Part H: Multidimensional Waves

Chapter 151. Drawing 3D Waves

To represent waves passing through a three-dimensional medium requires even more abstraction than two dimensions. Time and motion can still be shown by adding motion arrows to a snapshot. The full detail of how the disturbance varies in one cycle can be discarded in favor of showing only wave crests. But even so, each wave crest of a 3D wave is a surface passing through space, which is difficult to represent on paper.

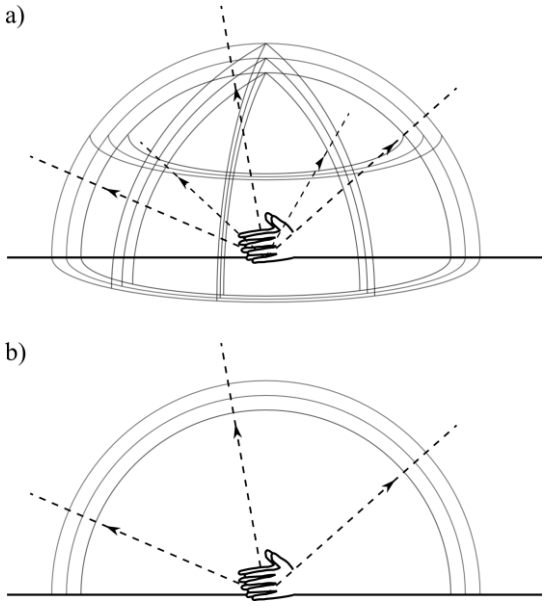

Figure 151.1

Three-dimensional wave emanating from point source, a three-crest wave packet. Dashed lines are rays. (a) 3D nature of the wave. (b) a plane through 3D space.

Sometimes it is important to draw the full shape of the wave crests, in a picture that is closer to a photograph than a graph. Figure 151.1(a) attempts to illustrate a wave packet from a point source with three wave crests, similar to the water ripple in Section 125a. With a point source and a uniform wave speed throughout the medium, each wave crest (that is, each local maximum of disturbance) moves away from the source as a section of a sphere. This is called a **spherical wave,** even when something blocks the path so that each wave crest is only part of a sphere. This can be quite difficult to draw. The wave crests must be shown as transparent, so that multiple crests can be seen. It is particularly difficult to draw rays that appear to be coming towards or going away from the viewer.

You might notice that the spherical wave crests look very similar to the shell of sound energy leaving the source in Chapter 53. Indeed, it is the disturbance of the medium, which is represented by the wave crests, which caries that energy. But the crests show additional information. For instance, for a periodic wave, the crests are spaced (along the rays) by one wavelength.

Because 3D drawings are difficult to make and interpret, the simpler **cross section** shown in Figure 151.1(b) is often used. Here we imagine a plane slicing through space, almost always one that includes the source. The drawing is then essentially the same as what is used for a 2D wave. It is up to the viewer to remember that the picture is only showing part of the situation. For instance, that the true wave crests are parts of spheres, not parts of circles.

The other especially important model for 3D waves is the **plane wave,** named after the fact that the wave crests have the shape of planes. Figure 151.2 shows an example, both with a 3D picture and in cross section.

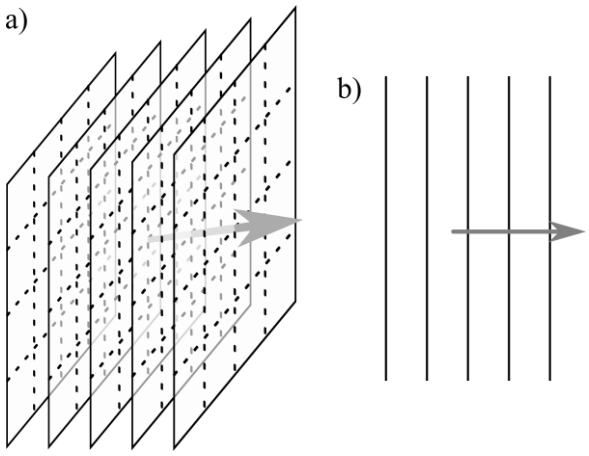

Figure 151.2

Three dimensional plane wave. (a) In perspective. (b) Cross section representation.

For three-dimensional waves, it remains true that rays are always perpendicular to the wave crests. Since the crests are surfaces, the exact meaning of this is a little different from the 2D case. But still, in cross-sectional diagrams like Figure 151.1(b) and Figure 151.2(b), the rays always cross the wave crest lines at right angles.

Since the rays are always perpendicular to the wave crests, the rays for a plane wave are all parallel to each other. The nice thing about plane waves is that the disturbance varies only along one direction (along the rays), and not in the direction of the other two dimensions. This means that plane waves are the 3D waves that are closest to a 1D wave, which also varies only along one direction. Since the energy of a plane wave is not spreading over a larger area as the wave moves, a plane wave can be truly periodic.

Chapter 8 says that very far from a point source, it can be a good model to say that the rays are parallel. We can now add some details to that model. The wave crests from a point source are technically curved, part of a sphere. But if the bit of wave that we are focused on is much smaller than the distance to the source, then the curvature of the wave crests is small enough to ignore, and we can model the wave as a plane wave.

Chapter 152. Doppler Effect

152a. Doppler Description

Once a sound source has launched a compression into the air, that compression moves through the air at the speed of sound, unaffected by any motion that the source may make afterwards. That is the underlying reason for the well-known Doppler effect, by which the sirens of passing emergency vehicles appear to change pitch as they go by. The Doppler effect actually encompasses two separate effects, which are usually combined into a single equation.

When Doppler proposed these formulae,[43,44] he was attempting to explain the color of light from binary stars. Ironically, his equations correctly describe all waves in a material medium, but ***not*** light! (His theory about the star colors was completely wrong as well, for completely different reasons.) The correct formula for the frequency shift of light would not be determined until Albert Einstein worked out the special theory of relativity. Nevertheless, it is called the relativistic Doppler effect. It is indeed incredibly important to modern astronomy, but the original Doppler equation applies for sound and all other waves besides light.

If a source is producing a sound with a particular frequency, then the **proper frequency** f is the frequency an observer would measure if neither the source nor the observer were moving. The proper frequency is a characteristic of the source itself; if the source is vibrating, the proper frequency is the frequency of its vibration. But if either the source or observer are moving, then the **perceived frequency** f' that the observer detects might be different. In fact, different observers might measure different perceived frequencies from the same source.

Specifically, it is the motion *relative to the medium* that's important. If a wind is blowing from your friend towards you, then even if you are both standing still on the ground, you are moving through the air towards your friend. When the medium is flowing, this can be confusing. It may help to think about which way each person would have to turn to feel the wind in their face.

1. When the source is moving through the medium *towards* the observer, the perceived frequency tends to be shifted up, higher than the proper frequency. The same consequence occurs when the observer is moving through the medium *towards* the source.

[43] Alec Eden, *The Search for Christian Doppler* (Wien: Springer-Verlag, 1992).
[44] Christian Doppler, *Ueber das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels* (Prague: Borrocsh und André, 1842).

2. When the source is moving through the medium *away from* the observer, or the observer is moving through the medium *away from* the source, the perceived frequency tends to be shifted down, lower than the proper frequency.
3. If all motions are ***directly*** towards or away from the other objects, then the result incorporating both source motion and observer motion can be calculated from the equation

$$f' = f\frac{s \pm s_o}{s \pm s_s} \quad . \tag{152.1}$$

The variables s, s_o, s_s are the speeds of sound, of the observer, and of the source respectively. The $\pm$ indicate to choose between $+$ or $-$ in accordance with the first two rules. The $\pm$ in the numerator is chosen based on the observer motion, and the $\pm$ in the denominator is chosen based on the source motion.

If the source or observer are moving obliquely, not directly towards or away from the other, then Eq. 152.1 isn't correct, but the qualitative rules 1 & 2 still work. The net frequency change $\Delta f = f' - f$ is called the **Doppler shift.**

Choosing the $\pm$ can be a bit tricky, especially for the source. For a moving observer, choosing the $+$ will shift the frequency up, which is natural enough. But if the source is moving towards the observer, we must choose the $-$ in the denominator in order to shift the frequency up.

In applying these rules, the motion of the source and the motion of the observer are considered separately, each relative to the medium. However, comparing the motion of the source and observer relative to each other also has significance. Suppose that two vehicles are traveling in the same direction at the same speed s_v. The rearward one is the source (perhaps an ambulance) and the forward one is the observer. Then the Doppler equation gives

$$f' = f\frac{s - s_v}{s - s_v} = f \quad . \tag{152.2}$$

The observer is moving away (top of fraction, shift down, choose $-$) and the source is moving towards (bottom of fraction, shift up, choose $-$), but the combined effect is to have no shift at all. In general, the ***overall*** shift will be up if the source and observer are getting closer together, and the ***overall*** shift will be down if they are separating.

152b. Extra: Doppler Reflection

One situation that can comes up is a source that is also the observer, with the sound reflecting off something to return from where it started. Application of the Doppler effect to measure speeds relies on this configuration. Either the source/observer or the reflecting object might be moving. The final equation for this situation is a little insidious: it looks just like Eq. 152.1, but it is hard to know which speeds to use, because the source and observer seem to be the same.

The key here is to treat the situation as two Doppler effects back-to-back. In the first Doppler effect, the sound origin emits a frequency f, and the reflecting object acts as the observer, receiving the sound at shifted frequency f'. In the second Doppler effect, the reflecting object becomes the source, sending the wave back out with the same frequency f'. The original source now becomes the observer of the final perceived frequency f'', shifted from f' by the same motions that caused the first shift, but now playing opposite roles.

For example, suppose a stationary Doppler detector is pointed at a baseball that is coming towards the detector. First the detector acts as the source of a sound, and the baseball "observes" that sound shifted up to frequency f'. The reflection of the sound then turns the baseball into a source, and the detector finally plays the part of observer when the reflected sound returns. In equations, using s_b for the speed of the baseball, this looks like

$$f' = f\frac{s + s_b}{s} \quad , \quad f'' = f'\frac{s}{s - s_b} \quad . \tag{152.3}$$

Simplify that by combining the equations, and the result looks very much like Eq. 152.1, with the baseball seeming to play the role of both source and observer. That is true enough, but the detector is also acting as both source and observer. If the detector is moving as well, one had better work with separate equations for f' and f''.

