beta X(k)}{\Delta} \right\rfloor \frac{\Delta}{\beta}\right) + \frac{\Delta}{4\beta}, & b = 0, \\ X(k) - \beta\left(X(k) - \left\lfloor \frac{\beta X(k)}{\Delta} \right\rfloor \frac{\Delta}{\beta}\right) + \frac{3\Delta}{4\beta}, & b = 1. \end{cases} \tag{3.25}$$

In the above embedding function, the quantization step-size Δ is scaled by $1/\beta$. However, the resultant distortion is compensated by the middle term in (3.25). Let us examine two extreme cases of β. If $\beta \to 1$, then (3.25) simplifies to (3.24), the standard dither modulation system. On the other hand, if $\beta \to 0$, then $X(k)$ is weakly quantized, while the watermark values become larger.

Further, the STDM differs from dither modulation and distortion-compensated dither modulation by incorporating a random vector, $\mathbf{v}$, and projection. We will use vector forms to represent watermark embedding and extraction functions for the STDM technique. Let $\mathbf{x}$ and $\mathbf{x}_{\mathrm{w}}$ be the vector form of $X(k)$ and $X_{\mathrm{w}}(k)$, then the STDM embedding function is given by

$$\mathbf{x}_{\mathrm{w}} = \begin{cases} \mathbf{x} - \left(\mathbf{x}^T\mathbf{v} - \left\lfloor \frac{\mathbf{x}^T\mathbf{v}}{\Delta} \right\rfloor \Delta\right)\mathbf{v} + \frac{\Delta\mathbf{v}}{4}, & b = 0, \\ \mathbf{x} - \left(\mathbf{x}^T\mathbf{v} - \left\lfloor \frac{\mathbf{x}^T\mathbf{v}}{\Delta} \right\rfloor \Delta\right)\mathbf{v} + \frac{3\Delta\mathbf{v}}{4}, & b = 1, \end{cases} \tag{3.26}$$

where the superscript ^{T} denotes vector or matrix transpose, and $\|\mathbf{v}\|_2^2 = 1$. Suppose the embedded watermark bit is "0", then the watermark extraction function yields

$$\begin{aligned} T &= \mathbf{x}_{\mathrm{w}}^T\mathbf{v} - \left\lfloor \frac{\mathbf{x}_{\mathrm{w}}^T\mathbf{v}}{\Delta} \right\rfloor \Delta \\ &= \left(\mathbf{x} - \left(\mathbf{x}^T\mathbf{v} - \left\lfloor \frac{\mathbf{x}^T\mathbf{v}}{\Delta} \right\rfloor \Delta\right)\mathbf{v} + \frac{\Delta\mathbf{v}}{4}\right)^T \mathbf{v} - \left\lfloor \frac{\mathbf{x}_{\mathrm{w}}^T\mathbf{v}}{\Delta} \right\rfloor \Delta \\ &= \left(\mathbf{x}^T\mathbf{v} - \mathbf{x}^T\mathbf{v} + \left\lfloor \frac{\mathbf{x}^T\mathbf{v}}{\Delta} \right\rfloor \Delta + \frac{\Delta}{4}\right) - \left\lfloor \frac{\mathbf{x}^T\mathbf{v}}{\Delta} \right\rfloor \Delta \\ &= \frac{\Delta}{4}, \end{aligned} \tag{3.27}$$

where $\left\lfloor \mathbf{x}_{\mathrm{w}}^T\mathbf{v}/\Delta \right\rfloor \Delta = \left\lfloor \mathbf{x}^T\mathbf{v}/\Delta \right\rfloor \Delta$, i.e., the quantization result of $\mathbf{x}_{\mathrm{w}}^T\mathbf{v}/\Delta$ is the same as the quantization result of $\mathbf{x}^T\mathbf{v}/\Delta$, because the watermark term is less than Δ and is discarded during quantization. In the above function, we observe that the correlation process is introduced in STDM watermark extraction, which is similar to that from SS based watermarking systems. Meanwhile, the result in (3.27) indicates that the threshold for comparison is also $\Delta/2$, the same as the threshold used in dither modulation.

3.1.4 Summary

As classical audio watermarking techniques, echo hiding, SS, and QIM have received tremendous research attentions during the research and development history of digital watermarking. This subsection has extracted the essential ingredients of these techniques, focusing on their corresponding watermark embedding and extraction

mechanisms. While echo hiding is originated from audiological discovery on human sensitivity towards echoes, SS and QIM techniques have more in-depth theoretical backgrounds, which could be seen from their corresponding information-theoretic analysis. In general, these fundamental techniques have provided excellent building blocks to develop audio watermarking systems with improved performance. In the following content of this section, we will introduce a series of advanced designs as well as recent novel perspectives on audio watermarking.

3.2 Advanced Designs

3.2.1 For Better Imperceptibility

We introduce two recent advanced designs of echo based audio watermarking systems for better imperceptibility in this subsection [12, 13]. Recall the generic echo kernel in (3.4)

$$h(n) = \delta(n) + \alpha w(n-d), \tag{3.28}$$

whose embedding distortion is controlled by α if we assume unit energy of echo filter $w(n)$. The proposed system in [12] first partitions the host audio signal frame into two channels, and then performs watermark embedding by adding the echoes in the first channel and subtracting the echoes from the second channel, a processing similar to patchwork embedding. However, the dual-channel scheme in [12] differs from the patchwork dual-channel scheme in the following ways. First, the partition of host signal samples is done by grouping even and odd samples respectively, instead of random selection in patchwork based method. Second, the partition is done on time domain host signal samples rather than mid-frequency bands of transform domain host signal. Define $s_{\text{even}}(n)$ and $s_{\text{odd}}(n)$ as

$$s_{\text{even}}(n) = [s(0), s(2), s(4), \ldots]^T, \tag{3.29}$$

$$s_{\text{odd}}(n) = [s(1), s(3), s(5), \ldots]^T, \tag{3.30}$$

then the dual-channel embedding function is given by

$$\begin{cases} s_{\text{w,even}}(n) = s_{\text{even}}(n) \otimes \left(\delta(n) + \frac{\alpha}{2}\tilde{p}(n-d)\right), \\ s_{\text{w,odd}}(n) = s_{\text{odd}}(n) \otimes \left(\delta(n) - \frac{\alpha}{2}\tilde{p}(n-d)\right), \end{cases} \tag{3.31}$$

where $\tilde{p}(n)$ is obtained by (3.9). The cepstrum of watermarked signal in the single-channel scheme using the MPN sequence is given by

$$\mathcal{C}_{X_{\text{w}}}(n) \approx \frac{\alpha}{2}\left[\tilde{p}(n-d) + \tilde{p}(-n-d)\right] + \mathcal{C}_X(n). \tag{3.32}$$

For the dual-channel scheme, the cepstra of the watermarked chennal signals are given by

$$\begin{cases} \mathcal{C}_{X_{\text{w,even}}}(n) \approx \frac{\alpha}{4}\left[\tilde{p}(n-d)+\tilde{p}(-n-d)\right]+\mathcal{C}_{X_{\text{even}}}(n), \\ \mathcal{C}_{X_{\text{w,odd}}}(n) \approx -\frac{\alpha}{4}\left[\tilde{p}(n-d)+\tilde{p}(-n-d)\right]+\mathcal{C}_{X_{\text{odd}}}(n). \end{cases} \tag{3.33}$$

From (3.33), the composite cepstrum is given by

$$\begin{aligned} \mathcal{C}(n) &= \mathcal{C}_{X_{\text{w.even}}} - \mathcal{C}_{X_{\text{w.odd}}} \\ &= \frac{\alpha}{2}\left[\tilde{p}(n-d)+\tilde{p}(-n-d)\right]+\mathcal{C}_{X_{\text{even}}} - \mathcal{C}_{X_{\text{odd}}}, \end{aligned} \tag{3.34}$$

where the key component term is identical to that from (3.32), but the interference term is weakened. This is because x_{even} and x_{odd} have similar values, especially for high sampling frequencies, which implies that $\mathcal{C}_{X_{\text{even}}}$ and $\mathcal{C}_{X_{\text{odd}}}$ have similar values too. We could observe two advantages of the proposed system in [12] over its single-channel counterpart in [11]. First, watermark imperceptibility is improved because the embedding strength is halved. Second, host signal interference rejection property is enhanced by the signal cancelling effect between even and odd host signal samples.

Recently, a further improvement of the imperceptibility of echo based audio watermarking system is proposed in [13], in which a filter design perspective is taken for the design of the echo kernel. In this way, the power spectrum of the echo kernel could be quantitatively tuned for optimal imperceptibility. The measurement of maximum power spectral margin (MPSM) is introduced here to characterize the global power spectral upper bound with respect to each frequency bin of the whole host audio clip. The block diagram of the systematic design of the echo filter $w(n)$ is shown in Fig. 3.5, where the frame subscript i is used for clarity of notations.

First, the host audio signal $s(n)$ is partitioned into 50% overlapped frames, and the power spectral density (PSD) function of each frame is normalized to dB sound pressure level (SPL) [25], i.e.,

$$P_i(k) = 90.302 + 20\log_{10}\left|\sum_{n=0}^{N-1}\text{Hann}(n)s_i(n)e^{-j2\pi kn/N}\right|, \tag{3.35}$$

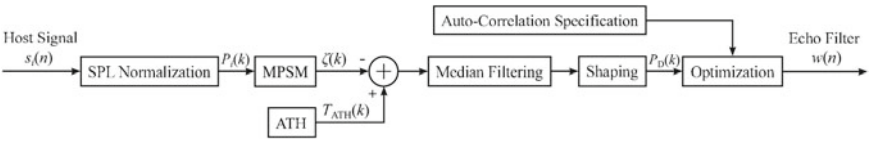


Fig. 3.5 Block diagram of systematic design of the echo filter [13]

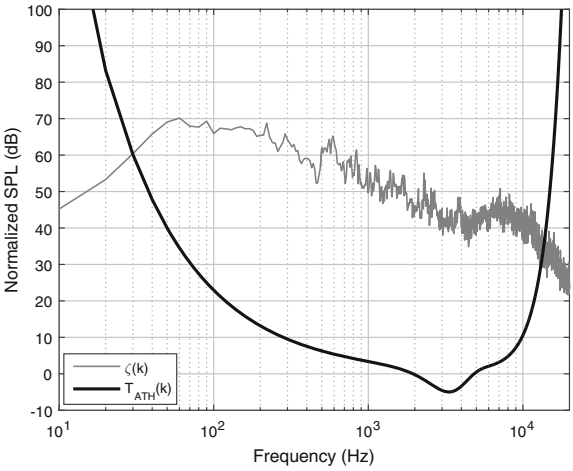


Fig. 3.6 An example of the MPSM of an audio clip

where Hann(n) is a Hanning window, N is the frame length, and the index k is bounded by K, the length of FFT. After that, the MPSM is obtained by selecting the maximum values across the frequency bins in each frame,

$$\zeta(k) = \max\left[P_0(k),\, P_1(k),\, P_2(k), \ldots\right]. \tag{3.36}$$

Meanwhile, the absolute threshold of hearing (ATH) of HAS is given by

$$T_{\text{ATH}}(k) = 3.64 f^{-0.8}(k) - 6.5 e^{-0.6(f(k)-3.3))^2} + 10^{-3} f^4(k) \tag{3.37}$$

in dB SPL, where $f(k) = k f_s/(1000\,\text{N})\,\text{Hz}$ with f_s being the sampling frequency. The desired PSD function, P_{D} is then obtained by the following tuning process

$$P_{\text{D}}(k) = \text{shap}\left\{\text{med}\left\{T_{\text{ATH}} - \zeta(k)\right\}\right\}, \tag{3.38}$$

which is median filtering followed by fine tuning. Examples of MPSM and the shaping process are provided in Figs. 3.6 and 3.7, respectively. Note that as shown in Fig. 3.7, the shaping process sets the desired PSD at extra low and high frequencies to a constant to facilitate more efficient optimization in subsequent procedures.

The imperceptibility and robustness of an echo based audio watermarking system depend on the echo filter $\alpha w(n)$ and the auto-correlation property of $w(n)$, respectively. Regarding the echo filter $\alpha w(n)$, although α generally controls the imperceptibility, the artifacts caused by adding echo watermarks are actually reflected in frequency domain with respect to HAS. Therefore, a series of echo based audio watermarking systems aim at shifting the PSD of $w(n)$ onto high frequency bands, a region HAS is less sensitive to. In [13], the PSD of $w(n)$ is not only pushed to extra high frequency bands, but also upper bounded by the ATH of HAS, yielding

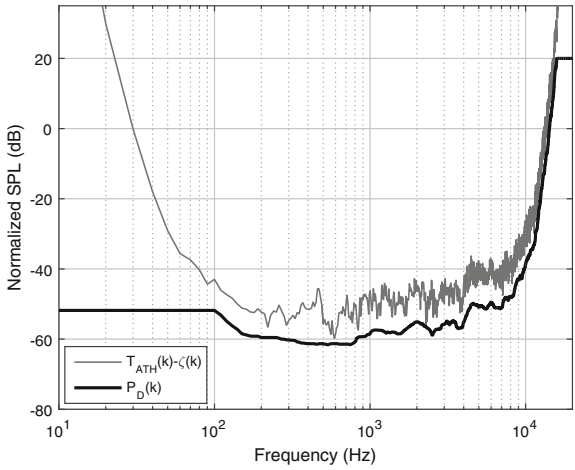


Fig. 3.7 An example of the shaping process and the corresponding output $P_{\mathrm{D}}(k)$

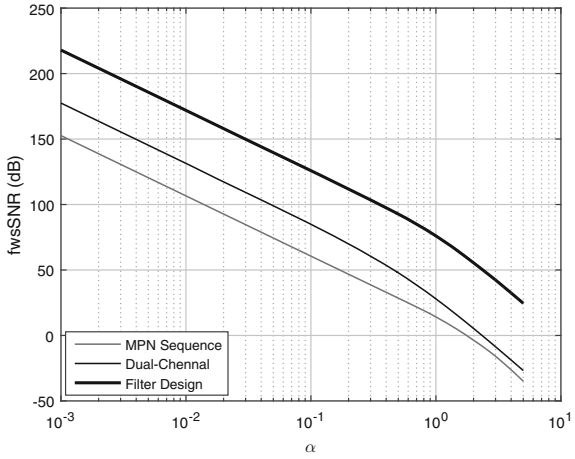


Fig. 3.8 Imperceptibility comparison among MPN sequence [11], dual-channel scheme [12], and filter design approach [13]. The fwsSNR curves are plotted against α, where $\tilde{p}(n)$ and $w(n)$ are normalized to have unit norm

substantially improved imperceptibility. The formulation of optimization problem to design $w(n)$ is generally given by

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

For simplicity, we omit the length derivations and realizations of the above problem formation, and refer the readers to [13] for more details. However, it should be noted that the designed echo filter $w(n)$ will let the PSD of the echo watermarks stay below the ATH of HAS, thank to the MPSM and shaping process. Therefore, the system proposed in [13] could achieve optimal imperceptibility among all the eisting works

on echo based audio watermarking. Figure 3.8 shows the imperceptibility comparison among the three echo based audio watermarking methods, MPN sequence [11], dual-channel scheme [12], and filter design approach [13].

3.2.2 For Improved Robustness

In this subsection, we address the most complicated problem of audio watermarking, i.e., robustness. The watermarked content may undergo unintentional or intentional attacks before a user is attempted to extract the watermarks. Due to the variety of attacks and uncontrollability of the watermarked content after distribution, it is very difficult for the watermarks to survive all kinds of attacks simultaneously. A comprehensive summary and classification of existing attacks are provided in [1]. While it tends to be a standard for modern watermarking systems to at least be robust against a series of normal (usually unintentional) attacks, such as A/D and D/A conversions, lossy compression, re-compression, and equalization, etc., it still remains a challenging problem for the system to deal with advanced attacks (usually intentional), especially a series of desynchronization attacks. In this subsection, we introduce several audio watermarking solutions robust against desynchronization attacks.

Although there has not been a rigorous definition for desynchronization attacks, we generally refer to them as the attacks that cause misalignments of watermark positions between the attacked and original watermarked copies. We introduce six kinds of such attacks here. (i) Cropping attack. This attack refers to the process of random cropping. However, the attacker may also be constrained to not crop the original content of the audio signal. (ii) Jittering. The jittering attack removes samples in an audio signal periodically. (iii) Time-shifting. This attack causes a time shift of the audio samples. A simple way of causing a time shift is to add a sequence of leading zeros to the audio signal. The next three desynchronization attacks are more difficult to deal with. (iv) Time-scale modification. This attack is also known as pitch-invariant time scaling, which alters the duration of the audio signal without altering the pitch and sampling frequency. (v) Time-invariant pitch scaling. In contrast to time-scale modification, pitch scaling alters the pitch of the audio signal while preserving duration of the signal. Sampling frequency is not altered either. (vi) Resampling. This attack modifies the duration and pitch of the audio signal at the same time. It is also known as speed scaling attack. All of these attacks could be easily implemented using handy audio processing software packages, such as Foobar2000, Audacity, and Adobe Audition. Extracting watermarks in copies having gone through such attacks is much more difficult and complicated. An illustrative example of the attacks iv, v and vi is depicted in Fig. 3.9, where a single sine wave is used as the signal under attack. It should be noted that time-scale modification and pitch scaling attacks are non-linear operations, which are most difficult to deal with.

