

Chapter 2

Principles of Psychoacoustics

Psychoacoustics is the science of sound perception, i.e., investigating the statistical relationships between acoustic stimuli and hearing sensations [51]. This study aims to build up the psychoacoustic model, a kind of quantitative model, which could closely match the hearing mechanism. A good understanding of the sensory response of the human auditory system (HAS) is essential to the development of psychoacoustic models for audio watermarking, where the perceptual quality of processed audio must be preserved to the greatest extent.

In this chapter, the basic structure and function of the auditory system, mainly the peripheral part, are illustrated for the comprehension of human hearing. Then, the hearing threshold and auditory masking phenomenon are analyzed to pave the way for deriving the psychoacoustic models. Finally, Psychoacoustic Model 1 in ISO/MPEG standard is implemented to be utilized in our audio watermarking scheme later on.

2.1 Physiology of the Auditory System

Hearing is the sense by which sound is perceived [52]. Human hearing is performed primarily by the auditory system, in which the peripheral part is of more relevance to our study. The peripheral auditory system (the ear, that portion of the auditory system not in the brain [53]) includes three components: the outer ear, the middle ear, and the inner ear, as illustrated in Fig. 2.1.

The whole process of capturing the sound through the ear to create neurological signals is an intricate and ingenious procedure. First, the sound wave travels through the auditory canal and causes the eardrum to vibrate. This vibration is transmitted via the ossicles of the middle ear to the oval window at the cochlea inlet. The movement of the oval window forces the fluid in the cochlea to flow, which results in the vibration of the basilar membrane that lies along the spiral cochlea. This motion causes the hair cells on the basilar membrane to be stimulated and to generate neural

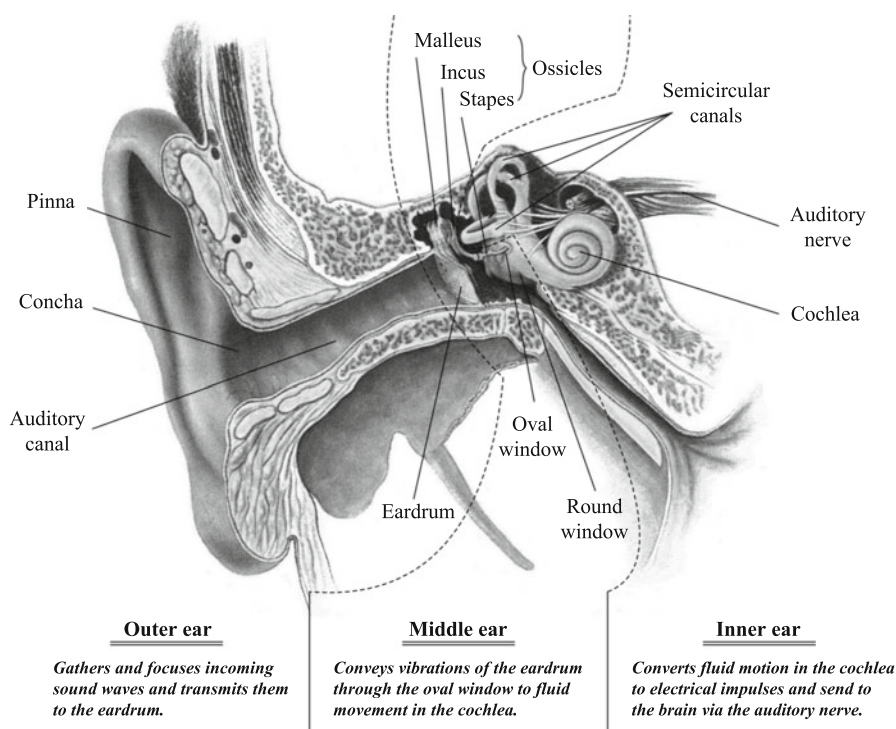


Fig. 2.1 Structure of the peripheral auditory system [57]

responses carrying the acoustic information. Then, the neural impulses are sent to the central auditory system through the auditory nerves to be interpreted by the brain [54, 55].

2.1.1 The Outer Ear

Sounds communicate the auditory system via the outer ear. The pinna and its deep center portion, the concha, constitute the externally visible part of the outer ear that serves focusing the sound waves at the entrance of the auditory canal (or auditory meatus). Since human pinna has no useful muscles, it is nearly immobile. Therefore, the head must be reoriented towards the direction of acoustical disturbance for a better collection and localization of sound. The auditory canal (usually 2–3 cm in length) is a tunnel through which the sound waves are conducted, and it is closed

with the eardrum (or tympanic membrane).¹ The eardrum is stretched tightly across the inner end of the auditory canal and is pulled slightly inward by structures in the middle ear [58]. Upon travelling through the auditory canal, sound waves impinge on the eardrum and cause it to vibrate. Then, these mechanical vibrations which respond to the pressure fluctuations of acoustic stimuli are passed along to the middle ear.

The outer ear plays an important role in human hearing. The pinna is of great relevance to sound localization, since it reflects the arriving sound in ways that depend on the angle of the source. The resonances occurring in the concha and auditory canal bring about an increase on sound pressure level (SPL) for frequencies between 1.5 kHz and 7 kHz. The extent of amplification depends on both the frequency and angle of the incident wave, as indicated in Fig. 2.2. For example, the gain is about 10–15 dB in the frequency range from 1.5 kHz to 7 kHz at an azimuthal angle of 45°. Moreover, the outer ear protects the eardrum and the middle ear against extraneous bodies and changes in humidity and temperature [59].

2.1.2 *The Middle Ear*

The eardrum vibrations are transferred through the middle ear to the inner ear. The middle ear is an air-filled chamber, bounding by the eardrum laterally and by the oval window of the cochlea medially. It contains three tiny bones known as the ossicles: the malleus (or hammer), incus (or anvil), and stapes (or stirrup). These three ossicles are interconnected sequentially and suspended in the middle ear cavity by ligaments and muscles. As shown in Fig. 2.1, the malleus is fused to the eardrum and articulates with the incus; the incus is connected to both the other bones; the stapes is attached to the incus and its footplate fits into the oval window of the cochlea. The oval window is a membrane-covered opening which leads from the middle ear to the vestibule of inner ear.

As an interface between the outer and inner ears, the middle ear has two functions. One function is to serve as an impedance-matching transformer that ensures an efficient transmission of sound energy. As we know, the outer and middle ear cavities are filled with air, while the inner ear is filled with fluid. So the passage of pressure waves from the outer ear to the inner ear involves a boundary between air and fluid, two mediums with different acoustic impedance.² In fact, approximately 99.9 % of sound energy incident on air/fluid boundary is reflected back within the air medium, so that only 0.1 % of the energy is transmitted to the fluid. It means that

¹In this sense, the auditory canal closed with the eardrum at its proximal end has a configuration as a resonator.

²Acoustic impedance is a constant related to the propagation of sound waves in an acoustic medium. Technically, sound waves encounter much less resistance when travelling in air than in fluid.

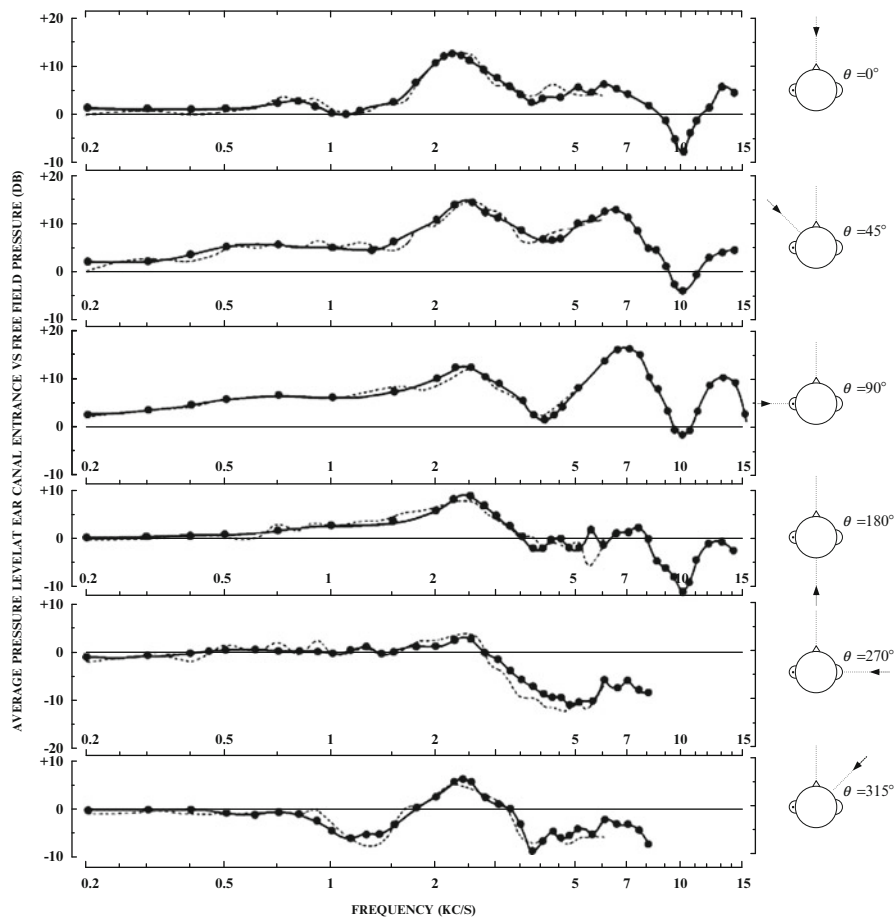


Fig. 2.2 Average pressure levels at auditory canal entrance versus free-field pressure, at six azimuthal angles of incidence [60]. *Notes:* (1) The sound pressure was measured with a probe tube located at the left ear of the subject. (2) A point source of sound was moved around a horizontal circle of radius 1 m with the subject’s head at the center. At $\theta = 0^\circ$, the subject was facing the source, and at $\theta = 90^\circ$, the source was normally incident at plane of left ear

if sound waves were to hit the oval window directly, the energy would undergo a loss of 30 dB before entering the cochlea. To minimize this reduction, the middle ear has two features to match up the low impedance at the eardrum with high impedance at the oval window. The first is related to the relative sizes of the eardrum and the stapes footplate which clings to the oval window. The effective area of the eardrum is about 55 mm^2 and that of the footplate is about 3.2 mm^2 ; thereupon they differ in size by a factor of 17 ($55\text{ mm}^2/3.2\text{ mm}^2 = 17$). So, if all the force exerted on the eardrum is transferred to the footplate, then the pressure (force per unit area) at the oval window is 17 times greater than at the eardrum. The second depends on the lever action of the

ossicular chain that amplifies the force of the incoming auditory signals. The lengths of the malleus and incus correspond to the distances from the pivot to the applied and resultant forces, respectively. Measurements indicate that the ossicles as a lever system increases the force at the eardrum by a factor of 1.3. Consequently, the combined effect of these actions effectively counteracts the reduction caused by the impedance mismatch [58]. Another function of the middle ear is to diminish the transmission of bone-conducted sound to the cochlea by muscle contraction. If these sounds were sent over to the cochlea, they would appear very loud that may be harmful to the inner ear [61].

2.1.3 *The Inner Ear*

The inner ear transduces the vibratory stimulation from the middle ear to neural impulses which are transmitted to the brain. The vestibular apparatus and the cochlea are the main parts in the inner ear. The vestibular apparatus is responsible for the sense of balance. It includes three semicircular canals and the vestibule. The cochlea is the central processor of the ear, where the organ of corti, the sensory organ of hearing, is located. The cochlea is a spiral-shaped bony tube structure of decreasing diameter, which coils up $2\frac{3}{4}$ times around a middle core containing the auditory nerve, as shown in Fig. 2.3a.³ The duct is filled with almost incompressible fluids and is enclosed by the oval window (the opening to the middle ear) and the round window (a membrane at the rear of the cochlea). When the stapes pushes back and forth on the oval window, the motion of the oval window causes the fluid to flow and impels the round window to move reciprocally, which lead to the variations of fluid pressure in the cochlea. The movements of the oval and round windows are indicated by the solid and dotted arrows in Fig. 2.3a.

