CHAPTER - 4

HUMAN HEARING AND AUDITORY MASKING

4.1 THE HUMAN EAR

The human auditory system consists of the ear, auditory nerve fibers, and a section of the brain. It converts sound waves into sensations perceived by the auditory cortex.

The ear is the outer peripheral system which converts acoustic energy (sound waves) into electrical impulses that are picked up by the auditory nerve. The ear itself is divided into three parts, the outer, middle, and inner ear, as shown in Fig.4.1.

![Fig.4.1: Structure of the human ear](image)

4.1.1 The Outer Ear

The outer ear consists of the pinna (the visible part of the ear), the meatus (ear canal), and terminates at the tympanic membrane (eardrum). The pinna collects sounds and aids in sound localization, that is to be more sensitive to sounds coming from the front of the listener.
The meatus is a tube which directs the sound to the tympanic membrane. The meatus acts as a quarter-wave resonator with a center frequency around 3000 Hz. This particular structure likely aids in the perception of obstruents\(^1\) which have much of their energy content in this frequency region.

### 4.1.2 The Middle Ear

The middle ear is considered to begin at the tympanic membrane and contains the ossicles, a set of three small bones. These bones are named *malleus* (hammer), *incus* (anvil), and *stapes* (stirrup). Acting primarily as levers performing an impedance matching transformation (from the air outside the eardrum to the fluid in the cochlea), they also protect against very strong sounds. The *acoustic reflex* activates middle ear muscles, to change the type of motion of the ossicles when low-frequency sounds with SPL above 85-90 dB reach the eardrum. Attenuating pressure transmission by up to 20 dB, the acoustic reflex is also activated during voicing in the speaker’s own vocal tract\(^90\). Due to their mass, the ossicles act as a low-pass filter with a cutoff frequency around 1000 Hz.

### 4.1.3 The Inner Ear

The inner ear is a bony structure comprised of the semicircular canals of the vestibula and the cochlea. The vestibule is the organ that helps balancing the body and has no apparent role in the hearing process. The cochlea is a cone-shaped spiral in which the auditory nerve terminates. It is the most complex part of the ear, wherein the mechanical pressure waves are converted into electrical pulses.

The cochlea is a tapered tube filled with a gelatinous fluid (*endolymph*). At its base this tube has a cross section of about 4 min\(^2\), and two membrane covered openings, the oval Window and the Round Window. The Oval Window is connected to the ossicles. The Round Window is free to move to equalize the pressure since the endolymph is incompressible.

---

1. Sounds produced by obstructing the air flow in the vocal tract, such as */s/* and */f/*.
2. Sounds produced by obstructing the air flow in the vocal tract, such as */s/* and */f/*.
The cochlea has two membranes running along its length, the Basilar Membrane (BM) and Reissner's Membrane. These two membranes divide the cochlea into three channels, as seen in Fig. 4.2.

![Fig. 4.2: Cross-section of the cochlea](image)

These channels are called the Scala Vestibuli, the Scala Media, and the Scala Tympani. Pressure waves travel from the Oval window through the Scala Vestibuli to the apex of the cochlea. A small opening (helicotrema) connects the Scala Vestibuli to the Scala Tympani. The sound pressure wave then travel back to the base through the Scala Tympani, terminating at the Round Window. Since the velocity of sound in the cochlea is about 1600 m/s, there is no appreciable phase delay.

### 4.1.4 The Basilar Membrane and the Hair Cells

Within the Basilar Membrane (BM), mechanical movements are transformed into nerve stimuli transmitted to the brain. The BM performs a crucial part of sound perception. It is narrow and stiff at the base of the cochlea, gradually tapering to a wide and pliable end at the apex of the cochlea. Each point on the cochlea can be viewed as a mass-spring system with a resonant frequency that decreases from base to apex. A frequency to place transformation is performed, such that if a pure tone is applied to the Oval Window, a section of the BM will vibrate. The amplitude of BM vibration is dependent on distance from the vocal window and the frequency of the
The BM vertical displacement is small near the oval window. Growing slowly, the vertical displacement reaches a maximum at a certain distance from the oval window. The amplitude of the vertical displacement then rapidly dies out in the direction of the helicotrema. The frequency of a signal that causes maximum displacement at a given point of the BM is called the Characteristic Frequency (CF).