An intuitive way, at the watermark embedder, to ensure certain synchronization of watermark bits, is to insert a sequence of synchronization bits before the embed-

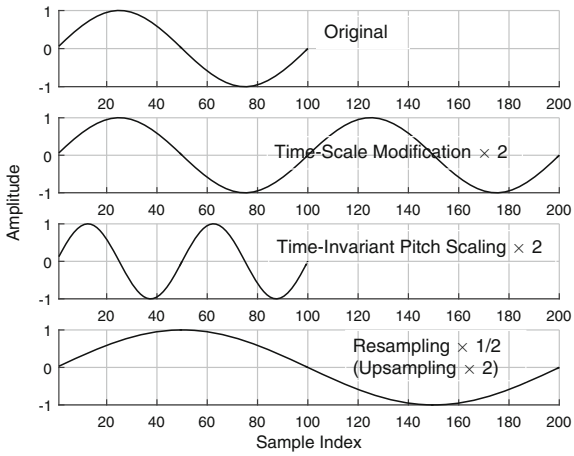


Fig. 3.9 Examples of different desynchronization attacks on a single sine wave

ding of watermark bits in each frame. This treatment has been incorporated in many existing works [24, 26–31]. Then during watermark extraction, only the estimated sequences whose leading bits are identical to the synchronization bits are considered to be successfully extracted watermarks. This can effectively reduce false positive rates. However, the effectiveness of such synchronization bits is limited. On one hand, the embedder has to vacate space to embed synchronization bits, which compromises embedding capacity. On the other hand, the synchronization bits are ineffective in dealing with advanced desynchronization attacks such as time-scale modification and time-invariant pitch scaling. Another attempt to deal with desynchronization attacks focuses on the receiver side without changing the watermark embedder. Specifically, the user is aware that the watermarked audio file may have undergone desynchronization attacks and tries to compensate the desynchronization in a way of exhaust search. In [32], the effectiveness of the template matching and exhaustive search methods, for image watermarking, is investigated, while only time shifting attack is considered. The same principle applies to audio.

A more effective way to deal with desynchronization attacks in audio watermarking could be considering the host audio signal as a single frame [22, 33] for watermark embedding. In [33], the multiplicative model for desynchronization attacks is changed to additive model via the mapping from linear scale to logarithmic scale. Therefore, both watermark embedding and extraction are performed in the logarithmic scale. Resynchronization is achieved via exhaustive search of a proposed tracking sequence concatenated with the DSSS sequence. A similar idea of exploiting logarithm scale is proposed in [22], with the incorporation of the patchwork technique. In order not to compromise embedding capacity, the synchronization bits in [22] are embedded as an overlay on the watermarked signal. During watermark extraction, the overlaid synchronization codes are first detected by correlating watermarked signal in logarithmic domain with the synchronization sequence in the same domain. The scaling factor is estimated by finding the peak value in the correlation function.

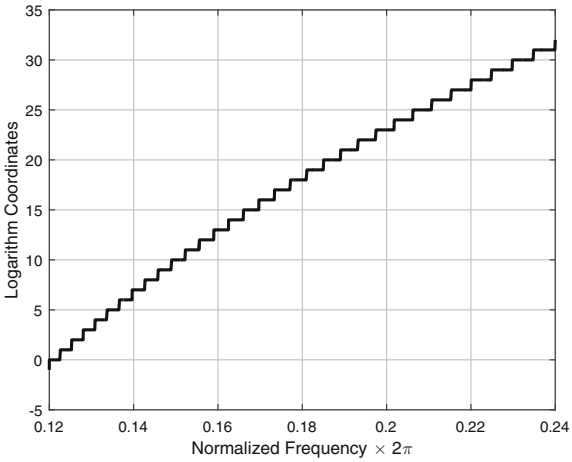


Fig. 3.10 The nonlinear mapping between normalized frequency and logarithm scale in [33]

Then, the watermarked signal is re-scaled according to the estimated scaling factor to restore the watermarked signal without desynchronization attacks (Fig. 3.10).

Apart from the above mentioned solutions that introduce extra components during either watermark embedding or extraction to facilitate watermark resynchronization, advanced desynchronization-resilient audio watermarking systems tend to investigate better features for watermark embedding, which is termed feature domain audio watermarking. Such type of features differ from log coordinate mapping (LCM) feature [33] and logarithmic DCT domain feature [22] by being robust against desynchronization attacks, i.e., the features remain unchanged even after desynchronization attacks, rather than being changed but detectable in [33] and [22]. One of such features is called robust audio feature in [34]. We will introduce the solution proposed in [34] in detail.

A robust audio segment extractor (RASE) is proposed in [34] to obtain the feature robust against desynchronization attacks. Let $y(n)$ be the unprocessed host audio samples with length N, then the RASE first calculates the first-order derivative of $y(n)$ via

$$y'(n) = y(n+1) - y(n-1), \tag{3.39}$$

where $n \in \{1, 2, \ldots, N-2\}$. Here, we assume $y'(0) = y'(N-1) = 0$. Then, the gradient signal $y'(n)$ is smoothed by a Gaussian filter $G(n)$,

$$G(n) = e^{-n^2/2\sigma^2}, \tag{3.40}$$

where σ is the standard deviation of the Gaussian function. The smoothed and squared gradient signal is then given by

$$z(n) = \big(y'(n) \otimes G(n)\big)^2. \tag{3.41}$$

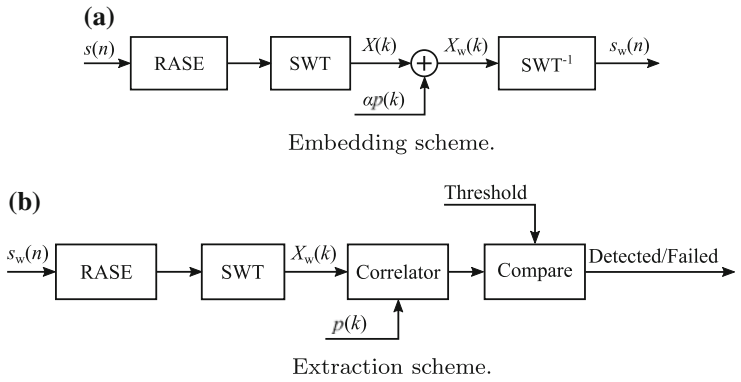


Fig. 3.11 Block diagram of the RASE based audio watermarking system [34]

After that, the obtained magnitude function $z(n)$ is reordered in descending order, denoted by z_{D}, i.e.,

$$[z_{\mathrm{D}}(0), z_{\mathrm{D}}(1), \ldots, z_{\mathrm{D}}(N-1)] = [z(l(0)), z(l(1)), \ldots, z(l(N-1))], \quad (3.42)$$

where $l(n)$ is the ordered index such that

$$z(l(0)) \geq z(l(1)) \geq \cdots \geq z(l(N-1)). \quad (3.43)$$

The set of order in $l(n)$ are called feature samples. Let the frame size be M, then these feature points are refined as follows to yield the robust segment indices for watermark embedding. Firstly, the first and last $M/2$ indices corresponding to both sides of the audio samples are removed. Secondly, for each leading index in $l(n)$, all the following indices which differ less than M from the leading index are removed. As a result, the refined $l(n)$ contains the indices of frame centers for watermark embedding. Finally, we obtain a series of M-sample frames, and each frame is centered at a refined leading index in $l(n)$. The block diagram of the RASE based audio watermarking system is shown in Fig. 3.11.

Thanks to the efficient feature extraction technique, the following watermark embedding procedure takes a simplified SS approach. Exploiting the shift invariant property, the stationary wavelet transform (SWT) is used instead of commonly used DWT. The embedding function for the ith frame is given by

$$X_{i,\mathrm{w}}(k) = X_{i,\mathrm{RASE}}(k) + \alpha \cdot p(k), \quad (3.44)$$

where $X_{i,\mathrm{RASE}}(k)$ consists of the robust host signal samples. Here, there are two differences from common SS watermark embedding. First, the spreading sequence $p(k)$ is drawn from Gaussian distribution instead of commonly used binary random sequence. Second, the watermark bits $b(i)$ are omitted in this scheme, meaning

Table 3.1 Comparison of robustness against desynchronization attacks

	LCM [33]	Localized [35]	RASE [34]	Patch [22]
Time scaling	± 20%	± 15%	± 50%	At least ± 20%
Pitch scaling	± 20%	Fail	± 50%	At least ± 20%
Resampling	± 25%	Fail	± 50%	50–200%

that the spreading sequence, $p(k)$, is the only embedded signal. The corresponding watermark extraction turned out to be a detection problem for $p(k)$. By neglecting embedding capacity, the spreading sequence is repeatedly embedded in the frames, and thus system robustness is substantially improved. The corresponding watermark detection is based on the following test statistic:

$$T_i = \sum_k X_{i,\mathrm{w}}(k)\,p(k) = \sum_k \left(X_{i,\mathrm{RASE}}(k)\,p(k) + \alpha \cdot p^2(k)\right). \tag{3.45}$$

The comparative results among several desynchronization-resilient audio watermarking systems are shown in Table. 3.1. It can be seen that single frame processing with logarithmic mapping [22, 33] can effectively deal with the most difficult desynchronization attacks. While these works could be considered as (logarithmic) feature domain audio watermarking, the more formal feature domain solution [34], which simply replaces the transform process in conventional transform domain watermarking system by feature extraction process, is believed to be a preferable approach for future research and development works on robust audio watermarking.

3.2.3 For Higher Capacity

The trade-off between embedding capacity and imperceptibility and the trade-off between embedding capacity and robustness are generally stated as follow. For a given cover audio clip, if one embeds more watermark bits into it, then the watermarks tend to be more perceptible and less robust, and vice versa. Let us take the SS technique as an example. In this frame based watermarking scheme, a single watermark bit is embedded into a single frame, yielding an embedding rate of 1 bit per frame. Therefore, the larger number of frames, the higher embedding capacity. However, using more frames will decrease the length of each frame, leading to shortened spreading sequence. It hence tends to weaken the asymptotic property of the spreading sequence, resulting in poorer correlation results. Consequently, the benchmark robustness is weakened. It is worth of noting here that in reversible audio watermarking, which will be detailed in Chap. 4, since robustness is not considered, the major research focus therein is the trade-off between imperceptibility and embedding capacity.

For robust audio watermarking, increasing embedding capacity without compromising imperceptibility and robustness is generally a difficult task. An intuitive solution to enlarging embedding capacity in frame based watermarking is to embed multiple watermark bits into a frame. By employing the code-division multiplexing (CDM) technique during watermark embedding, the original SS embedding function (3.11) alters to

$$X_{\mathrm{w}}(k) = X(k) + \alpha \sum_i b_i p_i(k), \tag{3.46}$$

where the index i does not exceed the frame length, and each watermark bit within the frame is assigned to a unique orthogonal spreading sequence, i.e.,

$$\sum_k p_i(k) p_j(k) = 0, \quad \forall i \neq j. \tag{3.47}$$

This is somewhat a straightforward treatment exploiting an advanced communication technique.

An alternative coding method in [36] associates a single spreading sequence with multiple watermark bits [36]. It differs from CDM by embedding one spreading sequence at a time, hence allowing more efficient control of imperceptibility. The block diagram of this system is shown in Fig. 3.12. Denote the vector form of the binary PN sequence, $p(k)$, $K \in \{0, \ldots, K-1\}$, as

$$\mathbf{p}_0 \triangleq \mathbf{p} = [p(0), p(1), \ldots, p(K-1)]^T. \tag{3.48}$$

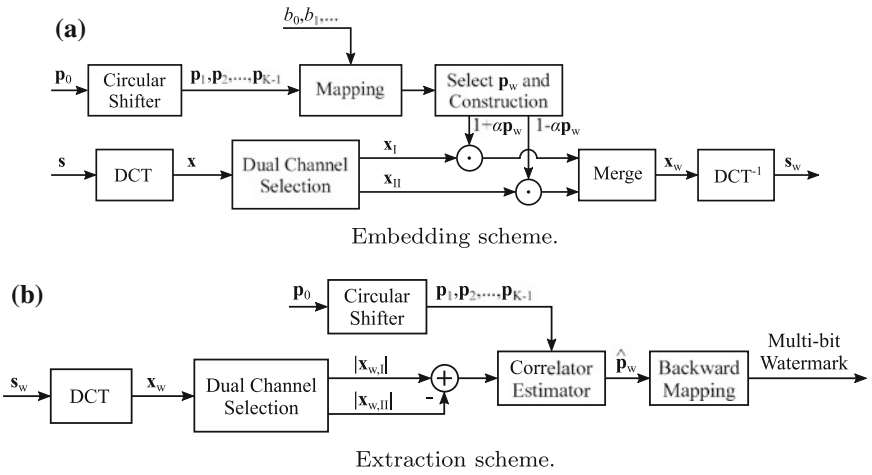


Fig. 3.12 Block diagram of a high capacity SS audio watermarking system [36]

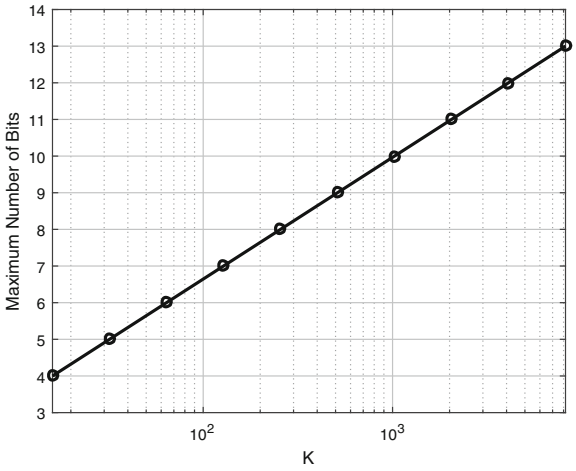


Fig. 3.13 Maximum number of watermark bits per frame as a function of K in [36]

Then, a series of $K-1$ PN sequences originates from $\mathbf{p}$ by introducing a circular shift once at a time, i.e.,

$$\mathbf{p}_1 = [p(K-1), p(0), p(1), \ldots, p(K-2)]^T, \tag{3.49}$$

$$\mathbf{p}_2 = [p(K-2), p(K-1), p(0), \ldots, p(K-3)]^T, \tag{3.50}$$

$$\vdots \tag{3.51}$$

$$\mathbf{p}_{K-1} = [p(1), p(2), p(3), \ldots, p(0)]^T. \tag{3.52}$$

Clearly, $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{K-1}$ are nearly mutual-orthogonal, and each of them has almost equal number of $+1$ s and -1 s. The length of a frame, K, determines the number of bits that could be represented by a single spreading sequence, and the maximum number of bits per spreading sequence is given by $\lfloor \log_2 K \rfloor$, which is shown in Fig. 3.13.

The subsequent watermark embedding mechanism takes a dual-channel approach similar to that in [12, 22], where the even and odd samples in a host audio frame are represented by

$$\mathbf{x}_{\mathrm{I}} = [X(0), X(2), \ldots, X(2K)]^T, \tag{3.53}$$

$$\mathbf{x}_{\mathrm{II}} = [X(1), X(3), \ldots, X(2K-1)]^T, \tag{3.54}$$

where signal length before dual channel selection is assumed to be $2K$. According to the to-be-embedded multi-bit watermark and the mapped spreading sequence, denoted by $\mathbf{p}_{\mathrm{w}}$, the watermark embedding functions are given by

$$\begin{cases} \mathbf{x}_{\mathrm{w,I}} = \mathbf{x}_{\mathrm{I}} + \alpha \mathbf{p}_{\mathrm{w}} \odot \mathbf{x}_{\mathrm{I}}, \\ \mathbf{x}_{\mathrm{w,II}} = \mathbf{x}_{\mathrm{II}} - \alpha \mathbf{p}_{\mathrm{w}} \odot \mathbf{x}_{\mathrm{II}}, \end{cases} \tag{3.55}$$

where $\odot$ represents element-wise product. After that, the two channel signals are merged into a single channel, followed by inverse DCT to obtain the watermarked signal in time domain.