Figure 2.3c shows the cross-section through one cochlea turn. Two membranes, Reissner's membrane and the basilar membrane, divide the cochlea along the spiral direction into three fluid-filled compartments: scala vestibuli, scala media, and scala tympani. The scala vestibuli and scala tympani are merged through a small opening called helicotrema at the apex, and they contain the same fluid (the perilymph) with most of the nervous system. The scala media is segregated from other scalae and contains a different fluid (the endolymph). On the scala media surface of basilar membrane (BM) lies the organ of corti. The changes of fluid pressure in the cochlea will cause the BM to deform, so that the hair cells⁴ on the organ of corti are

³Note that the cochlea is a cavity within the skull, not a structure by itself [58]. Hence the unraveled cochlea in Fig. 2.3b is impossible in practice, only for the sake of illustration.

⁴The hair cells including the outer and inner hair cells (OHC and IHC) are auditory receptors on the organ of corti.

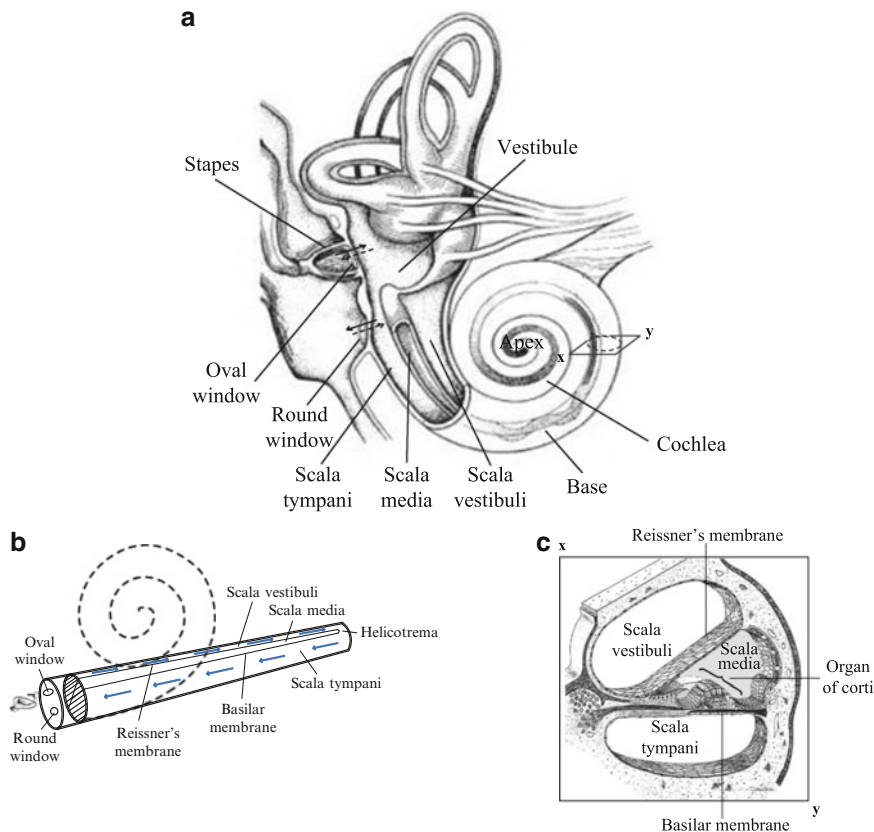


Fig. 2.3 Anatomy of the cochlea (a) Relative location of the cochlea in the inner ear [61] (b) Schematic of the unraveled cochlea (c) Cross-section through one cochlea turn [65]

stimulated to transduce the movement of the BM into neural impulses. Then the neural signals are carried over to the brain via auditory nerve, which ultimately lead to the perception of sound.

The basilar membrane extends along the spirals of the cochlea and is about 32 mm long. It is relatively narrower and stiffer at the base (near the windows), while it gets wider and more flexible at the apex (near the helicotrema). Accordingly, each location on the BM has different vibratory amplitude in response to sound of different frequencies, which means that each point resonates at a specific characteristic frequency (CF) [54]. As exemplified in Fig. 2.4a, for high-frequency tones, the maximum displacement of the BM occurs near the base, with tiny movement on the remainder of the membrane. For low-frequency tones, the vibration travels all the way along the BM, reaching its maximum close to the apex.⁵ Figure 2.4b

⁵There is one fact worth of attention, i.e., any location on the BM will respond to a wide range of tones that are lower than its CF. That's why low frequencies are less selective than high frequencies.

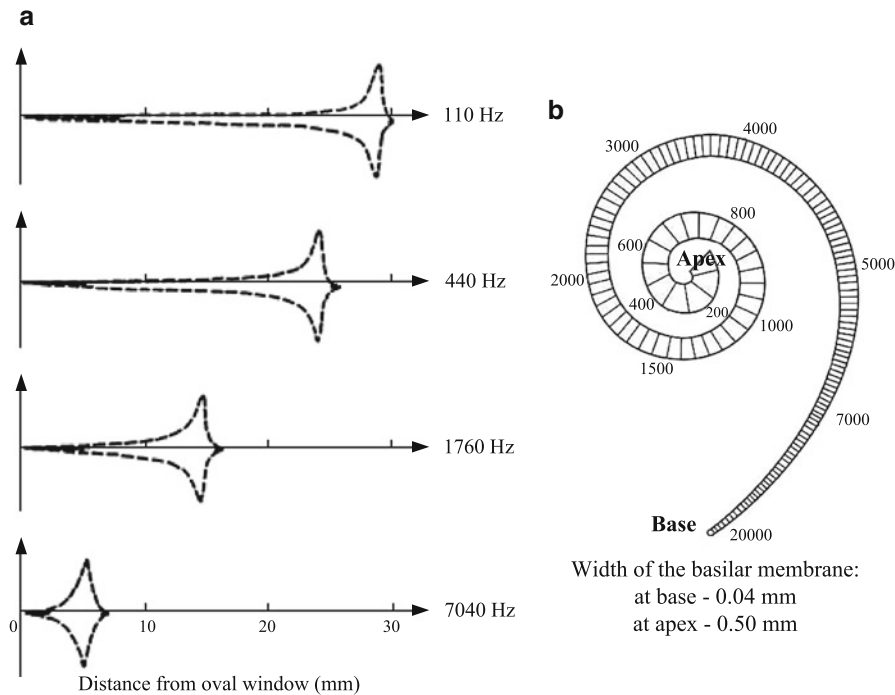


Fig. 2.4 Resonant properties of the basilar membrane (a) Envelopes of vibration patterns on the basilar membrane in response to sound of different frequencies [66] (b) Distribution of resonant frequencies along the basilar membrane [64]

summarizes the distribution of frequencies that produce maximum displacement at different positions along the basilar membrane. Note that the spacing of resonant frequencies is not linear to the frequency, but in a logarithmic scale approximately. It is called Bark scale or critical band rate corresponding to the concept of critical bands.

In this sense, the cochlea performs a transformation that maps sound frequencies onto certain locations along the basilar membrane, i.e., a “frequency-to-place” conversion [51]. It is of great importance to the comprehension of auditory masking. Since one frequency maximally excite only one particular point on the basilar membrane, the auditory system acts as a frequency analyzer which can distinguish the frequencies from each other. If two tones are different enough in frequency, the response of the BM to their combination is simply the addition of two individual

ones. That is, there are two vibration peaks along the BM, at the positions identical to where they would be if two tones were presented independently. However, if two tones are quite close in frequency, the basilar membrane would fail to separate the combination into two components, which results in the response with one fairly broad peak in displacement instead of two single peaks [58]. As for the interval how far two tones can be discriminated, it depends on critical bands and critical bandwidths discussed next.

2.2 Sound Perception Concepts

Sounds are rapid variations in pressure, which are propagated through the air away from acoustic stimulus. Our sense of hearing allows us to perceive sound waves of frequencies between about 20 Hz and 20 kHz. As discussed in the mechanism of human ear, perception of sound involves a complex chains of events to read the information from sound sources. Naturally, we are often surrounded with a mixture of various sounds and the perception of one sound is likely to be obscured by the presence of others. This phenomenon is called auditory masking, which is the fundamental of psychoacoustic modelling. Here, some basic terms related to auditory masking are introduced.

2.2.1 Sound Pressure Level and Loudness

Sound reaches human ear in the form of pressure waves varying in time, $s(t)$. Physically, the pressure p is defined as force per unit area, and the unit in MKS system is Pascal (Pa) where $1 \text{ Pa} = 1 \text{ N/m}^2$. Also, the intensity is defined as power per unit area and its unit is W/m^2 . In psychoacoustics, values of sound pressure vary from 10^{-5} Pa (ATH, absolute threshold of hearing) to 10^2 Pa (threshold of pain). To cover such a broad range, (SPL) is defined in logarithm units (dB) as

$$L_{\text{SPL}}/\text{dB} = 10 \log_{10} \left(\frac{p}{p_0} \right)^2 = 10 \log_{10} \left(\frac{I}{I_0} \right), \quad (2.1)$$

where L_{SPL} is the SPL of a stimulus, p is the pressure of stimulus in Pa, $p_0 = 20 \mu\text{Pa}$ is the reference pressure of a tone with frequency around 2 kHz, I is sound intensity of the stimulus, and $I_0 = 10^{-12} \text{ W/m}^2$ is the reference's intensity correspondingly [11].

The hearing sensation that relates to SPL is loudness of sound, expressed in *phon*. Note that loudness is a psychological, not a physical, attribute of sound. By definition, the loudness level of a 1 kHz tone is equal to its SPL in dB SPL [61]. The perceived loudness of sound depends upon its frequency as well as its intensity, as described by a series of equal-loudness contour in Fig. 2.5. Each equal-loudness

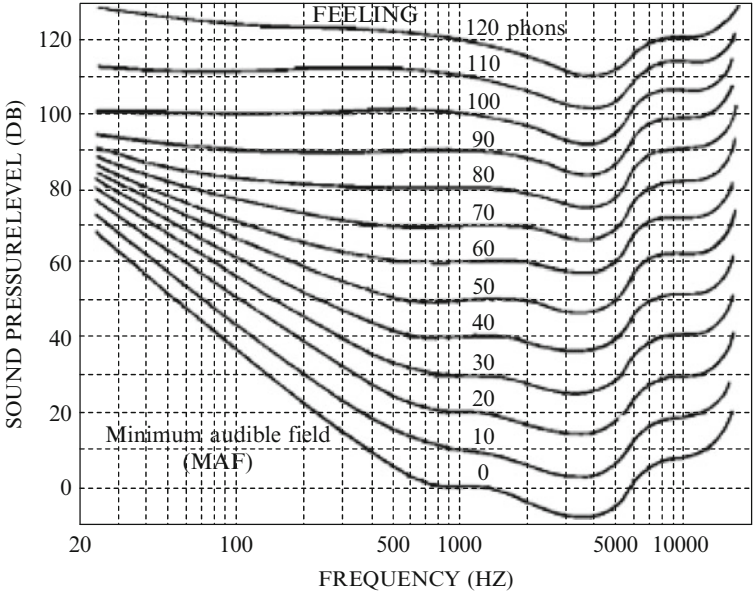


Fig. 2.5 Equal-loudness contours [69]

contour represents SPLs required at different frequencies in order that all tones on the contour are perceived equally loud [68]. The loudness of 20 phons contour at 100 Hz with 50 dB SPL is perceived similar to 1 kHz with 20 dB SPL. In Fig. 2.5, the deviation from the maximum sensitivity region of equal-loudness contours at high phons (i.e., 120 phons) is lower than those of low phons (i.e., 10 phons). This indicates that the sensitivity to frequency changing of HAS at low phons is relatively higher than high phons. Hence, complex sounds with identical frequency and phase components might sound different due to variations in loudness [58].

