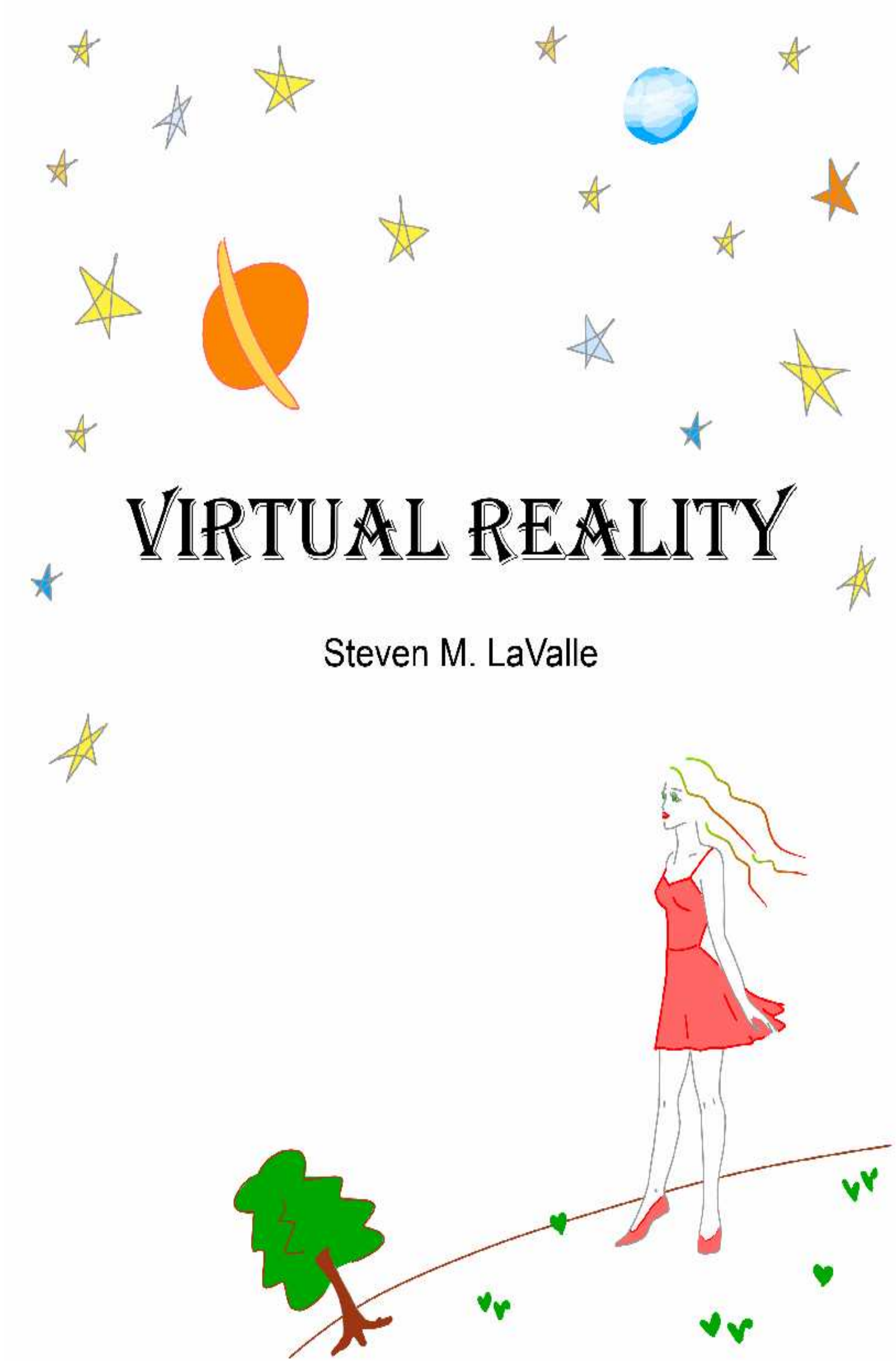


Chapter 11

Audio

Steven M. LaValle

University of Illinois

Available for downloading at **http://vr.cs.uiuc.edu/**

Chapter 11

Audio

Hearing is an important sense for VR and has been unfortunately neglected up until this chapter. Developers of VR systems tend to focus mainly on the vision part because it is our strongest sense; however, the audio component of VR is powerful and the technology exists to bring high fidelity audio experiences into VR. In the real world, audio is crucial to art, entertainment, and oral communication. As mentioned in Section 2.1, audio recording and reproduction can be considered as a VR experience by itself, with both a CAVE-like version (surround sound) and a headset version (wearing headphones). When combined consistently with the visual component, audio helps provide a compelling and comfortable VR experience.

Each section of this chapter is the auditory (or audio) complement to one of Chapters 4 through 7. The progression again goes from physics to physiology, and then from perception to rendering. Section 11.1 explains the physics of sound in terms of waves, propagation, and frequency analysis. Section 11.2 describes the parts of the human ear and their function. This naturally leads to auditory perception, which is the subject of Section 11.3. Section 11.4 concludes by presenting auditory rendering, which can produce sounds synthetically from models or reproduce captured sounds. When reading these sections, it is important to keep in mind the visual counterpart of each subject. The similarities make it easier to quickly understand and the differences lead to unusual engineering solutions.

11.1 The Physics of Sound

This section parallels many concepts from Chapter 4, which covered the basic physics of light. Sound wave propagation is similar in many ways to light, but with some key differences that have major perceptual and engineering consequences. Whereas light is a *transverse wave*, which oscillates in a direction perpendicular to its propagation, sound is a *longitudinal wave*, which oscillates in a direction parallel to its propagation. Figure 11.1 shows an example of this for a parallel wavefront.

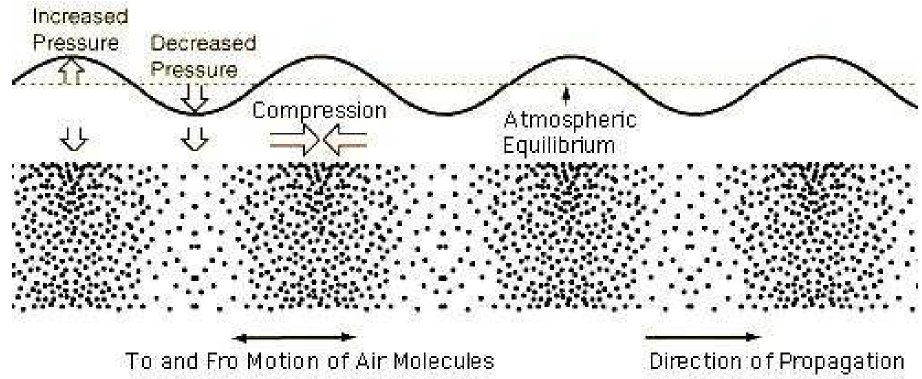


Figure 11.1: Sound is a longitudinal wave of compression and rarefaction of air molecules. The case of a pure tone is shown here, which leads to a sinusoidal pressure function. (Figure from Pond Science Institute.)

Sound corresponds to vibration in a medium, which is usually air, but could also be water, or any other gases, liquids, or solids. There is no sound in a vacuum, which is unlike light propagation. For sound, the molecules in the medium displace, causing variations in pressure that range from a *compression* extreme to a decompressed, *rarefaction* extreme. At a fixed point in space, the pressure varies as a function of time. Most importantly, this could be the pressure variation on a human eardrum, which is converted into a perceptual experience. The sound pressure level is frequently reported in *decibels* (abbreviated as *dB*), which is defined as

$$N_{db} = 20 * \log_{10}(p_e/p_r). \tag{11.1}$$

Above, p_e is the pressure level of the peak compression and p_r is a reference pressure level, which is usually taken as 2×10^{-7} newtons / square meter.

Sound waves are typically produced by vibrating solid materials, especially as they collide or interact with each other. A simple example is striking a large bell, which causes it to vibrate for many seconds. Materials may also be forced into sound vibration by sufficient air flow, as in the case of a flute. Human bodies are designed to produce sound by using lungs to force air through the vocal cords, which causes them to vibrate. This enables talking, singing, screaming, and so on.

Sound sources and attenuation As in the case of light, we can consider rays, for which each *sound ray* is perpendicular to the sound propagation wavefront. A point sound source can be defined, which produces emanating rays with equal power in all directions. This also results in power reduction at a quadratic rate as a function of distance from the source. Such a point source is useful for modeling, but cannot be easily achieved in the real world. Planar wavefronts can be achieved by vibrating a large, flat plate, which results in the acoustic equivalent of collimated light. An important distinction, however, is the *attenuation* of sound as it propagates through a medium. Due to energy lost in the vibration of molecules, the sound intensity decreases by a constant factor (or fixed percentage) for every unit of distance from the planar source; this is an example of *exponential decay*.