During watermark extraction, the task is to detect the spreading sequence used for watermark embedding and then reversely map to the corresponding watermark bits. To achieve this, we first apply the same forward DCT transform and channel selection to obtain $\mathbf{x}_{\mathrm{w,I}}$ and $\mathbf{x}_{\mathrm{w,II}}$. Then we form the following expression:

$$\left|\mathbf{x}_{\mathrm{w,I}}\right| - \left|\mathbf{x}_{\mathrm{w,II}}\right| = \left|\mathbf{x}_{\mathrm{I}} + \alpha \mathbf{p}_{\mathrm{w}} \odot \mathbf{x}_{\mathrm{I}}\right| - \left|\mathbf{x}_{\mathrm{II}} - \alpha \mathbf{p}_{\mathrm{w}} \odot \mathbf{x}_{\mathrm{II}}\right| \tag{3.56}$$

$$= \left(\left|\mathbf{x}_{\mathrm{I}}\right| - \left|\mathbf{x}_{\mathrm{II}}\right|\right) + \alpha \mathbf{p}_{\mathrm{w}} \odot \left(\left|\mathbf{x}_{\mathrm{I}}\right| + \left|\mathbf{x}_{\mathrm{II}}\right|\right), \tag{3.57}$$

where the derivation from (3.56) to (3.57) is based on $|\mathbf{1} \pm \alpha \mathbf{p}_{\mathrm{w}}| < 1$, with $0 < \alpha < 1$ and $\mathbf{1}$ is an all-one vector with an appropriate length. The operator $|\cdot|$ here denotes element-wise absolute value. We note from (3.57) that i) the host signal interference tends to be attenuated by subtracting the magnitudes of the odd channel from those of the even channel, because of the high correlation between adjacent samples in the host audio signal. In fact, the higher the sampling frequency, the higher the correlation between $\mathbf{x}_{\mathrm{w,I}}$ and $\mathbf{x}_{\mathrm{w,II}}$, and the better the host interference cancellation. ii) In contrast to the host interference attenuation, the term $(|\mathbf{x}_{\mathrm{I}}| + |\mathbf{x}_{\mathrm{II}}|)$ strengthens the parameter of interest, i.e., $\mathbf{p}_{\mathrm{w}}$. Therefore, the index of $\mathbf{p}_{\mathrm{w}}$, denoted by $\hat{i}$, could be estimated by

$$\hat{i} = \underset{i \in \{0,1,\ldots,K-1\}}{\arg\max} \mathbf{p}_i^T \left(\left|\mathbf{x}_{\mathrm{w,I}}\right| - \left|\mathbf{x}_{\mathrm{w,II}}\right|\right). \tag{3.58}$$

Thanks to the dual-channel setting and the watermark bit coding via the special mapping, the system could enjoy improved imperceptibility, robustness, and embedding capacity, as compared to its conventional counterparts.

3.3 Novel Perspectives

In the above subsections, we have introduced the classical audio watermarking techniques and a series of recent development for different designing purposes. In this subsection, we will briefly discuss several recent attempts that design audio watermarking systems from unique perspectives.

In [37], an alternative solution to increasing capacity while preserving the imperceptibility and 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. In [38, 39], based on the formant enhancement technique, a novel speech watermarking system is proposed. Formants are 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 bit

"+1" is embedded by enhancing the second sharpest one. The applicability of this method to speech tampering detection is also discussed in [39]. A unique perspective for audio watermarking is seen in [40], 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. Despite conventional imperceptibility control largely relies on masking modeling and ATH, a series of novel methods of further studies on HAS with the purpose of discovering alternatively available "spaces" for audio watermarking are presented in [41–43], where cochlear delay characteristics have been utilized to embed watermarks. Cochlear delay is the non-uniform delay 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. Further, limitation of requiring host signal during watermark extraction has been broken by the improved system proposed in [44]. Note that in [43], the authors also proposed another subjective imperceptibility measurement method called post hoc test with analysis of variance (ANOVA).

References

1. Hua G, Huang J, Shi YQ, Goh J, Thing VLL (2016) Twenty years of digital audio watermarking - a comprehensive review. Signal Process 128:222–242
2. Gruhl D, Bender W (1996) Echo hiding. In: Proceedings of information hiding workshop, Cambridge, U.K., pp 295–315
3. Cox IJ, Kilian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Trans Image Process 6(12):1673–1687
4. Chen B, Wornell GW (2001) Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. IEEE Trans Inf Theory 47(4):1423–1443
5. Oh HO, Seok JW, Hong JW, Youn DH (2001) New echo embedding technique for robust and imperceptible audio watermarking. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 1341–1344
6. Kim HJ, Choi YH (2003) A novel echo-hiding scheme with backward and forward kernels. IEEE Trans Circuits Syst Video Technol 13(8):885–889
7. Childers DG, Skinner DP, Kemerait RC (1977) The cepstrum: a guide to processing. Proc IEEE 65(10):1428–1443
8. Oppenheim AV, Schafer RW (2004) From frequency to quefrency: a history of the cepstrum. IEEE Signal Process Mag 21(5):95–106
9. Ko BS, Nishimura R, Suzuki Y (2005) Time-spread echo method for digital audio watermarking. IEEE Trans Multimed 7(2):212–221
10. Hua G, Goh J, Thing VLL (2015) Cepstral analysis for the application of echo-based audio watermark detection. IEEE Trans Inf Forensics Secur 10(9):1850–1861
11. Xiang Y, Peng D, Natgunanathan I, Zhou W (2011) Effective pseudonoise sequence and decoding function for imperceptibility and robustness enhancement in time-spread echo-based audio watermarking. IEEE Trans Multimed 13(1):2–13
12. Xiang Y, Natgunanathan I, Peng D, Zhou W, Yu S (2012) A dual-channel time-spread echo method for audio watermarking. IEEE Trans Inf Forensics Secur 7(2):383–392

13. Hua G, Goh J, Thing VLL (2015) Time-spread echo-based audio watermarking with optimized imperceptibility and robustness. IEEE/ACM Trans Audio Speech Lang Process 23(2):227–239
14. Hu P, Peng D, Yi Z, Xiang Y (2016) Robust time-spread echo watermarking using characteristics of host signals. Electron Lett 52(1):5–6
15. Kirovski D, Malvar HS (2003) Spread-spectrum watermarking of audio signals. IEEE Trans Signal Process 51(4):1020–1033
16. Malvar HS, Florencio DAF (2003) Improved spread spectrum: a new modulation technique for robust watermarking. IEEE Trans Signal Process 51(4):898–905
17. Valizadeh A, Wang ZJ (2011) Correlation-and-bit-aware spread spectrum embedding for data hiding. IEEE Trans Inf Forensics Secur 6(2):267–282 ISSN 1556–6013
18. Yeo IK, Kim HJ (2003) Modified patchwork algorithm: a novel audio watermarking scheme. IEEE Speech Audio Process 11(4):381–386
19. Kang H, Yamaguchi K, Kurkoski BM, Yamaguchi K, Kobayashi K (2008) Full-index-embedding patchwork algorithm for audio watermarking. IEICE Trans E91-D(11):2731–2734
20. Kalantari NK, Akhaee MA, Ahadi SM, Amindavar H (2009) Robust multiplicative patchwork method for audio watermarking. IEEE Trans Audio Speech Lang Process 17(6):1133–1141
21. Natgunanathan I, Xiang Y, Rong Y, Zhou W, Guo S (2012) Robust patchwork-based embedding and decoding scheme for digital audio watermarking. IEEE Trans Audio Speech Lang Process 20(8):2232–2239
22. Xiang Y, Natgunanathan I, Guo S, Zhou W, Nahavandi S (2014) Patchwork-based audio watermarking method robust to de-synchronization attacks. IEEE/ACM Trans Audio Speech Lang Process 22(9):1413–1423
23. Bender W, Gruhl D, Morimoto N, Lu A (1996) Techniques for data hiding. IBM Syst J 35(3.4):313–336
24. Arnold M (2000) Audio watermarking: features, applications and algorithms. In: IEEE international conference on multimedia and expo, 2000, (ICME 2000), vol 2. IEEE, pp 1013–1016
25. Spanias A, Painter T, Atti V (2007) Audio signal processing and coding. Wiley, New Jersey chapter 5
26. Lie WN, Chang LC (2006) Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification. IEEE Trans Multimed 8(1):46–59
27. Megas D, Serra-Ruiz J, Fallahpour M (2010) Efficient self-synchronised blind audio watermarking system based on time domain and fft amplitude modification. Signal Process 90(12):3078–3092
28. Lei B, Soon IY, Tan EL (2013) Robust svd-based audio watermarking scheme with differential evolution optimization. IEEE Trans Audio Speech Lang Process 21(11):2368–2377
29. Wang XY, Zhao H (2006) A novel synchronization invariant audio watermarking scheme based on DWT and DCT. IEEE Trans Signal Process 54(12):4835–4840
30. Wang XY, Niu PP, Yang HY (2009) A robust, digital-audio watermarking method. IEEE Multimed 16(3):60–69
31. Wang XY, Qi W, Niu PP (2007) A new adaptive digital audio watermarking based on support vector regression. IEEE Trans Audio Speech Lang Process 15(8):2270–2277
32. Barni M (2005) Effectiveness of exhaustive search and template matching against watermark desynchronization. IEEE Signal Process Lett 12(2):158–161
33. Kang X, Yang R, Huang J (2011) Geometric invariant audio watermarking based on an lcm feature. IEEE Trans Multimed 13(2):181–190
34. Pun CM, Yuan XC (2013) Robust segments detector for de-synchronization resilient audio watermarking. IEEE Trans Audio Speech Lang Process 21(11):2412–2424
35. Li W, Xue X, Lu P (2006) Localized audio watermarking technique robust against time-scalecale modification. IEEE Trans Multimed 8(1):60–69
36. Xiang Y, Natgunanathan I, Rong Y, Guo S (2015) Spread spectrum-based high embedding capacity watermarking method for audio signals. IEEE/ACM Trans Audio Speech Lang Process 23(12):2228–2237
37. Fallahpour M, Megas D (2015) Audio watermarking based on fibonacci numbers. IEEE/ACM Trans Audio Speech Lang Process 23(8):1273–1282 ISSN 2329–9290

38. Wang S, Unoki M (2015) Speech watermarking method based on formant tuning. IEICE Trans Inf Syst E98-D(1):29–37
39. Wang S, Miyauchi R, Unoki M, Kim NS (2015) Tampering detection scheme for speech signals using formant enhancement based watermarking. J Inf Hiding Multimed Signal Process 6(6):1264–1283
40. Nishimura R (2012) Audio watermarking using spatial masking and ambisonics. IEEE Trans Audio Speech Lang Process 20(9):2461–2469
41. Unoki M, Hamada D (2010) Method of digital-audio watermarking based on cochlear delay characteristics. Int J Inno Comput Inf Control 6(3(B)):1325–1346
42. Unoki M, Miyauchi R (2013) Multimedia information hiding technologies and methodologies for controlling data, Method of digital-audio watermarking based on cochlear delay characteristics. IGI Global, pp 42–70
43. Unoki M, Imabeppu K, Hamada D, Haniu A, Miyauchi R (2011) Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics. J Inf Hiding Multimed Signal Process 2(1):1–23
44. Unoki M, Miyauchi R (2015) Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay. IEICE Trans Inf Syst E98-D(1):38–48

Chapter 4
Reversible Audio Watermarking

Abstract This section introduces the reversible audio watermarking (RAW) techniques. RAW can be classified according to how the inherent redundancies in the audio signals are utilized. Three types of redundancies and the corresponding RAW algorithms are reviewed. The companding (compression and expansion) approach employs the redundancy in representing a single sample. The prediction error expansion approach uses the redundancy between adjacent samples in time domain. The cochlear delay-based approach exploits the redundancy in human perception of sound.

4.1 Introduction

Although the modifications to an audio signal during watermark embedding are imperceptible, they are still not acceptable in some applications as the slight changes of the host signal may lead to different interpretation of the audio content. For example, if the watermarked audio is presented to court and is used for legal purpose, then the original host audio is mandatory. Using RAW techniques, one may completely recover the original audio once the hidden watermark is retrieved.

For RAW, the following requirements need to be considered.

Reversibility After the watermark is extracted, the original host audio signal should be recovered. If such a requirement can be relaxed in some applications, then partial recovery is also an option. For partial recovery, after the recovery operation, the difference between the recovered host signal and the original host signal should be smaller than that between the watermarked host signal and the original host signal.

Distortion The embedding induced distortion should be as small as possible. For RAW, segmental SNR and ODG are often used to measure the distortion [1–3].

Payload The payload of RAW should be as high as possible. It is usually measured by the number of embedded information bits (net payload) per host sample (bps). A typical RAW algorithm may provide payload ranging from 0.1 to 1bps.

Y. Xiang et al., *Digital Audio Watermarking*, SpringerBriefs in Signal Processing, DOI 10.1007/978-981-10-4289-8_4

Robustness Robustness to signal processing, such as noise addition, low-pass filtering and compression, is usually not required for RAW. There are, however, recent researches on improving the robustness of reversible watermarking systems [4, 5]. Nevertheless, the strength of attacks considered in RAW is much smaller than those encountered in robust watermarking.

Trade-offs might need to be considered among the above requirements, depending on the application scenario. For example, under the restriction of perfect reversibility, there is a trade-off between embedding distortion and payload. If robustness is also required, then less space is left for distortion and payload trade-off. On the contrary, if only partial reversibility is needed, then more space will be provided for distortion and payload trade-off.

The development of RAW has been almost in parallel with reversible image watermarking. Actually, some reversible watermarking algorithms for images can be transformed to deal with audio signals after appropriate modifications. However, audio signals have their own features, which makes RAW itself not a trivial extension of reversible image watermarking. Some aspects related to audio features are as follows.

- Audio signals are one-dimensional (1D) signals, as compared to two-dimensional (2D) images. This affects the way the redundancies can be used. For example, for image, one may use the information of local texture to help predict the sample but such feature is absent from audio.
- The number of bits per sample (or bit depth) of an audio is different from that of an image. For uncompressed pulse-code modulation (PCM) format for audio, the bit depth is usually 16 bits, while the typical bit depth for image is 8. Furthermore, the histogram features of audios and images are different. The histogram of an audio signal is close to a Laplacian distribution but the type of histogram of an image varies widely, from uniform ones to peaky ones having even more than one peak.
- The perception of audios is different from that of images. Some psychoacoustic facts can be leveraged to utilize the perceptual redundancies, such as cochlear delay.
- For speech signal, a typical but special audio signal, there is a generation model (source model) that can be exploited. However, such generation model is not available for most image signals.

The inherent redundancies in audio signals are the space that can be utilized to hide watermark in a reversible way. So different RAW algorithms rely on different assumptions of the source models for the host signal. Different models reflect different redundancies. The three commonly used redundancies in RAW are summarized below.

Redundancy in representing a single sample One can see from Fig. 4.1, if an audio sample can be considered as the output from a discrete memoryless source, then the uneven histogram distribution reflects the redundancy in representing a single sample. In the extreme cases, some bins are not occupied, meaning that these

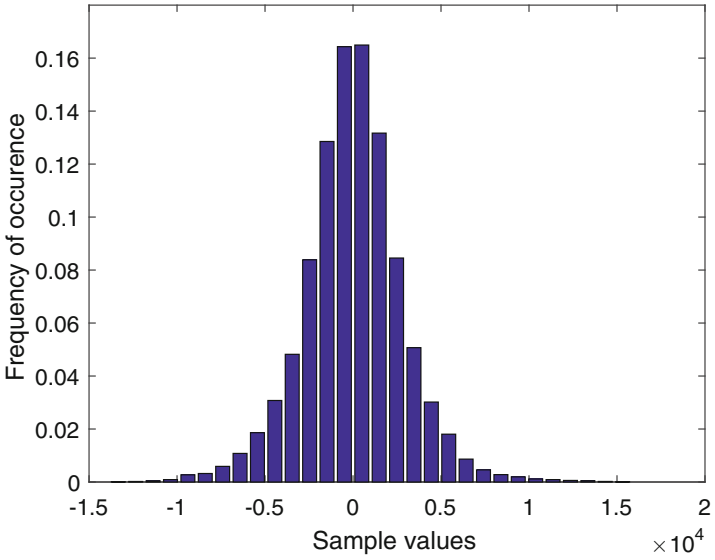


Fig. 4.1 One-dimensional histogram of an audio signal

bins can be used to embed watermark bits. In general, due to this uneven distribution, the output can be compressed to reserve space for reversible watermarking, which is the basis of lossless compression-based algorithms [6–8]. This is also the basis of histogram shift type reversible watermarking algorithm [9]. In RAW, the companding algorithm utilized this redundancy [2, 10]. This work is reviewed in Sect. 4.2.