2.2.2 Hearing Range and Threshold in Quiet

Human hearing spreads widely from 20 Hz to 20 kHz in frequency, as well as ranging from about 0 dB up to 120 dB in SPL. The most sensitive part is between 100 Hz and 8 kHz for human speech. Figure 2.6 shows hearing range of human, where different hearing thresholds are sketched in SPL curves as function of frequency.

The hearing threshold at the bottom is the threshold in quiet, or (ATH), which approximately corresponds to the baseline in Fig. 2.5. It decreases gradually from 20 Hz to 3 kHz and then increases sharply above 16 kHz. The threshold in quiet indicates, as a function of frequency, the minimum SPL of a pure tone to be audible in a noiseless environment. Thus under no circumstances the human ear can perceive

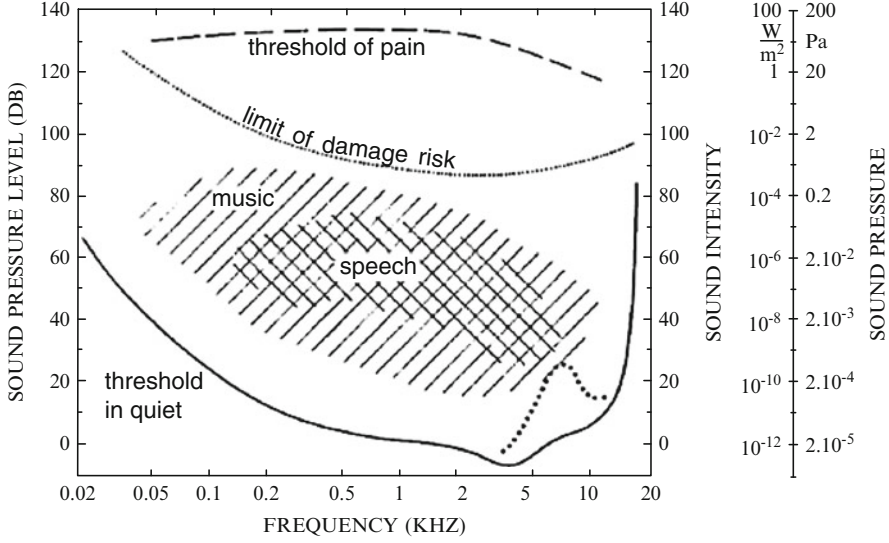


Fig. 2.6 Hearing range [11]

sounds at SPLs below that threshold. In other words, frequency components that fall below the threshold in quiet are insignificant to our perception of sound and unnecessary to be processed [51]. This property is crucial to the development of psychoacoustic model, where the threshold in quiet is approximated by the following frequency-dependent function:

Threshold in Quiet (f) / dB =

$$3.64 \left(\frac{f}{1000} \right)^{-0.8} - 6.5 \exp \left\{ -0.6 \left(\frac{f}{1000} - 3.3 \right)^2 \right\} + 10^{-3} \left(\frac{f}{1000} \right)^4 \quad (2.2)$$

as plotted on both linear and logarithmic scales in Fig. 2.7. Regarding Eq. (2.2), one point to note is that it only applies to the frequency range $20 \text{ Hz} \leq f \leq 20 \text{ kHz}$.

2.2.3 Critical Bandwidth

As discussed in Sect. 2.1.3, the cochlea performs a “frequency-to-place” conversion and each position on the basilar membrane responds to a limited range of frequencies. Accordingly, the peripheral auditory system acts as a spectrum analyzer, modelling as a bank of band-pass filters with overlapping passbands [61]. Empirically, the main hearing range between 20 Hz and 16 kHz is divided into 24 nonoverlapping critical bands, and the critical bandwidths (CB) are listed

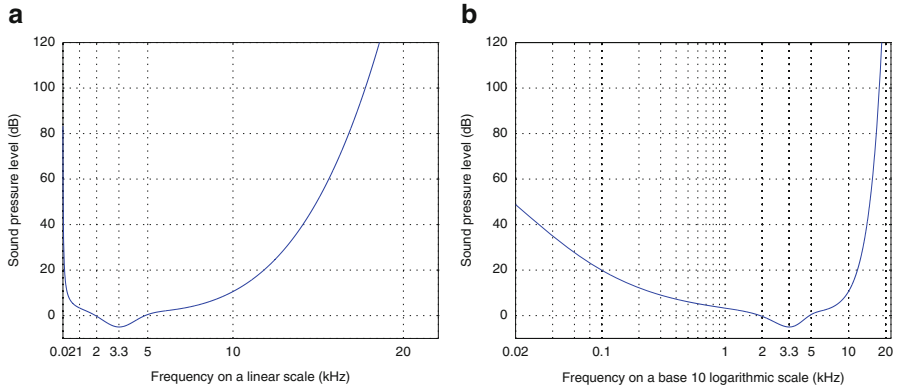


Fig. 2.7 Approximation for the threshold in quiet (a) Frequency on a linear scale (b) Frequency on a logarithmic scale

in Appendix D. We call it critical band rate scale and its unit is *Bark*. One Bark represents one critical band and corresponds to a distance along the basilar membrane of about 1.3 mm.⁶ Considering nonlinear spacing of resonant frequencies on the basilar membrane, it is expected that critical bandwidths are nonuniform, varying as a function of frequency. The following equation describes the dependence of Bark scale on frequency [11]:

$$z/\text{Bark} = 13 \arctan\left(\frac{0.76f_l}{1000}\right) + 3.5 \arctan\left(\frac{f_l}{7500}\right)^2, \quad (2.3)$$

where f_l is the lower frequency limit of critical bandwidth. For example, the threshold in quiet in Fig. 2.7 is plotted on Bark scale as shown in Fig. 2.8.

Note that each critical bandwidth only depends on the center frequency of the passband. It is demonstrated in Fig. 2.9, where the critical bandwidth at 2 kHz is measured. As shown in Fig. 2.9a, hearing threshold is flat about 33 dB until two tones are about 300 Hz away from each other, and then it drops off rapidly. A similar result is obtained from Fig. 2.9b, hearing threshold is rather flat about 46 dB until two noises are away from 300 Hz [51]. Consequently, the critical bandwidth is 300 Hz for a center frequency of 2 kHz. It is worth mentioning that the threshold in Fig. 2.9b is at 46 dB versus only 33 dB in a, which means narrowband noises reduce

⁶The whole length of 32 mm basilar membrane divided by 24 critical bands is 1.3 mm for each band.

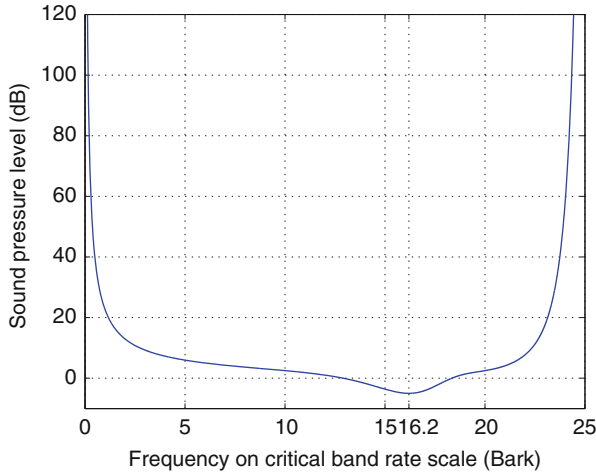


Fig. 2.8 Threshold in quiet on Bark scale

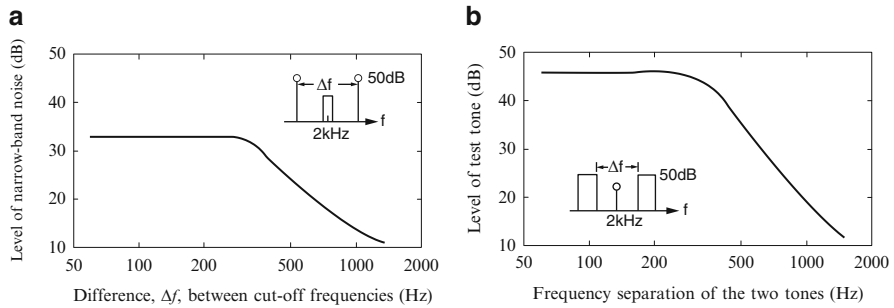


Fig. 2.9 Determination of the critical bandwidth [11] (a) The threshold for a narrowband noise 2 kHz centered between two tones of 50 dB as a function of the frequency separation between two tones (b) The threshold for a tone of 2 kHz centered between two narrowband noises of 50 dB as a function of the frequency separation between the cutoff frequencies of two noises

more audibility than tones. This fact is referred to “asymmetry of masking” and more details will be discussed in the next section.

On the basis of experimental data, an analytic expression is derived to better describe critical bandwidth Δf as a function of center frequency f_c [11]:

$$\Delta f/\text{Hz} = 25 + 75 \left[1 + 1.4 \left(\frac{f_c}{1000} \right)^2 \right]^{0.69}. \quad (2.4)$$

The concept of critical bandwidth contributes to the understanding of auditory masking, because CB around a masker denotes the frequency range over which the main masking effect operates. As demonstrated in Sect. 2.3.1, the masking curves distribute equally across the spectrum in Bark scale.

2.3 Auditory Masking

Due to the effect of auditory masking, the perception of one sound is related to not only its own frequency and intensity, but also its neighbor components. Auditory masking refers to the phenomenon that one faint but audible sound (the maskee) becomes inaudible in the presence of another louder audible sound (the masker). It has a great influence on hearing sensation and involves two types of masking, i.e., simultaneous masking and nonsimultaneous masking (including pre-masking and post-masking) as displayed in Fig. 2.10. Due to auditory masking, any signals below these curves cannot be heard. Therefore, by virtue of auditory masking, we can modify audio signals in a certain way without perceiving deterioration, as long as the modifications could be properly “masked.” This notion is the essence of audio watermarking [70, 71].