Propagation speed Sound waves propagate at 343.2 meters per second through air at 20° C (68° F). For comparison, light propagation is about 874,000 times faster. We have planes and cars that can surpass the speed of sound, but are nowhere near traveling at the speed of light. This is perhaps the most important difference between sound and light for making VR systems. The result is that human senses and engineered sensors easily measure differences in arrival times of sound waves, leading to stronger emphasis on temporal information.

Frequency and wavelength As in Chapter 4.1, the decomposition of waves into frequency components becomes important. For sound, the frequency is the number of compressions per second and is called *pitch*. The range is generally considered to be from 20 Hz to 20,000 Hz, which is based on human hearing, much in the same way that the frequency range for light is based on human vision. Vibrations above 20,000 Hz are called *ultrasound*, and are audible to some animals. Vibrations below 20 Hz are called *infrasound*.

Using (4.1) from Section 4.1 and the propagation speed $s = 343.2$, the wavelength of a sound wave can also be determined. At 20 Hz the wavelength is $\lambda = 343.2/20 = 17.1$m. At 20,000 Hz, it becomes $\lambda = 17.1$mm. The waves are the sizes of objects in our world. This causes the sound to interfere with objects in a complicated way that is difficult to model when trying to reproduce the behavior in VR. By comparison, light waves are tiny, ranging from 400 nm to 700 nm.

Doppler effect The sound pressure variations described above were for a fixed receiving point. If the point is moving away from the source, then the wavefronts will arrive at a reduced frequency. For example, if the receiver moves at 43.2m/s away from the source, then the waves would seem to be traveling at only $343.2 - 43.2 = 300$ meters per second. The received frequency shifts due to the relative motion between the source and receiver. This is known as the *Doppler effect*, and the frequency as measured at the receiver can be calculated as

$$f_r = \left(\frac{s + v_r}{s + v_s} \right) f_s, \tag{11.2}$$

in which s is the propagation speed in the medium, v_r is the velocity of the receiver, v_s is the velocity of the source, and f_s is the frequency of the source. In our example, $s = 343.2$, $v_r = -43.2$, and $v_s = 0$. The result is that a sound source with frequency $f_s = 1000$Hz would be perceived by the receiver as having

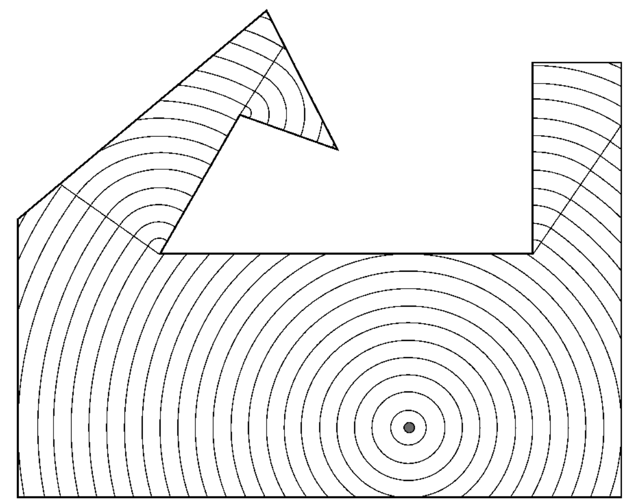


Figure 11.2: Waves can even bend around corners, due to *diffraction*. A top-down view of a room is shown. At each of the three interior corners, the propagating wavefront expands around it.

frequency $f_r \approx 876.7$. This is the reason why a siren seems to change pitch as a police car passes by. The Doppler effect also applies to light, but the effect is negligible in normal VR contexts (unless developers want to experiment with virtual time dilation, space travel, and so on).

Reflection and transmission As with light, wave propagation is strongly effected by propagation through media. Imagine a sound wave hitting an interior wall as someone yells from inside of a room. It may be helpful to think about a ray of sound approaching the wall. Due to reflection, much of the sound will bounce as if the wall were an acoustic mirror. However, some of the sound energy will penetrate the wall. Sounds propagates more quickly through more solid materials, resulting in a bending of the ray as it penetrates. This is refraction. Some of the sound escapes the far side of the wall and propagates through the air in an adjacent room, resulting in transmission. Thus, someone in the adjacent room can hear yelling. The total amount of energy contained in the sound waves before it hits the wall is split by reflection and transmission, with additional loss due to attenuation.

Diffraction Wavefronts can also bend around corners, which is called *diffraction*; see Figure 11.2. This would enable someone to hear a sound that is around the corner of a building, without relying on any reflection or transmission. More diffraction occurs for longer wavelengths; thus, a lower-pitched sound bends around

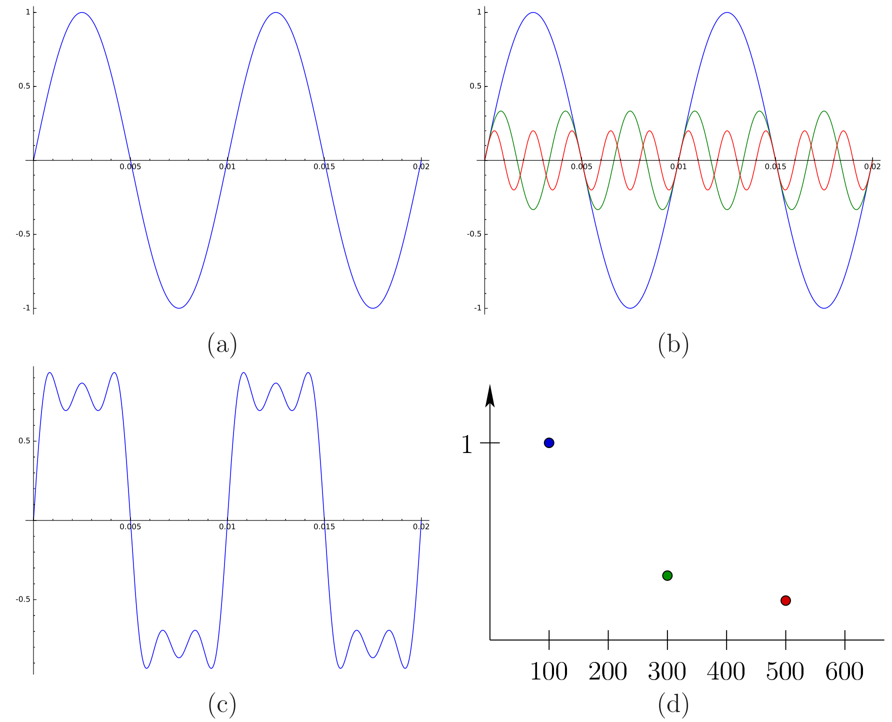


Figure 11.3: (a) A pure tone (sinusoid) of unit amplitude and frequency 100 Hz. (b) Three pure tones; in addition to the original blue, the green sinusoid has amplitude 1/3 and frequency 300 Hz, and the red one has amplitude 1/5 and frequency 500 Hz. (c) Directly adding the three pure tones approximates a square-like waveform. (d) In the frequency spectrum, there are three non-zero points, one for each pure tone.

corners more easily. This also explains why we are more concerned about acoustic diffraction in a room than light diffraction, although the latter is often important for lenses (recall the Fresnel lens drawback of Section 7.3).

Fourier analysis Spectral decompositions were important for characterizing light sources and reflections in Section 4.1. In the case of sound, they are even more important. A sinusoidal wave, as shown in Figure 11.3(a), corresponds to a *pure tone*, which has a single associated frequency; this is analogous to a color from the light spectrum. A more complex waveform, such the sound of a piano note, can be constructed from a combination of various pure tones. Figures 11.3(b) to 11.3(d) provide a simple example. This principle is derived from *Fourier analysis*, which enables any periodic function to be decomposed into sinusoids (pure tones

in our case) by simply adding them up. Each pure tone has a particular *frequency*, *amplitude* or scaling factor, and a possible timing for its peak, which is called its *phase*. By simply adding up a finite number of pure tones, virtually any useful waveform can be closely approximated. The higher-frequency, lower-amplitude sinusoids are often called *higher-order harmonics*; the largest amplitude wave is called the *fundamental frequency*. The plot of amplitude and phase as a function of frequency is obtained by applying the *Fourier transform*, which will be briefly covered in Section 11.4.