The vibration of the BM is picked up by the hair cells of the Organ of Corti. There are two classes of hair cells, the Inner Hair Cells (IHC) and Outer Hair Cells (OHC). About 90% of afferent (ascending) nerve fibers that carry information from the cochlea to the brain terminate at the IHC (Inner Hair Cell). Most of the efferent (descending) nerve fibers terminate at the OHC, which greatly outnumber the IHC. Empirical observations suggest that the OHC, with direct connection to the tectorial membrane, can change the vibration pattern of the BM, improving the frequency selectivity of the auditory system.

Measurements from afferent auditory nerves have shown further non-linearities in the auditory system. All IHCs show a spontaneous rate of firings in the absence of stimuli. As a stimulus (such as a tone burst at the CF for the IHC) is applied, the neuron responds with a high rate of firings, which after approximately 20 ms decreases to a steady rate. Once the stimulus is removed, the rate falls below the spontaneous rate for a short time before returning to the spontaneous rate.

4.2 MASKING

Human auditory masking is a highly complex process which is only partially understood, yet we experience the effects in everyday life. In noisy environments, such as an airport or a train station, noise seems to have a habit of lowering intelligibility just enough so that you miss the call for the flight or train you have to catch.
The American Standards Association (ASA) defines masking as the process or the amount (customarily measured in decibels) “by which the threshold of audibility is raised by the presence of another (masking) sound” [91]. Simply put, one sound cannot be heard because of another (typically louder) sound being present.

4.2.1 Threshold of Hearing

In order to be audible, sounds require a minimum pressure. Due in part to filtering in the outer and middle ear, this minimum pressure (considering for now a pure tone) varies considerably with frequency. The threshold of hearing (audibility) is unique from person to person and furthermore changes with a person’s age. To determine the effect of frequency on hearing ability, scientists played a sinusoidal tone at a very low power. The power was slowly raised until the subject could hear the tone. This level was the threshold at which the tone could be heard. The process was repeated for many frequencies in the human auditory range and with many subjects. As a result, the following plot was obtained (Fig. 4.3).

![Threshold of Hearing](image)

**Fig. 4.3: Threshold of hearing**
For signal processing purposes, the threshold is approximated by

$$T_v(f) = 3.64(f / 1000)^{-0.8} - 6.5e^{-0.6(f / 1000)^{3.3}} + 10^{-3}(f / 1000)^4 \text{ (dB SPL)},$$

which is measured in dB SPL, or dB relative to 20 μPa. It is assumed that the threshold of audibility is a result of the internal noise of the auditory system. Effectively, the internal noise is masking a very weak external signal. If a signal has any frequency components with power levels that fall below the absolute threshold of hearing, then these components can be discarded, as the average listener will be unable to hear those frequencies of the signal anyway.

### 4.2.2 Masking Effects

In the most-broad categories, masking effects can be classified as simultaneous or temporal. In simultaneous masking, the masking sound and the masked sound are present at the same time. Temporal masking refers to the effect of masking with a small time offset.

Due to the limited time resolution of the algorithm presented in the following chapters, temporal masking is of limited use, but can be used to hide preechoes. *Forward masking* where a sound is inaudible for a short time after the masker has been removed, can be between 5 ms and more than 150 ms. *Backward masking*, where a weak signal is inaudible before the onset of the masking signal, is usually below 5 ms [92][93][94].

In masking, we need to consider two kinds of sounds that can act as the masker. Noise-like sounds with a broad spectrum and little or no phase coherence can mask sounds with levels as little as 2-6 dB below the masker. Tone-like sounds need to be much louder, need as much as 18-24 dB higher amplitude to mask other tones or noise, partially due to phase distortion and the appearance of difference tones [95].