2.3.1 Simultaneous Masking

Simultaneous masking (or frequency masking) refers to masking between two sounds with close frequencies, where the low-level maskee is made inaudible by simultaneously occurring louder masker. Both masker and maskee can be sinusoidal

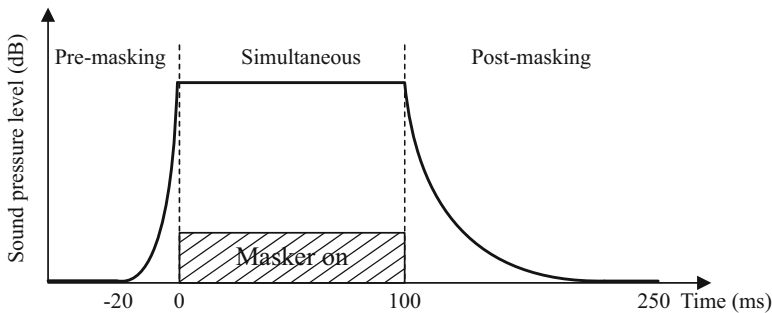


Fig. 2.10 Two types of masking: simultaneous and nonsimultaneous masking

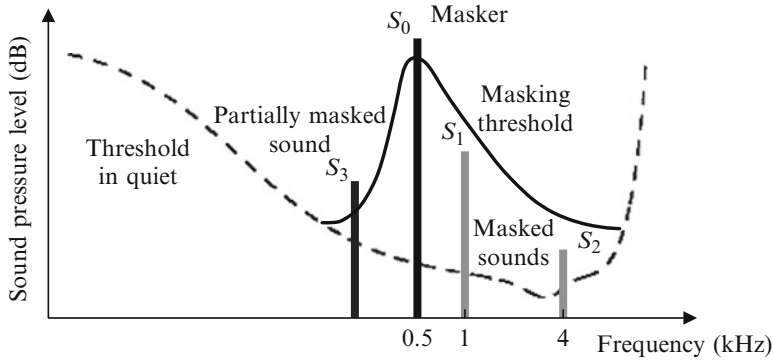


Fig. 2.11 Simultaneous masking

tone or narrowband noise.⁷ Figure 2.11 gives an example of simultaneous masking, where sound S_0 is the masker. Because of the presence of S_0 , the threshold in quiet is elevated to produce a new hearing threshold named as masking threshold. The masking threshold is a kind of limit for just noticeable distortion (JND) [72], which means that any sounds or frequency components below this threshold are masked by the presence of the masker. For instance, the weaker signal S_1 and S_2 are completely inaudible, as their SPLs are below the masking threshold. For the signal S_3 , it is partially masked and only the portion above the threshold is perceivable. Moreover, the effective masking ranges for the maskers at different frequencies are determined solely by critical bandwidths, as implied in Fig. 2.9. If the maskee lies in critical band of the masker, the maskee is more likely to be unperceived. The mechanism by which masking occurs is still uncertain [61]. In general, it is because the louder masker creates an excitation of sufficient strength on the basilar membrane. Then such an excitation prevents the detection of another excitation within the same critical band from a weaker sound [51].

The masking threshold depends on the characteristics of both masker and maskee. Considering two possibilities of each, there are four cases in simultaneous masking, that is, narrowband noise masking tone (NMT), tone masking tone (TMT), narrowband noise masking narrowband noise (NMN), and tone masking narrowband noise (TMN).

2.3.1.1 Narrowband Noise Masking Tone

Most often, the case of NMTs happens, where the masker is narrowband noise and the maskees are tones located in the same critical band. Figure 2.12 shows the masking thresholds for narrowband noise masker masking tones, where the noise is

⁷Here, narrowband means the bandwidth equal to or smaller than a critical band.

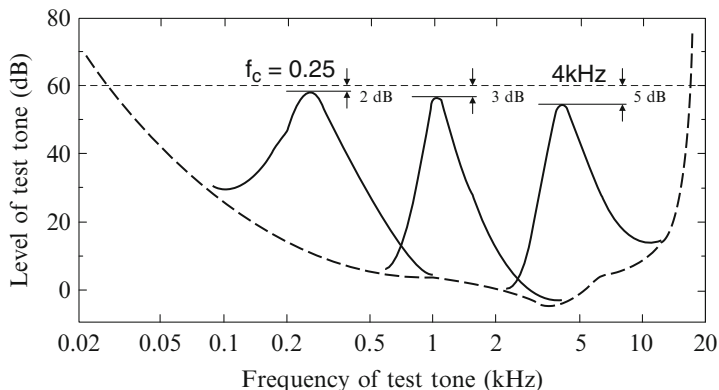


Fig. 2.12 Masking thresholds for a 60 dB narrowband noise masker centered at different frequencies [51]

at a SPL of 60 dB and centered at 0.25, 1, and 4 kHz separately. In the graph, solid lines represent masking thresholds, and the dashed line at the bottom is the threshold in quiet.⁸ The masking thresholds have a number of important features. For example, the form of curve varies with different maskers, but always reaches a maximum near the masker's center frequency. It means that the amount of masking is greatest when the maskee is located at the same frequency with the masker. The masking ability of a masker is indicated by the minimum signal to mask ratio (SMR), i.e., the minimum difference of SPL between the masker and its masking threshold. Therefore, higher SMR implies less masking. Another point is that low-frequency masker produces a broader masking threshold and provides more masking than high frequencies. Here, the 0.25, 1, and 4 kHz thresholds have a SMR of 2, 3, and 5 dB, respectively.

Figure 2.12 is sketched in normal frequency units, where the masking thresholds of different frequencies are dissimilar in shape. If graphed in Bark scale, all the masking thresholds look similar in shape as shown in Fig. 2.13.⁹ In this case, it is easier to model the masking threshold by the use of the so-called spreading function in Sect. 2.4.1.1. As a result, Bark scale is widely used in the area of auditory masking.

Moreover, the masking thresholds from a 1 kHz narrowband noise masker at different SPLs, L_{CB} , are outlined in Fig. 2.14. Although SPL of the masker is different, the minimum SMR remains constant at around 3 dB, corresponding to the value in Fig. 2.12. It means that the minimum SMR in NMT solely depends on the center frequency of masker. Also notice that the masking threshold becomes more asymmetric around the center frequency as the SPL increases. At frequencies lower than 1 kHz, all the curves have a steep rise. But at frequencies higher than 1 kHz,

⁸Hereafter, this rule does apply to all the graphs in Sect. 2.3.

⁹For illustration, all the curves are shifted upward to the masker's SPL (60 dB).

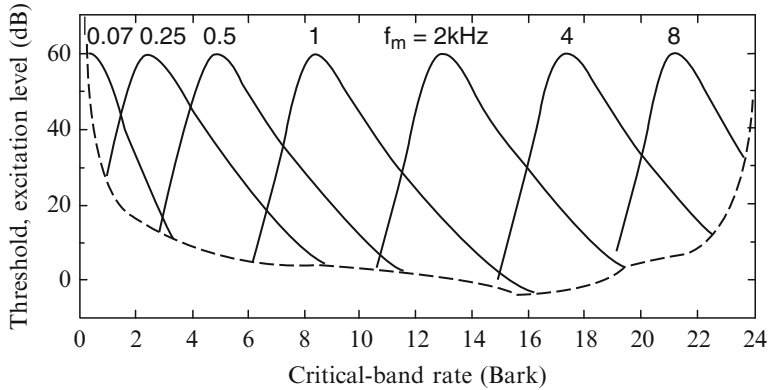


Fig. 2.13 Masking thresholds for a 60 dB narrowband noise masker centered at different frequencies in Bark scale [51]

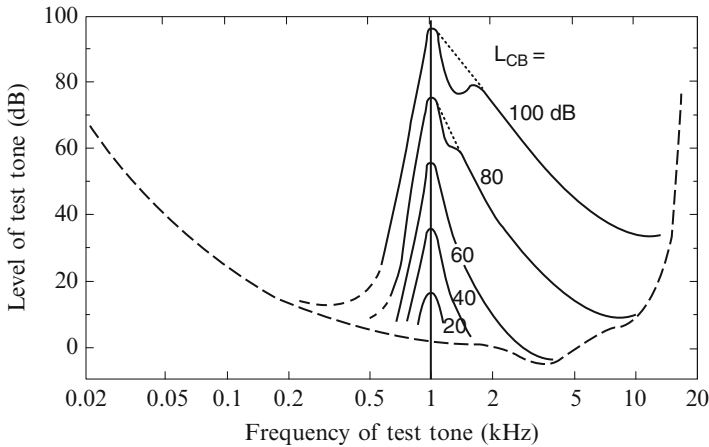


Fig. 2.14 Masking thresholds from a 1 kHz narrowband noise masker at different SPLs [51]

the slopes of maskers at higher SPLs decrease more gradually. Recall Fig. 2.4a; it is reasonable to expect that the masker is good at masking the tones whose frequencies are lower than its own frequency, rather than higher frequency tones [58]. To show the similarity in shape over all the masking thresholds, Fig. 2.15 plots the curves in Bark scale again.

2.3.1.2 Tone Masking Tone

The early work on auditory masking started from experiments on tones masking tones within the same critical band. Since both the masker and maskee are pure tones, their interference is likely to result in the occurrence of beats. Therefore,

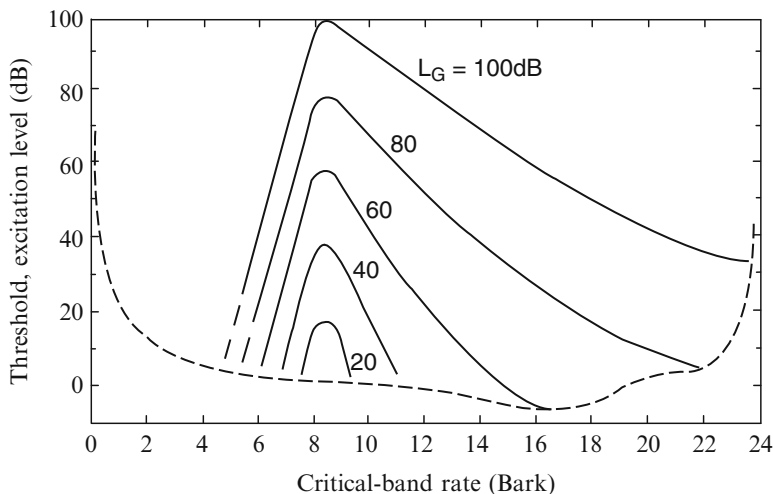


Fig. 2.15 Masking thresholds from a 1 kHz narrowband noise masker at different SPLs in Bark scale [51]

besides the masker and maskee, additional beating tones become audible and accordingly disturb the subjects. Figure 2.16 shows the masking thresholds from a 1 kHz tonal masker at different SPLs. During the course of approaching 1 kHz, the maskee was set 90° out of phase with the masker to prevent beating. Similar to Fig. 2.14, the masking thresholds spread also broader towards high frequencies than lower frequencies. However, an obvious difference lies in the minimum SMR, roughly 15 dB in Fig. 2.16 versus about 3 dB in Fig. 2.14. It indicates that the narrowband noise is a better masker than pure tone, referred as “asymmetry of masking” [73]. This fact actually has been demonstrated in Fig. 2.9 already. The masking threshold by narrowband noise masker in Fig. 2.9b is valued at 46 dB, higher than 33 dB by tonal masker in Fig. 2.9a. So in psychoacoustic modelling, we should identify the frequency components to be noise-like or tone-like and then calculate their masking thresholds separately.

2.3.1.3 Narrowband Noise or Tone Masking Narrowband Noise

In contrast to NMT and TMT, it is more difficult to characterize narrowband noise or tone masking narrowband noise. So far, relatively few studies in NMN and TMN are carried out. Under the case of NMN, the masking thresholds heavily rely on phase relationship between the masker and maskee. In other words, different relative phases between the masker and maskee would lead to different values of minimum SMRs. It is reported that measurements for wideband noise have minimum SMRs of about 26 dB [51, 73]. As for TMN, the minimum SMR tends to fluctuate between 20 and 30 dB [51].

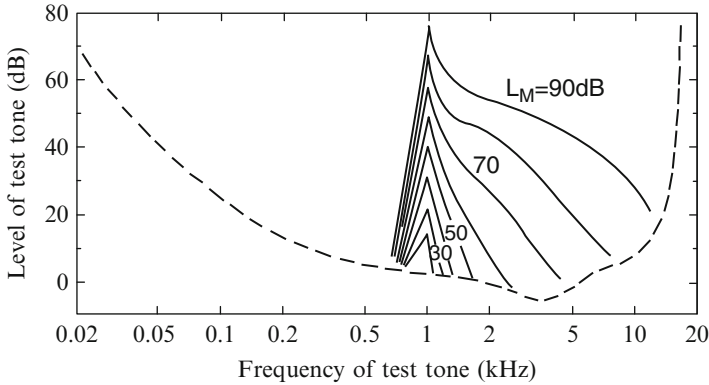


Fig. 2.16 Masking thresholds from a 1 kHz tonal masker at different SPLs [51]

2.3.2 *Nonsimultaneous Masking*

In addition to simultaneous masking, auditory masking can also take place when the maskee is present immediately preceding or following the masker. This is called nonsimultaneous masking or temporal masking. As exemplified in Fig. 2.10, one 200 ms masker masks a tone burst with very short duration relative to the masker.