Redundancy in correlation between samples Audio signal samples exhibit strong local correlation. Such redundancy can also be seen from the 2D or higher dimensional histogram. An example for 2D histogram is shown in Fig. 4.2. The histogram shift algorithm for higher dimensional histogram exploits such redundancy [11]. An alternative way to use this redundancy is to manipulate prediction error [12, 13], where the prediction error is expanded to allocate space for bit embedding. The RAW related this approach is presented in Sect. 4.3.

Redundancy from HAS The cochlear delay is an inherent feature of the cochlear in HAS. It is found that additional cochlear delay can be introduced into an audio signal without causing significant degradation. Such perceptual redundancy can be utilized in RAW and this work is reviewed in Sect. 4.4.

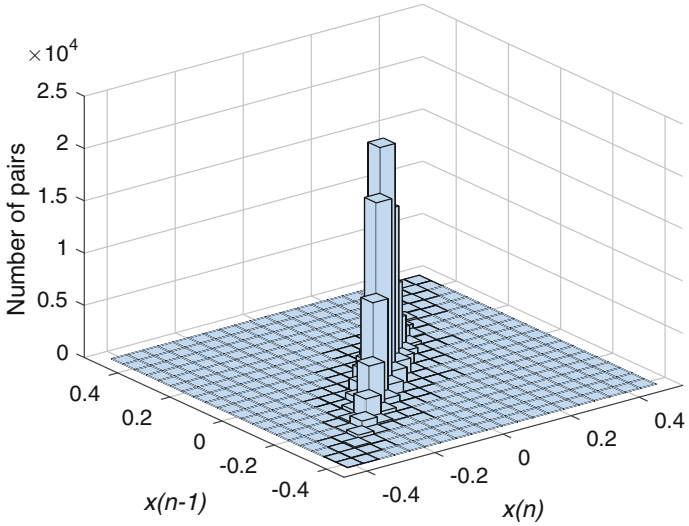


Fig. 4.2 Two-dimensional histogram of an audio signal

4.2 Companding-Based Algorithm

As outlined in Sect. 4.1, reversible watermarking exploits the redundancies in representing multimedia signals. The companding-based algorithm uses the redundancy in the first-order distribution of sample values [2, 10]. The samples are assumed to be generated from a discrete memoryless source, so the redundancy from the correlation between samples is ignored.

For most audio signals, the histogram is close to a Laplacian distribution. This is in contrast to the diverse histograms for image signals. Such a non-uniform distribution represents inherent redundancy that can be utilized for reversible watermarking. A special case is when certain bins of the histogram is not used in representing the audio signal, i.e., when the values corresponding to those bins do not appear in the audio samples. These bins can then be utilized by neighboring bins to hide watermark bits. For an audio signal, the histogram is concentrated around 0, and may not occupy the full range. So, companding (compression followed by expansion) can be utilized to allocate empty bins for watermark embedding. This is quite similar to histogram shift-based algorithm for reversible image watermarking. Both of them share the following features: (1) Only the first-order distribution is used, and (2) the histogram of the input is mapped to another histogram to allocate space for watermark embedding. However, the companding-based approach is different from the histogram shift-based approach in how this mapping is done.

4.2.1 Quantized Non-linear Companding

We assume that the bit depth (i.e., the number of bits needed to represent each sample) is D. For example, $D = 16$ is commonly used in PCM coding. Then, the range of samples is a set of integers $\mathbb{Z}_D = \left\{-2^{(D-1)}, \ldots, 0, \ldots, \ 2^{(D-1)} - 1\right\}$. In the trivial case where more than half of the histogram bins are not used, i.e., when $\max |x(n)| < 2^{D-2}$, less than 2^{D-1} bins are occupied and more than 2^{D-1} bins are empty. Thus, each occupied bin can be associated with one empty bin to hide one watermark bit. However, the condition $\max |x(n)| < 2^{D-2}$ generally does not hold. In this case, the original samples are compressed to the range $\mathbb{Z}_{D-1}$ using a nonlinear mapping C_Q. Note that after mapping, the set $\mathbb{Z}_{D-1}$ has smaller cardinality than that of $\mathbb{Z}_D$. So any mapping from $\mathbb{Z}_D$ to $\mathbb{Z}_{D-1}$ is lossy and non-revertible.

The compression function utilized by [2, 10] has the form

$$C_{\mathrm{Q}}\left(x(n)\right) = \mathrm{sgn}\left(x(n)\right) Q(\hat{x}(n)), \tag{4.1}$$

where

$$\hat{x}(n) = \left(2^{(D-2)} - 1\right)\left(\frac{|\tilde{x}(n)|^{\beta}}{|\tilde{x}(n)|^{\beta} + 1}\right)^{\frac{1}{\beta}}. \tag{4.2}$$

The function $Q(x)$ in (4.1) rounds to the nearest integer, and $\tilde{x}(n) = x(n)/2^{(D-2)}$ normalizes the input value. The parameter β is used to control the shape of the compression function. The companding error is

$$e(n) = x(n) - E_{\mathrm{Q}}\left(C_{\mathrm{Q}}\left(x(n)\right)\right), \tag{4.3}$$

where E_{Q} is the quantized expansion function. This error must be sent to the receiver in order to recover the original audio.

4.2.2 Reversible Watermarking Algorithm

The watermark embedding starts with calculating the companding error $e(n)$ for all samples. This error is then compressed by lossless compression $\mathcal{L}$ and concatenated with the payload $\mathbf{w}$ to form the watermark sequence

$$\mathbf{b} = \mathbf{w}||\mathcal{L}\left(\mathbf{e}\right).$$

The watermark sequence can be embedded by companding

$$x_{\mathrm{w}}(n) = 2C_{\mathrm{Q}}\left(x(n)\right) + b(n),$$

where the multiplication by a factor of 2 can be thought of as linear expansion.

At the decoder, the LSB of $\mathbf{x}_{\mathrm{w}}$ is first extracted to recover both the payload $\mathbf{w}$ and the compressed companding error $\mathcal{L}(\mathbf{e})$. After decompression, the companding error is recovered. Then, the compressed host sample can be recovered by

$$C_{\mathrm{Q}}(x(n)) = \frac{1}{2}\left(x_{\mathrm{w}}(n) - b(n)\right), \tag{4.4}$$

where the division by 2 can be realized by shifting towards LSB. Finally, the original host sample can be recovered by

$$x(n) = e(n) + E_{\mathrm{Q}}\left(C_{\mathrm{Q}}(x(n))\right) = e(n) + E_{\mathrm{Q}}\left(\frac{1}{2}\left(x_{\mathrm{w}}(n) - b(n)\right)\right). \tag{4.5}$$

4.3 Prediction Error Expansion

Unlike the companding-based or histogram shift-based approach, for prediction error expansion, the correlation between samples are exploited to hide data [3, 14–16]. The redundancy from correlation can be seen from the 2D histogram of audio signals, as shown in Fig. 4.2. The 1D histogram can be found by projecting the 2D histogram to each axis. Even though the corresponding 1D histogram may occupy all bins along their respective dimension, there are still empty bins available as revealed by the 2D histogram. As is done in lossless coding, such redundancy from time/spatial correlation can be utilized by processing prediction errors.

4.3.1 Prediction

Prediction of the current sample $x(n)$ is based on correlation between time domain samples. As a simple predictor, one may use the sample $x(n-1)$ as prediction of $x(n)$, i.e., $\hat{x}(n) = x(n-1)$. To better utilize the correlation and to better adaptive to the audio signal to be watermarked, an optimal predictor using more than one sample can be designed. The quantized linear predictor is of the following form:

$$\hat{x}(n) = \left\lfloor \sum_{i=1}^{P} c_i x(n-i) + \frac{1}{2} \right\rfloor. \tag{4.6}$$

Since the samples $x(k)$ with $k \geq n$ will be modified when embedding into $x(n+1)$, only the samples preceding $x(n)$ are utilized in the prediction. This predictor is causal and the samples preceding $x(n)$ are usually called *causal samples*.

In [14], a simplified form of (4.6) is used, where the coefficients are integers and hence rounding to nearest integer is no longer needed. The predictor in [14] is

$$\hat{x}(n) = c_1 x(n-1) - c_2 x(n-2) + c_3 x(n-3),$$

where $c_i \in \{0, 1, 2, 3\}$. Thus, there are totally $4^3 = 64$ different coefficient vectors $\mathbf{c} = [c_1, c_2, c_3]$. Five types of audio signals are considered including blues, classical, country, folk and popular music. For each type of music, the coefficient vector $\mathbf{c}$ is found by minimizing the overall energy of the prediction error. For example, for blues music, the optimal coefficient is found to be $\mathbf{c} = [1, 0, 0]$. So, the predictor is adaptive to the type of the music to be watermarked but the searching space of optimal coefficients is limited.

In general, the predictor should be optimized with respect to distortion and embedding rate. In [16], the coefficients of the quantized linear operator is optimized to maximize the following objective function:

$$F(\mathbf{c}) = 10 \log_{10} \left(\frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=1}^{N} [x(n) - x_{\mathrm{w}}(n)]^2} \right) \cdot \frac{\ell}{N}, \tag{4.7}$$

where ℓ is the number of payload bits embedded. Note that both $\mathbf{x}_{\mathrm{w}}$ and ℓ vary with $\mathbf{c}$. A differential evolution algorithm is designed to find the optimal $\mathbf{c}$. Note that this objective function can only be evaluated after the payload is embedded into the host. So, each evaluation of the objective function involves one pass of embedding. The computational complexity is extremely high.

As a compromise between computational complexity and performance, the predictor is usually optimized independently of the watermark embedding. In [3, 15], Burg method is applied to estimate the coefficients for an eight-order predictor. The optimal coefficients are either adaptive to the whole audio [15] or adaptive to smaller segments ranging from 0.25 to 2.0s [3].

4.3.2 Error Expansion

After obtaining the prediction error $e(n)$, one can expand $e(n)$ to embed watermark bit. In order to control the embedding distortion and embedding rate, an appropriate threshold T is chosen. If $|e(n)| \geq T + 1$, then the bins are shifted to reserve space for expansion:

$$e_{\mathrm{w}}(n) = \begin{cases} e(n) + T + 1, & \text{if } e(n) \geq T + 1, \\ e(n) - T, & \text{if } e(n) \leq -(T + 1). \end{cases} \tag{4.8}$$

Now there is an empty bin associated with each bin $|e(n)| \leq T$, so each of them can be expanded and the watermark bits are inserted into LSB after expansion

$$e_{\mathrm{w}}(n) = 2e(n) + b. \tag{4.9}$$

The expanded error is finally added to the predicted value to obtain the watermarked host: $x_w(n) = \hat{x}(n) + e_w(n)$.

At decoder, the prediction error can be calculated as $e_w(n) = x_w(n) - \hat{x}(n)$. Then, one must determine if the current sample was embedded with watermark or was shifted to reserve space. According to (4.8), if $e(n)$ was shifted then either $e_w(n) > 2T$ or $e_w(n) < -2T + 1$. While, from (4.9), if $e(n)$ was expanded, then $-2T \leq e_w \leq 2T + 1$. So, if $x_w(n)$ is embedded with watermark, then the watermark bit can be extracted as LSB of $e_w(n)$ as

$$b(n) = e_w(n) - 2\left\lfloor \frac{e_w(n)}{2} \right\rfloor. \tag{4.10}$$

The original host sample can be recovered from

$$x(n) = \frac{x_w(n) + \hat{x}(n) - b(n)}{2}. \tag{4.11}$$

For shifted samples, the original sample can be recovered as:

$$x(n) = \begin{cases} x_w - T - 1, & \text{if } e_w(n) > 0, \\ x_w + T, & \text{if } e_w(n) < 0. \end{cases} \tag{4.12}$$

In [14], the threshold is set as $T = \infty$. As a result, every $e(n)$ is expanded, as long as such expansion does not produce overflow or underflow. To prevent overflow or underflow, one must make sure that the watermarked sample is limited in the range $x_w(n) \in [-2^{(D-2)}, 2^{(D-1)} - 1]$, which in turn implies that the prediction error should satisfy:

$$-\frac{2^{(D-1)} + \hat{x}(n) + 1}{2} \leq e(n) \leq \frac{2^{(D-1)} - 1 - \hat{x}(n)}{2}. \tag{4.13}$$

While an expansion factor of 2 is utilized in (4.9), a smaller expansion factor can result in lower perceptual distortion. Hence, the algorithm in [3] uses a fractional expansion factor $\alpha \in (1, 2]$, in order to attain more flexible control of perceptual quality and embedding rate.

4.3.3 *Dealing with Overflow and Underflow*

For prediction error expansion-based RAW, when the condition (4.13) does not hold, one has to give up embedding into the current sample $x(n)$. So, the decoder has to decide whether the current $x_w(n)$ is a result of embedding or not. Since such a decision is based on the original sample $x(n)$ which is not available to the decoder before recovering the host signal, additional information needs to be transmitted from the embedder to the decoder. There are two commonly used techniques to convey this information to the decoder.

4.3.3.1 Location Map

The first technique is the *location map*. The location map $M(n)$ is a binary sequence which has the same dimension as that of the host signal, i.e., for audio signals, the location map is a 1D vector. This map indicates whether the current sample is embedded with watermark or not:

$$M(n) = \begin{cases} 1, \text{ if } x(n) \text{ is embedded with watermark} \\ 0, \text{ otherwise} \end{cases} \tag{4.14}$$

with $n = 0, \ldots, N-1$. It is then compressed by run-length coding or other lossless coding. Since the location map is needed during watermark recovery, it must be embedded into the host such that (1) watermark extraction does not need any parameters, and (2) watermark can be extracted before recovering the host. A feasible solution is to embed the location map to the end of the samples using LSB replacement. The original LSBs must be appended to the payload and embedded as ordinary watermarks. This technique is used in [14].

4.3.3.2 Flag Bit

An alternative to location map is using *flag bit* [13, 15]. As can be seen from (4.8) and (4.9), the maximum modification to a host sample is $T+1$ or $-T$. Let the whole dynamic range of the host sample be partitioned into the outer region Ω_{O} and inner region Ω_{I} as:

$$\Omega_{\mathrm{O}} = \left[-2^{D-1}, -2^{D-1} + T - 1\right] \cup \left[2^{D-1} - 1 - T, 2^{D-1} - 1\right], \tag{4.15}$$

$$\Omega_{\mathrm{I}} = \left[-2^{D-1} + T, 2^{D-1} - 2 - T\right]. \tag{4.16}$$

If $x(n) \in \Omega_{\mathrm{O}}$ and considering the worst case, then modifying $x(n)$ may lead to overflow/underflow. If overflow/underflow occurs, the embedder gives up modifying $x(n)$. At the receiver side, observing $x_{\mathrm{w}}(n)$, however, the decoder is unable to decide whether $x_{\mathrm{w}}(n) \in \Omega_{\mathrm{O}}$ is a result of the modification of $x(n)$ or is because the embedder gave up modifying $x(n)$. The flag bit is assigned a value 1 if the current sample is embedded with watermark, otherwise it is assigned with a value 0. Then the flag bit is inserted into the watermark sequence and is embedded into the previous embeddable sample. At the decoder, if $x_{\mathrm{w}}(n) \in \Omega_{\mathrm{O}}$, then one check the previous extracted bit. If it is 1, then the current sample was watermarked. Otherwise, it means that the current sample is not modified during embedding. All the flag bits can also be collected into a vector and embedded into the host using LSB replacement. For the same T and dynamic range of the audio signal, flag bits requires much less space than location map.

Overflow and underflow is a direct result of histogram mapping that are trying to allocate space for watermark embedding. From this perspective, both the companding error and over/underflow are caused by the many-to-one mapping in histogram mapping.

4.4 Cochlear Delay-Based Algorithm

The cochlear delay-based algorithm explores the redundancy of the signal to HAS. As outlined in Chap. 2, the HAS is insensitive to the phase delay that is similar to the inherent cochlear delay [17–20]. An all-pass filter can be used to introduce such phase delay. Since this filter is revertible, the modification to the host audio signal can be recovered by inverse-filtering. In this sense, an RAW algorithm can be designed. However, due to frame processing, inverse-filtering cannot recover the original audio exactly. So, this scheme is not a reversible watermarking in a strict sense but can be regarded as a reversible watermarking in a wide sense.

Let $\mathcal{E}$ denote the embedding operation, $\mathcal{D}$ and $\mathcal{R}$ denote the decoding and recovery operations, respectively. In addition, $D(x, y)$ denotes the distortion between signals x and y. A reversible watermarking scheme is *wide sense reversible* if

$$D\left(\mathbf{x}, \mathcal{R}\left(\mathcal{E}\left(\mathbf{x}, \mathbf{b}\right)\right)\right) < D\left(\mathbf{x}, \mathcal{E}\left(\mathbf{x}, \mathbf{b}\right)\right), \tag{4.17}$$

where $\mathbf{x}$ and $\mathbf{b}$ are the original host audio signal and the watermark bits, respectively. For the strict sense reversible watermarking, $D\left(\mathbf{x}, \mathcal{R}\left(\mathcal{E}\left(\mathbf{x}, \mathbf{b}\right)\right)\right) = 0$.