Chapter 153. Doppler Effect Derivation

Although care is required with the details, Eq. 152.1 for the Doppler effect can be derived using only the most basic ideas from this textbook.

Figure 153.1 shows a sound source (the black dot) with some specific frequency. The source emits a wave crest every period $T = 1/f$ so that the crests are one wavelength λ apart. These characteristics of the source are called the **proper period** and **proper frequency.** At the instant of the picture, the source is just about to emit another crest.

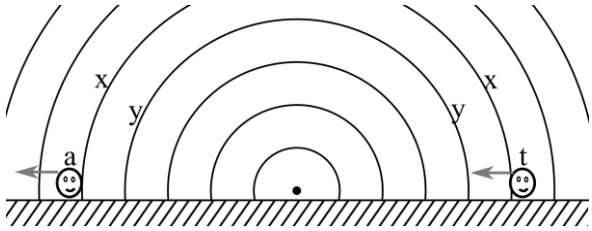

Figure 153.1

Observers moving near a sound source encounter the wave crests at different rates.

Observer t, who is moving **(t)**owards the source at a speed s_t, encounters the wave crests more frequently, and therefore perceives the sound to have a shorter **perceived period** T_t'. To understand this quantitatively requires only Eq. 5.2, along with its special application to waves Eq. 121.2. In Figure 153.1, observer t is just encountering wave crest x. The time for him to experience one cycle will equal the time for wave crest y and the observer to come together. The combined motion of observer and wave crest covers a total distance of one wavelength, which is given by

$$\lambda = sT_t' + s_t T_t' \quad , \tag{153.1}$$

where s is the speed of sound. We are usually most interested in relationships of frequencies. Using the variable f_t' for the **perceived frequency** and using Eq. 121.2 to introduce the proper frequency transforms the equation into

$$\frac{s}{f} = \lambda = (s + s_t)T_t' = (s + s_t)\frac{1}{f_t'} \quad , \tag{153.2}$$

$$f_t' = f\frac{s + s_t}{s} \quad , \tag{153.3}$$

On the other hand, observer a is moving **(a)**way from the source at speed s_a. She is also just receiving wave crest x in the figure. For crest y to catch her, the sound will have to travel one wavelength *plus* the distance she travels in that time,

$$sT_a' = \lambda + s_a T_a' \quad , \tag{153.4}$$

$$f_a' = f\frac{s - s_a}{s} \quad , \tag{153.5}$$

A different situation is shown in Figure 153.2, where it is now the source (still the black dot) that is moving at speed s_s while the observers are stationary. Once again, at the moment of the picture the source is just about to emit another crest. The open circles show where the source was exactly one, two, three, and four

periods in the past, and each of them is the center of their matching wave crest. The dashed wave crests were emitted earlier, at times when the source was at positions that are not shown.

The result of the moving source is that the wave crests are crowded together in front of the source and spread apart behind it (compared to the spacing in Figure 153.1). Unlike Figure 153.1, the actual wave in the air is modified by the motion. But still, the result is that the observers perceive a frequency different from the proper frequency.

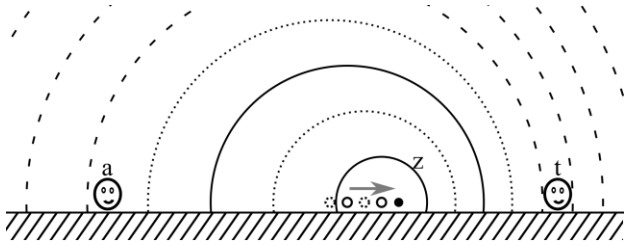

Figure 153.2

Moving source emits wave crests centered at different points. Solid and dotted crests came from corresponding source locations. Dashed crests were emitted earlier, source positions not shown.

This time observer t, **(t)**owards whom the source is moving, experiences a sound that has a shortened wavelength λ_t'. It is equal to the distance between the current source position and wave crest z, which equals [the distance crest z covered during the most recent period] minus [the distance that the source covered, from the previous open circle]. This comes together to the equation

$$\lambda_t' = sT - s_s T \quad . \tag{153.6}$$

The observer doesn't know that the wavelength is compressed, so we can use Eq. 121.2 to find the perceived frequency as follows, still using f for the proper frequency,

$$\frac{s}{f_t'} = \lambda_t' = (s - s_s)T = (s - s_s)\frac{1}{f} \quad , \tag{153.7}$$

$$f_t' = f\frac{s}{s - s_s} \quad . \tag{153.8}$$

Observer a also perceives a modified frequency and wavelength because the source is moving **(a)**way from her. Mathematically, this changes all the subtractions to additions in Eqs. 153.6–153.8.

Putting Eqs. 153.3, 153.5, and 153.8 all together creates Eq. 152.1, which handles all those scenarios. The $\pm$ choices account for the difference between motion towards or away. The observer and source, the algebra tells us, effect the numerator and the denominator respectively. Similar reasoning can be used to show that Eq. 152.1 can even handle situations when both the source and observer moving simultaneously.

Chapter 154. Sonic Boom and Shock Waves

In each cycle, a wave source emits a certain amount of energy, which we can associate with the wave crests in the pictures. When a wave source is moving, such as in Figure 153.2, the bunching up of the wavefronts means that the wave energy in that area is more densely packed as well. This is not what is usually meant by the Doppler effect, which specifically refers to the changes in perceived frequency. However, the two do go hand-in-hand. An observer in front of a moving sound source will hear a higher intensity sound than if the source were the same distance away and stationary. Behind the source, the intensity will be lower.

Equation 152.1 offers a chance to get into mathematical trouble. What if a sound source is moving towards you at the speed of sound? Following the rules yields an equation that divides by zero. This isn't a mistake. It's the result of trying to apply the equation outside the realm where it applies.

But we can associate that divide-by-zero with a physical event. Figure 154.1(a) shows the wave crests from a source that is moving at the same speed as the wave speed in the medium. Because the wave crests can never get away from the source, they are piling up at the source. It no longer makes sense to speak of the frequency there, on the right side of the figure, because there is no repeating wave. However, there is energy associated with all those wave crests. For sound waves, this is the origin of the **sonic boom** that occurs when high-speed aircraft exceed the speed of sound.

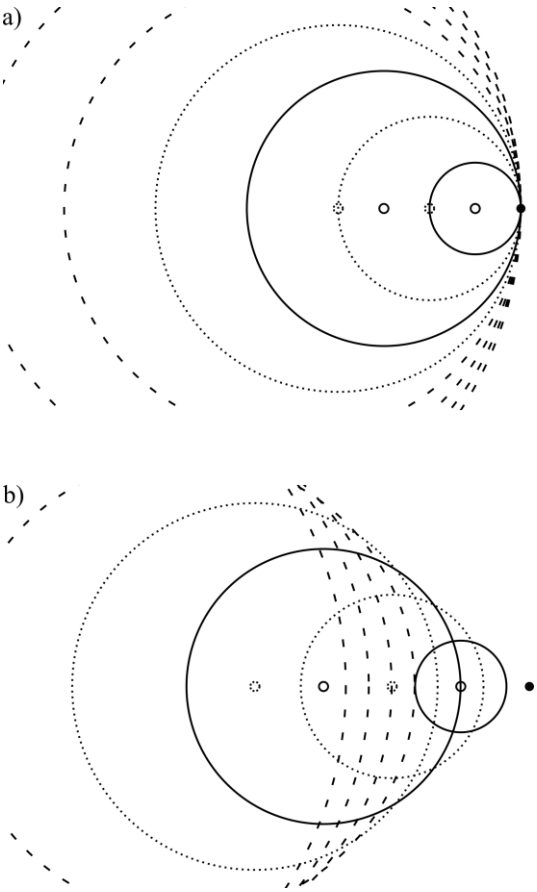

Figure 154.1

Wave crests from source moving (a) at the wave speed and (b) at 1.5 times the wave speed. Solid and dotted crests came from corresponding source locations. Dashed crests were emitted earlier, source positions not shown.

If the source moves faster than the speed of sound, then we get a picture like Figure 154.1(b). The source leaves behind a triangular "wake"—it looks similar to the wake behind a boat moving through water, although they are not really the same thing.[45] Along the front edge, the medium suddenly changes from having no wave to having a rather large disturbance. This is a **shock wave.** If this is a 3D wave, such as sound in air, then remember that each circle in our picture represents a sphere. The front edge thus makes the shape of a cone, called the **Mach cone** after physicist Ernst Waldfried Josef Wenzel Mach, who introduced the idea in 1887.

Chapter 155. Superposition in 2D and 3D

Nearly everything from Chapter 138, concerning the combination of waves in a one-dimensional medium, is also true in two and three dimensions. In the majority of cases, including all but absurdly loud sounds,

[45] Beyond this book: Wakes behind boats, called Kelvin wakes, are often confused with Mach cones because of the superficial resemblance. But they are very different and far more complicated. Kelvin wakes do not involve a periodic source, they occur at any boat speed, they have an opening angle nearly independent of boat speed, and they have a complicated internal structure.

the medium is linear. One consequence is that two waves pass through each other, without either one being altered. Also, when two waves overlap, the disturbance of the medium obeys superposition, being the sum of the disturbances from the individual waves.

But in multiple dimensions, two waves can move in much more varied ways than just comoving or countermoving. It also becomes much more difficult to represent the combined waves. Our wave crest pictures are best suited to simple waves, especially sinusoidal ones, while the very nature of wave superposition is to create complex disturbance patterns.

Even superpositions of relatively simple waves rapidly exceed our ability to pictorially represent them. Figure 155.1 shows a snapshot of 2D wave crests (perhaps a surface wave on water) for the superposition of a plane wave with a circular wave with the same wavelength. Part (a) is in grayscale, from white (highest) to black (lowest). Although the circular and plane contributions are somewhat visible, the full pattern is hard to describe. Figure 155.1(b) shows an attempt to represent the same wave with wave crests. But even the meaning of "wave crest" becomes unclear. For example, the gray-filled circles represent spots where the disturbance is at an isolated maximum, more like a mountain peak than a ridge. Is that a wave crest? The dotted segments show "crests" where the water height is higher than on either side of the dotted line, but along the dotted line the water surface is higher at the ends than in the middle, so those positions are not local maxima of disturbance.

