4 Introduction to psychoacoustics

One of the key elements in the development of reduced bit rate audio is the understanding and application of psychoacoustics. All of the current perceptual audio coders achieve high compression rates by exploiting the fact that signal information that cannot be detected by even a well-trained listener can be discarded. Several psychoacoustic principles, such as the absolute hearing threshold and masking, have been incorporated in codec design. In this chapter, we discuss some of the basic principles of psychoacoustics that are commonly used in perceptual audio coding.

4.1. Perception of loudness

While it is a relatively simple task to measure the sound pressure level of a sound source, it is quite a challenging task to relate the measured level to a perceived level by human listeners. The perceptual quantity that describes the magnitude of the source as experienced by a listener is called *loudness*. The difficulty in making a direct correlation between measured sound pressure level (SPL) and perceived loudness lies in the fact that there are multiple mechanisms responsible for producing the sensation. For example, it is common for people listening to reproduced music to observe changes in loudness when equalization is applied to the signal thus altering the balance between low and high frequencies.

There are several aspects of loudness that relate to perceptual audio codecs. One has to deal with the minimum threshold of hearing and how that varies with frequency. Another deals with how the hearing mechanism determines differences in loudness level when other parameters (e.g., spectrum or duration) are kept constant.

First, let us begin by defining the sound pressure level (SPL)

$$L = 20\log\left(\frac{p}{p_0}\right)\ \text{dB(SPL)}, \tag{4.1}$$

in which p is the pressure produced by a sound source and p_0 is the reference pressure of $20\,\mu$Pa that corresponds to the minimum audible threshold of human hearing for a 1 kHz tone. It is also useful to describe SPL in terms of sound intensity. Following the notation in [44] we can write the instantaneous intensity of sound as

$$I(t) = p(t)u(t), \tag{4.2}$$

in which $p(t)$ is the instantaneous sound pressure and $u(t)$ is the fluid velocity. For a plane wave, the fluid velocity is related to pressure by

$$u(t) = \frac{p(t)}{(\rho c)}, \tag{4.3}$$

in which ρ is the density (in the case of air $\rho = 1.29\,\mathrm{kg/m^3}$) and c is the speed of sound in air. From the above two equations, the intensity can be written as a function of pressure as:

$$I(t) = \frac{p^2(t)}{(\rho c)}. \tag{4.4}$$

Using equation (4.4) we can calculate the reference intensity corresponding to the reference pressure of 20 μPa.

$$I_0 = \frac{p_0^2}{(\rho c)} = \frac{p_0^2}{[(1.29\,\mathrm{kg/m^3})(343.6\,\mathrm{m/s})]} = 0.93 \times 10^{-12}\,\mathrm{W/m^2}, \tag{4.5}$$

which is often approximated (to account for temperature variations in the speed of sound) to $I_0 = 10^{-12}\,\mathrm{W/m^2}$. The sound pressure level of a sound source with intensity I, is then given by:

$$L = 10 \log\left(\frac{I}{I_0}\right) \mathrm{dB(SPL)}. \tag{4.6}$$

The dependence of loudness on SPL has been studied and is now an international standard [57]. It was found to relate the perceptual magnitude H to the physical sound intensity I by a power law

$$H = kI^{0.6}, \tag{4.7}$$

which holds over a range of SPL levels from about 40 dB SPL to 120 dB SPL and shows a slight deviation at lower SPL levels. Note that the power law exponent of 0.6 corresponds to a loudness doubling for every 10 dB increase in intensity. So a perceptual "doubling" is not the same as an electrical signal doubling.

Perhaps the most important characteristic of loudness perception for the design of audio codecs is this dependence of loudness on frequency. This relationship was investigated by Fletcher and Munson [34] over headphones and later Robinson and Dadson [106] for a free field (Figure 4.1). A different sound intensity is

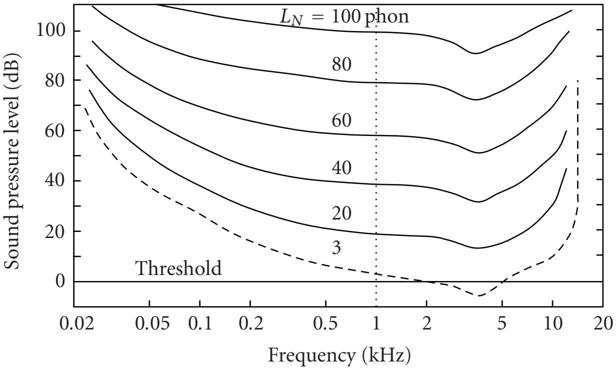

FIGURE 4.1. Equal loudness contours for pure tones. The curves show the sound pressure level required at each frequency to achieve a perceived loudness equal to a 1 kHz tone. Note that the curves tend to become flatter across frequency as the absolute loudness level is raised [148].