4.4.1 Watermark Embedding

Artificial cochlear delay can be introduced by a first-order all-pass filter

$$H(z) = \frac{-\beta + z^{-1}}{1 - \beta z^{-1}}, \tag{4.18}$$

where the parameter β can be optimized by minimizing the difference between the group delay of this filter and the cochlear delay characteristics. Using the least mean square (LMS) algorithm, the optimum β was found to be 0.795 if α is set as 0.1 [19].

The watermark embedding is frame-based, i.e., one bit is embedded into one frame. The watermark bits are embedded into the host signal by filtering the host signal with different cochlear delay filters. Let the length of the frames be L. For each frame, two cochlear filters, H_1 and H_0, are used to embed bit '1' and bit '0', respectively. Then the watermarked signal for the i-th frame is obtained as

$$x_{\mathrm{w}}(n) = -\beta_{b_i} x(n) + x(n-1) + \beta_{b_i} x_{\mathrm{w}}(n-1), \tag{4.19}$$

where $(i-1)L < n \leq iL$ and the parameter β_{b_i} is chosen by the watermark bit b_i. Due to this frame-based filtering, abrupt change may occur across the frame boundary. To avoid such effect, the last few (for example, 20) samples of each frame is determined by spline interpolation from neighboring samples.

As for the choice of β_0 and β_1, one must find a trade-off between watermark recovery and perceptual quality. First, both of the values should be close to the optimal value 0.795 to minimize perceptual distortion. Second, the difference between β_0 and β_1 should be large enough to facilitate watermark extraction. In [1], they were set as $\beta_0 = 0.795$ and $\beta_1 = 0.865$.

4.4.2 Watermark Extraction

Due to the usage of an all-pass filter of the form as in (4.18), one zero and one pole on the z-plane are introduced into the watermarked signal. So the watermark bits can be extracted by detecting the poles or zeros. Since the chirp z-transform can be used to evaluate z-transform on any arbitrary point on the z-plane, it can be used to extract watermark bits.

Let $A = A_0 e^{j2\pi\theta_0}$ be the starting point of the chirp z-transform, and $W = W_0 e^{j2\pi\phi_0}$ determines how the path spirals on the z-plane. Then the chirp z-transform for the i-th frame of the watermarked signal is

$$s_{i,\mathrm{w}}(k) = \sum_{n=0}^{L-1} x_{i,\mathrm{w}}(n) z_k^{-n}, \tag{4.20}$$

where $z_k = AW^{-k}$. To evaluate the z-transform on the two zeros $1/\beta_0$ and $1/\beta_1$, the starting point of the chirp z-transform can be chosen as $A = 1/\beta_0$ and $A = 1/\beta_1$, respectively. Let $s^0_{i,\mathrm{w}}(k)$ and $s^1_{i,\mathrm{w}}(k)$ be the corresponding chirp z-transform, then the extracted watermark can be obtained as

$$\hat{b}_i = \begin{cases} 0, \text{ if } s^0_{i,\mathrm{w}}(0) < s^1_{i,\mathrm{w}}(0), \\ 1, \text{ if } s^0_{i,\mathrm{w}}(0) \geq s^1_{i,\mathrm{w}}(0). \end{cases} \tag{4.21}$$

4.4.3 Host Recovery

For the i-th frame, after extracting the watermark bits $\hat{b}_i$, one may apply the inverse filter $H^{-1}_{\hat{b}_i}$ to the watermarked audio $\mathbf{x}_{i,\mathrm{w}}$ to 'recover' the original host audio. Since neighboring frames may be filtered by different inverse filters, abrupt change may occur across frame boundary. So, spline interpolation is used to smooth the last few samples of each frame. Due to the interpolation during embedding and host recovery, the inverse filtering cannot recover the original host audio exactly. Nevertheless, as

reported in [1], the recovered host has better perceptual quality than the watermarked host. For example, before recovery, the PEAQ value is below −1 (perceptible but not annoying) when the embedding rate is 8 bits/s. However, after host signal recovery, the PEAQ remains above −1 till 64 bits/s.

4.5 Remarks

Compared to reversible image watermarking, RAW, especially the reversible watermarking for speech signal, did not attract much attention to researchers. Several interesting research topics deserve further development.

The trade-off between reversibility and robustness is an interesting problem. Except for the work of cochlear delay-based algorithm, almost all reversible watermarking algorithms require perfect recovery of the original host after watermark extraction. It is desirable for some robust watermarking systems to have a certain level of reversibility after mild attacks, i.e., satisfying the condition of weak reversibility in (4.17). Related work from robust reversible watermarking for images can be extended to address this problem, such as [4, 5].

Reversible speech watermarking is another topic that may need more work. For speech signal, the speech generation model, i.e., the model for human articulatory system, can be utilized for better prediction. Such works are mature and widely used in speech coding. The use of perceptual redundancy can be further explored and combined with prediction error expansion algorithms. The current algorithm of this type for RAW is the cochlear delay approach, which provides only extremely low embedding rate, currently lower than 10 bps for ODG above −1.

Finally, the redundancies in time-frequency plane may be further utilized for RAW.

References

1. Unoki M, Miyauchi R (2008) Reversible watermarking for digital audio based on cochlear delay characteristics. In: International conference on intelligent information hiding and multimedia signal processing 2008, IIHMSP'08, pp 314–317
2. van der Veen M, Bruekers F, van Leest A, Cavin S (2003) High capacity reversible watermarking for audio. Proc SPIE 5020:1–11
3. Nishimura A (2016) Reversible audio data hiding based on variable error-expansion of linear prediction for segmental audio and G.711 speech. IEICE Trans Inf Syst 99–D(1):83–91
4. An L, Gao X, Li X, Tao D, Deng C, Li J (2012) Robust reversible watermarking via clustering and enhanced pixel-wise masking. IEEE Trans Image Process 21(8):3598–3611
5. Thabit R, Khoo BE (2014) Capacity improved robust lossless image watermarking. IET Image Process 8(11):662–670
6. Goljan M, Fridrich JJ, Du R (2001) Distortion-free data embedding for images. In: Proceedings of the 4th international workshop on information hiding, IHW'01, Springer, London, UK, pp 27–41

7. Kalker T, Willems FMJ (2003) Capacity bounds and constructions for reversible data-hiding. In: International society for optics and photonics electronic imaging 2003, pp 604–611
8. Celik MU, Sharma G, Tekalp AM, Saber E (2005) Lossless generalized-LSB data embedding. IEEE Trans Image Process 14(2):253–266
9. Ni ZC, Shi YQ, Ansari N, Su W (2006) Reversible data hiding. IEEE Trans Circuits Syst Video Technol 16(3):354–362
10. van der Veen M, van Leest A, Bruekers F (2003) Reversible audio watermarking. In: Audio engineering society convention 114
11. Li X, Li B, Yang B, Zeng T (2013) General framework to histogram-shifting-based reversible data hiding. IEEE Trans Image Process 22(6):2181–2191
12. Tian J (2003) Reversible data embedding using a difference expansion. IEEE Trans Circuits Syst Video Technol 13(8):890–896
13. Dragoi IC, Coltuc D (2014) Local-prediction-based difference expansion reversible watermarking. IEEE Trans Image Process 23(4):1779–1790
14. Yan D, Wang R (2008) Reversible data hiding for audio based on prediction error expansion. In: International conference on intelligent information hiding and multimedia signal processing, 2008, IIHMSP'08
15. Nishimura A (2011) Reversible audio data hiding using linear prediction and error expansion. In: Seventh international conference on intelligent information hiding and multimedia signal processing (IIH-MSP), 2011, pp 318–321
16. Wang F, Xie Z, Chen Z (2014) High capacity reversible watermarking for audio by histogram shifting and predicted error expansion. Sci World J 2014:1–7 Article ID 656251
17. Tanaka S, Unoki M, Aiba E, Tsuzaki M (2008) Judgment of perceptual synchrony between two pulses and verification of its relation to cochlear delay by an auditory model. Jpn Psychol Res 50(4):204–213
18. Unoki M, Imabeppu K, Hamada D, Haniu A, Miyauchi R (2011) Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics. J Inf Hiding Multimed Signal Process 2(1):1–23
19. Unoki M, Hamada D (2008) Audio watermarking method based on the cochlear delay characteristics. In: International conference on intelligent information hiding and multimedia signal processing, pp 616–619
20. Uppenkamp S, Fobel S, Patterson RD (2001) The effects of temporal asymmetry on the detection and perception of short chirps. Hear Res 158(12):71–83

Chapter 5
Audio Watermarking with Cryptography

Abstract Watermarking and cryptography are two closely related research subjects with a similar general objective of protecting important information in digital form. In terms of protecting multimedia content, encryption is applied to ensure the transmission of content information is secure, while watermarking is used for further protection after the content has been decrypted by authorized users. This chapter addresses encryption-related audio watermarking. First, we discuss several audio watermarking schemes with the incorporation of cryptography. After that, we will look into an alternative notion of leveraging the theory of compressive sensing for watermarking system design. It has been shown that the compressive sensing process is similar to an encryption process.

5.1 Watermark Embedding with Encryption

In this section, we introduce three audio watermarking schemes that work with encryption to enhance system security [1–3]. In a secure communication system, the sender encrypts the message before transmission to ensure security and privacy of the information to be distributed via the communication network. With a shared key at authorized or trusted receivers, the received signal could be effectively decrypted. Further, after the content has been decrypted, watermarking takes effect in protecting the decrypted content, serving as evidence once the content has been illegally manipulated (e.g., fragile watermarking) or distributed (e.g., robust watermarking). For simplicity and to emphasize the main focus of this book, the background of applied cryptography is not mentioned, and we refer the readers to [4] for more details. The encryption techniques considered here are all homomorphic encryptions such that the operations on ciphertext correspond with the same operations on plaintext.

The concept of partial encryption is exploited in [2] to develop an effective audio watermarking scheme for systems such as a paid on line music portal. The block diagram of partial encryption is shown in Fig. 5.1, where the host audio signal is partitioned into two portions, one for encryption and the other unchanged. After encryption, the two portions are merged to construct a partially encrypted audio

Y. Xiang et al., *Digital Audio Watermarking*, SpringerBriefs
in Signal Processing, DOI 10.1007/978-981-10-4289-8_5

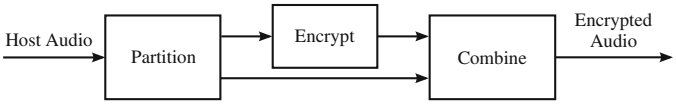


Fig. 5.1 Block diagram of partial encryption

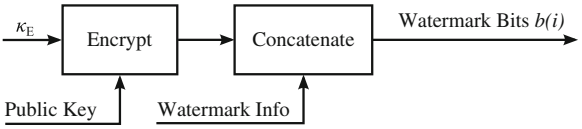
signal. Further, a system is called a commutative watermarking and encryption (CWE) system if watermarking and encryption are applied consecutively and interchangeably. In such a system, the same watermarked and encrypted audio signal is obtained regardless of either watermarking or encryption is applied first.

To ensure the commutative property, the discrete wavelet transform (DWT) coefficients of the host audio samples are partitioned, and watermarking and encryption are respectively applied to the non-overlapping portions of the coefficients. Let the DWT coefficients of a host audio signal be $X_{\mathrm{DWT}}(k)$, then the partial encryption process is given as follows:

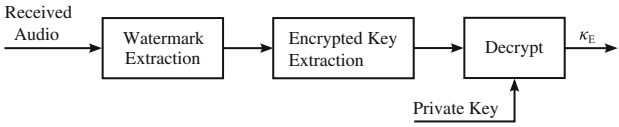
$$X_{\mathrm{I,E}}(k) = \mathbb{E}\{X_{\mathrm{I}}(k), \kappa_{\mathrm{E}}\}, \tag{5.1}$$

$$X_{\mathrm{E}}(k) = \mathrm{combine}\{X_{\mathrm{I,E}}(k), X_{\mathrm{II}}(k)\}, \tag{5.2}$$

where $\mathbb{E}\{\cdot\}$ denotes encryption, κ_{E} is the symmetric encryption key, and X_{I} and X_{II} are the two portions after partition such that $X_{\mathrm{I}} \cup X_{\mathrm{II}} = X_{\mathrm{DWT}}$, and $X_{\mathrm{I}} \cap X_{\mathrm{II}} = \emptyset$. $X_{\mathrm{E}}(k)$ is then used for watermark embedding. The key distribution framework is based on a client-server environment, in which registered clients have access to their respect unique keys for encryption. Unregistered users only have access to partially encrypted audio content with highly degraded audio quality. The key management mechanism as well as watermark construction are aggregated as shown in Fig. 5.2 via public-key cryptography. The registered users will receive a pair of public and private keys. The public key is used to encrypt the symmetric encryption key κ_{E}.



(a) Symmetric key processing and watermark construction.



(b) Symmetric key extraction.

Fig. 5.2 Key management system of the partial encryption audio watermarking system [2]

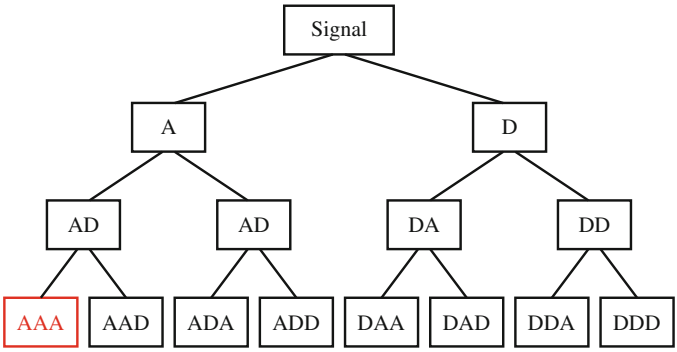


Fig. 5.3 3-level DWT. "A": approximation. "D": detail. Red coefficients are for watermarking

The encrypted key is then concatenated with watermark information to form the whole watermark sequence to be embedded into the host signal. At the decoder side, the watermark sequence is extracted and the portion for encrypted symmetric key is obtained. After that, the user uses the private key to decrypt the encryption key κ_{E} which is then used to decrypt the encrypted portion of host audio signal.

Details of the watermark embedding and encryption scheme are as follows:

1. Perform a 3-level DWT on the host audio signal. According to Fig. 5.3, $X_{\mathrm{DWT}}(k)$ is evenly divided into 8 groups, and we denote the lowest frequency group, "AAA", as $X(k)$, which is used for watermark embedding. Then, $X_{\mathrm{I}}(k)$ is chosen from the remaining 7 groups for encryption.
2. According to the length of $X(k)$, denoted by K, and the length of the constructed watermark bits $b(i)$, denoted by L, we partition $X(k)$ into frames of length $\lfloor K/L \rfloor$. For simplicity, assume K is evenly dividable by L, and denote

$$\mathbf{x} = [\mathbf{x}_0^T, \mathbf{x}_1^T, \ldots, \mathbf{x}_{L-1}^T]^T = [X(0), X(1), \ldots, X(K-1)]^T. \tag{5.3}$$

3. Obtain the sample mean of each frame, i.e.,

$$\bar{\mathbf{x}}_i = \frac{L\mathbf{x}_i^T\mathbf{1}}{K}, \tag{5.4}$$

and perform watermark embedding using the following function

$$\mathbf{x}_{i,\mathrm{w}} = \begin{cases} \mathbf{x}_i - \bar{\mathbf{x}}_i + \alpha\bar{\mathbf{x}}_i, & b(i) = +1, \\ \mathbf{x}_i - \bar{\mathbf{x}}_i - \alpha\bar{\mathbf{x}}_i, & b(i) = -1 \ \ (\text{or } 0), \end{cases} \tag{5.5}$$

followed by reconstructing the watermarked coefficients as

$$\mathbf{x}_{\mathrm{w}} = [\mathbf{x}_{0,\mathrm{w}}^T, \mathbf{x}_{1,\mathrm{w}}^T, \ldots, \mathbf{x}_{L-1,\mathrm{w}}^T]^T. \tag{5.6}$$

4. Randomly choose one of the remaining 7 DWT coefficient groups and perform encryption as described in (5.1). Then, reconstruct the watermarked and partially encrypted DWT coefficients.
5. Perform inverse discrete wavelet transform (IDWT) to obtain the watermarked and partially encrypted audio signal in time domain.

Note that watermark embedding and encryption could be carried out interchangeably without altering the final results, thanks to the independent design of the two processes.