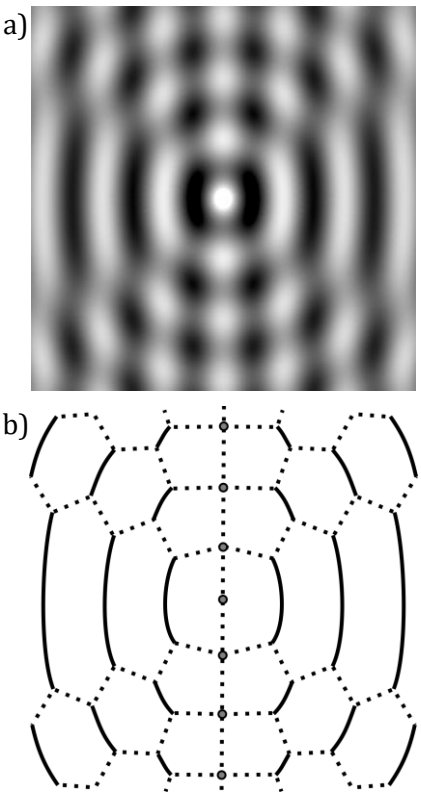

Figure 155.1

Superposition of a plane wave (moving left to right) and a point source at the center. (a) Grayscale representation. (b) Wave crests.

If all this weren't complicated enough, remember that this is only a snapshot. A movie showing how the surface changes in time would defy any simple description. In many places, it would be impossible to define rays.

But the primary usefulness of superposition is not that it allows us to combine waves. The real power of superposition is that it allows us to ***not*** combine waves. When faced with a situation involving several simple waves that mix together, we can take a different approach. Secure in the knowledge that we *could* combine them if necessary, each of the simple waves can be handled separately as it interacts with its environment. If the wave in Figure 155.1 were to do something, like reflect off a boundary, we can separately figure out where and how each simpler wave travels. Thanks to linearity and superposition, we can choose to leave the combining of the waves as the very last step, if indeed it is needed at all.

Chapter 156. Interference in 2D and 3D

Whenever two sinusoidal wave sources of the same frequency emit two waves that eventually meet somewhere, the phase relationship where the waves meet depends not only on the phase relationship between the sources, but also on a comparison of how far the two waves traveled. This was the alternative view described in Chapter 141, where two sounds were traveling along the same line. In this chapter, we apply the same idea when the sounds travel different paths from the two sources to the meeting point. To keep things from getting too complicated, we'll look at cases where the two sources are in phase with each

other, and where all wave travel is through the same medium so that the wavelength λ is the same everywhere.

Figure 156.1 illustrates a piece of this case, in a rather abstract way. The sources, A and B, are point sources, but the figure only shows the wave along a narrow path from the source. The waves from the two sources are mostly shown as if the other source did not exist; only in part (b) near point P is the superposition shown. And since the sources are point sources, the wave amplitude should decrease with distance traveled, but Figure 156.1 ignores that as well.

Part (a) represents the waves as graphed oscillations, plotted on a third axis in a 3D drawing. You could think of this as showing water surface waves, with the curves tracing the height of the water along a particular path. Or, more abstractly, the curves could be graphs of any sort of disturbance (for example, the density of air), so that the vertical axis (parallel to the source double-arrows) represents the disturbance and the two other dimensions represent spatial distances.

Part (b) represents the same waves with a grayscale. You could imagine this to show air density, with dark showing compression and light showing rarefaction. Or, more abstractly, the grayscale can represent any sort of disturbance (for example, water surface height). The graphs in part (a) represent a line down the center of the rectangles in part (b). Because the rectangles in part (b) show more than a single line through the medium, near the sources it is possible to see that the sources are point sources making circular waves.

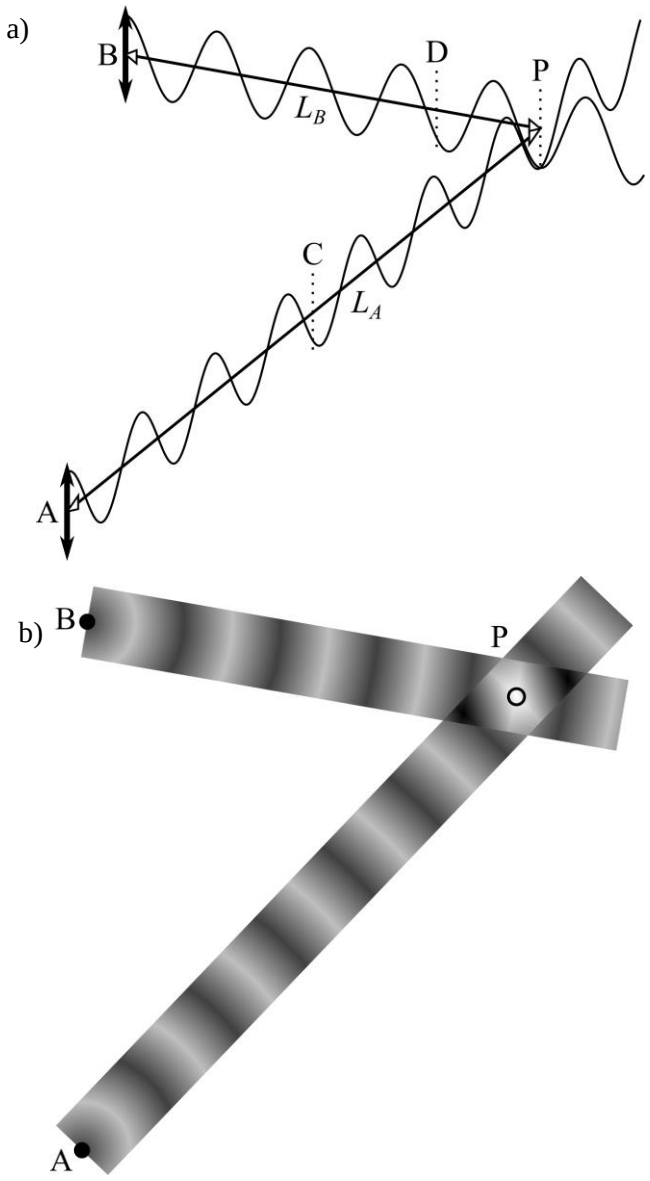

Figure 156.1

Two in-phase sources emit waves of same wavelength, meeting at point P. (a) Perspective 3D drawing, representing waves as disturbance graphs. (b) Disturbance shown with grayscale, medium gray for equilibrium state and white and black as disturbances away from equilibrium. Both parts ignore amplitude reduction that normally occurs as waves spread from sources.

Figure 156.1(a) is best suited for showing the phase relationships between the individual contributing waves. At the instant in time illustrated, both of the sources A and B (and the waves at the sources) are at reduced phase 0°. At P, where the two waves meet, both are at reduced phase 180°. Those particular numbers aren't very important; as time advances, the waves will move away from their sources, and at all locations the phases will increase. But the phases at the sources will always equal each other. Similarly, the wave phases at P will always match — the waves are in phase at point P. Figure 156.1(b) highlights the way that the waves combine. Because they are in phase at P, they reinforce each other, resulting in stronger blacks and whites.

What is special about position P, that puts the waves in phase there? P is $L_A = 6.5\lambda$ away from source A, and $L_B = 4.5\lambda$ away from source B. But once again those particular numbers aren't so important; the key is that wave A had to travel farther by $L = L_A - L_B = 2\lambda$, exactly two cycles extra.

The situation is essentially the same as for the comoving waves in Figure 141.1. The sources are in phase. Moving an equal distance away from the sources (such as positions C and D in Figure 156.1(a)), the two waves at those two points are still in phase. Thus, it is only the extra distance on the longer path to position P that accounts for the phase difference at position P. Eq. 141.3 still yields the non-reduced phase difference,

$$\frac{\Delta\phi}{L_A - L_B} = \frac{360°}{\lambda} \quad . \qquad (156.1)$$

Figure 156.1 only shows interference at one position, but more commonly the superposed waves can be directly observed over a large region. Figure 156.2(a) shows a real-life example in which the waves radiate in all directions from point sources. This removes the idealizations of Figure 156.1. The waves are ripples on the surface of water in a glass-bottomed tray.[46] On the right side of the picture, only the wave crests from source B are seen; that area is so far from source A that its wave amplitude is unobservably small.

Figure 156.2(b) and (c) show two patterns that can be seen in part (a). Part (b) shows the wave crests from the two individual sources, which can certainly still be identified in the region where the waves combine. However, these circles are *not* the subject of this chapter. Part (b) is provided as an example of what to ignore.

Figure 156.2(c) shows the **interference pattern** that is visible in part (a). In the region roughly between the sources, this pattern appears as bright lines. Further from the sources, each line in part (c) traces through a set of crests and troughs in part (a) (the triangles at the edge help to pick them out). These lines are the places where the two waves meet in phase, as at Figure 156.1 point P. They are called **constructive interference bands.**