There are two kinds of nonsimultaneous masking: (1) pre-masking or backward masking, occurring just before the onset of masker, and (2) post-masking or forward masking, occurring after the removal of masker. In general, the physiological basis of nonsimultaneous masking is that the auditory system requires a certain integration time to build the perception of sound, where louder sounds require longer integration intervals than softer ones [51].

2.3.2.1 **Pre-masking**

Pre-masking is somewhat unexpected since it happens before the presence of masker. As seen from Fig. 2.10, the duration of pre-masking is quite short (about 20 ms), whereas it is most effective only in 1–2 ms before the onset of masker [73]. It is suggested that the duration of masker might affect the time that pre-masking lasts. Up to now, however, no experimental results could specify such a relation.

Pre-masking has less masking capacity than post-masking and simultaneous masking; nevertheless, it plays a significant role in the compensation of pre-noise or pre-echo distortion [51].

2.3.2.2 Post-masking

Post-masking is better understood compared to pre-masking. It reflects a moderate decrease of the masking level after the masker is halted. As displayed in Fig. 2.10, post-masking level decays gradually to zero after a longer period of time (about 150 ms). Therefore, post-masking exhibits a higher masking capacity which is beneficial to most applications. Experimental studies have revealed that post-masking depends on the intensity and duration of the masker as well as relative frequency of the masker and maskee [51].

2.4 Psychoacoustic Model

The knowledge of auditory masking provides the foundation for developing psychoacoustic models. In psychoacoustic modelling, we use empirically determined masking models to analyze which frequency components contribute more to the masking threshold and how much “noise” can be mixed in without being perceived. This notion is applicable to audio watermarking, of which the imperceptibility is one prerequisite. Typically, in some audio watermarking techniques such as spread spectrum watermarking [74, 75] and wavelet domain watermarking [7, 76], the watermark signal is added to the host signal as a faint additive noise. To keep the watermarks inaudible, we often utilize the minimum masking threshold (MMT) calculated from psychoacoustic model to shape the amplitude of watermark signal.

2.4.1 *Modelling the Effect of Simultaneous Masking*

Modelling the effect of simultaneous masking is one major task of psychoacoustic model. In general, there are a series of steps involved. Firstly, the input audio signal is analyzed to classify its noise-like and tone-like frequency components, due to the phenomenon of “asymmetry of masking.” Secondly, the so-called spreading functions are derived to mimic the excitation patterns of noise-like and tone-like maskers, respectively. Thirdly, after shifted down by a certain amount for each masker, all the individual masking thresholds as well as ATH are added up in some manner to obtain a global masking threshold, an estimation on the concurrent masking effect. Finally, we take the lowest level of global masking threshold in each frequency band to obtain the (MMT), which represents the most sensitive limit.

2.4.1.1 Models for the Spreading of Masking

Models for the spreading of masking are developed to delineate excitation patterns of the maskers. As noticed from two examples of excitation patterns in Figs. 2.13

and 2.15, the shape of curves are quite similar and also easy to describe in Bark scale, because Bark scale is linearly related to basilar membrane distances. Accordingly, we define spreading function $SF(dz)$ as a function of the difference between the maskee and masker frequencies in Bark scale, $dz/\text{Bark} = z(f_{\text{maskee}}) - z(f_{\text{masker}})$. Apparently, $dz \geq 0$ when the masker is located at a lower frequency than the maskee, and $dz < 0$ when the masker is located at a higher frequency than the maskee.

There are a number of spreading functions introduced to imitate the characteristics of maskers. For instance, two-slope spread function is the simplest one that uses a triangular function:

$$10 \log_{10} SF(dz) / \text{dB} = \begin{cases} [-27 + 0.37 \max \{L_M - 40, 0\}] dz, & dz \geq 0 \\ 27dz, & dz < 0, \end{cases} \quad (2.5)$$

where L_M is SPL of the masker.

Another popular spreading function is proposed by Schroeder and expressed as the following analytical function:

$$10 \log_{10} SF(dz) / \text{dB} = 15.81 + 7.5(dz + 0.474) - 17.5 \sqrt{1 + (dz + 0.474)^2}. \quad (2.6)$$

After slight modification on Schroeder's spreading function, spreading function as Eq. (2.7) is adopted in ISO/IEC MPEG¹⁰ Psychoacoustic Model 2.

$$\begin{aligned} 10 \log_{10} SF(dz) / \text{dB} \\ = 15.8111389 + 7.5(1.05dz + 0.474) - 17.5 \sqrt{1 + (1.05dz + 0.474)^2} \\ + 8 \min \left(0, \left[(1.05dz - 0.5)^2 - 2(1.05dz - 0.5) \right] \right). \end{aligned} \quad (2.7)$$

It should be noted that the two spreading functions Eqs. (2.6) and (2.7) are independent of the masker's SPL, which is advantageous to reduction in computation when generating overall masking threshold.

The spreading function utilized in ISO/IEC MPEG Psychoacoustic Model 1 is different from Psychoacoustic Model 2:

$$10 \log_{10} SF(dz) / \text{dB} = \begin{cases} 17dz - 0.4L_M + 11, & -3 \leq dz < -1 \\ (0.4L_M + 6) dz, & -1 \leq dz < 0 \\ -17dz, & 0 \leq dz < 1 \\ -17dz + 0.15L_M(dz - 1), & 1 \leq dz < 8 \end{cases}. \quad (2.8)$$

¹⁰ISO: International Organization for Standardization; IEC: International Electrotechnical Committee; MPEG: Moving Picture Experts Group.

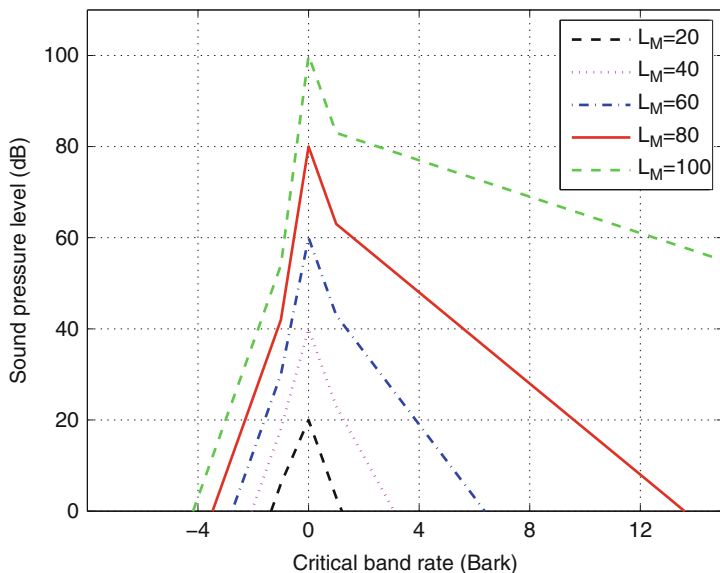


Fig. 2.17 Spreading function in ISO/IEC Psychoacoustic Model 1

Figure 2.17 shows spreading functions in Model 1 for different levels of the masker. It is seen that the higher SPL the masker has, the more asymmetric the curve looks. Specifically, higher frequencies exhibit more masking than lower frequencies when the level of masker is high. This two-piece linear spreading function is a good approximation to the masking thresholds of TMT in Fig. 2.16.

In addition, four models described above for spreading functions, i.e., two-slope SF, Schroeder SF, Psychoacoustic Model 1 SF, and Model 2 SF, are compared at a level of 80 dB in Fig. 2.18. Among these four models, two-slope spreading function is the most conservative one, and Model 1 spreading function allows for more upward spreading of masking than others [51].

2.4.1.2 Implementation of Psychoacoustic Model 1

In different application scenarios, psychoacoustic model can be implemented in different ways to satisfy the criteria required. ISO/IEC MPEG-1 Standard [77] utilizes two informative psychoacoustic models, Psychoacoustic Model 1 and 2, to determine the MMT for inaudibility. Typically, Model 1 is applied to MPEG Layers I and II and Model 2 to MPEG Layer III. Both models are commonly in use and well performed. Psychoacoustic Model 1 proposed a low-complication method to analyze spectral data and output SMR, whereas Psychoacoustic Model 2 performs a more detailed analysis at the expense of greater computational complexity

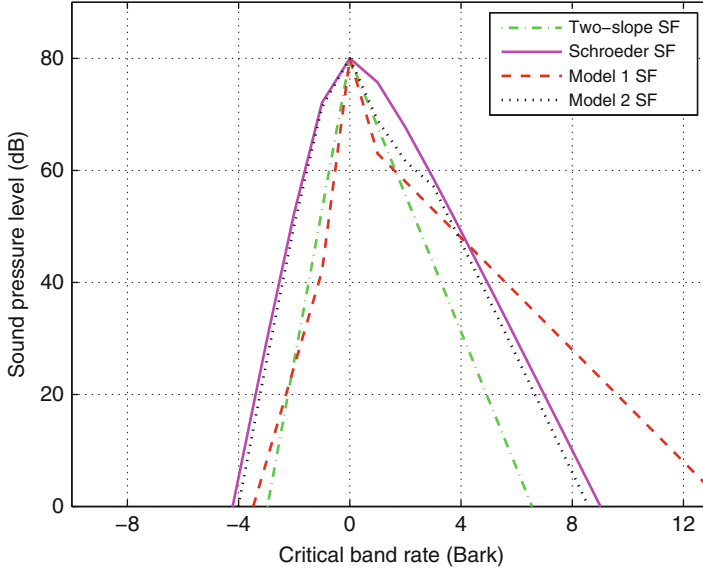


Fig. 2.18 Comparison of four spreading functions relative to an 80 dB masker

[31, 78, 79]. Hence, Psychoacoustic Model 1 for Layer I is later employed in our audio watermarking scheme in consideration of its higher efficiency.

In our case, the input to Psychoacoustic Model 1 is one frame of audio signal and the corresponding output is its MMT. The whole procedure of implementation consists of six steps [72, 73, 77, 80]:

1. FFT analysis and SPL normalization
2. Identification of tonal and nontonal maskers
3. Decimation of invalid tonal and nontonal maskers
4. Calculation of individual masking thresholds
5. Calculation of global masking threshold
6. Determination of the MMT

The details of each step are expounded as follows:

• **STEP 1: FFT analysis and SPL normalization**

For an accurate analysis of frequency components, fast Fourier transform (FFT) is performed to obtain a high-resolution spectral estimate of incoming frame $x(n)$. In Psychoacoustic Model 1, the input frame has a size of $N = 512$ points. To minimize the leakage effect, $x(n)$ is multiplied with a modified Hanning window $w(n)$ defined by

$$w(n) = \sqrt{\frac{8}{3}} \text{hann}(N) = \sqrt{\frac{8}{3}} \cdot \frac{1}{2} \left[1 - \cos\left(\frac{2\pi n}{N}\right) \right] \quad 0 \leq n \leq N-1, \quad (2.9)$$

where $\text{hann}(N) = \frac{1}{2} [1 - \cos(\frac{2\pi n}{N})]$ is the N -point Hanning window. Factor $\sqrt{\frac{8}{3}}$ is a gain to compensate the average power of $w(n)$, so that $\langle w(n)^2 \rangle \equiv \frac{1}{N} \sum_{n=0}^{N-1} [w(n)^2] = 1$. Then, power spectral density (PSD) of $x(n)$ is computed as

$$\text{PSD}(k) / \text{dB} = 10 \log_{10} \left| \frac{1}{N} \left[\sum_{n=0}^{N-1} x(n) w(n) \exp \left(-j \frac{2\pi n k}{N} \right) \right] \right|^2 \quad 0 \leq k < \frac{N}{2}. \quad (2.10)$$

After that, PSD estimate $\text{PSD}(k)$ is normalized to a SPL level of 96 dB, i.e., the maximal is limited to 96 dB.

$$\begin{aligned} P(k) / \text{dB} &= 96 - \max \{ \text{PSD}(k) \} + \text{PSD}(k) \\ &= \Delta_P + \text{PSD}(k), \end{aligned} \quad (2.11)$$

where $\Delta_P = 96 - \max \{ \text{PSD}(k) \}$. It is because we have no prior knowledge regarding actual playback levels, the absolute pressure level of a sound can only be specified by comparing to a reference. To this end, a sinusoid with amplitude equal to half of PCM quantizer spacing ($A_0 = \frac{\Delta}{2}$) is defined as having a SPL of 0 dB, i.e., $20 \log_{10} (A_0/A_0) = 0$ dB. Consequently, for 16-bit PCM data, a sinusoid with amplitude equal to the overload level of quantizer ($A_{\max} = \frac{(2^{16}-1)\Delta}{2}$) would have a SPL of about 96 dB, i.e., $20 \log_{10} (A_{\max}/A_0) = 20 \log_{10} (2^{16} - 1) \approx 96$ dB [51].