Masking also is somewhat dependent of the absolute level of the masker. Fig. 4.4 shows the amount of masking provided by a 1 kHz tone at
various absolute sound pressure levels $L_M$. It can be seen that the slope of the upwards part of the masking curve varies with level.

It should be noted that these curves are only averages, and vary from person to person. To illustrate, the dotted lines in Fig. 4.4 show the masking provided by a 60 dB pure tone at 1 kHz for two persons at the extremes of the sample set.

### 4.2.3 Critical Bands and the Bark Scale

The frequency selectivity of masking effects is described in terms of *Critical Bands* (CB). In general, a CB is the bandwidth around a center frequency which marks a (sudden) change in subjective response. For example, the perceived loudness of narrowband noise of fixed power density is independent of bandwidth as long as the noise is confined within a CB. If the bandwidth of the noise is further increased, the perceived loudness will also increase.

![Masking Effect of Tones](image)

*Fig. 4.4: Masking curves for 1 kHz masking tone*
While the exact mechanism behind this abrupt change in frequency selectivity is not known, at least some of it can be explained in Basilar Membrane (BM) and Inner Hair Cell (IHC) behavior. As discussed above, the BM is not a perfect frequency discriminator but each point on the BM responds to a range of frequencies. This behavior is modeled as a bank of overlapping bandpass filters, called auditory filters. The shape of these filters is not exactly known, and can change with signal level, hence they are not linear. However, this non-linearity is usually ignored. A more important property of the auditory filters is that their bandwidth changes with frequency.

Moore [95] describes CB as a measure of the ‘effective bandwidth’ of the auditory filters, though it must noted that the actual width of the CB is narrower than the corresponding auditory filter.

The actual width of Critical Bands is still in dispute. According to Zwicker the bandwidth of Critical Bands is relatively constant below 500 Hz, but above that it increases approximately in proportion with frequency. Moore’s measurements (to distinguish from the traditional CB, called Effective Rectangular Band, ERB) indicated narrower band-widths, and found changes in bandwidth even below 500 Hz. Both claim to correspond to fixed distances on the BM, 1.3 mm for Zwicker’s CB and 0.9 mm Moore’s ERB.

Aside them masking, the concept of auditory filtering and Critical Bands has many implications, and is the single most dominant concept in auditory theory [35]. Thus, an absolute frequency scale based on the original (as used by Zwicker) CB measurements is in common use. This scale is called the Bark scale, and the common function to convert from Hz to Bark is

\[ z(f) = 13 \arctan(0.00076f) + 3.5 \arctan \left( \frac{f}{7500} \right)^2. \]  

and the bandwidth (in Hz) of a CB at any frequency is given by

\[ BW_c(f) = 25 + 75 \left[ 1 + 1.4(f / 1000)^2 \right]^{1/69} \]
The bandwidth in Bark of a CB at any frequency is (by definition) 1. This “normalization” of Critical Bands in frequency domain allows for simpler calculation of auditory effects, such as the spread of masking, which is the amount of masking provided by signals outside the immediate critical band.

4.2.4 Excitation Patterns and the Masking Threshold

By modeling the auditory system as a filter bank, the excitation in dB at each point of the BM can be calculated. This Excitation Pattern is used in some algorithms as a first step to calculating the Masking Threshold, which indicates the threshold of hearing in the presence of a signal. However, there are many ways of calculating the excitation pattern. This is mostly due to differing models of auditory filters, from relatively crude non-overlapping rectangular filters to more complex shapes such as Roex(p) and Gammatone Filters [35]. Furthermore, there is still much dispute about how adjacent critical bands interact, both how excitations add up, or the shape of spreading functions which describe the spread of masking.

Psychoacoustic modeling has long-since been an integral part of audio compression. It exploits properties of the human auditory system to remove the redundancies inherent in audio signals that the human ear cannot perceive. More powerful signals at certain frequencies ‘mask’ less powerful signals at nearby frequencies by de-sensitizing the human ear’s basilar membrane (which is responsible for resolving the frequency components of a signal).