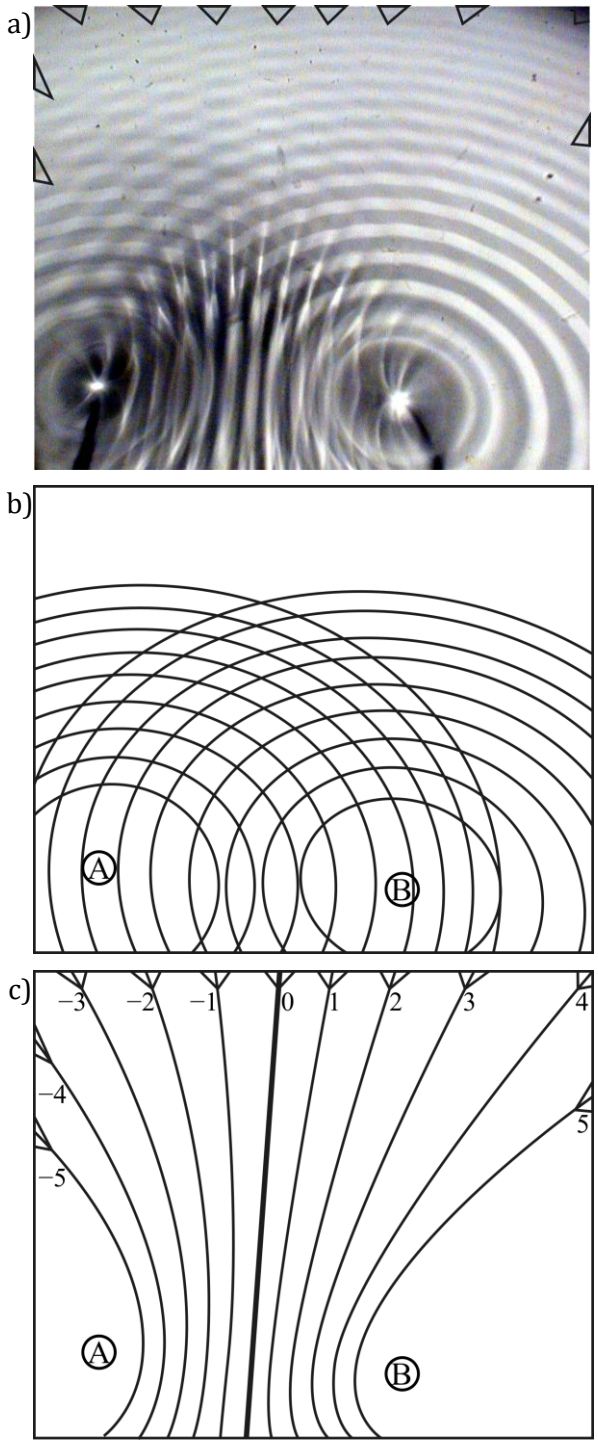

Figure 156.2

Two in-phase sources emit waves, forming an interference pattern. With perspective, circular waves appear as ellipses. (a) Photograph. (b & c) lines for certain shapes in (a).

[46] To make the water surface waves visible, light shines from above through the waves and onto a viewing screen below. For a wave without interference, such as the right side of the picture, bright bands show the water wave crests and the dark bands show the troughs.

The waves will meet in phase wherever their non-reduced phase difference is a multiple of 360°,

$$\Delta\phi = \phi_B - \phi_A = (360°)\, n \quad , n \text{ some integer,} \tag{156.2}$$

which means that the reduced phase difference is zero. The number n could be any integer, from negative to positive; it specifies how many cycles further away source B is than source A. Substituting this into Eq. 156.1, we find that these lines are the locations that satisfy the equation

$$L_A - L_B = n\lambda \quad , n \text{ some integer.} \tag{156.3}$$

Each value of n gives one constructive interference band. The numeric labels in Figure 156.2(c) indicate n. The absolute value of n is called the **order** of the band. The $n = 0$ band is straight, being the perpendicular bisector of a line between the sources. That is not clear in Figure 156.2 because the photo was taken from an angle, so Figure 156.3 shows an interference pattern from directly above. The shapes of the other bands are **hyperbolas.** You might recall that the function $y = 1/x$ makes an example of a hyperbola. In advanced geometry, equations like Eq. 156.3 are considered to define what makes a hyperbola. Figure 156.3 shows as dashed lines the asymptotes, which are the straight lines that the hyperbolas approach at large distances from the center.

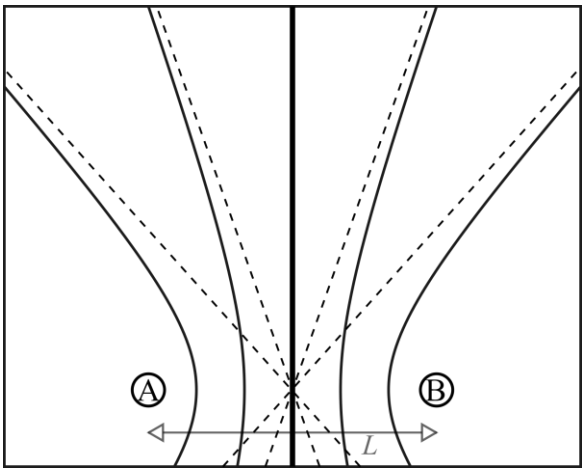

Figure 156.3
Two source interference pattern, with up to second order constructive bands, from directly above. Dashed lines are not part of pattern; they show asymptotes of curves.

Although this chapter looks at wider possibilities than Chapter 141, let's continue to use the variable L as the distance between the two sources. L is the largest possible difference of distances to any observation point. That is, L is the largest possible value of $L_A - L_B$. This means that there is a largest possible value of n, given by

$$n_{\max} = \left\lfloor \frac{L}{\lambda} \right\rfloor \quad , \tag{156.4}$$

where the brackets $\lfloor \quad \rfloor$ mean "round down to an integer." This in turn means that there are a limited number of constructive bands: $n = 0$ in the middle and $n_{\max}$ on either side, for a total number of bands given by

$$N = 2n_{\max} + 1 \quad . \tag{156.5}$$

If the two sources are moved closer together, then the number of constructive bands gets smaller. However, they continue to encompass a wide range of directions by spreading farther apart from each other.

In between the constructive bands are the **destructive interference bands,** where the waves meet out of phase. Because of the different distances to the sources, the amplitudes of the two waves are usually not equal, so that the destructive interference is not complete. Nevertheless, in the destructive bands the oscillations of the medium are smaller than in the neighboring constructive bands. In the region between the sources in Figure 156.2(a), the destructive bands are dark lines. Further from the sources, they are lines along which neither crests nor troughs appear.

To find the destructive bands mathematically, we modify Eq. 156.3 very slightly to

$$L_A - L_B = m\lambda \quad , m \text{ some half integer.} \tag{156.6}$$

The number m is a **half integer,** which refers to any integer plus one half. It is that extra half wavelength which sets the two contributing waves out of phase. (Notice that dividing an integer in half does not necessarily give a half integer. Dividing $(\text{odd})/2$ yields a half integer, but dividing $(\text{even})/2$ yields another integer.)

In physics, this interference pattern is considered a hallmark of waves. Thomas Young's observation of this pattern with light, in the early 19th century, was crucial in convincing the scientific world that light was a wave, even though the idea had been proposed by many before him. In the early days of quantum mechanics, when it was suggested that atomic particles had wave-like characteristics, experimentalists turned to two-source interference to test the proposition. Interference of sound waves is a little less groundbreaking, but it is not just an academic exercise. For a stereo sound system with two (or more) speakers, interference may result in **dead spots** where the sound volume is very low for particular frequencies.

So far, interference patterns have been described in a two-dimensional medium. What about those stereo speakers, radiating sound in three dimensions? The underlying ideas, about waves traveling from sources and meeting at locations in the medium, stay the same. All the equations and terms in this chapter work for 3D space as well, but the interpretation of them changes a bit. One band, specified by a particular value for the distance difference $L_A - L_B$, takes the form of a surface. Imagine Figure 156.3 rotated about the axis that connects the two sources. The central constructive band forms a plane, which is still the perpendicular bisector of the line between the sources. The other constructive bands form shapes like bowls, with one of the sources inside.

One final point should be made, especially in case you search elsewhere for information about interference. Many illustrations of interference start with a single source, producing a wave that hits a wall with two small holes. The waves emanating out of two holes then interfere with each other. (The equivalence between two holes and two sources will be particularly clear if you have read about Huygens wavelets in Chapter 157.) The reason to mention this is that the single source is not really an intrinsic part of two-point interference. It is just a convenient way to create two in-phase sources. In fact, it's the only practical method for light waves.

Chapter 157. Huygens Wavelets

In 1678, Christiaan Huygens came up with an unusual way to understand the passage of waves through a medium.[47] (Huygens was concerned with waves of light, but his ideas apply to any wave.) To set things up, as a wave passes through a medium (of any dimensionality), we choose to split the medium into two parts, with the wave progressing from one part into the other. The boundary between the parts, which we choose, can be any shape — it doesn't have to be a plane, nor does it have to take the shape of a wave crest, nor is there any other restriction. This imaginary boundary might coincide with a physical boundary between two media, but that is optional. Regardless, we will consider the transmission of the wave through the imaginary boundary.

Huygens' idea is that we can consider the medium at the boundary to be a vibrating object. For 2D and 3D waves, all the little bits of the medium at the boundary are many vibrating objects. **Huygens' principle** says that the propagation of the wave can be understood in this way: the incoming wave ends at the boundary, causing the boundary object(s) to vibrate, and in turn the vibrating object(s) are the source of the outgoing wave. If the boundary is physically real, that "outgoing" wave includes both a transmitted and a reflected wave.

Applications of Huygens' principle range from the almost trivially obvious to the very mathematically complex. On the simpler side, consider a rope with a knot and a wave traveling along the rope from left to

[47] Christiaan Huygens, *Traité de la Lumiere* (Leyden: Pierre Vander, 1690).

right. You might say that the wave passes through the knot, and as a result the knot moves. By Huygens' principle, you could instead say that the wave from the left terminates at the knot, causing the knot to move, and then the motion of the knot creates a new wave on the right. This is not to say that the first description is wrong. Huygens' principle just provides second, alternative viewpoint that is equally valid.

Why consider this more complicated alternative? The key insight is that the wave on the right side of the rope depends *only* on the motion of the knot. It does not depend on why the knot moves. If instead you moved the knot in the same way with your hand, then the resulting wave on the right side would be exactly the same. (Of course, moving the knot with your hand also would create a leftward moving wave in the left half of the rope; but that part is not what Huygens' principle is trying to explain.)

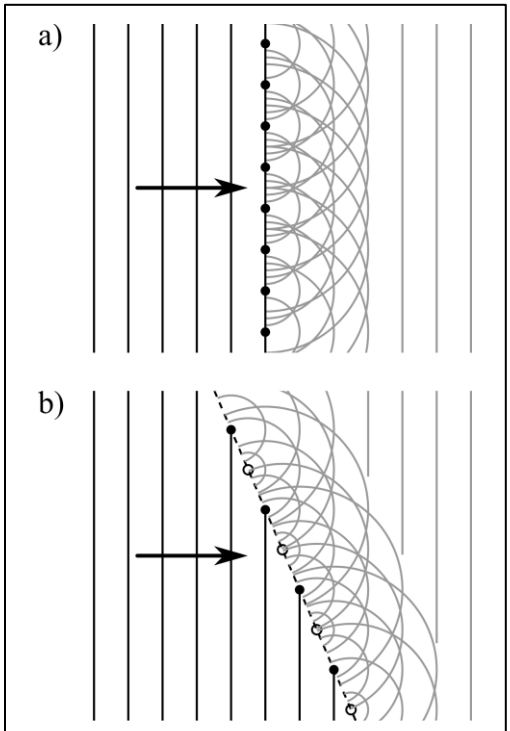

Figure 157.1

Huygens' wavelets for a plane wave, with the Huygens boundary (a) parallel to and (b) angled from wave crests. Outgoing crests are gray.

