Chapter 11

Audio

Steven M. LaValle
University of Illinois

Copyright Steven M. LaValle 2015, 2016

Available for downloading at http://vr.cs.uiuc.edu/
Chapter 11
Audio

<table>
<thead>
<tr>
<th>Chapter Status</th>
<th>Taken from Virtual Reality, S. M. LaValle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>This online chapter is not the final version! Check <a href="http://vr.cs.uiuc.edu/">http://vr.cs.uiuc.edu/</a> for information on the latest draft version.</td>
</tr>
<tr>
<td></td>
<td>This draft was compiled on June 27, 2016.</td>
</tr>
</tbody>
</table>

Hearing is an important sense and has been unfortunately neglected up until this chapter. VR developers tend to focus mainly on the vision part because it is our strongest sense; however, the auditory component of VR is powerful and important as well. By itself, 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 set version (wearing headphones). When combined consistently the visual component, audio becomes crucial to providing a compelling and comfortable VR experience.

Each section of this chapter is the auditory 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 are fascinating.

11.1 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 large 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 wave front.

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 differs from light. 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. This could be the pressure variation on a human eardrum, which is converted into a perceptual experience. The sound pressure level frequently reported in decibels (abbreviated as dB), which is defined as

\[ N_{db} = 20 \times \log_{10}(P_t/P_r), \]

in which \( P_t \) is the pressure level of the peak compression and \( P_r \) is a reference pressure level, which is usually taken as \( 2 \times 10^{-5} \) 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.
forcing them to vibrate. This enables talking, singing, screaming, and so on.

### Sound sources and attenuation

As in the case of light, we can consider sound wavefronts and rays, in which each ray is perpendicular to the wavefront. A point sound source can be considered, which 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 increases 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. Human senses and engineered sensors easily measure differences in arrival times of sound waves, leading to 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 \( s = 343.2 \) m/s, the wavelength of a sound wave can also be determined. At 20 Hz, the wavelength is \( \lambda = \frac{343.2}{20} = 17.1 \text{m} \). At 20,000 Hz, it becomes \( \lambda = 17.1 \text{mm} \). These are much longer than for light!

### Doppler effect

Recall from Figure ?? how the sound pressure variation is obtained for a fixed receiving distance. If the point is moving away from the source, then the wavefronts will arrive at a reduced frequency. If the receiver moves at 42.2 m/s away from the source, then the waves would seem to be traveling at only \( 343.2 \times 0.422 \approx 300 \text{m/s} \). The received frequency shifts due to the relative motion between the source and receive. This is known as the Doppler effect, and the frequency as measured at the receiver can be calculated as

\[
    f_r = \left(1 + \frac{v_r}{s + v_s}\right) f_s,
\]

in which \( s \) is the 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 \( f_s = 1000 \text{Hz} \) would be perceived by the receiver as having 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 we want an experience with time dilation, space travel, and so on).

### Interactions between media

As with light, wave propagation is strongly affected 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 propagate more quickly through most 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. The amount of sound that is diffracted depends on the wavelength relative to the length of the corner. More diffraction occurs for a longer wavelength; thus, a lower-pitched sound bends around 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 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, is considered to be a combination of pure tones of various amplitudes, frequencies, and time shifts. Figure 11.3(b) to 11.3(d) provides an 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 amplitude or scaling factor, and a possible timing for its peak, which is called its phase shift. By simply adding up a finite
11.1. PHYSICS OF SOUND

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.

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; an efficient algorithm for computing it is called the FFT, for Fast Fourier Transform. Thus, fast numerical methods exist for converting an audio signal into a spectral decomposition. Likewise, a spectral decomposition can be quickly converted back into an audio signal.

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 make our heads too large. 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 amplitudes and phase shifts. As will be seen shortly, each ear is more like a single-pixel camera operating at tens of thousands of FPS, rather than capturing a large image at a slower frame rate. The interesting information for sound is distributed time, whereas it is mostly distributed over space for vision.

11.2 The Physiology of Human Hearing

The anatomy of the human ear is shown in Figure 11.4. Recall from Section 5.3 the complications of eye movements. Although cats and some other animals can rotate their ears, humans cannot, which simplifies some of the VR engineering problems. The ear is divided into outer, middle, and inner parts, based on the flow of waves.
11.2. THE PHYSIOLOGY OF HUMAN HEARING

Figure 11.4: The physiology of the human auditory system.

**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. 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 from the middle ear. Its covers only 55mm² of area. It 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 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 from 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 maleus (hammer), incus (anvil), and stapes (stirrup), and they 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 much amplitude as possible. This avoids the tendency of a higher impedance material to reflect the sound away, as in voices reflecting over the surface of a lake.

**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. It accomplishes this in a beautiful way that performs a spectral decomposition in the process so that the neural impulses encode amplitudes of frequency components.

Figure 11.5 illustrates its operation. As see 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
11.2. THE PHYSIOLOGY OF HUMAN HEARING

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.)

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 become sensitive to a 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 mechanoreceptors convert displacements of hairs into neural impulses. The hairs are displaced as the basilar membrane vibrates because the ends 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 lower 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 [1]. If pure tones at two different frequencies are presented to the ear, the basilar membrane produces a third tone, which is sometimes audible [?]. Also, the neural impulses that result from mechanoreceptor output are not linearly proportional to the frequency amplitude. Furthermore, the detection one tone may cause detections of nearby tones (in terms of frequency) to be inhibited [?], much like lateral inhibition in horizontal cells (recall from Section 5.2). Using concepts that will be defined and explained in Section 11.4.1, the result is that the human auditory processing is considered to be a nonlinear filter because it does not perform a perfect Fourier decomposition. Otherwise, it would be equivalent to a linear filter.