An example of the initial and normalized PSD estimates as well as the threshold in quiet are shown in Fig. 2.19, where the frequencies of two graphs are plotted on linear and Bark scales, respectively. Note that in psychoacoustic models, an offset depending on the overall bit rate is employed for the threshold in quiet. It is equal to -12 dB for bit rates no less than 96 kbits/s and 0 dB for bit rates less than 96 kbits/s per channel [77]. Sound tracks used in our experiments are of CD quality, whose bit rates are normally greater than 96 kbits/s. Therefore, by comparing Fig. 2.19 to Figs. 2.7 and 2.8, the threshold in quiet in Fig. 2.19 is shifted downward by 12 dB.

• **STEP 2:** Identification of tonal and nontonal maskers

On account of “asymmetry of masking,” it is required to discern frequency components as tonal (i.e., sinusoidal) and nontonal (i.e., noise-like) maskers. Tonal maskers are selected from local maxima of normalized PSD estimate, $P(k)$. A local maxima refers to the maximum PSD within its two neighbors:

$$P(k) \geq P(k+1) \text{ and } P(k) \geq P(k-1) \quad (2.12)$$

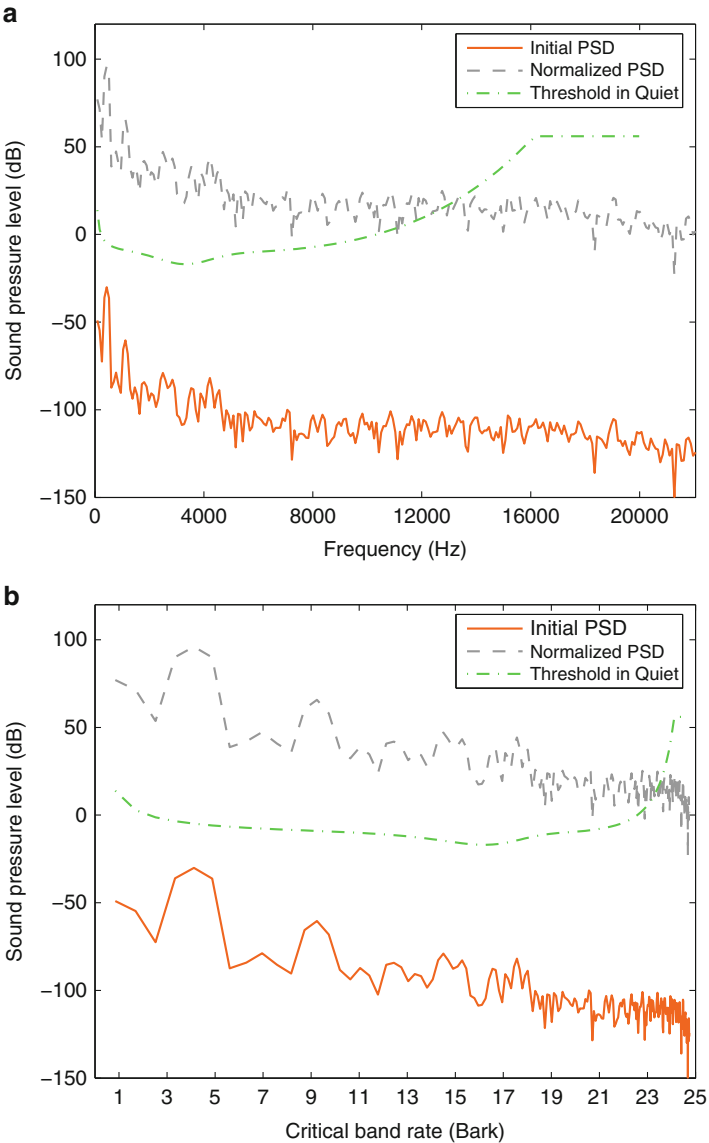


Fig. 2.19 Initial and normalized PSD estimates (a) Frequency on linear scale (b) Frequency on Bark scale

If the value of a local maxima is at least 7 dB greater than that of its neighboring components within a certain Bark range D_k , such a maxima will be marked as a tonal masker. All the tonal components comprise the “tonal” set, S_{TM} :

$$S_{\text{TM}} = \{P(k) \mid [P(k) - P(k \pm D_k)] \geq 7 \text{ dB}\}, \quad (2.13)$$

where D_k varies with different frequency indices.¹¹

$$D_k \in \begin{cases} \{\pm 2\}, & 2 < k < 63 \Leftrightarrow \frac{2F_s}{N} \sim \frac{63F_s}{N} \text{ kHz} \\ \{\pm 2, \pm 3\}, & 63 \leq k < 127 \Leftrightarrow \frac{63F_s}{N} \sim \frac{127F_s}{N} \text{ kHz} \\ \{\pm 2, \pm 3, \dots, \pm 6\}, & 127 \leq k \leq 250 \Leftrightarrow \frac{127F_s}{N} \sim \frac{250F_s}{N} \text{ kHz} \end{cases}.$$

One point to note is that [77] did not specify the value of D_k for $251 \leq k \leq 256$, because the maskers within this range are already dominated by the threshold in quiet (as seen in Fig. 2.19) and have no contribution to masking threshold. Actually, it is the first criterion for decimation in Step 3.

As the effect of masking is additive in the logarithmic domain, the SPL of each tonal component is calculated by

$$P_{\text{TM}}(k) / \text{dB} = 10 \log_{10} \left[10^{\frac{P(k-1)}{10}} + 10^{\frac{P(k)}{10}} + 10^{\frac{P(k+1)}{10}} \right]. \quad (2.14)$$

In addition, the remaining components within each critical band¹² are treated to be nontonal. So we sum up their intensities as the SPL of a single nontonal masker for each critical band, P_{NM} :

$$P_{\text{NM}}(\bar{k}) / \text{dB} = 10 \log_{10} \sum_j \left[10^{\frac{P(j)}{10}} \right] \quad \forall P(j) \notin S_{\text{TM}}, \quad (2.15)$$

where that \bar{k} is the frequency index nearest to the geometric mean¹³ of each critical band. Correspondingly, all the nontonal components are put into the “nontonal” set, S_{NM} .

¹¹The frequency edges are calculated based on the sampling frequency F_s .

¹²Critical band boundaries vary with the Layer and sampling frequency. ISO/IEC IS 11172-3 [77] has tabulated such parameters in Table D.2a–f. In our case, Table D.2b for Layer I at a sampling frequency of 44.1 kHz is adopted.

¹³The geometric mean of a data set $[a_1, a_2, \dots, a_M]$ is defined as $\left(\prod_{m=1}^M a_m \right)^{1/M}$. It is sometimes

called the log-average, i.e., $\left(\prod_{m=1}^M a_m \right)^{1/M} = 10^{\frac{1}{M} \sum_{m=1}^M \log_{10}(a_m)}$.

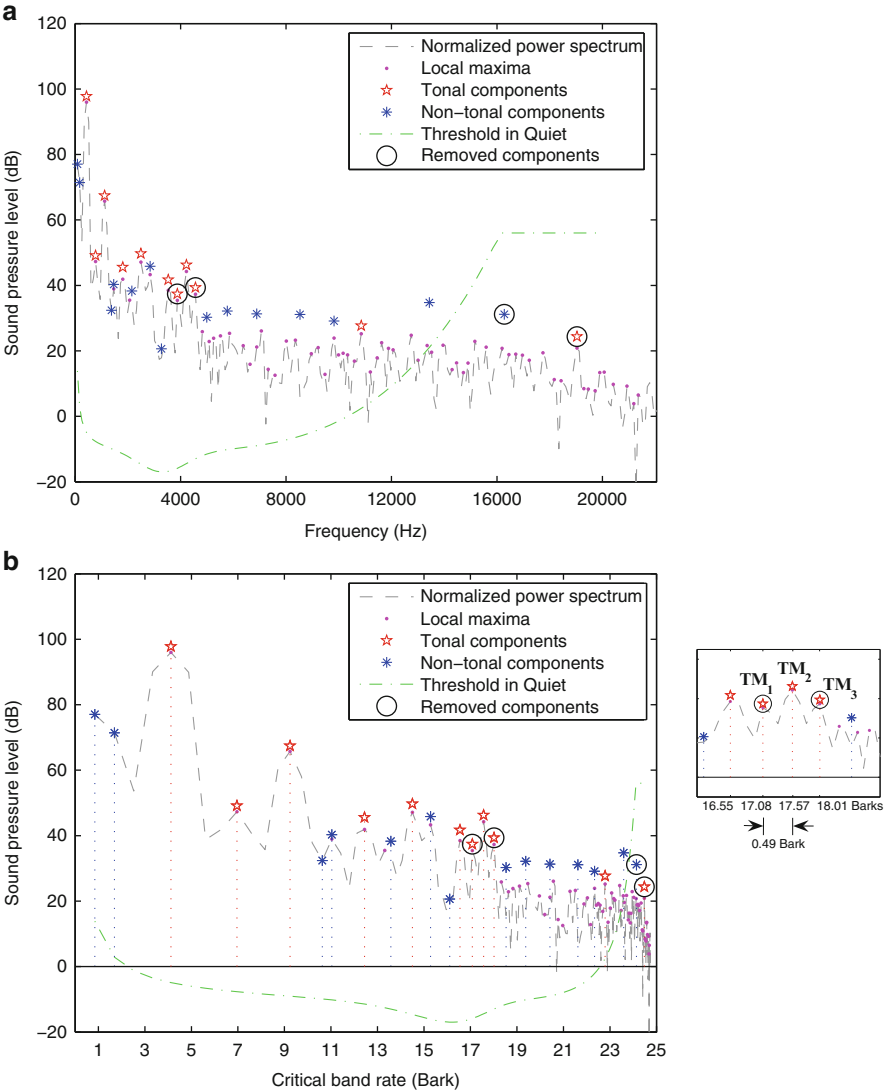


Fig. 2.20 Tonal and nontonal maskers (a) Frequency on a linear scale (b) Frequency on Bark scale

Tonal and nontonal maskers are denoted by pentagram and asterisk symbols in Fig. 2.20, respectively. Particularly, the associated critical band for each masker is indicated in the graph on Bark scale.