A straightforward conclusion is that if the incoming wave is periodic, then the outgoing wave is periodic as well, with the same period and frequency. The Huygens vibrator at the boundary receives the frequency from the incident wave and passes it on to the transmitted wave. If the boundary should be a physical boundary between two media, as described for 1D waves in Chapter 149, then some of the incident wave energy may be reflected back into the original medium. The Huygens vibrator also passes that same frequency on to the reflected wave.

Mathematical complexity appears when applying this principle to 2D or 3D waves. The idea is illustrated in Figure 157.1,[48] which serves equally well as a 2D wave or a cross section of a 3D wave. Waves are represented by drawing the wave crests. The Huygens boundary chosen, with dots along it, is a line (for 2D media) or a surface (for 3D media). The boundary is now an infinity of points, all vibrating, represented in the figure by a sampling of eight dots. The dots are not necessarily in phase, as each one vibrates in time with the wave as it reaches them; the open dots in Figure 157.1(b) are out of phase with the filled dots.

Each of the infinite boundary vibrators creates a small outgoing circular (2D) or spherical (3D) wave, called a **Huygens wavelet.** In the figure, the wavelet from each dot is shown for three wavelengths. In the transmission region, the outgoing wave is the superposition of all the Huygens wavelets. The superposition is the same as described in Chapter 155, but because there are an infinity of wavelets calculus is needed to do the mathematics properly. Even with only eight Huygens wavelets in Figure 157.1, instead of infinity, you can start to see the outgoing wave crests developing from them. When the calculus is done, the wavelet interference recreates the truly straight crests which we know must come out on the right side of the boundary.[49]

[48] For the purist: Figure 157.1 actually contains an error, having to do with the wavelets' phase relative to the plane wave, as explained in Eugene Hecht, *Optics,* 2nd ed. (Reading, Massachusetts: Addison-Wesley, 1987), 438. This issue is best ignored, except in much more advanced treatments of the topic.

[49] It is worth noting that when Huygens proposed this model, calculus had hardly been invented, and it certainly had not developed to the extent that we now use it to implement his principle. It is a testament to Huygens' imagination that he could propose this model before the mathematical underpinnings existed.

As with all such snapshot figures, keep in mind that this is just a moment in time. As time progresses, the wave slides in from the left. The wavelets that are shown were created over the three periods just before the snapshot.

So far, this chapter has described an awful lot of work to recreate conclusions that we already knew. Why bother? This illustrates an important aspect in the development of new physical models. New models can provide a new perspective which allows us to discover new relationships; some will be described in the following paragraphs. But a new model must not contradict preceding, successful models, at least not to the extent that they have been successful. This is called the **correspondence principle.** It is therefore important to convince ourselves that the new model agrees with the older one under appropriate conditions. (The correspondence principle is especially significant in the context of quantum mechanics. Quantum mechanics sometimes makes such outlandish predictions that it is especially important to be sure that it does not contradict the successful parts of classical physics.)

Huygens' principle is behind many of the other relationships in this book, although an explanation of how are frequently omitted.

As already noticed in the 1D case, Huygens' principle continues to tell us that for periodic waves, the outgoing frequency is equal to the incoming frequency, even if there is a change of medium.

Huygens' principle can provide an underlying reason for the rule, given in Section 125a, that rays must be perpendicular to the wave crests.

When a 2D or 3D wave is transmitted through a boundary between two media, Huygens' principle leads to the conclusion that the wave crests of the incident wave must connect to the wave crests of the transmitted wave. When an incoming wave crest hits the boundary, the Huygens vibrator is at its maximum of oscillation, and thus is at that moment creating a crest of the outgoing wave. This explains the refraction relationships described in Chapter 161.

When a 2D or 3D wave is reflected from a boundary between two media, the consequences are the same as for transmission except for one addition. Just as described in Chapter 149 for 1D waves, the reflected wave may be inverted if the boundary is more fixed than free. Upright reflection means the reflected and incident wave crests connect, just as for the transmitted wave. Inverted reflection means that the reflected wave troughs meet the incident wave crests at the boundary, and vice versa.

Chapter 158. Reflectors

Chapter 8 describes that waves reflect off boundaries such that the angle between the reflected ray (sound path) and the boundary is the same as the angle between the incident ray and the boundary. When sounds and other waves reflect off curved surfaces, some useful effects can result. This book qualitatively describes the basics, but none of the resulting mathematics is included here.

When waves encounter a **concave** surface, the reflections tend to gather the rays into a smaller region. Examples are shown in Figure 158.1(a&b). Sometimes a set of rays, all originating somewhere else, are bent such that they nearly come together at a single place. Such a place is called an **image point** or **focus.** Notice that once the rays pass through the image point/focus, they leave as if the image point were a new point source of a wave.

Many different sets of rays can form a focus, but there is on that is especially important. It results from a set of parallel incident rays (a plane wave) which are also parallel to the curved surface's axis of symmetry, as seen in Figure 158.1. This focus is called the **focal point** of the reflector.

This focusing works the best for a surface with the shape of a parabola, as in Figure 158.1(a). A perfect focal point is formed, with the axial rays all converging to exactly the same point. A parabola is a shape familiar to most people, because it is also the shape of the arc made by a projectile thrown through the air. The equivalent shape for a 3D plane wave is a paraboloid, a bowl for which every cross section is a parabola.

While a parabolic reflector is the best shape for making a focus, it is only perfect for rays that enter parallel to the parabola axis. Usually it is easier to manufacture a reflector that is an arc of a circle (2D waves) or a section of a sphere (3D waves), as shown in Figure 158.1(b). Plane waves that pass close by the circle's center (the dot) are reflected towards an imperfect focal point that is half way between the circle center and the reflecting surface. The top and bottom rays in in Figure 158.1(b) are too far from the centerline, and are clearly not headed for the focal point.

This effect is put to good use in many ways. **Parabolic microphones** put a regular microphone at the focus of a parabolic reflector. By gathering sound energy from a large area, these can detect very soft sounds coming from a very specific direction. Many antennae for electromagnetic waves, including satellite dishes, use this same focusing to receive a weak signal. Reversing the direction of all the rays, these reflectors can turn a point source into a **beam** (a plane wave of well-defined area), such as in vehicle headlights. On the lighter side, **whispering galleries** have two reflectors facing each other (often built into the architecture). A whisper from the focal point of one reflector is turned into a beam with an intensity too soft for humans to hear, but the whisper is refocused to audible intensity at the focal point of the other reflector.

Figure 158.1(c) shows a plane wave reflecting off a **concave** reflecting parabola, which results in the reflected rays diverging away from one another. One might say that there is no focusing going on here. However, the reflected rays do all appear to be emanating from a single point, as indicated by the dotted lines that extend the rays behind the reflector. Because of this similarity to a focus, such a point is called a **virtual image point.** In fact, this is the case where the choice of the word "image" makes the most sense. Like the (virtual) image in your bathroom mirror, it looks as if there is a source there, even though there really is nothing behind the reflector.

The term "focus" is not normally used for a virtual image point, but the special virtual image point that results from an axial plane wave is still called the **focal point** of the convex reflector. To further distinguish the two types of image points, a focus (such as happens with a concave reflector) can be called a **real image point.**

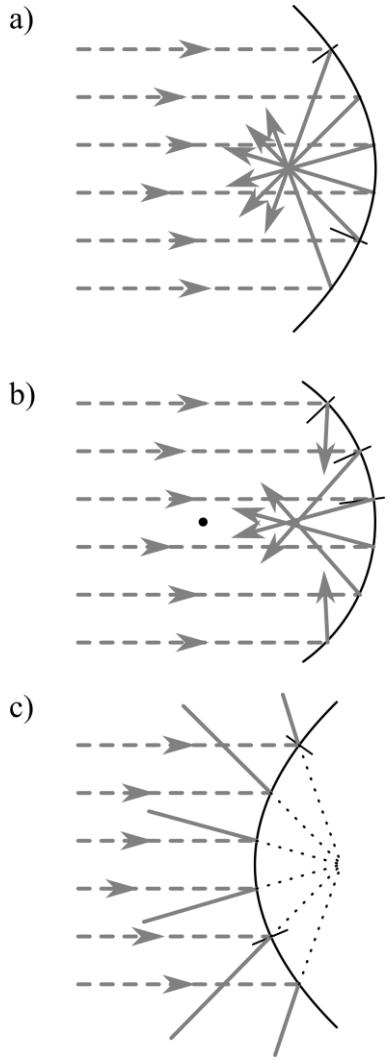

Figure 158.1
Reflection of plane wave (dashed rays) from curved surfaces: (a) Concave parabola. (b) Concave arc of a circle; black dot marks center of circle. (c) Convex parabola; dotted lines extend reflected rays back behind reflector. Several reflection points show surface normal lines.

Chapter 159. Reflection

The equal angles rule for reflection is fairly easy to accept intuitively, because it is exactly the same rule followed by simple objects (for instance, a rubber ball with no spin) bouncing off a wall. But the reason behind the behavior is quite different for balls and for waves.

As a sound reflects off a hard, smooth wall, the reflected wave must have the same frequency and wavelength as the incident wave. Also, the wave crests of the reflected wave must connect with the wave crests of the incident wave. This is because, as in Chapter 148, the hard surface allows the air density to be free to vary as the air compresses against the wall or sucks away from it. This leads to upright reflection, where an incident compression makes a reflected compression.

Figure 159.1 illustrates the only way to satisfy both of these requirements: to have the incident and reflected wave crests making the same angle with the boundary. Notice how the reflected wave crests meet the incident wave crests at the boundary. If the reflected wave crests made any other angle with the boundary, then they could not both meet the incident wave crests and have the same wavelength.