Although psychoacoustic testing is common for monophonic signals, little is known about the psychoacoustic properties of multi-channel audio stimuli.

4.2.5 Psychoacoustic/Perceptual Coding

Audio codecs utilizing a psychoacoustic model generally calculate per critical band the amount of noise that can be masked [96][97]. The samples
are then quantized to the lowest bit-level allowed so that the introduced noise is still imperceptible. Utilizing this system therefore reduces the overall bit-rate of the audio signal and provides compression. This technique is sometimes referred to as *frequency-domain error confinement* [98].

4.2.5.1 The Sensitivity Threshold

There are several psychoacoustic phenomena of importance to the compression of speech signals. These phenomena are experienced every day by most human beings and mostly go unnoticed.

The ear’s sensitivity varies in a non-linear fashion over the human-audible range. The ear is more sensitive to sounds at certain frequencies and less sensitive to others. A good demonstrator of this is shown in Fig. 4.6.

![Threshold in Quiet](image)

**Fig. 4.5: The Human Ear’s Sensitivity Threshold**

Fig. 4.5 shows the threshold above which the amplitude of a signal must be in order for that sound to be distinguished. As can be observed, the ear is most sensitive to the sound around the 3 kHz mark. This is due to the fact that the most important sounds a human ear has to distinguish and be able to discern from other sounds is speech, and due to the process of human evolution the ear has become most sensitive to these frequencies.

Another factor to be observed is the ear’s sensitivity to very low and high frequencies. The ear has trouble discerning very low frequencies due to all the interference and vibrations they cause, and high frequencies are difficult for the ear to perceive as they carry so little energy as to not stimulate the ear correctly.
4.2.6 Simultaneous Masking

The main contributor to the compression of a signal is ‘*simultaneous masking*’. Simultaneous masking refers to the instant when signals (of similar or non-similar frequency) occur at the same time.

Hairs within the inner ear are sensitive to different frequencies. When they are stimulated by a powerful sound at a certain frequency the ear is unable to detect a soft sound present at the same pitch. The hairs around this area are also interfered with causing the receptors around this area to become less sensitive to certain frequencies, thereby not allowing the ear to discern differences between certain pitches.

This effect is evident throughout every-day life; the classic example being that a person can talk on the telephone to another person and hear a clock ticking in the background only when the other person is not talking.

![Fig. 4.6: The Conventional Masking Diagram](image)

The conventional masking diagram shown in Fig. 4.6 demonstrates how a 50 dB tone at just below 500 Hz can makes a signal of about 25 dB at above 1 kHz. This effect is best described by picturing the sensitivity threshold (on above diagram) changing its characteristics due to an external signal. With the
application of a singular tone the sensitivity threshold curve is raised at a
certain point, thereby making it necessary for tones around this point to be
more powerful in order to be perceived.

Psychoacoustic models, in the general case, are built through the
subjective testing of impartial individuals by their ability to detect sounds in
the presence of another. In order to build a viable model many different
aspects of hearing must be tested. For instance, a model should give an
understanding of how pure tones mask pure tones, how pure tones mask
noise, how noise masks pure tone and how noise masks noise. Furthermore,
the testing does not have to be confined to purely singular tones and so tones
of different frequency could be used to see how they mask singular tones and
vice versa. For each frequency tested, several masking-tone volume levels
should be used and along with the addition of four other simultaneous audio
channels. Thus models can easily become quite complex.

The quality of a psychoacoustic model is therefore depending on the
amount of testing. Obviously, the more variables accounted for the better the
overall model.

4.3 Conclusion: In this chapter the human auditory system has been
discussed very elaborately and also the masking method which is defined as a
process or the amount by which threshold of audibility is raised by the
presence of another (masking) sound has been discussed. The various masking
methods for example, simultaneous masking method and psychoacoustic
model, their effects and parameters are also been discussed in this chapter.