required at each frequency in order to produce the same level of perceived loudness. Human listeners are much less sensitive to low frequencies and thus a much higher SPL level is required to produce the same perceived loudness as that of a sound at high frequencies. The frequency-dependent intensity (or SPL) levels that give rise to the same perceived loudness are known as *equal loudness levels*. In fact, because these levels also depend on the absolute intensity, they form a family of curves known as the *equal loudness contours* (Figure 4.1). Each curve corresponds to the SPL at which a 1 kHz tone is perceived to be equally loud as a tone at another frequency. For example, on the 20 phon curve, a sound at 50 Hz must be 30 dB louder in sound pressure level in order to be perceived as having the same loudness as the reference tone at 1 kHz. Although the loudness level is measured in dB, it is distinguished from SPL and it is designated by a unit called the *phon*. The lowest curve in the figure is called the *minimum audible field (MAF)* and represents the hearing threshold averaged over a large population sample.

At low levels, the equal loudness curves resemble the shape of the MAF curve, but as the absolute loudness level increases the curves become flatter. This shows that the difference in balance between low and high frequencies decreases at high levels.

4.2. Masking

Another basic characteristic of human hearing that is exploited in perceptual audio coding is that of masking. It is relatively simple to observe this phenomenon in everyday life by simply noticing that a very loud sound (the masker signal) can prevent another sound (the maskee signal) from being heard. For example, a person near running water has difficulty hearing another person talking to them because the water acts as the masker signal. From studies of the ear physiology, it is known that the physical mechanism for producing sound cues in the brain begins with electrical discharges at nerve endings on the basilar membrane. While

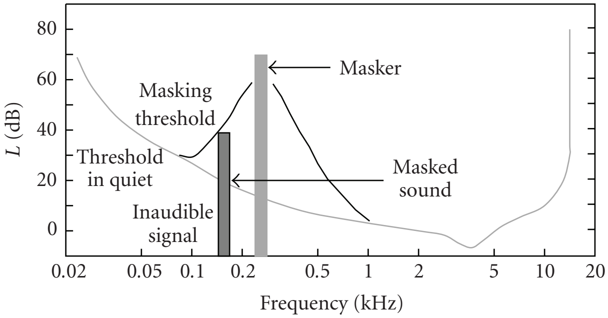

Figure 4.2. The masker signal causes a threshold shift that locally raises the minimum audible threshold. Any other signal with frequency components within the raised threshold range will not be audible and will therefore be ignored by a perceptual codec.

this discharge is occurring, the nerves involved cannot be stimulated by another sound. Thus if another sound happens to require some subset of the same nerves, it will be *masked*. The first published observation of this phenomenon is by Mayer [91] in 1876. As described by Fletcher [34], Mayer observed that low-frequency sounds had a very different masking effect from high-frequency sounds. A more general definition of masking includes not just the case when the maskee signal is not audible, but also when the masker produces a reduction in the perceived loudness of the maskee. This phenomenon is called *partial masking* [111].

The first systematic experiments to determine the effects of masking and the role of parameters such as spectrum, level, and duration of the masking signal were conducted by Fletcher (for a complete review see [34]). Starting with pure tones, a masker tone was generated and kept at a constant level while signal tones of different frequencies were increased in intensity from their threshold of audibility level until they just became perceptible in the presence of the masker. The level difference from the audibility threshold without the masker represents the threshold shift (Figure 4.2). This effect of masking is particularly relevant to perceptual audio codecs particularly because the threshold shift varies linearly with the masker level [45].

Fletcher's experiments verified Mayer's observations that high frequencies are masked more than low frequencies. This implies that for complex sounds that contain multiple tones, masking will be manifest as an apparent "boost" in low frequencies because the high frequencies are more easily masked (and thus not perceived as intense). What is interesting about this low-frequency boost is that it only occurs in monaural masking experiments. If the masker and maskee signals are presented to each ear separately, this apparent interference does not occur.