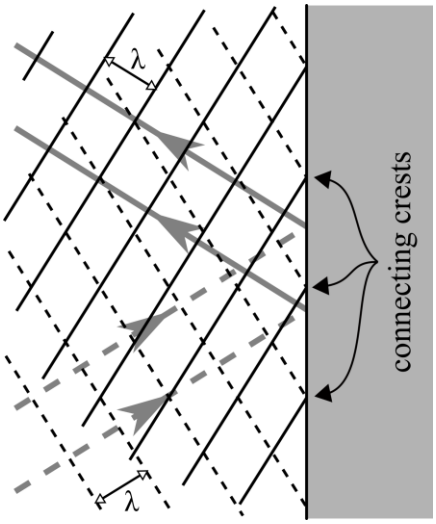

Figure 159.1

Plane wave reflecting off a hard surface. Wave crests are black, rays (two for each wave) are gray. Dashed lines are incident wave, solid lines are reflected wave.

It is this wave crest relationship that really determines the angle of reflection. The rays are just along for the ride, but the result is that they, too, make the same angle with the boundary. Practically speaking, it is usually easier to follow and understand the rays. All those wave crests in Figure 159.1 are hard to take in.

Chapter 160. Hard and Porous Surfaces

Reflection as described in Chapter 159 is the 2D or 3D equivalent of an upright density reflection. It is possible to have a sound reflection where the compression wave inverts, however. That happens if the density is mostly fixed at the boundary, which conversely means that the air displacement is mostly free at the boundary. An example would be if sound in air were to encounter a helium balloon. Since the sound speed is higher in helium than in air, that situation is on the left side of Figure 149.1, where wave displacements are reflected upright. Density, behaving in the opposite way, would reflect inverted.

Another reflecting surface that mostly fixes the density by freeing displacement is a porous or extremely rough surface. Examples are the leaf litter on a forest floor, or the grass of a lawn. The air trapped in the rough surface provides a sort of cushion, which tends to fix the density there, and result in inverted compression waves.

Happily, as long as the wavelength is significantly longer than the spacing of roughness features, none of this affects the basic conclusion of Chapter 159. With inversion, the crests of the reflected wave must meet the troughs of the incident wave. But the troughs are still spaced by one wavelength, just like the crests. So, the reflection angle still must match the incidence angle. The direction of sound reflection is the same as for hard, smooth boundaries.

Chapter 161. Refraction at a Boundary

161a. Wave Refraction

Chapter 8 describes that, when sound (or any 2D or 3D wave) is transmitted from one medium to another, it usually changes its direction of travel. This change of direction is called **refraction**. The **law of refraction,** which governs exactly which direction it takes, is commonly known as **Snell's Law**, although it is now well established that Willebrord Snellius neither discovered the rule first nor published it first.[50] Although the law was first obtained in reference to light, it applies to all waves, and this chapter will present a version most easily applied to sound.

It is easier to follow the rays than the wave crests. Figure 161.1 focuses on the ray in the center, while the rays on the left and right are there as a reminder that the rays are only symbolic of the motion of a wider wave. Small reflected rays are also in the figure, as a reminder that transmission through the boundary is not all that happens.

Traditionally, directions of the rays are specified as angles from the **surface normal,** a line that is perpendicular to the boundary between the media at the point where a ray meets that boundary. In Figure 161.1 the surface normal is the vertical dash-dot line. The normal and the boundary together divide the plane of the picture into four quadrants. As in both examples in Figure 161.1, the refracted ray is *always* in the quadrant diagonally opposite the quadrant of the incident ray. That is, the refracted ray never reverses the wave's direction along the boundary, which would put the refracted ray in the quadrant labeled with the "No Entry" symbol. The transmitted ray cannot become parallel with the surface normal either, unless the incident ray is parallel with the normal as well.

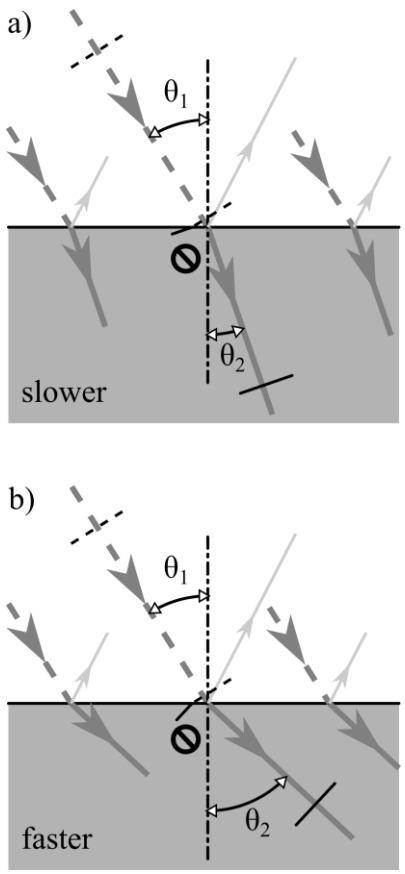

Figure 161.1

Refraction of wave rays incident from above. Wave speed in gray medium is (a) slower and (b) faster than in white medium. Bits of wave crests are included on middle rays.

The angles labeled with θ in the figure measure the directions of the rays. The relationship between them is

$$\frac{\sin\theta_1}{s_1} = \frac{\sin\theta_2}{s_2} \quad , \qquad (161.1)$$

where s_1 and s_2 are the wave speeds in the respective media. A nice thing about this relationship is that it does not depend on which ray is incident or transmitted. It is only necessary to match up the speeds and angles.

An important qualitative distinction depends on the relative wave speeds in the two media. Figure 161.1(a) shows the case where the refracted wave speed is slower than the incident wave speed. The refracted ray then bends *towards* the normal. In Figure 161.1(b), where the refracted wave speed is faster than the incident wave speed, the refracted wave bends *away from* the normal. Both cases are described by the following rule.

> As the wave is refracted, its ray bends towards the medium with the slower wave speed.

[50] Roshdi Rashed, "A Pioneer in Anaclastics: Ibn Sahl on Burning Mirrors and Lenses," *Isis* 81 (1990): 478.

An alternate way to remember which way the rays refract is to consider a short part of a wave crest as it passes through the boundary. In Figure 161.1(a) the left end of the wave crest hits the second medium first, and must slow down. As the right end of the crest continues at the higher speed, the whole crest swings clockwise. The ray, which must remain perpendicular to the wave crest, swings clockwise with it. In Figure 161.1(b) the opposite happens. The left end of the wave crest speeds up, swinging the crest counterclockwise.

161b. Extra: Total Reflection

When a wave approaches a boundary with a medium on the other side that would give a higher wave speed, it can happen that no transmission is possible at all. Figure 161.2 shows several examples of reflection and refraction where the wave speed in the second medium is double the wave speed in the first. Because of refraction, as the incident angle increases, the refracted angle increases faster.

In Figure 161.2(c) the incident angle is so large that the refracted angle is 90°, meaning that the refracted wave can't get in to

the second medium. If the incident angle is equal to or greater than this **critical angle** then no transmission is possible, and all of the wave energy is reflected, which is called **total reflection**.

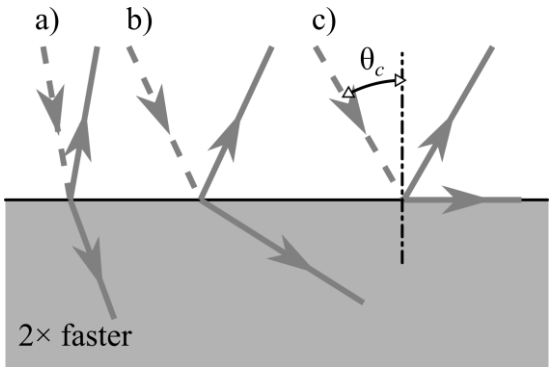

Figure 161.2

Reflection and refraction at boundary. Wave speed in gray medium is double speed in white medium.

The critical angle can be determined from Eq. 161.1 and the fact that $\sin 90° = 1$, giving the equation

$$\sin\theta_c = \frac{s_1}{s_2} \quad . \tag{161.2}$$

This phenomenon is sometimes called **total internal reflection** because for light waves it tends to happen when the light is inside a transparent substance, trying to leave. For sound, however, there is no particular reason to expect the slow medium to be "internal" to anything. In fact, because sound speeds vary so widely, this phenomenon is much more likely than for light. When sound in air meets with a solid or liquid surface, the critical angle is in the vicinity of 10°. Unless the sound wave hits the surface nearly head on, it will totally reflect from the surface.

Chapter 162. Wavelength Change upon Transmission

Although refraction is easier to envision by focusing on wave rays, the underlying reasons for the law of refraction come from the behavior of the wave crests. Just as described for one-dimensional waves in Chapter 149, the transmitted wavelength will be stretched or contract. The extent of that stretching is most easily determined for periodic waves, for which we can use Eq. 121.2. Since the incident and transmitted waves must have the same frequency, we rearrange that equation to

$$\frac{s_1}{\lambda_1} = f = \frac{s_2}{\lambda_2} \quad , \tag{162.1}$$

which is to say that the wavelengths of the two waves are proportional to the wave speeds.

But, as a wave transmits through a boundary, the wave crests of the transmitted wave must connect with the wave crests of the incident wave, for reasons detailed in Chapter 157. Figure 162.1 illustrates that the only way to satisfy all these requirements is to have the incident and transmitted wave crests making different angles with the boundary. They must be refracted.

Notice how the transmitted wave crests meet the incident wave crests at the boundary. If the refracted wave crests made any other angle with the boundary, then they could not both meet the incident wave crests and have the proper wavelength. In the case illustrated, the slower medium results in a shorter wavelength, so that the wave crests must be closer to parallel with the boundary. If the medium of the transmitted wave had a faster wave speed, then all those relationships would be reversed.

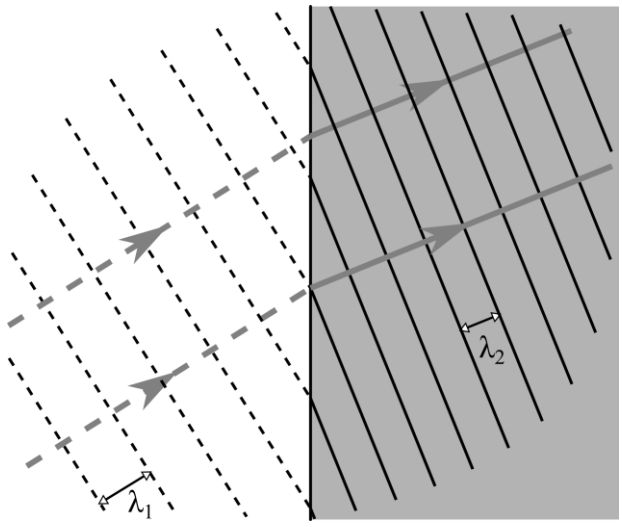

Figure 162.1

Refraction at boundary where wave speed slows, showing whole wave as series of wave crests.

