

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

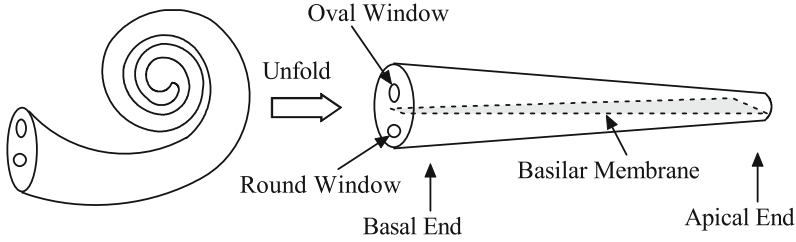


Fig. 2.1 Illustrative diagram of the cochlear in inner ear

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:

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

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]

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

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} :

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

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

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

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]:

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

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.

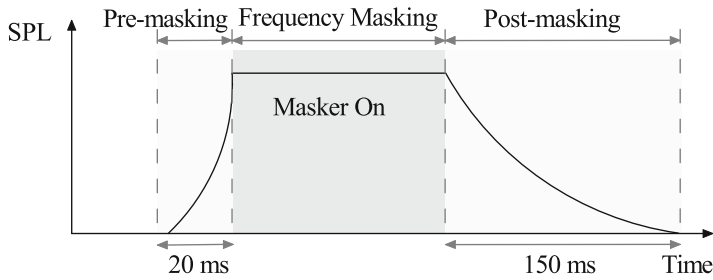


Fig. 2.2 Time masking and frequency masking

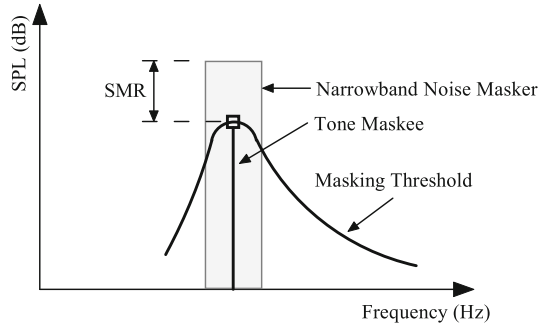
2.1.3 Masking Effects

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:

$$s_w(n) = s(n) + \alpha s(n - d), \quad (2.6)$$

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].

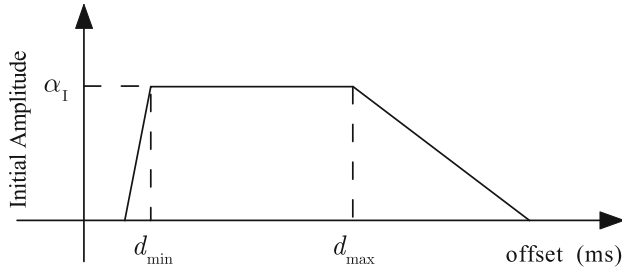


Fig. 2.4 The imperceptible region for a single echo, where $\alpha_I = 0.31$, $d_{\min} = 0.9$ ms, and $d_{\max} = 3.4$ ms

- 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 several critical bands are especially important.

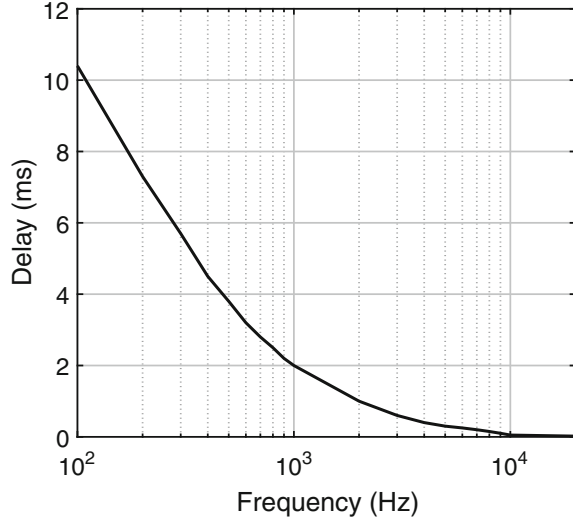
These experimental results suggest that, for an echo kernel to be imperceptible, the frequency response should be as flat as possible in the lower Bark bands. Besides, large fluctuations should be pushed to higher Bark bands. In this respect, the echo kernel with only one single delay is not optimum. 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

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

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

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

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

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

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 permuted 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 20s. 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 $SDG = S_W - S_{HR}$. The last column of Table 2.2 shows the SDG values versus the impairment scale. So, $SDG = 0$ means the watermark is imperceptible, while $SDG = -4.0$ 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.

¹The 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 $s_w(n)$ 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:

$$\text{SNR} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} [s(n) - s_w(n)]^2}, \quad (2.10)$$

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 $s_w(n)$. 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 $\text{SNR}(m)$ be the SNR for the m th segment. Then, the time-domain segmental SNR is calculated as:

$$\text{segSNR} = \frac{1}{M} \sum_{m=0}^{M-1} \max \{ \min \{ 35, \text{SNR}(m) \}, -10 \}, \quad (2.11)$$

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 -10dB .