The watermark extraction and signal decryption scheme is shown below:

1. Perform 3-level DWT on the watermarked and partially encrypted audio signal.
2. From the DWT coefficients, find the portion corresponding to "AAA", and construct the frames
$$\mathbf{x}_{\mathrm{w}} = [\mathbf{x}_{0,\mathrm{w}}^T, \mathbf{x}_{1,\mathrm{w}}^T, \ldots, \mathbf{x}_{L-1,\mathrm{w}}^T]^T. \tag{5.7}$$
3. Extract each watermark bit via
$$\hat{b}(i) = \begin{cases} +1, & \bar{\mathbf{x}}_{i,\mathrm{w}} > 0, \\ -1 \ \ (\text{or } 0), & \bar{\mathbf{x}}_{i,\mathrm{w}} < 0. \end{cases} \tag{5.8}$$
4. According to the extracted watermark bits, obtain the encrypted symmetric encryption key and the index of the DWT coefficient group that has been encrypted.
5. Decrypt the encrypted DWT coefficients and restore the watermarked DWT coefficients without encryption.
6. Perform IDWT to obtain watermarked signal with better audio quality.

In the above system, the distortions of audio signal are caused by two processes, i.e., watermark embedding and encryption. Since the watermarks are embedded in terms of modifying the means of samples, quality degradation could be kept in a very low level. On the other hand, encryption will yield a set of totally different coefficients and make the audio content corresponding to these coefficients changed significantly. Further, the coefficients used for encryption are selected randomly to ensure system security, and all the DWT coefficients are determined solely by the host audio signal. Thus, it is relatively more difficult to control audio quality loss for the encryption process. However, it is in fact preferred that the encryption process could lead to more quality degradation to prevent unregistered users from accessing the high quality audio.

In order to have more control of quality degradation caused by encryption, the selection of coefficients for encryption could be modified. Specifically, the 3-level DWT coefficients, except the "AAA" group, are treated as a single vector, and the coefficients for encryption start from the highest frequency group "DDD", which is governed by a ratio parameter. Since the "AAA" portion takes $1/8$ of the DWT coefficients, the ratio of coefficients for encryption could be set between 0 to $7/8$. The ratio parameter is kept secret to ensure security. In this way, the larger the ratio is, the more encryption distortion is expected, and vice versa.

In [1], an alternative robust audio watermarking system for partially encrypted-compressed audio signals is proposed, where the RC4 based encryption technique is used [4, 5] and the MP3 compression system is considered. The general processing flow at the watermark embedder is as follows. First, the uncompressed host audio samples are framed and passed through a filter bank and the psychoacoustic model, which provide the controlling information for bit allocation and quantization. Then, some of the quantized coefficients are selected for encryption using the RC4 technique while the unselected coefficients are kept unchanged. The encrypted quantized and original quantized coefficients are then restored to their corresponding original positions in the audio samples, followed by Huffman coding and bitstream formatting. A term called sample-pair difference in the encrypted quantized samples is privately shared between the embedder and the receiver to ensure reliable host signal interference rejection during watermark extraction. The watermarks are then embedded in the encrypted coefficients of the host audio data. Watermark extraction can be performed either in encrypted domain or decrypted domain, thank to the homomorphic encryption.

The different between [1, 2] is that the embedding in the former is carried out in time domain instead of DWT domain, and the watermarks are embedded in the encrypted samples instead of non-encrypted samples in the later. Another watermarking system that combines encryption and watermarking is reported in [3], where the system is actually designed for image watermarking. However, the watermarks are audio samples encrypted by a chaotic map. In the next section, we introduce some works on watermarking using over-complete transform dictionaries. They bring the theory of compressive sensing to the context of watermarking, making them quite different from conventional frameworks.

5.2 Over-Complete Dictionary Based Watermarking

5.2.1 Compressive Sensing and Encryption

Before proceeding to introduce watermarking systems based on over-complete dictionaries, we first briefly introduce the basic framework of compressive sensing, which has attracted a tremendous amount of research attentions over the past decade. Consider an underdetermined linear system of equations

$$\mathbf{s} = \mathbf{D}\mathbf{x}, \tag{5.9}$$

where $\mathbf{s} \in \mathbb{R}^{M\times 1}$, $\mathbf{x} \in \mathbb{R}^{N\times 1}$, and $\mathbf{D} \in \mathbb{R}^{M\times N}$. Here, $M < N$, thus $\mathbf{D}$ is an over-complete dictionary, also known as sensing matrix. We also assume that $\mathbf{D}$ has unit norm columns, i.e., $\|\mathbf{d}_i\|_2 = 1, \forall i \in \{0, 1, \ldots, N-1\}$, and $\text{rank}\{\mathbf{D}\} = M$. In the context of compressive sensing [6], the process of representing $\mathbf{x}$ by the observed signal $\mathbf{s}$ is called sensing, while the reverse process, i.e., finding $\mathbf{x}$ based on the

observation and the sensing matrix (dictionary) $\mathbf{D}$, is called sparse signal recovery, modeled by the following formulation

$$\min_{\mathbf{x}} \quad \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{D}\mathbf{x} = \mathbf{s}. \tag{5.10}$$

Here, an important property of $\mathbf{x}$, i.e., sparsity, is introduced to indicate whether the recovery of $\mathbf{x}$ holds as an equality or approximation. Generally, the more zeros $\mathbf{x}$ has, the sparser $\mathbf{x}$ is, leading to higher possibility that the equality would hold. In addition, we further assume that $\mathbf{D}$ consists of independent and identically distributed (i.i.d.) Gaussian variables. More about determining the dictionary could be found in [7], which is not in the scope of this book. The methods to solve (5.10) can be categorized into greedy searching [8–10] and ℓ_1-norm regularized [11–13] ones. There also exists another important category of methods termed sparse Bayesian methods. We direct interested readers to [14, 15] for more information. The greedy algorithms are generally faster but the sparsity of the solutions is not optimized, while the solutions based on the ℓ_1-norm regularized formulation are generally more sparse but require more computation power.

In the meantime of the development of compressive sensing technique, it has been shown that the compressive sensing framework actually leads to an encryption scheme, where the sensing matrix $\mathbf{D}$ serves as the encryption key [16]. In one of the pioneering works on security analysis of compressive sensing schemes, the authors in [17, 18] have pointed out that compressive sensing could not achieve perfect secrecy but can guarantee computation secrecy. Recently, a comprehensive review of compressive sensing technique used in information security field is reported in [19]. Compared to conventional sensing which is governed by Nyquist sampling theory, compressive sensing brings improved security to the system.

Now, we link compressive sensing to watermarking. For the process described by (5.9), it could not only be considered as a sensing processing but also viewed as an over-complete dictionary based transform. If $\mathbf{s}$ represents the host signal frame, then $\mathbf{x}$ is the transform domain host signal, whose length is expanded from M to N. Consider a conventional watermarking system in transform domain, the forward transform is performed based on an orthonormal basis, e.g., the discrete Fourier transform (DFT), discrete cosine transform (DCT), or DWT matrices are orthonormal (square) matrices. Compared to conventional transform matrices, the over-complete matrices lead to two different situations. First, the transform is more secure because of the uncertainty and randomness of $\mathbf{D}$. In contrast, a conventional transform always indicates a structured transform matrix which is publicly known. Second, the use of over-complete dictionaries expands the dimension of the host signal from M to N, which is likely to lead to other properties that do not exist in conventional systems. Therefore, the forward transform of host signal for an over-complete dictionary based watermark embedder becomes the problem of (5.10). Unfortunately, since natural multimedia signals are general non-sparse, sparse signal recovery algorithms developed in compressive sensing context may not be directly effective for watermarking systems. To solve this problem, the designer has to create special algorithms suitable

for the specific scenario and some good design examples are shown in [20, 21]. Note that although these techniques are originally designed for image watermarking, they can be directly applied to audio data by replacing image samples with audio samples.

5.2.2 *Design Example 1—Preliminary Study*

We consider a spread spectrum (SS) like signal model for over-complete dictionary based watermarking. Recall the generic SS embedding function (3.11) and rewrite it in vector form

$$\mathbf{x}_{\mathrm{w}} = \mathbf{x} + \alpha b \mathbf{p}, \tag{5.11}$$

where $\mathbf{x}$ is the transform domain samples based on an orthonormal transform matrix. Note that (5.11) represents transform domain processing. If we denote such an orthonormal matrix by $\mathbf{H}$ such that $\mathbf{H}^T = \mathbf{H}^{-1}$ and $\mathbf{s} = \mathbf{H}\mathbf{x}$, then the original domain embedding function is obtained by multiplying $\mathbf{H}$ to the left of both sides of (5.11), i.e.,

$$\mathbf{s}_{\mathrm{w}} = \mathbf{H}\mathbf{x}_{\mathrm{w}} = \mathbf{H}\mathbf{x} + \alpha b \mathbf{H}\mathbf{p} = \mathbf{s} + \alpha b \mathbf{H}\mathbf{p}, \tag{5.12}$$

in an M-dimensional space. In an over-complete dictionary based watermarking system, the transform dictionary $\mathbf{H}$ is replaced by $\mathbf{D}$, introducing a new dimension $N > M$. Thus, the spreading sequence $\mathbf{p}$ should also be expanded to validate the additive embedding. We denote such a spreading sequence by $\mathbf{p}_{\mathbf{D}}$. Therefore, the embedding function using $\mathbf{D}$ and $\mathbf{p}_{\mathbf{D}}$ is giben by

$$\mathbf{s}_{\mathrm{w}} = \mathbf{D}\mathbf{x}_{\mathrm{w}} = \mathbf{D}(\mathbf{x} + \alpha b \mathbf{p}_{\mathbf{D}}). \tag{5.13}$$

However, during watermark extraction, the recovery of the portion $(\mathbf{x} + \alpha b \mathbf{p}_{\mathbf{D}})$ must be reliable so that b exists and could be detected. This is somewhat straightforward in conventional scheme, where one only needs to multiply $\mathbf{H}^T$ left to $\mathbf{s}_{\mathrm{w}}$. However, since $\mathbf{D}$ does not strictly have an inverse, the extraction of b in (5.13) is not easy.

The first attempt to enable the traceability of $(\mathbf{x} + \alpha b \mathbf{p}_{\mathbf{D}})$ is via a minimum norm formulation, which leads to a closed-form solution and the concept of watermark projection. Let the watermark extraction process start with the following optimization problem

$$\min_{\mathbf{y}_{\ell_2}} \left\| \mathbf{y}_{\ell_2} \right\|_2^2 \quad \text{s.t.} \quad \mathbf{s}_{\mathrm{w}} = \mathbf{D}\mathbf{y}_{\ell_2}, \tag{5.14}$$

which can be neatly solved via pseudo inverse, i.e.,

$$\begin{aligned} \mathbf{y}_{\ell_2} &= \mathbf{D}^T(\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{s}_{\mathrm{w}} \triangleq \mathbf{D}^{\dagger}\mathbf{s}_{\mathrm{w}} \\ &= \mathbf{D}^{\dagger}\mathbf{D}(\mathbf{x} + \alpha b \mathbf{p}_{\mathbf{D}}) \\ &= \mathbf{D}^{\dagger}\mathbf{D}\mathbf{x} + \alpha b \mathbf{D}^{\dagger}\mathbf{D}\mathbf{p}_{\mathbf{D}}, \end{aligned} \tag{5.15}$$

which is the sum of the projections of $\mathbf{x}$ and $\mathbf{p_D}$ onto the row space of $\mathbf{D}$ respectively. We further define a decomposition pair of $\mathbf{p_D}$ as

$$\mathbf{p}_{\mathrm{P}} = \mathbf{D}^{\dagger}\mathbf{D}\mathbf{p}_{\mathbf{D}}, \tag{5.16}$$

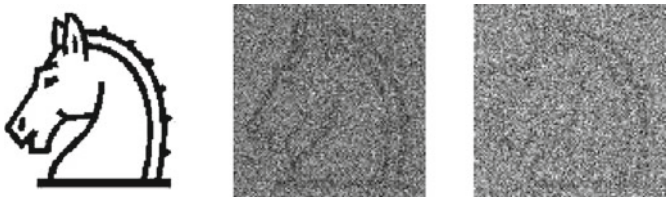
$$\mathbf{p}_{\mathrm{O}} = \mathbf{p}_{\mathbf{D}} - \mathbf{p}_{\mathrm{P}}, \tag{5.17}$$

which are projection (P) and orthogonal (O) components of $\mathbf{p_D}$ respectively. It is then indicated that only if $\mathbf{p_D}$ also lies within the row space of $\mathbf{D}$, i.e., $\mathbf{p}_{\mathrm{P}} = \mathbf{p_D}$, can the full knowledge of $\mathbf{p_D}$ be contained in $\mathbf{y}_{\ell_2}$. However, the two most common forms of $\mathbf{p_D}$, i.e., a pseudorandom sequence or a vectorized pattern signal, are generated without considering the above condition. In fact, none of the existing watermarking systems considers the relationship between the transform dictionary and the watermark, because this is a trivial problem when $\mathbf{H}$ is an orthonormal dictionary. The hidden information bit is extracted via

$$\hat{b} = \mathrm{sgn}\left\langle \mathbf{p}_{\mathrm{P}}, \mathbf{D}^{\dagger}\mathbf{s}_{\mathrm{w}} \right\rangle = \mathrm{sgn}\left(\langle \mathbf{p}_{\mathrm{P}}, \mathbf{x} \rangle + \alpha b \left\| \mathbf{p}_{\mathrm{P}} \right\|_2^2 \right), \tag{5.18}$$

where $\mathbf{D}^{\dagger}\mathbf{D}\mathbf{x} = \mathbf{x}$. Comparing conventional SS watermark embedding using $\mathbf{H}$ and the alternative system using $\mathbf{D}$, we see very similar performance in terms of embedding distortion and detection robustness, which are characterized by the second and first terms at the right hand side of (3.12) and (5.18) respectively. However, system security could be substantially improved thanks to the utilization of random overcomplete dictionary. An illustrative example of watermark projection is shown in Fig. 5.4, where we can see that the larger the dimension expansion from M to N the more incomprehensible the projected watermark pattern could be.

Now, we can think of two privacy preserving scenarios according to whether the watermark bits are modulated by $\mathbf{p_D}$ or not. If watermark modulation is considered, then the projection is carried out on the spreading sequence as been discussed in the aforementioned content. The original spreading sequence is $\mathbf{p_D}$ while the equivalent spreading sequence (without orthogonal redundant portion) is $\mathbf{p}_{\mathrm{P}}$. In this way, the attackers only have the access to $\mathbf{p}_{\mathrm{P}}$ but not $\mathbf{p_D}$, and $\mathbf{p_D}$ could serve as a key to enhance system security. The second scenario does not implement watermark modulation, which means that the samples of $\mathbf{p_D}$ are directly the watermark bits, and the whole



(a) Original $\mathbf{p}$. (b) $\mathbf{p}_{\mathrm{P}}$, $N \approx 3M$. (c) $\mathbf{p}_{\mathrm{P}}$, $N \approx 5M$.

Fig. 5.4 An illustration of watermark projection and comprehensibility

watermark sequence is embedded repeatedly in each frame. In this way, it becomes very difficult for the adversaries to recover the original watermarks even if they could manage to recover the projected version $\mathbf{p}_{\mathrm{P}}$. A use case is described as follows. A data owner wants to share some data with others. The owner embeds the identification pattern's projection data as shown in Fig. 5.4(b) or (c) into the host data before sharing. In the meantime, for some users, the owner wants to authorize them the copyright but not willing to let them know the embedded information. To ensure privacy, the owner passes $\mathbf{s}_{\mathrm{w}}$, $\mathbf{D}$, and $\mathbf{p}_{\mathrm{P}}$, to them. Therefore, these users can use the provided information to claim that the copies are authorized, but they only know the incomprehensible $\mathbf{p}_{\mathrm{P}}$ instead of $\mathbf{p}_{\mathbf{D}}$. For the other users, the owner has no problem of sharing the mark pattern, thus, $\mathbf{s}_{\mathrm{w}}$, $\mathbf{D}$, and $\mathbf{p}_{\mathbf{D}}$, are passed to these users. In this example, we can see that the dictionary $\mathbf{D}$ has become a component of the copyright mark, which is not applicable in conventional system. More importantly, the decomposed watermark $\mathbf{p}_{\mathrm{P}}$ enables an effective way to deal with the first group of users, which is also not possible in conventional cases. Since the embedding and extraction of watermarks could be performed off-line, it is reasonable to conclude that the privacy preserving feature is more important than computational efficiency.

Next, we discuss the system design in a compressive sensing framework. Different from utilizing projection and pseudo inverse, the over-complete dictionary based transform could be carried out by minimizing the ℓ_0 norm, which forms a sparse signal recovery problem. However, it has been shown in [20] that combining SS and sparsity in watermarking system is not an easy task. To shed some light on such a framework, a single support modification fragile watermarking system is proposed. We will simply denote $\mathbf{p}_{\mathbf{D}}$ by $\mathbf{p}$, since in the following content $\mathbf{H}$ is not considered.