Where are the lenses? At this point, the most obvious omission in comparison to Chapter 4 is the acoustic equivalent of lenses. As stated above, refraction occurs for sound. Why is it that human ears do not focus sounds onto a spatial image in the same way as the eyes? One problem is the long wavelengths in comparison to light. Recall from Section 5.1 that the photoreceptor density in the fovea is close to the wavelength of visible light. It is likely that an "ear fovea" would have to be several meters across or more, which would makes our heads too large. Another problem is that low-frequency sound waves interact with objects in the world in a more complicated way. Thus, rather than forming an image, our ears instead work by performing Fourier analysis to sift out the structure of sound waves in terms of sinusoids of various frequencies, amplitudes, and phases. Each ear is more like a single-pixel camera operating at tens of thousands of "frames per second", rather than capturing a large image at a slower frame rate. The emphasis for hearing is the distribution over *time*, whereas the emphasis is mainly on *space* for vision. Nevertheless, both time and space are important for both hearing and vision.

11.2 The Physiology of Human Hearing

Human ears convert sound pressure waves into neural impulses, which ultimately lead to a perceptual experience. The anatomy of the human ear is shown in Figure 11.4. The ear is divided into outer, middle, and inner parts, based on the flow of sound waves. Recall from Section 5.3 the complications of eye movements. Although cats and some other animals can rotate their ears, humans cannot, which simplifies this part of the VR engineering problem.

Outer ear The floppy part of the ear that protrudes from the human head is called the *pinna*. It mainly serves as a funnel for collecting sound waves and guiding them into the *ear canal*. It has the effect of amplifying sounds in the 1500 to 7500Hz frequency range [27]. It also performs subtle filtering of the sound, causing some variation in the high-frequency range that depends on the incoming direction of the sound source. This provides a powerful cue regarding the direction of a sound source.

After traveling down the ear canal, the sound waves cause the *eardrum* to vibrate. The eardrum is a cone-shaped membrane that separates the outer ear

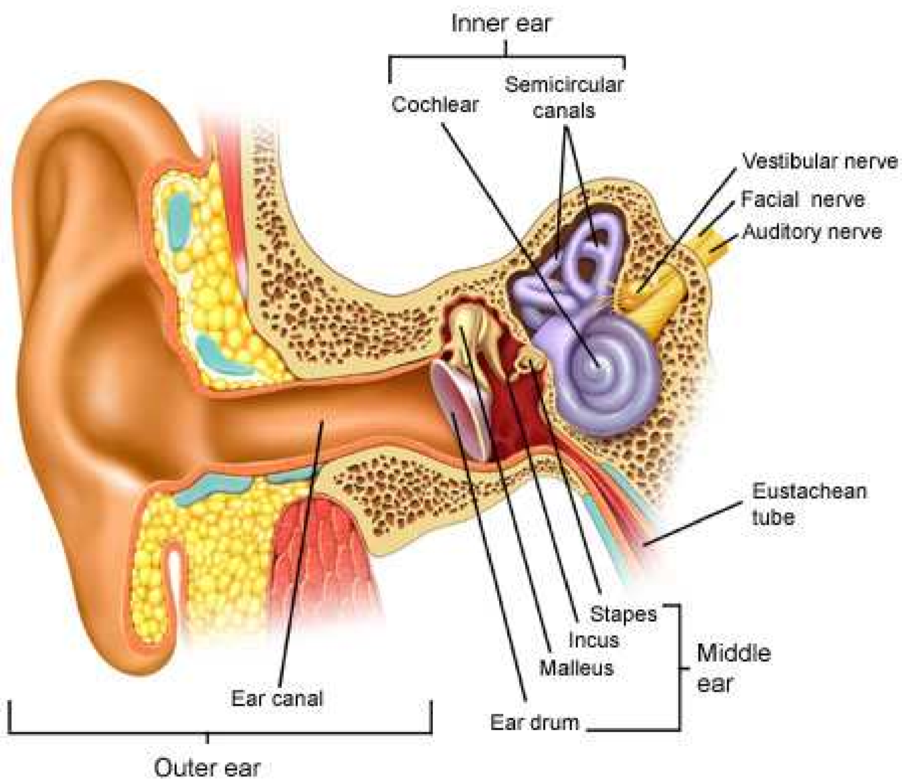
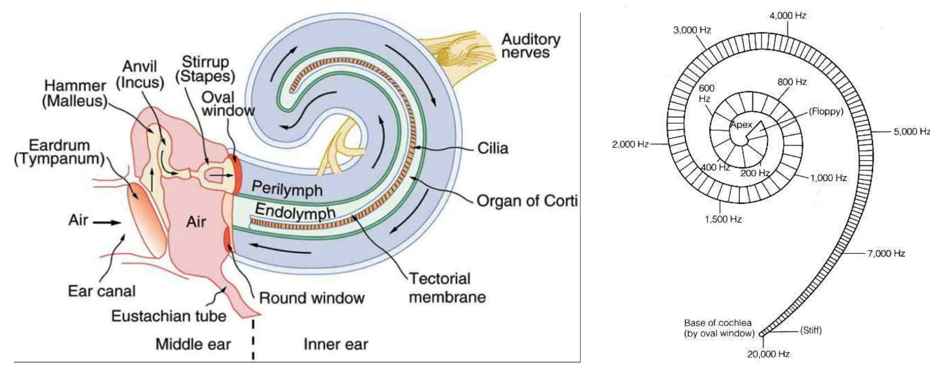


Figure 11.4: The physiology of the human auditory system.

from the middle ear. Its covers only 55mm^2 of area. If this were a camera, it would have a resolution of one pixel at this point because no additional spatial information exists other than what can be inferred from the membrane vibrations.

Middle ear The main function of the middle ear is to convert vibrating air molecules in the outer ear into vibrating liquid in the inner ear. This is accomplished by bones that connect the eardrum to the inner ear. The air and the liquid of the inner ear have differing *impedance*, which is the resistance to vibration. The bones are called the malleus (hammer), incus (anvil), and stapes (stirrup), and they are connected in series via muscles and ligaments that allow relative movement. The purpose of the bones is to match the impedance so that the pressure waves are transmitted to the inner ear with as little power loss as possible. This avoids the tendency of a higher impedance material to reflect the sound away. An example of this is voices reflecting over the surface of a lake, rather than being transmitted into the water.



(a) (b)

Figure 11.5: The operation of the cochlea: (a) The perilymph transmits waves that are forced by the oval window through a tube that extends the length of the cochlea and back again, to the round window. (b) Because of varying thickness and stiffness, the central spine (basilar membrane) is sensitive to particular frequencies of vibration; this causes the mechanoreceptors, and ultimately auditory perception, to be frequency sensitive.

Inner ear The inner ear contains both the vestibular organs, which were covered in Section 8.2, and the *cochlea*, which is the sense organ for hearing. The cochlea converts sound energy into neural impulses via mechanoreceptors. This is accomplished in a beautiful way that performs a spectral decomposition in the process so that the neural impulses encode amplitudes and phases of frequency components.

Figure 11.5 illustrates its operation. As seen in Figure 11.5(a), eardrum vibration is converted into oscillations of the *oval window* at the base of the cochlea. A tube that contains a liquid called *perilymph* runs from the oval window to the *round window* at the other end. The *basilar membrane* is a structure that runs through the center of the cochlea, which roughly doubles the length of the tube containing perilymph. The first part of the tube is called the *scala vestibuli*, and the second part is called the *scala tympani*. As the oval window vibrates, waves travel down the tube, which causes the basilar membrane to displace. The membrane is thin and stiff near the base (near the oval and round windows) and gradually becomes soft and floppy at the furthest away point, called the *apex*; see Figure 11.5(b). This causes each point on the membrane to vibrate only over a particular, narrow range of frequencies.