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 $X_{m,w}(k)$. 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:

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

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

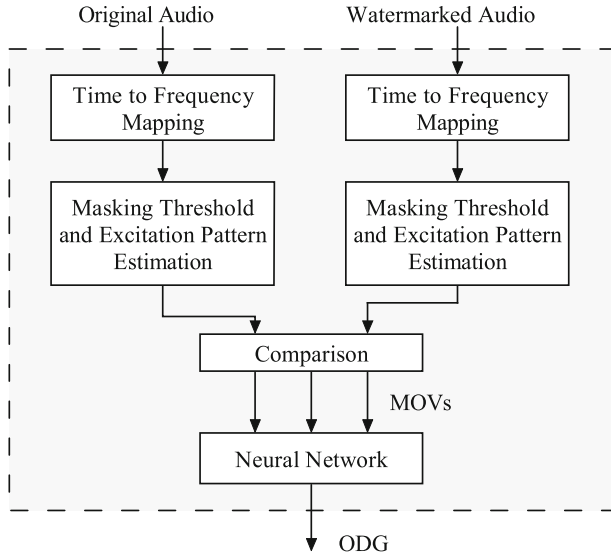
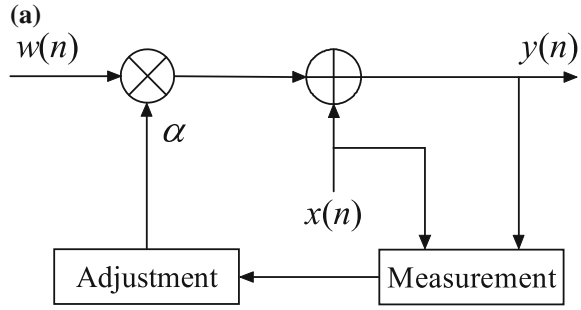


Fig. 2.6 Block diagram of PEAQ [26]

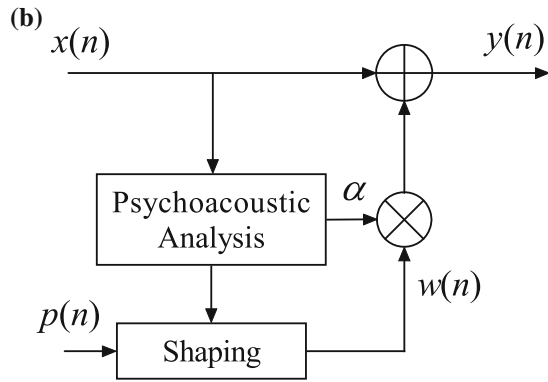
performing psychoacoustic analysis of the frequency components using masking threshold method and excitation pattern method. For the masking threshold method, the masking threshold of the original audio signal is analyzed and compared to the spectrum of the watermarked audio signal. The inaudible components, 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

Heuristic control.



Analytic 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:

$$P_{\text{TM}}(k) = 10 \log_{10} [10^{0.1P(k-1)} + 10^{0.1P(k)} + 10^{0.1P(k+1)}]. \quad (2.13)$$

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

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

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

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

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:

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

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

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

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

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

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

where N_{TM} and N_{NM} are the numbers of tonal maskers and noise maskers, respectively, and $T_{\text{ATH}}(i)$ 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

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

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

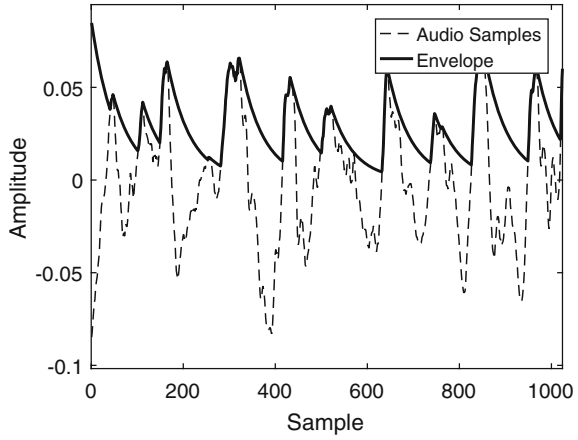
For the *perceptual filtering* approach [40], an all-pole filter $H(f)$ is designed to approximate the masking threshold $M(f)$ within each frame:

$$\min_{H(f)} \int_{-1/2}^{1/2} [\sigma^2 |H(f)|^2 - M(f)]^2 df, \quad (2.20)$$

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 squared and normalized $t(n)$ to achieve time domain shaping [40]:

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

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

Digital Audio Watermarking

Fundamentals, Techniques and Challenges

Xiang, Y.; Hua, G.; Yan, B.

2017, XII, 90 p. 32 illus., 3 illus. in color., Softcover

ISBN: 978-981-10-4288-1