Let $\mathbf{k} \triangleq [k_0, k_1, \ldots, k_{K-1}]$ be the set of column indices of $\mathbf{D}$ that correspond to the nonzero elements of $\mathbf{x}_{\ell_0}$. Then $\mathbf{s} = \mathbf{D}\mathbf{x}_{\ell_0}$ can be rewritten in a compact form

$$\mathbf{s} = \Phi \dot{\mathbf{y}}, \tag{5.19}$$

where

$$\Phi \triangleq [\mathbf{d}_{k_0}, \mathbf{d}_{k_1}, \ldots, \mathbf{d}_{k_{K-1}}] \in \mathbb{R}^{M \times K}, \tag{5.20}$$

and $\dot{\mathbf{y}} \in \mathbb{R}^{K \times 1}$ is composed by selecting the nonzero elements of $\mathbf{x}_{\ell_0}$. Here, Φ is a tall matrix with full column rank. Otherwise (5.19) can be further compressed. The columns of $\mathbf{D}$ that have not participated in the above linear combination are denoted by Ψ,

$$\Psi \triangleq [\mathbf{d}_{l_0}, \mathbf{d}_{l_1}, \ldots, \mathbf{d}_{l_{N-K-1}}] \in \mathbb{R}^{M \times (N-K)}, \tag{5.21}$$

where $\mathbf{l} \triangleq [l_0, l_1, \ldots, l_{N-K-1}]$ does not overlap with $\mathbf{k}$. Considering Φ, Ψ, and $\mathbf{D}$ as different sets of $\mathbf{d}_i$, it follows $\Phi \cup \Psi = \mathbf{D}$ and $\Phi \cap \Psi = \emptyset$. Because $\mathrm{rank}(\mathbf{D}) = M$, $\mathrm{rank}(\Phi) = K$, and $K < M$, it is indicated that the column space of Ψ intersects with the null space of Φ^T. We use $\Delta \in \mathbb{R}^{M \times (M-K)}$ to denote the orthonormal basis of the null space of Φ^T such that $\Phi^T \Delta = \mathbf{0}$. Using the above parameters, the watermark

embedding and detection algorithms for a single support modification system can be designed as Algorithms 1 and 2 respectively, which are shown on the next page.

In Algorithm 1, a single watermark chip b is inserted into index l_p which has not been used to represent $\mathbf{s}$. In this setup, conventional modulation sequences, i.e., random realizations of $\mathbf{p}$ are not needed. Instead, αb is absorbed in $\mathbf{p}$ such that

$$\mathbf{p} = [0, 0, \ldots, p_{l_p}(= \alpha b), \ldots, 0]^T. \tag{5.22}$$

Note that the atom at the selected index has the strongest projection in Δ (Step 5). In fact, multiple chips can be inserted as long as the indices for insertion have not been used to represent $\mathbf{s}$. In addition, the insertion formula, i.e., Step 6 of Algorithm 1 can also take other forms. It is then suggested that the flexibility of choosing insertion locations and inserting formula can improve the system performance, but here, we focus on the simplest design to illustrate a simple, feasible, and effective system in sparse domain, whose crucial component is the detection algorithm.

In Algorithm 2, the watermarked signal $\mathbf{s}_{\mathrm{w}}$ is first projected into the column space of Δ via Step 3. Due to $\Delta^T \Phi = \mathbf{0}$, we can have the following explicit expression

$$\mathbf{s}_{\mathrm{w,P}} = \Delta^{\dagger}\Delta^T \mathbf{D}(\mathbf{x}_{\ell_0} + \mathbf{p}) = \Delta^{\dagger}\Delta^T \mathbf{D}\mathbf{p}, \tag{5.23}$$

meaning that the projection procedure removes the host interference and forms a new underdertermined linear system with a modified dictionary $\Delta^{\dagger}\Delta^T \mathbf{D}$, whose restricted

Algorithm 1: Single Support Modification - Embedding

Input: $\mathbf{s}$, $\mathbf{D}$, α, $b = \pm 1$
Output: $\tilde{\mathbf{s}}_{\mathrm{w}}$

1 Initialization: $\mathbf{p} = [p_0, p_1, \ldots, p_{N-1}]^T \leftarrow \mathbf{0}$;
2 $\mathbf{x}_{\ell_0} \leftarrow \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{s} = \mathbf{D}\mathbf{x}.$;
3 $\Phi \leftarrow [\mathbf{d}_{k_0}, \ldots, \mathbf{d}_{k_{K-1}}]$, $\Psi \leftarrow [\mathbf{d}_{l_0}, \ldots, \mathbf{d}_{l_{N-K-1}}]$;
4 $\Delta \leftarrow \mathrm{null}(\Phi^T) \quad \text{s.t.} \quad \Delta^T\Delta = \mathbf{I}$;
5 $l_w \leftarrow \arg\max_{l_i} \|\Delta^{\dagger}\Delta^T \mathbf{d}_{l_i}\|_2^2, \quad i = 0, 1, \ldots, N-K-1$;
6 $p_{l_p} \leftarrow \alpha b$;
7 $\mathbf{s}_{\mathrm{w}} \leftarrow \mathbf{D}(\mathbf{x}_{\ell_0} + \mathbf{p})$.

Algorithm 2: Single Support Modification - Detection

Input: $\mathbf{s}_{\mathrm{w}}$, $\mathbf{D}$, $\mathbf{k}$
Output: $\hat{b}$

1 $\Phi \leftarrow [\mathbf{d}_{k_0}, \ldots, \mathbf{d}_{k_{K-1}}]$;
2 $\Delta \leftarrow \mathrm{null}(\Phi^T) \quad \text{s.t.} \quad \Delta^T\Delta = \mathbf{I}$;
3 $\mathbf{s}_{\mathrm{w,P}} \leftarrow \Delta^{\dagger}\Delta^T \mathbf{s}_{\mathrm{w}}$;
4 $\hat{\mathbf{p}} \leftarrow \arg\min_{\mathbf{y}} \|\mathbf{y}\|_0 \quad \text{s.t.} \quad \mathbf{s}_{\mathrm{w,P}} = \Delta^{\dagger}\Delta^T \mathbf{D}\mathbf{y}$;
5 $\hat{b} \leftarrow \mathrm{sgn}(\hat{p}_j), \quad j = \arg\min_i i \quad \text{s.t.} \quad \hat{p}_i \neq 0.$

isometry property (RIP) is rigorously discussed in [22]. Note that the problem of identifying $\mathbf{p}$ from $\tilde{\mathbf{s}}_{\mathrm{w,P}}$ can always be efficiently solved because $\mathrm{card}(\mathbf{p}) = 1$, which satisfies the mutual incoherence property (MIP) condition [9]. Here the RIP and MIP conditions are not provided and we refer interested readers to the references for more information. Therefore, we can take the simplest method to implement Step 4 of Algorithm 2, e.g., orthogonal matching pursuit (OMP) [9] with single iteration. Mathematically, the single iteration OMP first calculates

$$\begin{aligned}\mathbf{s}_{\mathrm{w,P}}^T \Delta^\dagger \Delta^T \mathbf{D} &= \mathbf{p}^T \mathbf{D}^T \Delta^\dagger \Delta^T \Delta^\dagger \Delta^T \mathbf{D} \\ &= \alpha b \mathbf{d}_{l_w}^T \Delta^\dagger \Delta^T [\mathbf{d}_0, \ldots, \mathbf{d}_{l_w}, \ldots, \mathbf{d}_{N-1}],\end{aligned} \tag{5.24}$$

where $(\Delta^\dagger \Delta^T)(\Delta^\dagger \Delta^T) \ldots (\Delta^\dagger \Delta^T) = \Delta^\dagger \Delta^T$ is the property of a projection matrix, and it is easy to detect l_p since

$$\max(\mathbf{s}_{\mathrm{w,P}}^T \Delta^\dagger \Delta^T \mathbf{D}) = p_{l_p} \mathbf{d}_{l_p}^T \Delta^\dagger \Delta^T \mathbf{d}_{l_p}. \tag{5.25}$$

Thus the strongest supporting atom is $\hat{\mathbf{d}} \triangleq \Delta^\dagger \Delta^T \mathbf{d}_{l_p}$. Then, the OMP algorithm projects $\tilde{\mathbf{s}}_{\mathrm{w,P}}$ onto this atom and obtains the single support value of $\hat{\mathbf{p}}$, and we have $\hat{p} = \mathrm{sgn}(b)$, an exact recovery of the information bit. Such system can be considered as a (semi-) informed system, where partial information about the host signal, i.e., $\mathbf{k}$, is needed during the detection phase. Note that this is inapplicable in conventional cases where the exact complete dictionary $\mathbf{H}$ does not have unused atoms in representing $\mathbf{s}$. Note that if we also make l_p available at detection phase, then a more robust detection can be achieved because possible error in detecting l_p can be avoided in noisy conditions.

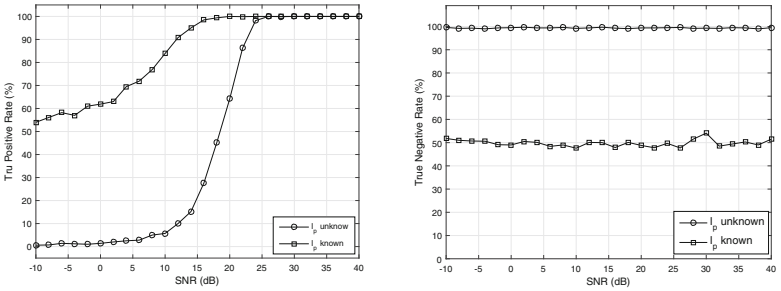
In the noise-free situation, host interference is rejected by the use of Δ, and successful detection is guaranteed for an arbitrary $\alpha > 0$. However, in a noisy condition, (5.23) becomes

$$\mathbf{s}_{\mathrm{w,P}} = \Delta^\dagger \Delta^T [\mathbf{D}(\mathbf{x}_{\ell_0} + \mathbf{p}) + \mathbf{v}] = \Delta^\dagger \Delta^T \mathbf{D}\mathbf{p} + \mathbf{v}_{\mathrm{P}}, \tag{5.26}$$

where $\mathbf{v}$ is additive white Gaussian noise (AWGN) and $\mathbf{v}_{\mathrm{P}}$ is its projection on Δ. If the detection is performed without the knowledge of l_p, then Step 4 of Algorithm 2 will be affected by $\mathbf{v}_{\mathrm{P}}$, and l_p could be wrongly identified. However, if l_p is known, then the scalar projection of (5.26) onto $\hat{\mathbf{d}}$ yields

$$\hat{b} = \mathrm{sgn}\left(\alpha b + \frac{\mathbf{d}_{l_p}^T \Delta^\dagger \Delta^T \mathbf{v}}{\mathbf{d}_{l_p}^T \Delta^\dagger \Delta^T \mathbf{d}_{l_p}}\right), \tag{5.27}$$

where the term at the right hand side in the parenthesis is the interference. Figure 5.5 shows the synthetic analysis results of the system in noisy environments obtained by averaging 1000 realizations of AWGN, where the host signal $\mathbf{s}$ and dictionary $\mathbf{D}$ are both generated by i.i.d. Gaussian random variables with normal distribution,



(a) Sensitivity. (b) Specificity.

Fig. 5.5 Sensitivity (**a**) and specificity (**b**) of the single support modification system under different SNR values, where $M = 128, N = 1024, \alpha = 0.5, K \approx 100$, and the resultant embedding distortion is 27 ± 0.5 dB

having $K \approx 100$. The advantage of using l_p during the detection process is evidenced in Fig. 5.5(a). However, we can observe from Fig. 5.5(b) that performing detection without using l_p has nearly 100% accuracy in detecting non-watermarked signals. In other words, it is almost free from false positive decisions. In contrast, conducting detection using l_p yields a random guess when examining signals that have not been watermarked. Generally, Fig. 5.5(b) reflects the system performance when watermark is not embedded. In this case, signal at the known index is not marked at all (αb does not exist in (5.27) anymore), and the corresponding value is solely governed by noise. Therefore, the detection performance turns out to be like random guesses with around 50% correct rate. In contrast, if the index is unknown, then Steps 4 and 5 of Algorithm 2 are solely governed by noise, and the strongest support becomes random. Thus, the possibility of the strongest support index being coincident with the known index becomes extremely low (less than $1/(N-K)$), and the true negative rate approaches 100%.

5.2.3 *Design Example 2—Dictionary Decomposition*

Apart from the projection based and single support modification system mentioned above, which utilizes random dictionaries, it is also possible to construct structured dictionaries with specific atoms. This is intensively studied in [21, 23, 24]. We will briefly introduce the main idea of such a system in this subsection. Let us first denote the underdetermined linear system (5.9) equivalently but in a decomposition form

$$\mathbf{s} = \sum_{k=0}^{K-1} X(k)\mathbf{d}_{\gamma_k} + \mathbf{r}_K, \tag{5.28}$$

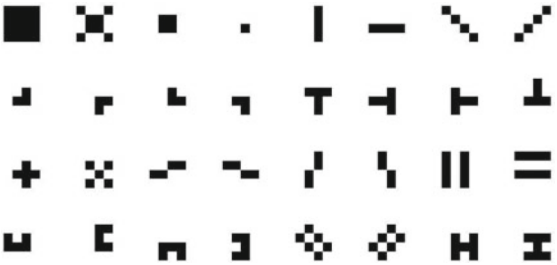


Fig. 5.6 Example of a subset of atoms used in [21]

where $X(k)$ represents the transform domain samples (decomposition coefficients), γ_k is the decomposition path indexing the columns of $\mathbf{D}$ that have been used to represent $\mathbf{s}$, and $\mathbf{r}_K$ is the residual when $\mathbf{s}$ is decomposed using K iterations. It is also important to note that (5.28) must operate in integer domain, i.e., quantized sample level, in order to validate the corresponding watermarking technique. An example of a constructed dictionary subset is shown in Fig. 5.6 for the image watermarking system proposed in [21]. Note that such a dictionary could be similarly constructed for audio samples in integer space with vectorized patterns. For example, the samples of an 8-bit depth quantized image take the integer values from 0 to 255, while 16-bit depth quantization corresponds to 0 to $2^{16} - 1$. Watermark embedding should not exceed the range of the space. The decomposition algorithm must satisfy two conditions:

- *Decomposition path stability:* The decomposition path used for decompose $\mathbf{s}$ must be identical to the decomposition path obtained when decomposing watermarked signal. If the path is unstable, the watermark extractor will not be able to read the embedded watermarks correctly.
- *Coefficient stability:* During watermark embedding, the decomposition coefficients are the samples to be watermarked. This means that the coefficients $X(k)$ will be altered in some way. Therefore, during watermark extraction, if we do not change the decomposition path, the system must ensure that the modified coefficients are recoverable.

The above conditions are indeed fundamental but in fact very difficult to satisfy. One solution to the above problem is designed as follows. During signal decomposition, at iteration k, a sequence of coefficient candidates, denoted by $c(n)$, $n \in 0, 1, \ldots, N-1$, satisfying

$$c(n) = \arg\max_x x \quad \text{s.t.} \quad \mathbf{r}_{k-1} - x\mathbf{d}_n \geq \mathbf{0}, \tag{5.29}$$

are obtained, which correspond to the maximum scalar applicable to each atom without causing negative residual samples. Then, the coefficient is determined by the one yielding the highest decomposition energy, or equivalently the lowest residual energy, i.e.,

$$\gamma_k = \arg\min_n \|\mathbf{r}_{k-1} - c(n)\mathbf{d}_n\|_2^2, \tag{5.30}$$

and

$$X(k) = c(\gamma_k). \tag{5.31}$$

Substitute (5.31) into (5.28), we obtain the transform domain expression of the host signal. Note that the decomposition only reduces positive valued samples and does not incur any negative values, thus all the parameters lie in the quantization integer space. Now, suppose that a single coefficient, $X(i)$, is modified to $X_{\mathrm{w}}(i)$ when embedding the watermark, which is similar to the single support modification discussed in the previous subsection. Then, the watermarked signal can be decomposed as follows if the same decomposition algorithm is applied,

$$\mathbf{s}_{\mathrm{w}} = \sum_{k=0, k\neq i}^{K-1} X(k)\mathbf{d}_{\gamma_k} + X_{\mathrm{w}}(i)\mathbf{d}_{\gamma_i} + \mathbf{r}_K, \tag{5.32}$$

which indicates both decomposition path and coefficient stability. The proof of such stability is detailed in [21]. Note that the system also works with multiple-bit embedding. The embedding function is given by

$$X_{\mathrm{w}}(i) = \begin{cases} X(i) + 1, & b \neq \mathrm{LSB}\{X(i)\} \text{ and } (\mu > 0 \text{ or } X(i) = 0), \\ X(i) - 1, & b \neq \mathrm{LSB}\{X(i)\} \text{ and } (\mu < 0 \text{ or } X(i) = 2^{N_{\mathrm{bit}}}), \\ X(i), & b = \mathrm{LSB}\{X(i)\}, \end{cases} \tag{5.33}$$

where b is the information bit to be embedded, μ is an i.i.d. random variable with uniform distribution $\mu \sim \mathcal{U}(-1, 1)$, and N_{bit} is the quantization depth. The above system is originally designed for steganography, which is concerned with imperceptibility and undetectability in its specific context. Therefore, the system is not robust against attacks. In fact, the system is highly sensitive to any modification of the watermarked signal, because of the vulnerability of the decomposition path and coefficient stability. However, the system could have substantially improved security in terms of undetectability even at high embedding capacity.