It is this wave crest relationship that really determines the angle of refraction. The rays are just along for the ride, but the result is that they, too, bend at the boundary. Practically speaking, it is usually easier to follow and understand the rays.

A little trigonometry along with Eq. 162.1 is all you need to derive Eq. 161.1. This book will leave that derivation as a small puzzle for the reader. But a broader point is that, once again, as so often happens in physics, a few relatively straightforward rules have combined into a not-so-simple result.

Chapter 163. Continuous Refraction

Chapter 161 describes how waves are refracted at boundaries where the wave speed changes. But sometimes the wave speed in a medium changes slowly over a large distance. If the wave speed change is gradual enough, then no reflected waves will result when the wave encounters the change. But no matter how gradual, the wave speed change implies a change in wavelength, which in turn requires refraction, or a bending of the rays. The reason is essentially the same as described in Chapter 162 for abrupt boundaries.

Because this refraction happens continuously, it cannot be described by a simple equation. Real examples involve a variety of ways for the wave speed to vary through the medium, each one resulting in different shapes for the rays. But the same qualitative description still holds true in all cases.

> As the wave is refracted, its ray bends towards the medium with the slower wave speed.

If the waves speed variation is gradual, then so is the bending. It is useful to imagine **normal lines** that point as directly as possible from the regions of higher wave speed to the regions of lower wave speed; those lines play a role similar to the surface normal in boundary refraction. A ray can never bend to the extent that it points directly parallel to the normal lines, unless it starts out parallel to them. And a ray certainly can never bend to the extent that it reverses direction to cross a normal line more than once.

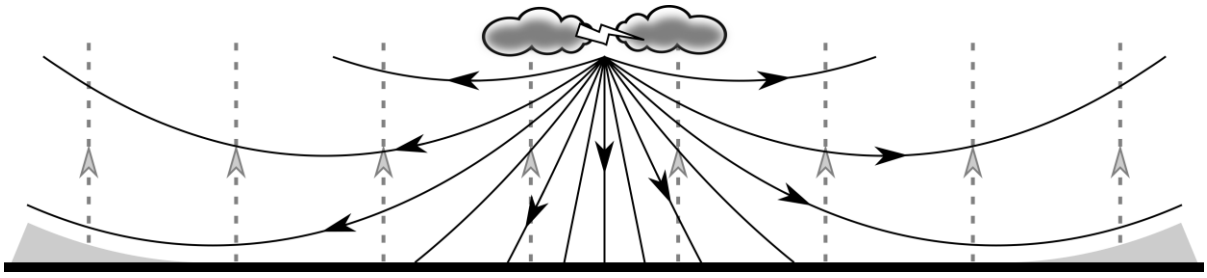

Figure 163.1

Bent rays of cloud thunder (solid lines) under normal atmospheric conditions, with colder air further from ground. Dashed gray are normal lines, indicating direction of wave speed decrease. Gray triangles near ground are sound shadow zones. Picture is accurate for linear variation of sound speed between ground and clouds at altitude 10 km. Horizontal scale is compressed by a factor of two, to make the curves more visible.

Atmospheric Refraction

One example occurs naturally in the atmosphere. Under normal conditions, the air within about ten kilometers of the ground gets colder with higher altitude. According to Section 7b, this means that the sound speed drops with altitude. It works out to be about a $4\ \mathrm{m/s}$ change for each kilometer of altitude.[51] This means that sounds which travel large distances tend to refract away from the ground. Figure 163.1 illustrates what happens to thunder made by a cloud-to-cloud lightning strike. Dashed gray lines in the figure show the normal lines, passing vertically upwards directly from higher sound speed regions to lower sound speeds. The rays radiate from the source, and the normal lines help to visualize how they are refracted. However, the rays that are oriented nearly parallel to the normal lines are hardly affected. In fact, the sound that is passing straight down from the lightning is speeding up as it approaches the ground; the normal lines do not indicate some sort of "wind" that impedes the passage of the wave.

An interesting consequence of this wave propagation pattern is that observers on the ground and sufficiently far away will not hear the thunder at all, even though they can see the lightning. The gray triangles in Figure 163.1 indicate **sound shadows,** where the sound never reaches because the ground is "in the way" (considering the curved path the sound must take).

The SOFAR Channel

In the deep ocean, another continuous refraction effect occurs for underwater sound (that is, 3D compression waves in the water, not 2D water surface waves). Near the surface the water is warmer, while descending into the ocean depths means moving to colder water. By itself, it turns out that decreases the water's bulk modulus, and thus decreases the speed of sound. But going deeper also means that the pressure steadily increases, which by itself increases the bulk modulus and sound speed. Over large distances, these changes can't be neglected. In roughly the top 1 km the temperature effect is stronger, but below that the temperature is essentially constant and the pressure effect dominates.[52]

The end result is that descending into the ocean, the sound speed at first decreases moderately rapidly, then reaches a minimum value at a depth of about 1 km, and then slowly increases for greater depths. The depth of minimum sound speed varies considerably throughout the world, generally being deeper near the equator where the surface warming is strongest.[53] The sound speed changes are only about $20\ \mathrm{m/s}$, or roughly

[51] "Speed of sound, temperature and pressure at various elevations," *The Engineering ToolBox,* n.d., http://www.engineeringtoolbox.com/elevation-speed-sound-air-d_1534.html (March 2015).
[52] W. H. Munk, P. Worcester, and C. Wunsch, *Ocean Acoustic Tomography* (Cambridge: Cambridge University Press, 1995).
[53] W. H. Munk and A. M. G. Forbes, "Global Ocean Warming: An Acoustic Measure?" *Journal of Physical Oceanography* 19 (1989): 1765.

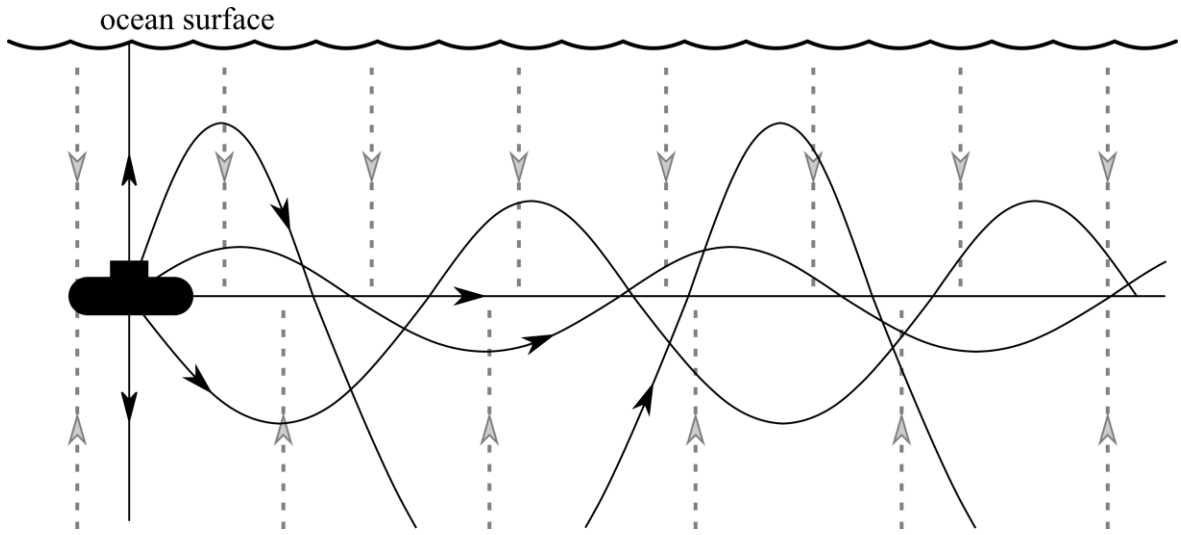

Figure 163.2

Rays (solid lines) in the SOFAR channel. Dashed gray are normal lines, indicating direction of wave speed decrease. Horizontal scale is compressed by a factor of roughly twenty; to-scale rays would all be within 10° of horizontal. Rays are only sketches; exact rays depend on how sound speed changes with depth.

1.5% of the average water sound speed, but this is enough to cause significant refraction over large distances.

Figure 163.2 gives a sketch of the result, showing a sound leaving a submarine positioned at the depth of minimum sound speed. The sound is continually refracted back towards this special depth, creating the so-called SOund Fixing And Ranging channel, or SOFAR channel. It's important to make clear that this figure is a sketch, to show the qualitative ideas. Also, the figure is highly compressed horizontally; in a to-scale picture, only sound trajectories within about 12° of horizontal would be bent enough to avoid hitting the ocean surface or the ocean floor, where most of the sound energy would be dissipated.

To be clear, the curved lines here do ***not*** represent anything vibrating, nor do they represent disturbances that make a wave. Rather, they are rays that trace the paths of sound waves themselves. A more detailed picture would include wave crests, drawn perpendicular to the rays. But you can probably guess that such a picture would become extremely complicated and hard to understand. It is easier to consider sounds leaving the submarine in different directions to be separate wave packets following rays.

By keeping these sound paths from hitting the surface or ocean floor, the SOFAR channel prevents these sounds from spreading out as much as they would in three dimensions. Since the sound energy spreads out less, the sound carries much further than it otherwise would. It is believed that the SOFAR channel enables whales to communicate over very large distances, and it certainly has consequences for submarine operation.

Oddly enough, the sounds that take the longest paths are the ones that cover horizontal distances the fastest. The sounds that travel directly along the center of the channel are always at the slower speed, and despite the fact that they travel the shortest distance, they arrive at a distant listener last.

Similar Phenomena

Rays also bend through the air when there is a wind, and some references describe this as refraction. However, in the effects described in this chapter, the sound speed variation does not depend on its direction of travel. In the wind effect, the sound speed is only altered along the direction of the wind, and most authorities agree that this is not included in the term refraction. It is certainly an interesting effect, but it will not be included in this book.