• **STEP 3:** Decimation of invalid tonal and nontonal maskers

On considering their possible contributions to masking threshold, the sets of tonal and nontonal maskers are examined according to two criteria as follows:

One rule is that any tonal and nontonal maskers below the threshold in quiet are removed. That is, only the maskers that satisfy Eq. (2.16) are retained, where $ATH(k)$ is the SPL of threshold in quiet at frequency index k :

$$P_{TM, NM}(k) \geq ATH(k). \quad (2.16)$$

For example, one of each tonal and nontonal maskers between 24 and 25 Barks is discarded, as shown in Fig. 2.20b.

The other rule is to simplify any group of maskers occurring within a distance of 0.5 Bark: only the masker with the highest SPL is preserved and the rest are eliminated.

$$P_{TM, NM}(k) = \arg \max_{k_0 \in [-0.5, 0.5]} P_{TM, NM}(k + k_0). \quad (2.17)$$

For example, two pairs of tonal maskers between 17 and 19 Barks, $\{TM_1, TM_2\}$ and $\{TM_2, TM_3\}$, are inspected. As shown in an enlarged drawing on the right of Fig. 2.20b, the distance between $\{TM_1, TM_2\}$ is 0.49 Bark, and TM_1 has a lower SPL than TM_2 . Therefore, TM_2 is preserved, whereas TM_1 is removed. Similarly, we dispose of TM_3 but retain TM_2 for $\{TM_2, TM_3\}$.

$$\begin{aligned} TM_2 &\leftarrow \{TM_1, TM_2\} \left| \begin{array}{l} \text{Distance : } 17.57 - 17.08 = 0.49 \text{ Bark} \\ \text{SPL : } P_{TM_1} < P_{TM_2} \end{array} \right. \\ TM_2 &\leftarrow \{TM_2, TM_3\} \left| \begin{array}{l} \text{Distance : } 18.01 - 17.57 = 0.44 \text{ Bark} \\ \text{SPL : } P_{TM_2} > P_{TM_3} \end{array} \right. \end{aligned}$$

In Fig. 2.20, the invalid tonal and nontonal maskers being decimated are denoted by a circle.

- **STEP 4:** Calculation of individual masking thresholds

After eliminating invalid maskers, individual masking threshold is computed for each tonal and nontonal masker. An individual masking threshold $L(j, i)$ refers to the masker at frequency index j contributing to masking effect on the maskee at frequency index i . It corresponds to $L[z(j), z(i)]$, where $z(j)$ and $z(i)$ are the masker and maskee's frequencies in Bark scale. In MPEG psychoacoustic models, only a subset of samples over the whole spectrum are considered to be maskees and involved in the calculation of global masking threshold. The number and frequencies of maskees also depend on the Layer and sampling frequency, as tabulated from Table D.1a–f in [77]. In our case, Table D.1b for Layer I at a sampling frequency of 44.1 kHz is adopted, where 106 maskees are taken into account.

The individual masking thresholds for tonal and nontonal maskers, $L_{TM}[z(j), z(i)]$ and $L_{NM}[z(j), z(i)]$, are calculated by

$$L_{TM}[z(j), z(i)] / \text{dB} = P_{TM}[z(j)] + \Delta_{TM}[z(j)] + SF[z(j), z(i)] \quad (2.18)$$

$$L_{\text{NM}}[z(j), z(i)]/\text{dB} = P_{\text{NM}}[z(j)] + \Delta_{\text{NM}}[z(j)] + \text{SF}[z(j), z(i)], \quad (2.19)$$

where $P_{\text{TM}}[z(j)]$ and $P_{\text{NM}}[z(j)]$ are the SPLs of tonal and nontonal maskers at a Bark scale of $z(j)$, respectively. The term Δ_X is called masking index, an offset between the excitation pattern and actual masking threshold. As mentioned in Sect. 2.3.1, the excitation pattern needs to be shifted by an appropriate amount in order to obtain the masking curve relative to the masker. Because tonal and nontonal maskers have different masking capability, i.e., the noise is a better masker than pure tone, the masking indices of tonal and nontonal maskers are defined separately as follows [77]:

$$\Delta_{\text{TM}}[z(j)] = -6.025 - 0.275z(j) \quad (2.20)$$

$$\Delta_{\text{NM}}[z(j)] = -2.025 - 0.175z(j). \quad (2.21)$$

The term $\text{SF}[z(j), z(i)]$ is the spreading function discussed already in Sect. 2.4.1.1. Psychoacoustic Model 1 employs spreading function in Eq. (2.8), rewritten in the following expression:

$$10 \log_{10} \text{SF}(dz)/\text{dB} = \begin{cases} 17dz - 0.4P_X[z(j)] + 11, & -3 \leq dz < -1 \\ (0.4P_X[z(j)] + 6) dz, & -1 \leq dz < 0 \\ -17dz, & 0 \leq dz < 1 \\ -17dz + 0.15P_X[z(j)](dz - 1), & 1 \leq dz < 8 \end{cases}, \quad (2.22)$$

where dz is the distance from the maskee to masker, $dz = z(i) - z(j)$, as defined in Sect. 2.4.1.1. $P_X[z(j)]$ refers to $P_{\text{TM}}[z(j)]$ in the case of tonal masker, otherwise $P_{\text{NM}}[z(j)]$ for nontonal masker. Notice that for reasons of implementation complexity, the masking is no longer considered if $dz < -3$ Bark or $dz \geq 8$ Bark and thereby $L_{\text{TM}}[z(j), z(i)]$ and $L_{\text{NM}}[z(j), z(i)]$ are set to $-\infty$ dB outside the above ranges [77].

Figure 2.21 shows the individual masking thresholds for both tonal and nontonal maskers survived from the decimation.

- **STEP 5:** Calculation of global masking threshold

The global masking threshold is the combination of individual masking thresholds and the threshold in quiet. Since the mixture of masking is additive, the global masking threshold at frequency index i is calculated according to

$$L_G(i)/\text{dB} = 10 \log_{10} \left[10^{\frac{\text{ATH}(i)}{10}} + \sum_{j=1}^{N_{\text{TM}}} 10^{\frac{L_{\text{TM}}[z(j), z(i)]}{10}} + \sum_{j=1}^{N_{\text{NM}}} 10^{\frac{L_{\text{NM}}[z(j), z(i)]}{10}} \right], \quad (2.23)$$

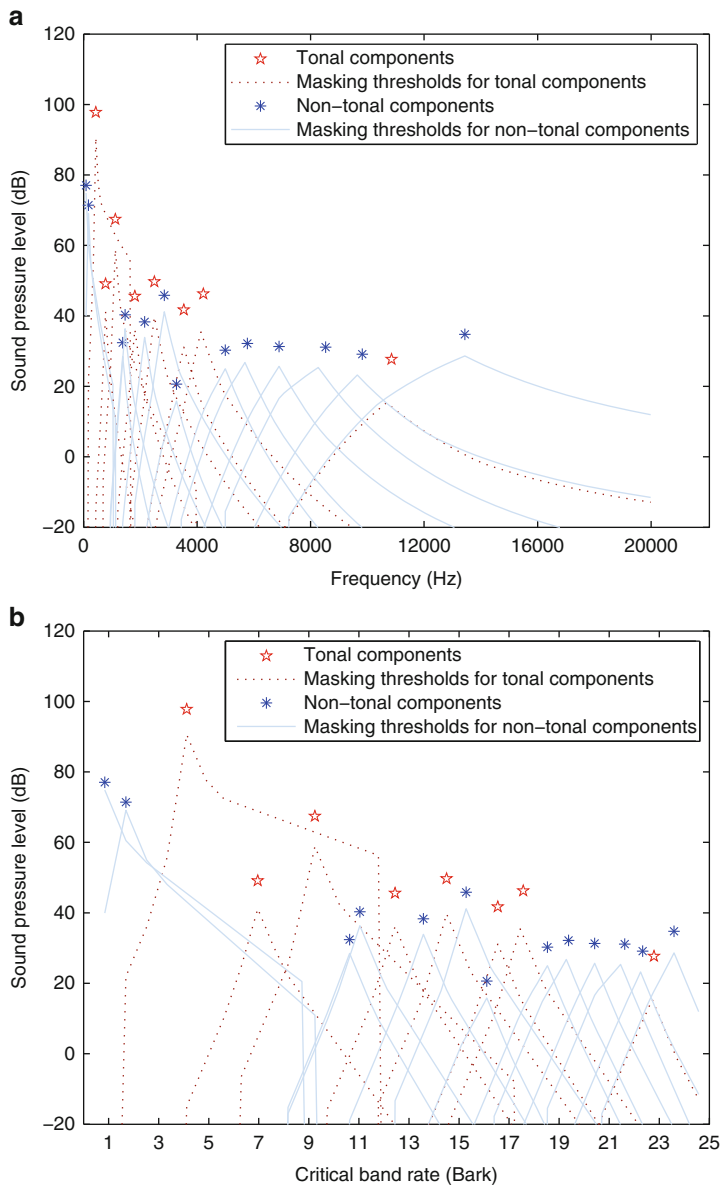


Fig. 2.21 Individual masking thresholds (a) Frequency on linear scale (b) Frequency on Bark scale

where $ATH(i)$ is the SPL of threshold in quiet at frequency index i , N_{TM} and N_{NM} are the number of tonal and nontonal maskers, and $L_{TM}[z(j), :]$ and $L_{NM}[z(j), :]$ are their corresponding individual masking thresholds.

The global masking threshold is denoted by a bold dashed black line in Fig. 2.22.

- **STEP 6:** Determination of the MMT

The MMT is derived from the global masking threshold. As mentioned in Step 4, the global masking threshold L_G is computed on only a subset of samples (here 106 samples) over the frequency spectrum, i.e., $1 \leq i \leq 106$. Then these spectral subsamples are mapped onto 32 uniform subbands, as shown in Fig. 2.23. Each subband contains $\frac{N/2}{32} = \frac{512/2}{32} = 8$ samples. Therefore, the minimum masking level in the n th subband ($1 \leq n \leq 32$) is determined by the following expression:

$$L_{\text{Min}}(n) / \text{dB} = \min_{f_{id}(i) \in \text{subband } n} L_G(i), \quad (2.24)$$

where $f_{id}(i)$ is the frequency index corresponding to the i th subsample. After spreading every $L_{\text{Min}}(n)$ ($1 \leq n \leq 32$) over its subband with 8 samples, we get the MMT L_{MMT} :

$$L_{\text{MMT}}(m) = L_{\text{Min}}(n) \quad m = [8(n-1) + 1] : 8n. \quad (2.25)$$

2.4.1.3 Comparison Between Psychoacoustic Model 1 and Model 2

The general idea of implementation on Psychoacoustic Model 2 is similar to Model 1. However, the concrete operations of calculating MMT in Psychoacoustic Model 2 are quite different from that of Model 1, as depicted in the following steps [78,81]:

- **STEP 1:** FFT analysis and calculation of complex spectrum

The input to Model 2 is a set of 1,024 samples, twice longer than 512-point frame in Model 1. Before performing FFT, a Hanning window is applied as well.

- **STEP 2:** Definition of threshold calculation partitions and spreading function

The notion of “threshold calculation partitions” is a significant difference in Model 2. Instead of identifying the tonal and nontonal maskers in each critical band in Model 1, Model 2 groups the frequency lines into so-called threshold calculation partitions. Such partitions are also of nonlinear widths, but with finer frequency resolution than critical band. Each partition has a width of either one FFT line (at low frequencies) or 1/3 critical band (at high frequencies), whichever is wider [78]. According to this criterion, there are 57 partitions at a sampling frequency of 44.1 kHz by calculation. The result complies with Table D.3b in [77].