Mechanoreceptors The basilar membrane is surrounded by a larger and complicated structure called the *organ of Corti*, which additionally contains mechanoreceptors that are similar to those shown in Section 8.2. See Figure 11.6. The

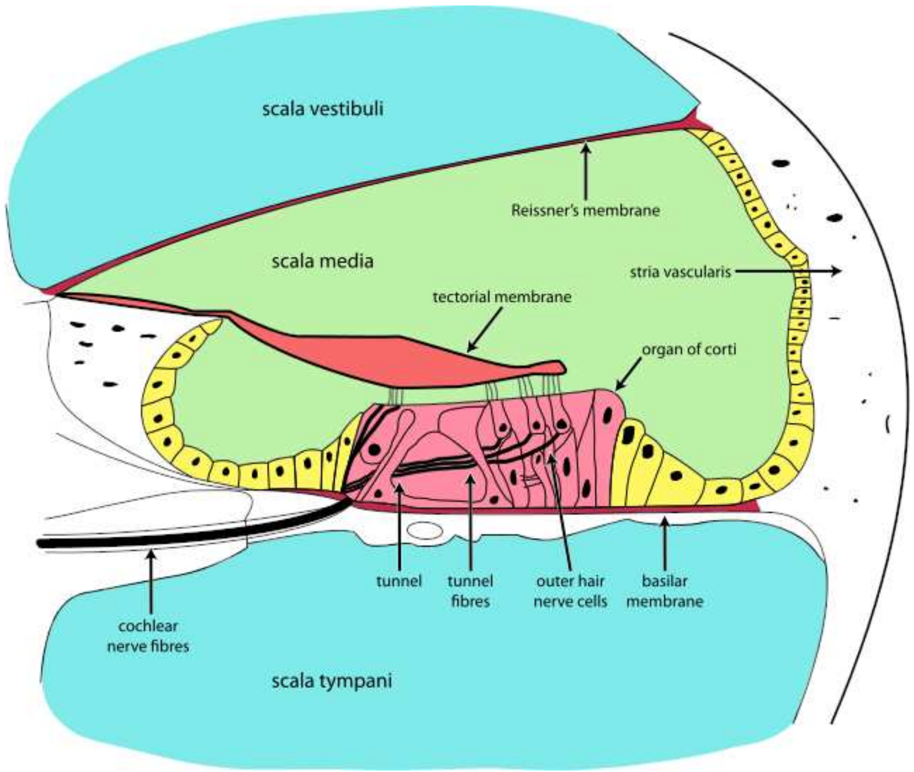


Figure 11.6: A cross section of the *organ of Corti*. The basilar and tectorial membranes move relative to each other, causing the hairs in the mechanoreceptors to bend. (Figure from multiple Wikipedia users.)

mechanoreceptors convert displacements of hairs into neural impulses. The hairs are displaced as the basilar membrane vibrates because the ends of some are attached to the *tectorial membrane*. The relative motions of the basilar and tectorial membranes causes a shearing action that moves the hairs. Each ear contains around 20,000 mechanoreceptors, which is considerably less than the 100 million photoreceptors in the eye.

Spectral decomposition By exploiting the frequency-based sensitivity of the basilar membrane, the brain effectively has access to a spectral decomposition of the incoming sound waves. It is similar to, but not exactly the same as, the Fourier decomposition which discussed in Section 11.1. Several differences are mentioned in Chapter 4 of [7]. If pure tones at two different frequencies are simultaneously presented to the ear, then the basilar membrane produces a third tone, which is sometimes audible [5]. Also, the neural impulses that result from

Figure 11.7: Due to the *precedence effect*, an auditory illusion occurs if the head is placed between stereo speakers so that one is much closer than the other. If they output the same sound at the same time, then the person perceives the sound arriving from the closer speaker, rather than perceiving an echo.

mechanoreceptor output are not linearly proportional to the frequency amplitude. Furthermore, the detection one of tone may cause detections of nearby tones (in terms of frequency) to be inhibited [17], much like lateral inhibition in horizontal cells (recall from Section 5.2). Section 11.4.1 will clarify how these differences make the ear more complex in terms of filtering.

Auditory pathways The neural pulses are routed from the left and right cochleae up to the highest level, which is the *primary auditory cortex* in the brain. As usual, hierarchical processing occurs as the signals are combined through neural structures. This enables multiple frequencies and phase shifts to be analyzed. An early structure called the *superior olive* receives signals from both ears so that differences in amplitude and phase can be processed. This will become important in Section 11.3 for determining the location of an audio source. At the highest level, the primary auditory cortex is mapped out *tonotopically* (locations are based on frequency), much in the same way as topographic mapping of the visual cortex.

11.3 Auditory Perception

Now that we have seen the hardware for hearing, the next part is to understand how we perceive sound. In the visual case, we saw that perceptual experiences are often surprising because they are based on adaptation, missing data, assumptions filled in by neural structures, and many other factors. The same is true for auditory experiences. Furthermore, *auditory illusions* exist in the same way as optical illusions. The McGurk effect from Section 6.4 was an example that used vision to induce incorrect auditory perception.

Precedence effect A more common auditory illusion is the *precedence effect*, in which only one sound is perceived if two nearly identical sounds arrive at slightly different times; see Figure 11.7. Sounds often reflect from surfaces, causing *reverberation*, which is the delayed arrival at the ears of many "copies" of the sound due to the different propagation paths that were taken from reflections,

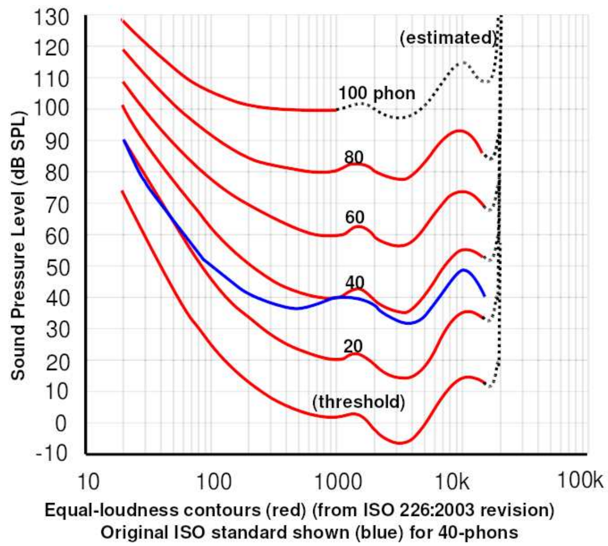


Figure 11.8: Contours of equal loudness perception as a function of frequency.

transmissions, and diffraction. Rather than hearing a jumble, people perceive a single sound. This is based on the first arrival, which usually has the largest amplitude. An echo is perceived if the timing difference is larger than the *echo threshold* (in one study it ranged from 3 to 61ms [25]). Other auditory illusions involve incorrect localization (*Franssen effect* and *Glissando illusion* [3]), illusory continuity of tones [24], and forever increasing tones (*Shepard tone illusion* [18]).

Psychoacoustics and loudness perception The area of psychophysics, which was introduced in Section 2.3, becomes specialized to *psychoacoustics* for the case of auditory perception. Stevens' law of perceived stimulus magnitude and Weber's law of just noticeable differences (JNDs) appear throughout the subject. For example, the exponent for Stevens law (recall (2.1)), for perceived loudness of a 3000 Hz pure tone is $x = 0.67$ [20]. This roughly means that if a sound increases to a much higher pressure level, we perceive it as only a bit louder. A more complicated example from psychoacoustics is shown in Figure 11.8, which are contours that correspond to equal loudness perception as a function of frequency. In other words, as the frequency varies, at what levels are the sounds perceived to be the same loudness? This requires careful design of experiments with human subjects, a problem that is common throughout VR development as well; see Section 12.4.

Pitch perception When considering perception, the frequency of a sound wave is referred to as *pitch*. Perceptual psychologists have studied the ability of people to detect a targeted pitch in spite of confusion from sounds consisting of other wavelengths and phases. One fundamental observation is that the auditory perception system performs *critical band masking* to effectively block out waves that have frequencies outside of a particular range of interest. Another well-studied problem is the perception of differences in pitch (or frequency). For example, for a pure tone at 1000 Hz, could someone distinguish it from a tone at 1010 Hz? This is an example of JND. It turns out that for frequencies below 1000 Hz, humans can detect a change of frequency that is less than 1 Hz. The discrimination ability decreases as the frequency increases. At 10,000 Hz, the JND is about 100 Hz. In terms of percentages, this means that pitch perception is better than a 0.1% difference at low frequencies, but increases to 1.0% for higher frequencies.

Also regarding pitch perception, a surprising auditory illusion occurs when the fundamental frequency is removed from a complex waveform. Recall from Figure 11.3 that a square wave can be approximately represented by adding sinusoids of smaller and smaller amplitudes, but higher frequencies. It turns out that people perceive the tone of the fundamental frequency, even when it is removed, and only the higher-order harmonics remain; several theories for this are summarized in Chapter 5 of [7].

Localization One of the main areas of psychoacoustics is *localization*, which means estimating the location of a sound source by hearing it. This is crucial for many VR experiences. For example, if people are socializing, then their voices should seem to come from the mouths of corresponding avatars. In other words, the auditory and visual cues should match. Any kind of sound effect, such as a car or zombie approaching, should also have matched cues.