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), must 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 limited data, assumptions implied by neural structures, adaptation, and many other factors. The same is true for auditory experiences. In fact, auditory illusions exist as well. The McGurk effect from Section 6.4 was an example that used vision to induce an incorrect auditory perception.
11.3. AUDITORY PERCEPTION

S. M. LaValle: Virtual Reality

same loudness? This requires careful design of experiments with human subjects, a problem that is common throughout VR development as well; see Section 12.3.

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 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) as a function of frequency. In other words, for a pure tone at 1000 Hz, could you 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. 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 [1].

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 virtual people. 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.
11.3. AUDITORY PERCEPTION

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 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 examples will now be given: 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 we are usually not consciously aware of this distortion, our auditory complex 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 makes a higher-pitched popping sound. 4) Finally, a powerful monaural cue is provide 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. It is also used by some people 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 first 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 source. Note that sound travels 21.5cm in about 0.6ms, which means that surprisingly small differences are used for localization.

Suppose the brain measures that the difference in arrival times is 0.3ms. What is the set of possible 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 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...
11.4. AUDITORY RENDERING

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 hyperboloid, but approximately looks like a cone.

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 cue as 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, 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 issues of self motion versus the motion of objects emerge based on the auditory input.

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 [8] 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 throughout 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 numerically. 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 \( \Delta t \) be the sampling interval. The sampling rate or (sampling frequency) is \( 1/\Delta t \) Hz. For example, with 1000 Hz sampling frequency, Deltat 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

\(^1\)The values are also discretized, and are represented using floating-point numbers. This level of detail will be ignored.
11.4. AUDITORY RENDERING

can refer to the \( k \)th 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; see Figure ???. A linear filter is a special a special kind of filter that satisfies two algebraic properties. If \( x \) represents an entire signal, then let \( F(x) \) represent the result after running it through the filter.

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 through 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 \( c \). For example, this means that if we double the sound amplitude, the output sound from the filter doubles its amplitude as well.

A linear filter generally takes the form

\[
y[k] = c_0 x[k] + c_1 x[k - 1] + c_2 x[k - 2] + c_3 x[k - 3] + \cdots + c_n x[k - n],
\]

(11.3)
in which each \( c_i \) is a constant, and \( n + 1 \) is the number of samples involved in the filter. 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:

\[
y[k] = \frac{1}{3} x[k] + \frac{1}{3} x[k - 1] + \frac{1}{3} x[k - 2].
\]

(11.4)

This one is an example of an exponential smoothing (also called exponentially weighted moving average):

\[
y[k] = \frac{1}{2} x[k] + \frac{1}{4} x[k - 1] + \frac{1}{8} x[k - 2] + \frac{1}{16} x[k - 3].
\]

(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 \( (FIR) \). 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 \( k[x] = 0 \) 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.

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

(11.6)

GIVE TRANSFORM. continuous? Provide DFT. Explain. Inverse? FFT?

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 the linear filter by applying the Laplace transform to the finite impulse response []).

11.4.2 Acoustic modeling

The geometric modeling concepts from Section ?? 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 in audio. Due to high visual acuity, geometric models designed for visual rendering may have a high level of
11.4. AUDITORY RENDERING

(a) (b)

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

detail. Recall from Section 5.4 that humans can distinguish 60 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 [?]. 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 old analog TV sounds like this. In practical settings, the sound 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 magnitude and phase. The magnitude 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 [?] 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 diffraction pattern for sound may be hard to 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 be a combination of 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 discussed in [?].

11.4.3 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 wave propagation. The cheaper way is to shoot visibility rays and characterize the dominant interactions between sound sources, surfaces, and ears. Both cases will be discussed here.

Numerical wave propagation Numerical: PDEs, FEM, Helmholtz wave equation, expensive; still need surface scattering models.

Spatial impulse response rendering.

Visibility-based wave propagation Combinatorial: Visibility, ray tracing, simple reflections, fast but sloppy

11.4.4 Auralization (rendering)

This is analogous to shading. Need to figure out what should go into each ear.

Account for scattering due to pinna, head, hair, body?

Solution: HRTF

Give basic math. Figure showing directions of sources and head.

Artificial heads

HRTF + movement

4. Display
11.4. AUDITORY RENDERING

Tracking issues If the user turns his head, 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. Therefore, 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 later is analogous to the IPD, the distance between pupils for the case of vision.

A bit about displays (a.k.a. speakers)

Further Reading

Virtual Auditory Displays, M. Vorlander, B. Shinn-Cunningham, 2015, Handbook of Virtual Environments, 2nd Ed.

Give references for Fourier analysis, FFTs.

Yost, Fundamentals of Hearing, 2000; see Mather Chapter 4, 5.

A Beam Tracing Approach to Acoustic Modeling for Interactive Virtual Environments, Thomas Funkhouser et al., 1998.

The room acoustic rendering equation S Siltanen, T Lokki, S Kiminki, L Savioja The Journal of the Acoustical Society of America 122 (3), 1624-1635


Effects of center frequency and bandwidth on echo threshold and buildup of echo suppression, Xuefeng Yang and D. Wesley Grantham


Monaural cues Zahorik, 2002.


Integrating Real-Time Room Acoustics Simulation into a CAD Modeling Software to Enhance the Architectural Design Process, Sönke Pelzer, Lukas Aspöck, Dirk Schröder, and Michael Vorländer, Buildings 2014, 2, 113-138

Merimaa, Spatial Impulse Response Rendering

Virtual sound source positioning using vector base amplitude panning, V Pulkki, Journal of the Audio Engineering Society 45 (6), 456-466

Spatial sound reproduction with directional audio coding, V Pulkki, Journal of the Audio Engineering Society 55 (6), 503-516
Bibliography