5.3 Remarks

In this section, we have addressed several audio watermarking system designs with the consideration of cryptography to enhance system performance, especially for security. The first part focuses on combining watermark embedding and encryption, in which two disjoint parts of the host signal are watermarked and encrypted respectively. The encryption not only enhances system security, but also causes more quality degradation than the conventional watermarking systems. This is an intended treatment, and as a result, unauthorized users could only have access to quality-degraded version of the watermarked signal.

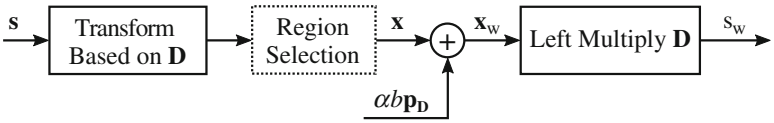


Fig. 5.7 An ideal SS like over-complete dictionary based embedding system

While combining watermarking and encryption is somewhat intuitive and modular (watermarking and encryption are interchangeable), the second part of this chapter investigates a more difficult task, i.e., watermarking in over-complete dictionary based transform domain, in which the over-complete dictionary based transform could be considered as an encryption process. It can be seen from both [20, 21] that the systems have to deal with the problem of ensuring watermarked signal recovery in dimension expanded transform domain, which does not exist in the conventional orthonormal dictionary based systems. The projection based system proposed in [20] is in fact rather similar to a conventional SS based system, because the signal portion perpendicular to the row space of the dictionary is abandoned during the processing. Therefore, whether an SS like system operating in over-complete dictionary based transform domain could be obtained is still unknown to us. Recall the single support modification system in [20], which only modifies a single host signal coefficient to ensure stability during recovery. It is obvious that the system is not applicable to the applications demanding robustness. Similarly, the dimension expansion achieved in [21] only helps in improving system security, but it does not contribute to robustness or imperceptibility. In fact, the decomposition algorithm at most iterates M times, which does not exceed the dimension of the host signal in the original domain, because the decomposition algorithm must eliminate a support once a time to satisfy the stability requirement.

In summary, we consider an SS like watermarking system fully operating in dimension expanded space[1] as a worthy research subject. The ideal system structure, although not realized yet, is depicted in Fig. 5.7.

References

1. Subramanyam AV, Emmanuel S (2012) Audio watermarking in partially compressed-encrypted domain. In 2012 IEEE international conference on systems, man, and cybernetics (SMC), pp 2867–2872
2. Datta K, Gupta IS (2013) Partial encryption and watermarking scheme for audio files with controlled degradation of quality. Multimedia Tools Appl 64(3):649–669

[1] It is called sparse domain if the signal transformed into this domain is sparse. However, since most multimedia signals are not really sparse, we avoid abusing the word "sparse" in the context of this book, and refer to it as dimension expanded domain, or over-complete dictionary based transform domain.

3. Hamdi B, Hassene S (2013) A new approach combining speech chaotic encryption with fragile image watermarking for audio securing and intrusion detection. In 2013 international conference on electrical engineering and software applications (ICEESA), pp 1–6
4. Schneier B (1996) Applied cryptography. Wiley, New York
5. Castelluccia C, Mykletun E, Tsudik G (2005) Efficient aggregation of encrypted data in wireless sensor networks. In Proceedings of the second annual international conference on mobile and ubiquitous systems: networking and services (MobiQuitous), pp 109–117
6. Candes EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE Sign Process Mag 25(2):21–30 ISSN 1053-5888
7. Aharon M, Elad M, Bruckstein A (2006) K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Sign Process 54(11):4311–4322
8. Mallat SG, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. IEEE Trans Sign Process 41(12):3397–3415
9. Tropp JA (2004) Greed is good: algorithmic results for sparse approximation. IEEE Trans Inf Theor 50(10):2231–2242
10. Dai W, Milenkovic O (2009) Subspace pursuit for compressive sensing signal reconstruction. IEEE Trans Inf Theor 55(5):2230–2249
11. Candes EJ, Tao T (2005) Decoding by linear programming. IEEE Trans Inf Theor 51(12):4203–4215
12. Kim SJ, Koh K, Lustig M, Boyd S, Gorinevsky D (2007) An interior-point method for large-scale ℓ_1-regularized least squares. IEEE J Sel Topics Sign Process 1(4):606–617
13. Mohimani H, Babaie-Zadeh M, Jutten C (2009) A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm. IEEE Trans Sign Process 57(1):289–301
14. Babacan S, Molina R, Katsaggelos A (2010) Bayesian compressive sensing using laplace priors. IEEE Trans Image Process 19(1):53–63
15. Ji S, Xue Y, Carin L (2008) Bayesian compressive sensing. IEEE Trans Sign Process 56(6):2346–2356
16. Candes EJ, Tao T (2006) Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans Inf Theor 52(12):5406–5425
17. Rachlin Y, Baron D (2008) The secrecy of compressed sensing measurements. In Proceedings of 46th Annual Allerton Conference on Communication Control Computer, Urbana-Champaign, IL, pp 813–817
18. Orsdemir A, Altun HO, Sharma G, Bocko MF (2008) On the security and robustness of encryption via compressed sensing. In proceedings of IEEE military communication conference on (MILCOM), pp 1–8, San Diego, CA, 2008
19. Zhang Y, Zhang LY, Zhou J, Liu L, Chen F, He X (2016) A review of compressive sensing in information security field. IEEE Access 4:2507–2519
20. Hua G, Xiang Y, Bi G (2016b) When compressive sensing meets data hiding. IEEE Sign Process Lett 23(4):473–477
21. Cancelli G, Barni M (2009) MPSteg-Color: Data hiding through redundant basis decomposition. IEEE Trans Inf Forensics Secur 4(3):346–358 Sep
22. Chang L-H, Wu J-Y (2014) An improved rip-based performance guarantee for sparse signal recovery via orthogonal matching pursuit. IEEE Trans Inf Theor 60(9):5702–5715
23. Cancelli G, Barni M, Menegaz G (2006) MPSteg: Hiding a message in the matching pursuit domain. Proc SPIE 6072:1–9
24. Cancelli G, Barni M (2007) MPSteg-Color: A new steganographic technique for color images. In 9th International Workshop in Information Hiding (IH 2007), vol 4567, Saint Malo, France, pp 1–15

Chapter 6
Conclusion and Future Work

Abstract In Chaps. 1–5, we have comprehensively reviewed the fundamental concepts and designing techniques related to digital audio watermarking. Specifically, the general background and fundamental concepts are provided in Chap. 1. The issues regarding the problem on imperceptibility for audio watermarking are discussed in Chap. 2. Typical audio watermark embedding and extraction techniques, from classical ones to their latest variations and developments, are detailed in Chap. 3. Furthermore, two emerging audio watermarking topics, i.e., reversible audio watermarking and cryptography-aided audio watermarking, are investigated in Chaps. 4 and 5, respectively. In this chapter, we discuss the limitations, challenges, and future work on audio watermarking.

6.1 Limitations and Challenges

The limitations of existing audio watermarking systems are closely associated with the performance criteria, i.e., imperceptibility, robustness, security, capacity, and computational complexity. First, we note that it is unnecessary for all audio watermarking systems to simultaneously satisfy all of the performance criteria. In fact, this is currently not possible, and more importantly, none of the existing applications has such a requirement. Therefore, the current limitations commonly exist in those systems which trade off a few criteria. For imperceptibility, the criterion that nearly all audio watermarking systems have to account for, the current state-of-the-art methods are psychoacoustic model based analytical tuning or automated heuristic tuning according to audio quality measurements. The former suffers from watermark removal by lossy compression, and the latter suffers from inefficiency. For fragile or semi-fragile audio watermarking, e.g., reversible or cryptography-aided audio watermarking, the channel between transmitter and receiver is secure, thus robustness does not need to be considered during system design. In this way, limitations mainly lie in the trade-off between embedding capacity and imperceptibility. Generally, if watermark embedding is not performed at bit level (such as reversible embedding methods), the commonly applied frame based embedding methods are limited by low embedding rate at 1 bit per frame. Although recent proposals such

Y. Xiang et al., *Digital Audio Watermarking*, SpringerBriefs in Signal Processing, DOI 10.1007/978-981-10-4289-8_6

as [1] can offer certain improvements on embedding capacity, the resultant one is still relatively low, especially when compared with image watermarking. In applications such as broadcast monitoring and second screen, computational efficiency becomes one of the major criteria.

In addition, security has been an important but less studied criterion for watermarking [2, 3]. Due to the nature of watermarking which indicates the difference between watermarking and cryptography, watermarking systems could not achieve the security level of encryption. To date, most existing systems use a pseudo-random noise (PN) sequence to add security to the systems. In Chap. 5, we have shown how cryptography could be incorporated in conventional watermarking system to enhance system security. It is worth of noting that the use of over-complete dictionary for forward and inverse transforms could be a promising solution to this problem.

Lastly, the most complicated performance criterion, robustness, is required in most situations, but it is also the most difficult criterion to satisfy. The application that requires the most stringent robustness would be copyright marking system, which has to deal with unexpected attacks. Although the development of watermarking techniques enables the designers to tackle normal processing attacks (e.g., amplitude scaling, noise addition, re-compression, etc.), robustness against malicious attacks, especially a series of desynchronization attacks, is still an open challenge. There have been several works dedicated to dealing with desynchronization attacks, e.g., [4–7], which have presented several possible solutions with certain assumptions. The ultimate challenge for dealing with desynchronization attacks is believed to be the problem of finding a good feature set for watermark embedding. In other words, watermarks are preferable to be embedded in desynchronization invariant feature samples, instead of conventional transform domain samples. However, the freedom of the adversaries on manipulating an illegally obtained audio copy enables more complicated attacks, including but not limited to collision attack and combinations of attacks. In view of this, incorporation of authentication (tampering detection) mechanisms in audio watermarking systems could be one possible remedy when facing advanced attacks.

6.2 Future Work

According to the above-mentioned limitations and challenges, a few potential future works are summarized as follows.

6.2.1 Watermark Embedding Domain

Despite the variety of audio watermarking system designs, a fundamental and important question has not yet been answered, that is, should we perform a transform before embedding watermarks? In other words, what is the basic differences between time

domain methods and transform domain methods? Intuitively, one would consider frequency domain methods as a better choice 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 level of imperceptibility. For robustness, security and capacity, even less clues have been revealed to indicate which domain is a better choice. Furthermore, while many transforms have been proposed, the features or advantages of a specific transform over other transforms are missing in the literature. To answer these questions, a systematic and fair comparison setup using appropriate imperceptibility and robustness evaluation tools needs to be designed.

6.2.2 Pattern or Random Watermarks

Although 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., [8–11]. An example of pattern and random watermarks is provided in Fig. 6.1. Note that the detection schemes for systems with random watermarks first extract the watermark bits which are then used to calculate the bit error rate (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

(a) Embedding scheme. (b) Extraction scheme.

Fig. 6.1 An example of pattern (*left*) and random (*right*) watermarks

watermark pattern in [9] 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 audio watermarking systems.

6.2.3 Handling Deynchronization

Desynchronization attacks are the most difficult attacks in audio watermarking systems. This is intrinsically because the pitch-invariant time scaling, time-invariant pitch scaling, and resampling attacks apply nonlinear operations to the host signal which could not be sufficiently restored by linear operations. Also note that inserting synchronization bits [6, 9–14] (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 [5, 15]. The essential contributions of the two works lie in the proposals of using the histogram and robust audio segment extractor (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 [5] is highly effective in dealing with desynchronization attacks, although only the simplest realization of the spread spectrum (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.

6.2.4 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. Combining echo-based systems with the concept of constant frame number and a localized embedding scheme would achieve this goal. Specifically, one may set a constant frame number 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 jittering, time-invariant pitch scaling, pitch-invariant time scaling, and resampling attacks. Furthermore, if constant frame number is combined with localized watermark embedding, then the resultant system could also be robust against cropping and time shifting attacks, making a system simultaneously robust against a series of desynchronization attacks. In addition, time domain desynchronization-invariant features which are compatible with echo kernel and linear filtering process are worth of investigations.

6.2.5 Controlling Imperceptibility

Generally, imperceptibility is very likely to be compromised when robustness is 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 Chap. 2 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 objective difference grade (ODG) based heuristic tuning could be currently the optimal under practical considerations, because it enables the system being both automated and effective. Further, alternative solutions that discover available embedding spaces other than using psychoacoustic models (e.g., harmonics [16]) are also a potential research focus that could lead to further imperceptibility improvement.

6.2.6 Time-Frequency Domain Approach

It has been indicated from Chap. 2 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 an audio signal is 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.

6.3 Conclusion

The above discussions have provided a few concepts and general examples to reveal possible strategies to deal with the current challenges in developing practically robust audio watermarking systems. The great potentials for further improvements have been revealed, which call for future research attentions. 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 a watermarking system than for designers to protect it. It is still unknown whether a globally robust audio watermarking system could be designed.

Finally, the reversible audio watermarking introduced in Chap. 4 and the over-complete dictionary based watermarking system in Chap. 5 are among the emerging topics in this area. They are worth of further research attention for creating novel applications as well as system frameworks.

References

1. Xiang Y, Natgunanathan I, Rong Y, Guo S (2015) Spread spectrum-based high embedding capacity watermarking method for audio signals. IEEE/ACM Trans Audio Speech Lang Process 23(12):2228–2237
2. Cox IJ, Doërr G, Furon T (2006) Digital watermarking: watermarking is not cryptography. In: Proceedings of 5th international workshop, IWDW 2006, Jeju Island, Korea, November 8–10, 2006. Springer, Heidelberg, pp 1–15
3. Bas P, Furon T (2013) A new measure of watermarking security: the effective key length. IEEE Trans Inf Forensics Secur 8(8):1306–1317
4. Kang X, Yang R, Huang J (2011) Geometric invariant audio watermarking based on an lcm feature. IEEE Trans Multimed 13(2):181–190
5. Pun CM, Yuan XC (2013) Robust segments detector for de-synchronization resilient audio watermarking. IEEE Trans Audio Speech Lang Process 21(11):2412–2424
6. Megas D, Serra-Ruiz J, Fallahpour M (2010) Efficient self-synchronised blind audio watermarking system based on time domain and fft amplitude modification. Signal Process 90(12):3078–3092
7. Xiang Y, Natgunanathan I, Guo S, Zhou W, Nahavandi S (2014) Patchwork-based audio watermarking method robust to de-synchronization attacks. IEEE/ACM Trans Audio Speech Lang Process 22(9):1413–1423
8. Khaldi K, Boudraa AO (2013) Audio watermarking via EMD. IEEE Trans Audio Speech Lang Process 21(3):675–680
9. Lei B, Soon IY, Tan EL (2013) Robust svd-based audio watermarking scheme with differential evolution optimization. IEEE Trans Audio Speech Lang Process 21(11):2368–2377
10. Wang XY, Qi W, Niu PP (2007) A new adaptive digital audio watermarking based on support vector regression. IEEE Trans Audio Speech Lang Process 15(8):2270–2277
11. Wang XY, Niu PP, Yang HY (2009) A robust, digital-audio watermarking method. IEEE Multimed 16(3):60–69
12. Lie WN, Chang LC (2006) Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification. IEEE Trans Multimed 8(1):46–59
13. Wang XY, Zhao H (2006) A novel synchronization invariant audio watermarking scheme based on DWT and DCT. IEEE Trans Signal Process 54(12):4835–4840
14. Arnold M (2000) Audio watermarking: features, applications and algorithms. In: IEEE international conference on multimedia and expo 2000, (ICME 2000), vol 2. IEEE, pp 1013–1016
15. Xiang S, Huang J (2007) Histogram-based audio watermarking against time-scale modification and cropping attacks. IEEE Trans Multimed 9(7):1357–1372
16. Geyzel Z (2014) Audio watermarking