The JND concept is applied for localization to obtain the *minimum audible angle* (*MAA*), which is the minimum amount of angular variation that can be detected by a human listener. A spherical coordinate system is usually used for localization, in which the listener's head is at the origin; see Figure 11.9. The angle in the horizontal plane between the forward direction and the source is called the *azimuth*, which extends from -180 to 180 degrees. The angle corresponding to deviation of the source from the horizontal plane is called the *elevation*, which extends from -90 to 90 degrees. The third coordinate is the *radius* or *distance* from the origin (head center) to the source. The MAA depends on both frequency and the direction of the source. Figure 11.10 shows a plot of the MAA as a function of frequency, at several values for azimuth. The amount of variation is surprising. At some frequencies and locations, the MAA is down to 1 degree; however, at other combinations, localization is extremely bad.

Monaural cues Auditory localization is analogous to depth and scale perception for vision, which was covered in Section 6.1. Since humans have a pair of

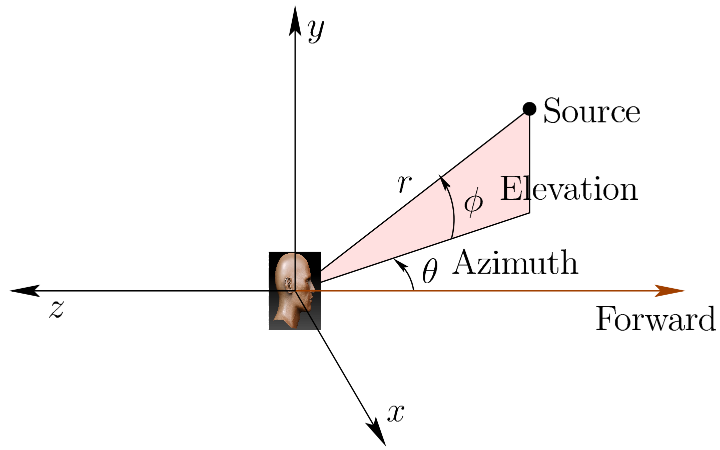


Figure 11.9: Spherical coordinates are used for the source point in auditory localization. Suppose the head is centered on the origin and facing in the $-z$ direction. The *azimuth* θ is the angle with respect to the forward direction after projecting the source into the xz plane. The *elevation* ϕ is the interior angle formed by a vertical triangle that connects the origin to the source and to the projection of the source into the plane. The radius r is the distance from the origin to the source.

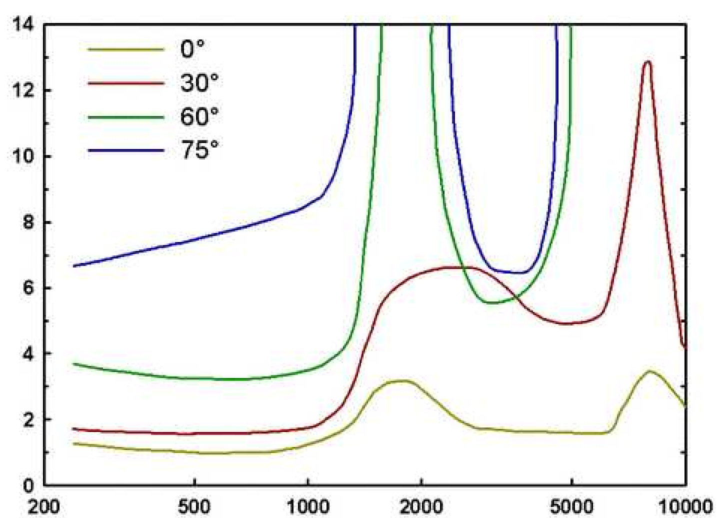


Figure 11.10: Plots of the *minimum audible angle* (*MAA*) as a function of frequency. Each plot corresponds to a different azimuth angle.

ears, localization cues can be divided into ones that use a single ear and others that require both ears. This is analogous to monocular and binocular cues for vision. A *monaural cue* relies on sounds reaching a single ear to constrain the set of possible sound sources. Several monaural cues are [28]:

1. The pinna is shaped asymmetrically so that incoming sound is distorted in a way that depends on the direction from which it arrives, especially the elevation. Although people are not consciously aware of this distortion, the auditory system uses it for localization.

2. The amplitude of a sound decreases quadratically with distance. If it is a familiar sound, then its distance can be estimated from the perceived amplitude. Familiarity affects the power of this cue in the same way that familiarity with an object allows depth and scale perception to be separated.

3. For distant sounds, a distortion of the frequency spectrum occurs because higher-frequency components attenuate more quickly than low-frequency components. For example, distant thunder is perceived as a deep rumble, but nearby thunder includes a higher-pitched popping sound.

4. Finally, a powerful monaural cue is provided by the reverberations entering the ear as the sounds bounce around; this is especially strong in a room. Even though the precedence effect prevents us perceiving these reverberations, the brain nevertheless uses the information for localization. This cue alone is called *echolocation*, which is used naturally by some animals, including bats. Some people can perform this by making clicking sounds or other sharp noises; this allows *acoustic wayfinding* for blind people.

Binaural cues If both ears become involved, then a *binaural cue* for localization results. The simplest case is the *interaural level difference* (*ILD*), which is the difference in sound magnitude as heard by each ear. For example, one ear may be facing a sound source, while the other is in the *acoustic shadow* (the shadow caused by an object in front of a sound source is similar the shadow from a light source). The closer ear would receive a much stronger vibration than the other.

Another binaural cue is *interaural time difference* (*ITD*), which is closely related to the TDOA sensing approach described in Section 9.3. The distance between the two ears is approximately 21.5cm, which results in different arrival times of the sound from a source. Note that sound travels 21.5cm in about 0.6ms, which means that surprisingly small differences are used for localization.

Suppose that the brain measures the difference in arrival times as 0.3ms. What is the set of possible places where the source could have originated? This can be solved by setting up algebraic equations, which results in a conical surface known as a hyperboloid. If it is not known which sound came first, then the set of possible places is a hyperboloid of two disjoint sheets. Since the brain knows which one came first, the two sheets are narrowed down to one hyperboloid sheet, which is

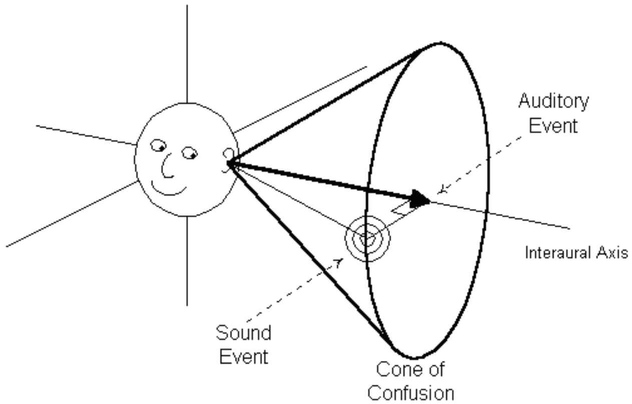


Figure 11.11: The *cone of confusion* is the set of locations where a point source might lie after using the ITD binaural cue. It is technically a hyberboloid, but approximately looks like a cone.

called the *cone of confusion*; see Figure 11.11 (in most cases, it approximately looks like a cone, even though it is hyperboloid). Uncertainty within this cone can be partly resolved, however, by using the distortions of the pinna.

The power of motion More importantly, humans resolve much ambiguity by simply moving their heads. Just as head movement allows the powerful vision depth cue of parallax, it also provides better auditory localization. In fact, *auditory parallax* even provides another localization cue because nearby audio sources change their azimuth and elevation faster than distant ones. With regard to ITD, imagine having a different cone of confusion for every head pose, all within a short time. By integrating other senses, the relative head poses can be estimated, which roughly allows for an intersection of multiple cones of confusion to be made, until the sound source is precisely pinpointed. Finally, recall that the motion of a source relative to the receiver causes the Doppler effect. As in the case of vision, the issue of perceived self motion versus the motion of objects emerges based on the auditory input arises. This could contribute to vection (recall Section 8.2).