4.2.1. Frequency masking

The masking of complex sounds that consist of multiple pure tones can be explained by "complex addition" of the effects produced by individual tones. That is

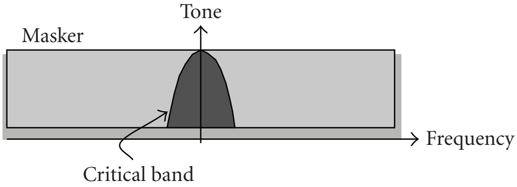

Figure 4.3. With a broadband masker, only the energy in the critical band around the pure tone gives rise to masking.

to say, the masking not only combines the effects of each tone, but also presents the effects of the sum and difference of the individual tones.

The spectrum of the masker signal also plays a critical role in masking. A pure tone (single line spectrum) can act as a masker. As Mayer showed [91], low frequency tones can mask higher frequencies; however, high frequencies are not good maskers of lower frequencies. Later studies by Wegel and Lane [133] verified this and went further to show that as the intensity of the masker signal is raised, masking spreads in the direction of higher frequencies, but not towards lower frequencies. The effect is called *upward spread of masking*. Furthermore, the width of the masking pattern at low frequencies is much wider than the width at higher frequencies. This implies that low frequency maskers have an effect over a much wider range of frequencies.

The human ear has the ability to act as a spectrum analyzer and analyze the frequency content in a signal using filter banks known as the auditory filters. The width of these auditory filters varies with the frequency around which they are centered. In Fletcher's work [28], it was shown that the masking threshold of a tone increased with the bandwidth of the masking noise. However, after a certain width of the masking noise there was no further increase in the masking effect. This suggested that each auditory filter has a certain critical bandwidth and Fletcher theorized at the threshold and the signal power and noise power would be equal within the critical bandwidth range. He defined a critical band as the ratio of signal to noise power, which when expressed in dB corresponds to the difference between the signal and the masking noise (Figure 4.3).

4.2.2. Temporal masking

So far we have discussed masking that arises due to the audibility threshold shift that a human perceives when a masking sound is present at the same time. It is also possible to create a masking effect when the masker and signal have occurred at different times. This is called *temporal masking* and can manifest itself as *backward masking* in which the signal is generated before the masker, or *forward masking* in which the signal comes after the masker (Figure 4.4).

In backward masking, the masker causes a rise in the threshold for signals that arrived *before* it. This may seem counterintuitive, but it has been demonstrated for time gaps on the order of 25 ms [31]. The level of the masker must be significantly higher than the signal for backward masking. Although the signal arrives at the ear

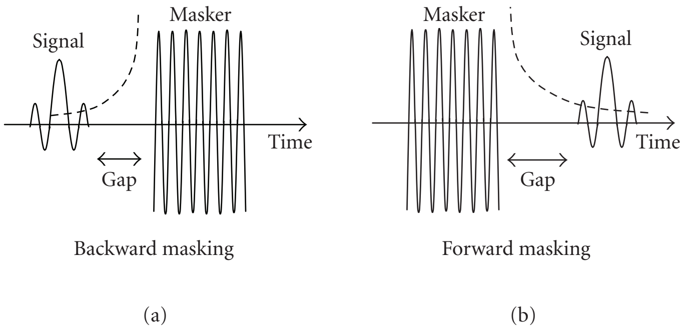

FIGURE 4.4. Temporal masking can occur in both the forward and backward directions.

earlier, the higher loudness masker is processed first thus giving rise to backward masking.

The most common type of temporal masking is forward masking. This effect happens over much longer time intervals between the signal and the masker that can reach 200 ms [31]. The amount of masking as well as the decay of the masking effect are directly proportional to the time gap between the masker and the signal.

4.2.3. Interaural masking

Most of the early masking experiments were monaural in nature with both the masker and the signal rendered in the same ear. Masking is also produced in a binaural situation with the masker in one ear and the signal in the other. As the intensity of the masker is raised it reaches a level at which it can mask a signal at the opposite (contralateral) ear. Although the reasons are not fully understood, it is hypothesized that the masker interacts with the signal much later in the hearing signal path, within the central nervous system at a point where binaural signal representation is maintained [149].

Interaural masking is a much weaker effect compared to monaural masking. It rises by a fraction of a dB as the masker level rises and decreases when there is a gap between the masker and the signal. Contrary to monaural masking, however, the effect is stronger for higher frequencies than lower frequencies. Zwislocki [149] showed that this type of masking correlates highly with the firing rate of neurons of the lower auditory nervous system.