The spreading function in Model 2 is described by Eq. (2.7) and one specific spreading function is defined for each partition. Note that $10 \log_{10} \text{SF}(dz)$ in Eq. (2.7) is level-independent and thereby suitable for alleviating the computational burden of convolution in Step 4.

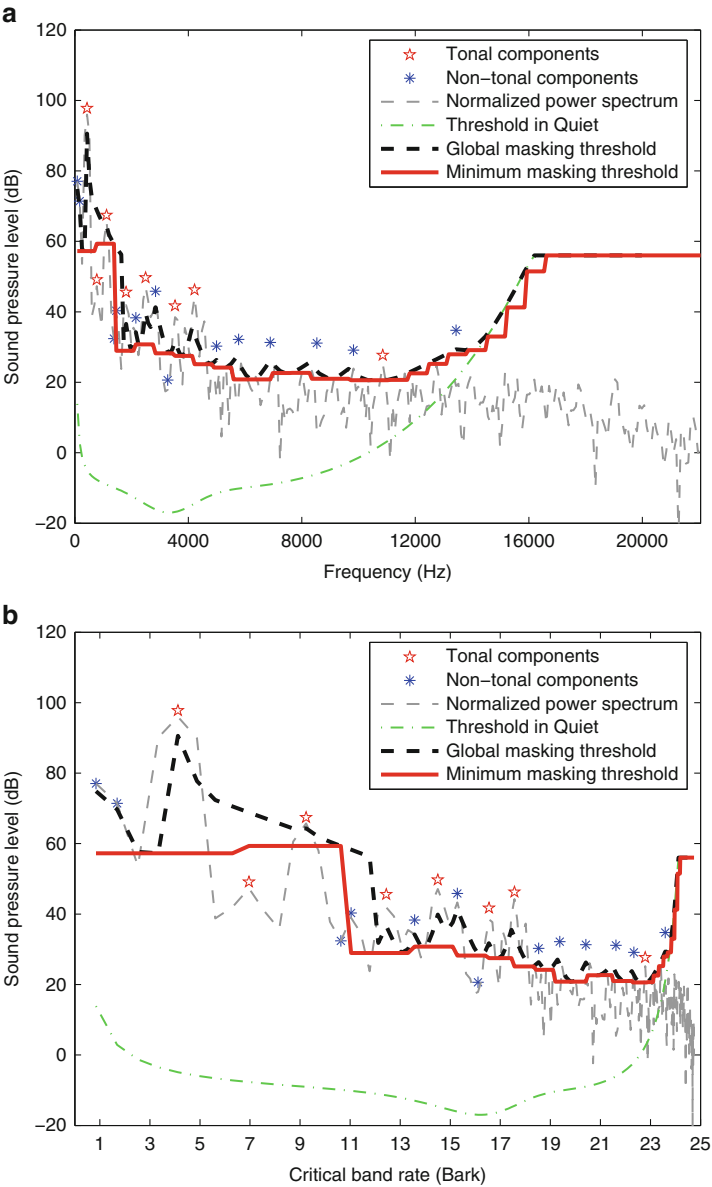


Fig. 2.22 Global masking threshold and minimum masking threshold (a) Frequency on linear scale (b) Frequency on Bark scale

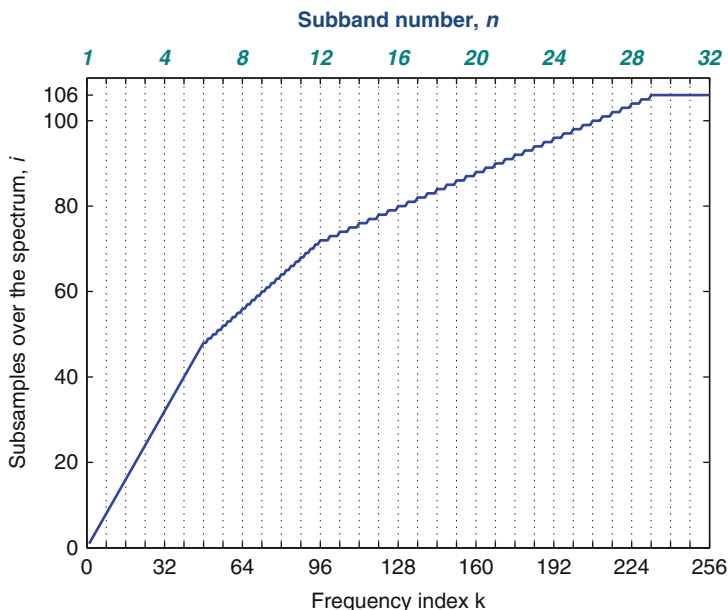


Fig. 2.23 Mapping between spectral subsamples and subbands

- **STEP 3:** Calculation of unpredictability measure and weighted partition energy

Rather than selecting the relevant tonal and nontonal maskers in each critical band, Model 2 introduces the property of unpredictability to describe how predictable (tonal -like) the frequency component is. Unpredictability measure depends on the magnitude and phase of complex spectrum. After weighting the energy of each frequency line with unpredictability measure, we sum them up as the weighted energy of each partition.

- **STEP 4:** Convolution of weighted partition energy and spreading function

As the behavior of simultaneous masking, the partition spreads its weighted energy into the adjacent partitions. The overall masking effect is computed by the convolution of spreading functions and weighted energy of each partition.

- **STEP 5:** Calculation of tonality index and SMR

Tonality index is a measure in Model 2, which is not used in Model 1. It denotes the relative tonality of the maskers in each partition. The value of tonality index is limited to the range of 0 (high unpredictability and noise-like) and 1 (low unpredictability and tonal). Based on tonality index as well as an attenuation shift factor between NMT and TMN, the SMR of each partition is calculated.

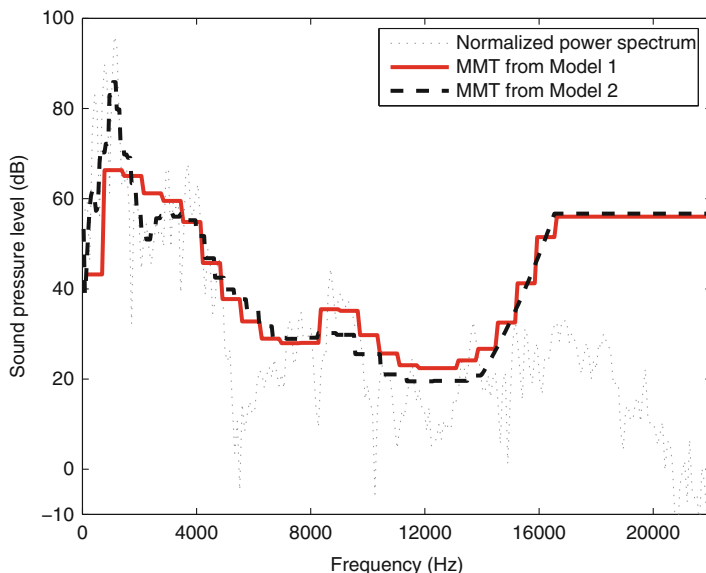


Fig. 2.24 Comparison of MMTs from Psychoacoustic Model 1 and 2

- **STEP 6:** Determination of the MMT

After obtaining SMR, the masking level of each partition is calculated by multiplying SMR to the inverse of signal energy, and then it spreads evenly over the frequency line(s) within the partition. Finally, the MMT is determined by taking the bigger value between the masking level and the threshold in quiet.

Figure 2.24 illustrates a comparison of the MMTs from Psychoacoustic Model 1 and 2. In view of the overall trend, MMT from Model 1 is analogous to that from Model 2, although a bit less accurate at low frequencies. Generally, the difference in masking effect of two psychoacoustic models is not evident [78]. On the other hand, as the price of high precision, Model 2 involves more calculations such as finer resolution of partitions, unpredictability measure, and the convolution process. Consequently, it slows down the speed of execution, which is against the requirement of audio watermarking. Therefore, we prefer Psychoacoustic Model 1 for our application.

2.4.2 Modelling the Effect of Nonsimultaneous Masking

In addition to simultaneous masking, the effect of nonsimultaneous masking is also well exploited for developing perceptual models.

In [51], a time-sliding window is adopted in modelling the effect of nonsimultaneous masking. To resemble pre- and post-masking curves in Fig. 2.10, a weighting

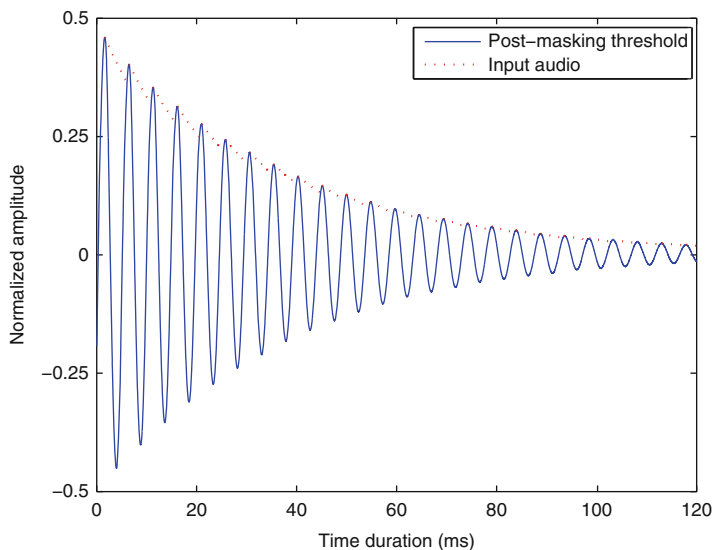


Fig. 2.25 Modelling the effect of post-masking

function of time is designed to be in a shape of bulge: a larger weight on components near the center of window, but gradual attenuation on components near the edges. Generally, it is assumed that such temporal smoothing is applied to signal spectrum, resulting in a smoothed output signal in time domain.

Different from [51], the modified envelope of input audio was used to approximate the effect of post-masking in [71]. In particular, the estimated masking curve increases with the envelope of signal and decays as an exponential function $e^{-\alpha t}$. The decay constant α ($\alpha \geq 0$) controls decaying rate as required, where $\alpha = 1.2 \times 10^{-3}$ in Fig. 2.25.

2.5 Summary

The ultimate aim of this chapter is to establish a psychoacoustic model that emulates the HAS. Accordingly, audio watermarking techniques are able to analyze the host audio signal in order to determine how the watermarks can be rendered as inaudible as possible.

The chapter started with the physiology of the peripheral auditory system including the outer, middle, and inner ears. The outer ear collects sound waves in the air and channels them to interior parts of the ear; the middle ear transforms the acoustical vibration of sound waves into mechanical vibration and passes them onto the inner ear; the inner ear transduces mechanical energy into nerve impulses that are transmitted to the brain. Then, some fundamental concepts of psychoacoustics

such as SPL, loudness, human hearing range, threshold in quiet, and critical bandwidth were introduced. The notions of two types of auditory masking, i.e., simultaneous and nonsimultaneous masking, were also explained. In simultaneous masking, it is noted that the masking ability of narrowband noise is superior to pure tone. Based on the acquired knowledge, the ways of constructing the models for simultaneous and nonsimultaneous masking effects are investigated respectively, particularly simultaneous masking. After reviewing several models for the spreading of masking, we described the details of implementing Psychoacoustic Model 1 in ISO/MPEG standard, followed by a comparison with Model 2. On balance, two psychoacoustic models have similar perceptual quality, but Model 2 requires more computation than Model 1. Consequently, we adopted Psychoacoustic Model 1 in the audio watermarking scheme we developed in this book.

Audio Watermark

A Comprehensive Foundation Using MATLAB

Lin, Y.; Abdulla, W.H.

2015, XX, 199 p. 62 illus., 38 illus. in color., Hardcover

ISBN: 978-3-319-07973-8