11.4 Auditory Rendering

We now arrive at the problem of producing sounds for the virtual world, and sending them to aural displays (speakers) so that the user perceives them as they were designed for the VR experience. They should be consistent with visual cues and with past auditory experiences in the real world. Whether recorded sounds, synthetic sounds, or a combination, the virtual pressure waves and their rendering

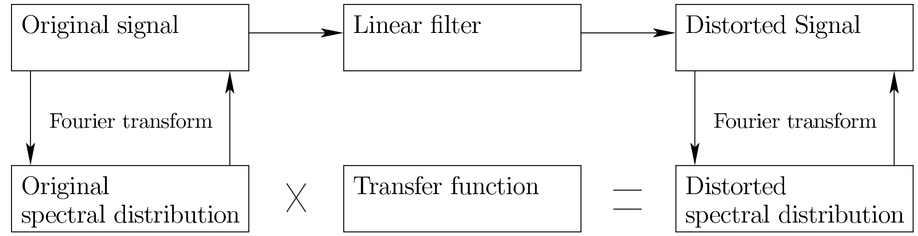


Figure 11.12: An overview of a linear filter and its relationship to Fourier analysis. The top row of blocks corresponds to the time domain, whereas the bottom row is the frequency (or spectral) domain.

to speakers should sufficiently fool the user's brain.

11.4.1 Basic signal processing

The importance of frequency components in sound waves should be clear by now. This remains true for the engineering problem of synthesizing sounds for VR, which falls under the area of *signal processing*. A brief overview is given here; see [1, 6] for further reading. As the core of this subject is the characterization or design of *filters* that transform or distort signals. In our case the signals are sound waves that could be fully synthesized, captured using microphones, or some combination. (Recall that both synthetic and captured models exist for the visual case as well.)

Figure 11.12 shows the overall scheme, which will be presented over this section. The original signal appears in the upper left. First, follow the path from left to right. The signal enters a black box labeled *linear filter* and becomes distorted, as shown in the right. What is a linear filter? Some background concepts are needed before returning to that question.

Sampling rates Signal processing formulations exist for both *continuous-time*, which makes nice formulations and mathematical proofs, and *discrete-time*, which has an uglier appearance, but corresponds directly to the way computers process signals. Because of its practical value, we will focus on the discrete-time case.

Start with a signal as a function of time, with values represented as $x(t)$. Using digital processing, it will be sampled at regular time intervals. Let Δt be the sampling interval. The *sampling rate* or (*sampling frequency*) roughly $1/\Delta t$ Hz. For example, with 1000 Hz sampling frequency, Δt is one millisecond. According to the *Nyquist-Shannon sampling theorem*, the sampling rate should be at least two times the highest frequency component in the signal. Since the highest frequency component for audio is 20,000 Hz, this suggests that the sampling rate should be at least 40,000 Hz. By no coincidence, the sampling rate of CDs and DVDs are 44,100 Hz and 48,000 Hz, respectively.

By sampling the signal, an array of values is produced.[1] At 1000 Hz, the array would contain a thousand values for every second. Using an index variable k, we can refer to the kth sample as $x[k]$, which corresponds to $x(k\Delta t)$. Arbitrarily, the first sample is $x[0] = x(0)$.

Linear filters In the context of signal processing, a *filter* is a transformation that maps one signal to another. Each signal is a function of time, and the filter is like a black box that receives the one signal as input, and produces another as output. If x represents an entire signal (over all times), then let $F(x)$ represent the resulting signal after running it through the filter.

A *linear filter* is a special kind of filter that satisfies two algebraic properties. The first algebraic property is *additivity*, which means that if two signals are added and sent through the filter, the result should be the same as if they were each sent through the filter independently, and then the resulting transformed signals were added. Using notation, this is $F(x+x') = F(x)+F(x')$ for any two signals x and x'. For example, if two different sounds are sent into the filter, the result should be the same whether they are combined before or after the filtering. This concept will become useful as multiple sinusoids are sent through the filter.

The second algebraic property is *homogeneity*, which means that if the signal is scaled by a constant factor before being sent though the filter, the result would be the same as if it were scaled by the same factor afterwards. Using notation, this means that $cF(x) = F(cx)$ for every constant c and signal x. For example, this means that if we double the sound amplitude, then the output sound from the filter doubles its amplitude as well.

A linear filter generally takes the form

$$y[k] = c_0x[k] + c_1x[k-1] + c_2x[k-2] + c_3x[k-3] + \cdots + c_nx[k-n], \qquad (11.3)$$

in which each c_i is a constant, and $n+1$ is the number of samples involved in the filter. One may consider the case in which n tends to infinity, but it will not be pursued here. Not surprisingly, (11.3) is a linear equation. This particular form is a *causal filter* because the samples on the left occur no later than the sample $y[k]$. A non-causal filter would require dependency on future samples, which is reasonable for a recorded signal, but not for live sampling (the future is unpredictable!).

Here are some examples of linear filters (special cases of (11.3)). This one takes a moving average of the last three samples:

$$y[k] = \frac{1}{3}x[k] + \frac{1}{3}x[k-1] + \frac{1}{3}x[k-2]. \qquad (11.4)$$

[1]The values are also discretized, and are represented using floating-point numbers. This level of discretization will be ignored.

Alternatively, this is an example of *exponential smoothing* (also called *exponentially weighted moving average*):

$$y[k] = \frac{1}{2}x[k] + \frac{1}{4}x[k-1] + \frac{1}{8}x[k-2] + \frac{1}{16}x[k-3]. \qquad (11.5)$$

Finite impulse response An important and useful result is that the behavior of a linear filter can be fully characterized in terms of its *finite impulse response* (*FIR*). The filter in (11.3) is often called an *FIR filter*. A *finite impulse* is a signal for which $x[0] = 1$ and $x[k] = 0$ for all $k > 0$. Any other signal can be expressed as a linear combination of time-shifted finite impulses. If a finite impulse is shifted, for example $x[2] = 1$, with $x[k] = 0$ for all other $k \neq 2$, then a linear filter produces the same result, but it is just delayed two steps later. A finite impulse can be rescaled due to filter linearity, with the output simply being rescaled. The results of sending scaled and shifted impulses through the filter are also obtained directly due to linearity.

Nonlinear filters Any (causal) filter that does not follow the form (11.3) is called a *nonlinear filter*. Recall from Section 11.2, that the operation of the human auditory system is almost a linear filter, but exhibits characteristics that make it into a nonlinear filter. Linear filters are preferred because of their close connection to spectral analysis, or frequency components, of the signal. Even if the human auditory system contains some nonlinear behavior, analysis based on linear filters is nevertheless valuable.

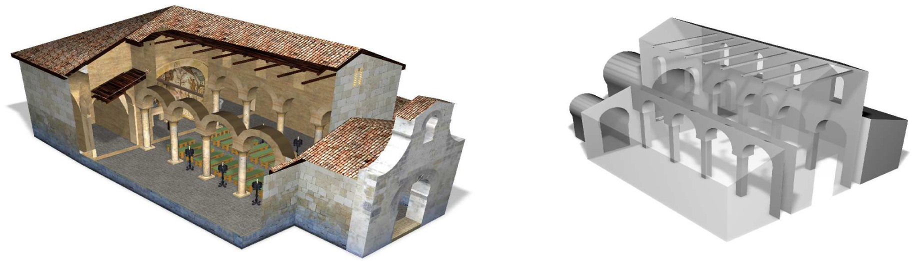
Returning to Fourier analysis Now consider the bottom part of Figure 11.12. The operation of a linear filter is easy to understand and compute in the *frequency domain*. This is the function obtained by performing the Fourier transform on the signal, which provides an amplitude for every combination of frequency and phase. This transform was briefly introduced in Section 11.1 and illustrated in Figure 11.3. Formally, it is defined for discrete-time systems as

$$X(f) = \sum_{k=-\infty}^{\infty} x[k]e^{-i2\pi fk}, \qquad (11.6)$$

in which $X(f)$ is the resulting spectral distribution, which is a function of the frequency f. The exponent involves $i = \sqrt{-1}$ and is related to sinusoids through Euler's formula:

$$e^{-i2\pi fk} = \cos(-2\pi fk) + i\sin(-2\pi fk). \qquad (11.7)$$

Unit complex numbers are used as an algebraic trick to represent the phase. The *inverse Fourier transform* is similar in form and converts the spectral distribution back into the time domain. These calculations are quickly performed in practice by using the *Fast Fourier Transform* (*FFT*) [1, 6].



(a) (b)

Figure 11.13: An audio model is much simpler. (From Pelzer, Aspock, Schroder, and Vorlander, 2014, [11])

Transfer function In some cases, a linear filter is designed by expressing how it modifies the spectral distribution. It could amplify some frequencies, while suppressing others. In this case, the filter is defined in terms of a *transfer function*, which is applied as follows: 1) transforming the original signal using the Fourier transform, 2) multiplying the result by the transfer function to obtain the distorted spectral distribution, and then 3) applying the inverse Fourier transform to obtain the result as a function of time. The transfer function can be calculated from the linear filter by applying the discrete Laplace transform (called *z-transform*) to the finite impulse response [1, 6].

11.4.2 Acoustic modeling

The geometric modeling concepts from Section 3.1 apply to the auditory side of VR, in addition to the visual side. In fact, the same models could be used for both. Walls that reflect light in the virtual world also reflect sound waves. Therefore, both could be represented by the same triangular mesh. This is fine in theory, but fine levels of detail or spatial resolution do not matter as much for audio. Due to high visual acuity, geometric models designed for visual rendering may have a high level of detail. Recall from Section 5.4 that humans can distinguish 30 stripes or more per degree of viewing angle. In the case of sound waves, small structures are essentially invisible to sound. One recommendation is that the acoustic model needs to have a spatial resolution of only 0.5m [22]. Figure 11.13 shows an example. Thus, any small corrugations, door knobs, or other fine structures can be simplified away. It remains an open challenge to automatically convert a 3D model designed for visual rendering into one optimized for auditory rendering.

Now consider a sound source in the virtual environment. This could, for example, be a “magical” point that emits sound waves or a vibrating planar surface. The equivalent of white light is called *white noise*, which in theory contains equal

weight of all frequencies in the audible spectrum. Pure static from an analog TV or radio is an approximate example of this. In practical settings, the sound of interest has a high concentration among specific frequencies, rather than being uniformly distributed.

How does the sound interact with the surface? This is analogous to the shading problem from Section 7.1. In the case of light, diffuse and specular reflections occur with a dependency on color. In the case of sound, the same two possibilities exist, again with a dependency on the wavelength (or equivalently, the frequency). For a large, smooth, flat surface, a specular reflection of sound waves occurs, with the outgoing angle being equal to the incoming angle. The reflected sound usually has a different amplitude and phase. The amplitude may be decreased by a constant factor due to absorption of sound into the material. The factor usually depends on the wavelength (or frequency). The back of [22] contains coefficients of absorption, given with different frequencies, for many common materials.

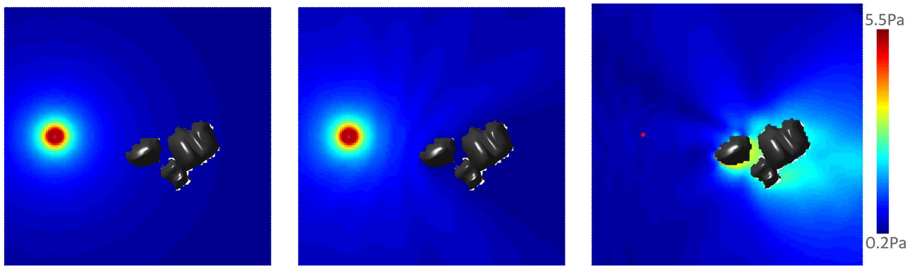
In the case of smaller objects, or surfaces with repeated structures, such as bricks or corrugations, the sound waves may scatter in a way that is difficult to characterize. This is similar to diffuse reflection of light, but the scattering pattern for sound may be hard to model and calculate. One unfortunate problem is that the scattering behavior depends on the wavelength. If the wavelength is much smaller or much larger than the size of the structure (entire object or corrugation), then the sound waves will mainly reflect. If the wavelength is close to the structure size, then significant, complicated scattering may occur.

At the extreme end of modeling burdens, a *bidirectional scattering distribution function* (*BSDF*) could be constructed. The BSDF could be estimated from equivalent materials in the real world by a combination of a speaker placed in different locations and a microphone array to measure the scattering in a particular direction. This might work well for flat materials that are large with respect to the wavelength, but it will still not handle the vast variety of complicated structures and patterns that can appear on a surface.

Capturing sound Sounds could also be captured in the real world using microphones and then brought into the physical world. For example, the matched zone might contain microphones that become speakers at the equivalent poses in the real world. As in the case of video capture, making a system that fully captures the sound field is challenging. Simple but effective techniques based on interpolation of sounds captured by multiple microphones are proposed in [12].

11.4.3 Auralization

Propagation of sound in the virtual world As in visual rendering, there are two main ways to handle the propagation of waves. The most expensive way is based on simulating the physics as accurately as possible, which involves computing numerical solutions to partial differential equations that precisely model



(a) (b) (c)

Figure 11.14: Computed results for sound propagation by numerically solving the Helmholtz wave equation (taken from [8]): (a) The pressure magnitude before obstacle interaction is considered. (b) The pressure after taking into account scattering. (c) The scattering component, which is the pressure from (b) minus the pressure from (a).

wave propagation. The cheaper way is to shoot visibility rays and characterize the dominant interactions between sound sources, surfaces, and ears. The choice between the two methods also depends on the particular setting; some systems involve both kinds of computations [8, 22]. If the waves are large relative to the objects in the environment, then numerical methods are preferred. In other words, the frequencies are low and the geometric models have a high level of detail. At higher frequencies or with larger, simpler models, visibility-based methods are preferable.

Numerical wave propagation The *Helmholtz wave equation* expresses constraints at every point in $\mathbb{R}^3$ in terms of partial derivatives of the pressure function. Its frequency-dependent form is

$$\nabla^2 p + \frac{\omega^2}{s^2} p = 0, \tag{11.8}$$

in which p is the sound pressure, ∇^2 is the Laplacian operator from calculus, and ω is related to the frequency f as $\omega = 2\pi f$.

Closed-form solutions to (11.8) do not exist, except in trivial cases. Therefore, numerical computations are performed by iteratively updating values over the space; a brief survey of methods in the context of auditory rendering appears in [8]. The wave equation is defined over the obstacle-free portion of the virtual world. The edge of this space becomes complicated, leading to *boundary conditions*. One or more parts of the boundary correspond to sound sources, which can be considered as vibrating objects or obstacles that force energy into the world. At these locations, the 0 in (11.8) is replaced by a *forcing function*. At the other boundaries, the wave may undergo some combination of absorption, reflection,

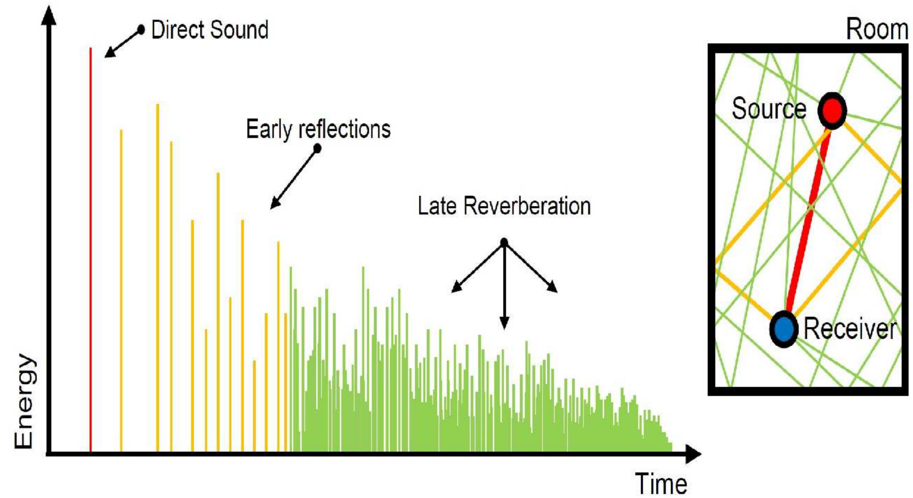


Figure 11.15: Reverberations. (From Pelzer, Aspock, Schroder, and Vorlander, 2014, [11])

scattering, and diffraction. These are extremely difficult to model; see [15] for details. In some rendering applications, these boundary interactions may simplified and handled with simple *Dirichlet boundary conditions* and *Neumann boundary conditions* [26]. If the virtual world is unbounded, then an additional *Sommerfield radiation condition* is needed. For detailed models and equations for sound propagation in a variety of settings, see [15]. An example of a numerically computed sound field is shown in Figure 11.14.

Visibility-based wave propagation The alternative to numerical computations, which gradually propagate the pressure numbers through the space, is visibility-based methods, which consider the paths of sound rays that emanate from the source and bounce between obstacles. The methods involve determining ray intersections with the geometric model primitives, which is analogous to ray tracing operations from Section 7.1.

It is insightful to look at the impulse response of a sound source in a virtual world. If the environment is considered as a linear filter, then the impulse response provides a complete characterization for any other sound signal [9, 11, 14]. Figure 11.15 shows the simple case of the impulse response for reflections in a rectangular room. Visibility-based methods are particularly good at simulating the reverberations, which are important to reproduce for perceptual reasons. More generally, visibility-based methods may consider rays that correspond to all of the cases of reflection, absorption, scattering, and diffraction. Due to the high computational cost of characterizing all rays, *stochastic ray tracing* offers a practical alternative by randomly sampling rays and their interactions with materials [22]; this falls under the general family of Monte Carlo methods, which are used, for example, to approximate solutions to high-dimensional integration and optimization problems.

Entering the ear Sound that is generated in the virtual world must be transmitted to each ear in the physical world. It is as if a virtual microphone positioned in the virtual world captures the simulated sound waves. These are then converted into audio output through a speaker that is positioned in front of the ear. Recall from Section 11.3 that humans are able to localize sound sources from auditory cues. How would this occur for VR if all of the sound emanates from a fixed speaker? The ILD and ITD cues could be simulated by ensuring that each ear receives the appropriate sound magnitude and phase to that differences in amplitude and timing are correct. This implies that the physical head must be reproduced at some level of detail in the virtual world so that these differences are correctly calculated. For example, the distance between the ears and size of the head may become important.

HRTFs This solution would still be insufficient to resolve ambiguity within the cone of confusion. Recall from Section 11.3 that the pinna shape distorts sounds in a direction-dependent way. To fully take into account the pinna and other parts of the head that may distort the incoming sound, the solution is to develop a *head-related transfer function* (*HRTF*). The idea is to treat this distortion as a linear filter, which can be characterized in terms of its transfer function (recall Figure 11.12). This is accomplished by placing a human subject into an anechoic chamber and placing sound sources at different locations in the space surrounding the head. At each location, an impulse is generated on a speaker, and the impulse response is recorded with a small microphone placed inside of the ear canal of a human or dummy. The locations are selected by incrementally varying the distance, azimuth, and elevation; recall the coordinates for localization from Figure 11.10. In many cases, a *far-field approximation* may be appropriate, in which case a large value is fixed for the distance. This results in an HRTF that depends on only the azimuth and elevation.

It is, of course, impractical to build an HRTF for every user. There is significant motivation to use a single HRTF that represents the "average listener"; however, the difficulty is that it might not be sufficient in some applications because it is not designed for individual users (see Section 6.3.2 of [22]). One compromise might be to offer a small selection of HRTFs to users, to account for variation among the population, but they may be incapable of picking the one most suitable for their particular pinnae and head. Another issue is that the transfer function may depend on factors that frequently change, such as wearing a hat, putting on a jacket with a hood or large collar, or getting a haircut. Recall that adaptation occurs throughout human perception and nearly all aspects of VR. If people adapt to frequent changes in the vicinity of their heads in the

real world, then perhaps they would also adapt to an HRTF that is not perfect. Significant research questions remain in this area.

Tracking issues The final challenge is to ensure that the physical and virtual ears align in the matched zone. If the user turns her head, then the sound should be adjusted accordingly. If the sound emanates from a fixed source, then it should be perceived as fixed while turning the head. This is another example of the perception of stationarity. Accordingly, tracking of the ear pose (position and orientation) is needed to determine the appropriate "viewpoint". This is equivalent to head tracking with simple position and orientation offsets for the right and left ears. As for vision, there are two choices. The head orientation alone may be tracked, with the full pose of each ear determined by a head model (recall Figure 9.8). Alternatively, the full head pose may be tracked, directly providing the pose of each ear through offset transforms. To optimize performance, user-specific parameters can provide a perfect match: The distance along the z axis from the eyes to the ears and the distance between ears. The latter is analogous to the IPD, the distance between pupils for the case of vision.

Further Reading

For mathematical and computational foundations of acoustics, see [15, 21]. Physiology and psychoacoustics are covered in [10, 27] and Chapters 4 and 5 of [7]. Localization is covered thoroughly in [2]. The cone of confusion is discussed in [19]. Echo thresholds are covered in [16, 25].

Some basic signal processing texts are [1, 6]. For an overview of auditory displays, see [23]. Convenient placement of audio sound sources from a psychophysical perspective is covered in [12]. Auditory rendering is covered in detail in the book [22]. Some key articles on auditory rendering are [4, 9, 11, 13, 14]

Bibliography

[1] A. Antoniou. *Digital Signal Processing: Signals, Systems, and Filters*. McGraw-Hill Education, Columbus, OH, 2005.

[2] J. Blauert. *Spatial Hearing: Psychophysics of Human Sound Localization*. MIT Press, Boston, MA, 1996.

[3] D. Deutsch, T. Hamaoui, and T. Henthorn. The glissando illusion and handedness. *Neuropsychologia*, 45:2981–2988, 2007.

[4] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings ACM Annual Conference on Computer Graphics and Interactive Techniques*, pages 21–32, 1998.

[5] D. O. Kim, C. E. Molnar, and J. W. Matthews. Cochlear mechanics: Nonlinear behaviour in two-tone responses as reflected in cochlear-new-fibre responses and in ear-canal sound pressure. *Journal of the Acoustical Society of America*, 67(5):1704–1721, 1980.

[6] R. G. Lyons. *Understanding Digital Signal Processing, 3rd Ed.* Prentice-Hall, Englewood Cliffs, NJ, 2010.

[7] G. Mather. *Foundations of Sensation and Perception*. Psychology Press, Hove, UK, 2008.

[8] R. Mehra, N. Raghuvanshi, L. Antani, A. Chandak, S. Curtis, and D. Manocha. Wave-based sound propagation in large open scenes using an equivalent source formulation. *ACM Transactions on Graphics*, 32(2), 2013.

[9] J. Merimaa and V. Pulkki. Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, 2005.

[10] B. Moore. *An Introduction to the Psychology of Hearing, 6th Ed.* Brill, Somerville, MA, 2012.

[11] Sönke Pelzer, Lukas Aspöck, Dirk Schröder, and Michael Vorländer. Integrating real-time room acoustics simulation into a cad modeling software to enhance the architectural design process. *Buildings*, 2:1103–1138, 2014.

[12] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.

[13] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[14] V. Pulkki and J. Merimaa. Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests. *Journal of the Audio Engineering Society*, 54(1/2):3–20, 2006.

[15] S. W. Rienstra and A. Hirschberg. *An Introduction to Acoustics*. Endhoven University of Technology, 2016. Available at http://www.win.tue.nl/∼sjoerdr/papers/boek.pdf.

[16] P. Robinson, A. Walther, C. Faller, and J. Braasch. Echo thresholds for reflections from acoustically diffusive architectural surfaces. *Journal of the Acoustical Society of America*, 134(4):2755–2764, 2013.

[17] M. B. Sachs and N. Y. S. Kiang. Two-tone inhibition in auditory nerve fibres. *Journal of the Acoustical Society of America*, 43:1120–1128, 1968.

[18] R. N. Shepard. Circularity in judgements of relative pitch. *Journal of the Acoustical Society of America*, 36(12):2346–2453, 1964.

[19] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *Journal of the Acoustical Society of America*, 107(3):1627–1636, 2002.

[20] S. S. Stevenson. On the psychophysical law. *Psychological Review*, 64(3):153–181, 1957.

[21] L. L. Thompson and P. M. Pinsky. Acoustics. *Encyclopedia of Computational Mechanics*, 2(22), 2004.

[22] M. Vorländer. *Auralization.* Springer-Verlag, Berlin, 2010.

[23] M. Vorländer and B. Shinn-Cunningham. Virtual auditory displays. In K. S. Hale and K. M. Stanney, editors, *Handbook of Virtual Environments, 2nd Edition.* CRC Press, Boca Raton, FL, 2015.

[24] R. M. Warren, J. M. Wrightson, and J. Puretz. Illusory continuity of tonal and infratonal periodic sounds. *Journal of the Acoustical Society of America*, 84(4):1338–1142, 1964.

[25] X. Yang and W. Grantham. Effects of center frequency and bandwidth on echo threshold and buildup of echo suppression. *Journal of the Acoustical Society of America*, 95(5):2917, 1994.

[26] H. Yeh, R. Mehra, Z. Ren, L. Antani, M. C. Lin, and D. Manocha. Wave-ray coupling for interactive sound propagation in large complex scenes. *ACM Transactions on Graphics*, 32(6), 2013.

[27] W. A. Yost. *Fundamentals of Hearing: An Introduction, 5th Ed.* Emerald Group, Somerville, MA, 2006.

[28] P. Zahorik. Assessing auditory distance perception using virtual acoustics. *Journal of the Acoustical Society of America*, 111(4):1832–1846, 2002.