Chapter 164. Diffraction

164a. Sound Around Corners

Chapter 8 describes that sound can travel around corners. The detailed prediction of this effect is one of the most difficult areas of wave mechanics. Although it was first examined in a published account in 1665 by Francesco Grimaldi,[54] who gave the phenomenon the name **diffraction,** significant advances have continued even into the late 20th century.[55] This book will settle for a relatively descriptive treatment, although some math will still sneak through. Three situations are described in particular: sound passing through a small hole, sound passing around a small obstruction, and sound diffracting around an edge.

Advanced treatments of diffraction rely heavily on the idea of Huygens wavelets, covered in Chapter 157. Briefly, as a wave passes an obstruction, every place where there is *not* an obstruction acts as a source of a new wave. The end result is the superposition of all those new waves. This bears a resemblance to the interference described in Chapter 156, in which a pattern is created by the superposition of two waves. If you have read that chapter, look for the similarities in the behavior of diffracted waves.

164b. Extra: Sound Around an Edge

When sound, or any wave, passes by the edge of an obstruction that is many wavelengths in cross-sectional size, the wave diffracts around the edge into the region behind the object. Let's call a ray that just grazes the obstruction's edge a **shadow ray.** If there were no such thing as diffraction, such rays would mark the boundary between the region with waves and the region without waves. The **shadow region** is the region on one side of the shadow rays and behind the obstruction.

Describing what happens is a little difficult because of the unusual way the amplitude varies. There are also some surprising effects just outside the shadow region. Figure 164.1 illustrates what happens for an incident plane wave, with separate parts focused on (a) the non-shadow region and (b) the shadow region. If this is to represent a 3D wave, then the obstruction in Figure 164.1 would be a wall with a long straight edge perpendicular to the page. Other large obstructions would give qualitatively similar results.

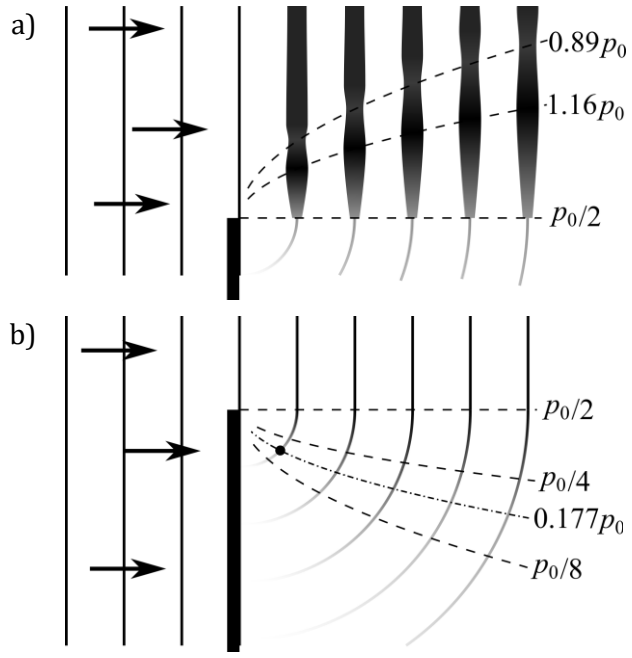

Figure 164.1
Diffraction of plane wave past the edge of large obstruction. (a) Focus on region above the edge. (b) Focus on region below the edge. Amplitude of the incident wave is p_0.

Along a shadow ray, the amplitude of the diffracted wave is half of the incident wave amplitude. This isn't a big surprise: from the perspective of those points, half of the wave has been eliminated.

In the shadow region, circular wave crests travel away from the obstruction edge, the same shape as if the edge were a point source (for 2D waves) or a line source (for 3D waves). The amplitude of those waves, however, is quite unusual. In Figure 164.1(b) the amplitude is approximately shown by the grayscale of

[54] Francesco Maria Grimaldi, *Physico-mathesis de lumine, coloribus, et iride* (Bologna: Victorio Bonati, 1665), 1–11.
[55] Eugene Hecht, *Optics*, 2nd ed. (Reading, MA: Addison-Wesley, 1987), 392–394, 464.

the wave crests. A few lines of constant amplitude are also shown by dashed lines, which have the shape of parabolas with apexes at the edge. The black circle is a key point, located one wavelength from the edge at an angle of 45° from the shadow ray. Along the parabola through that point, the dash-dot line in Figure 164.1(b), the amplitude is 17.7% of the incident amplitude. If you imagine increasing the wavelength, then that point is pushed further from the edge, which in turn makes the 17.7% parabola wider. Thus, although the picture is far from simple, we do have a simple rule that longer wavelength waves diffract around an edge more than shorter wavelengths.

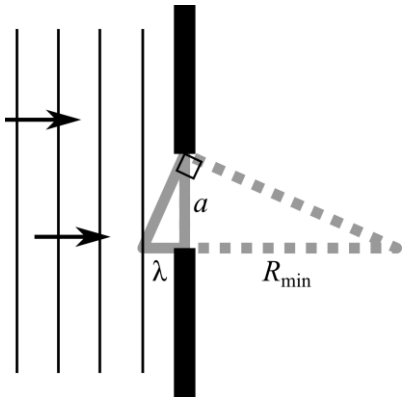

Figure 165.1
Plane wave approaches hole of width $a = 2.25\,\lambda$. Diffraction pattern will appear a distance $R_{\min}$ beyond hole.

In Figure 164.1(a), focusing on the non-shadow region, the amplitude variations are quite subtle. To make them easier to see, the amplitude is show both by a grayscale and by the width of the wave crest lines. The odd thing is that there is a line (again a parabola) along which the amplitude is actually *larger* than that of the incident wave! In fact, the amplitude oscillates with distance from the shadow ray, settling eventually (far from the shadow region) to be equal to the incident amplitude. For light waves, these variations are quite visible. For sound waves, however, the resulting intensity variations are near the limit of detectability by the human ear.

Chapter 165. Diffraction Through a Hole

When a wave, such as sound, passes through a small hole, immediately on the other side of the hole the wave shapes require advanced math to predict. But looking further away from the hole, a simple pattern appears, called a **Fraunhofer diffraction pattern**. What is meant by "small" and "further away" here? Define the variable a to be the smallest dimension of the hole, λ to be the wavelength of the wave, and R to be the distance from the hole to the point of observation. To see a Fraunhofer pattern, you have to look at least a distance $R_{\min}$ away from the hole as given by the equation

$$\frac{R_{\min}}{a} = \frac{a}{\lambda} \quad . \tag{165.1}$$

To see what this implies, consider Figure 165.1. On the side of the hole with the **incident wave** (the left side here) a right triangle is drawn with sides of length a and λ. The hypotenuse of that triangle forms one side of a larger right triangle. This large triangle is divided into two smaller ones by the hole width, and all three are similar triangles. $R_{\min}$ is the horizontal side of the smaller triangle that is entirely on the side of the hole with the **transmitted wave.**

With Figure 165.1 as a guide, we see that a "small" hole is one with a smallest dimension that is less than a few (perhaps 5) wavelengths. Larger holes will still create the Fraunhofer pattern, but only at a rather large distance. In fact, close behind such larger holes (roughly $R < R_{\min}/10$), the wave will pass through the hole's center unaffected, while near the edges the wave will distort as described in Section 164b. In between that region and $R_{\min}$, things get really messy as the wave pattern transitions from unaffected to a Fraunhofer pattern.

Far enough from a hole, then, the transmitted wave takes the form of a Fraunhofer diffraction pattern as illustrated in Figure 165.2. Most of the transmitted wave is in a diffraction beam, with a wedge or cone shape, that radiates from the center of the hole like a point source. The wave crests maintain the same shape (that is, how the amplitude depends on angle) as they spread out. The width of this beam can be described by the angle θ_1 from the centerline to a line where there is no wave amplitude at all. Figure 165.2 shows

the case for a 2D wave passing through a gap, or a 3D wave passing through a long narrow slit, for which this angle satisfies

$$\sin\theta_1 = \frac{\lambda}{a} \quad . \tag{165.2}$$

A 3D wave passing through a circular hole spreads out a bit more, governed by

$$\sin\theta_1 = 1.22\frac{\lambda}{a} \quad . \tag{165.3}$$

Table 165.1 gives several specific examples values.

Table 165.1

Diffraction pattern widths for narrow rectangular and for circular holes.

Hole size a/λ	θ_1 for Slit	θ_1 for Circle
5	11.5°	14.1°
4	14.5°	17.8°
3	19.5°	24.0°
2	30.0°	37.6°
1	90.0°	> 90°

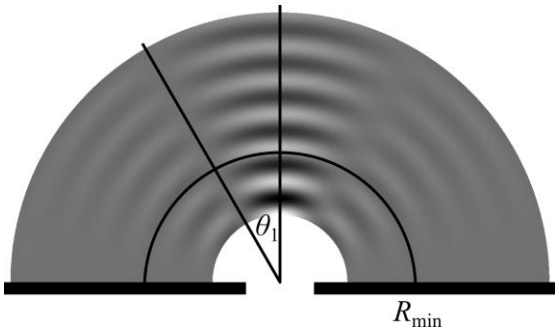

Figure 165.2

Fraunhofer diffraction pattern from hole with width that is double the wavelength. Grayscale shows wave disturbance, with medium gray indicating equilibrium state. At distance less than R_{min} true pattern will differ somewhat from this model.

About 90% of the energy in the transmitted diffraction wave is in the central beam. But in Figure 165.2 it is evident that there is some wave action outside of the θ_1 angle. If you have read Chapter 156, you may notice the similarity to the constructive and destructive bands of two source interference. Indeed, a narrower diffraction pattern would show multiple bands of wave action on each side of the central beam. However, there are significant differences from two source interference as well. For instance, the central beam is twice as wide as the other bands, the central beam has a much larger amplitude than other bands, and there are no hyperbolic shapes.

Notice that as the hole gets smaller, the intensity of the transmitted wave or sound gets smaller for two separate reasons. First, a smaller hole simply lets through less power from the incident wave. In addition, a smaller hole also causes the power to be spread over a wider angle, and thus a larger area. Holes less than one wavelength wide are almost the same as point sources, because the central "beam" spreads so wide that the diffracted sound intensity is nearly the same in all directions.

Chapter 166. Diffraction Around an Object

A rule called **Babinet's principle** says that when a wave passes by an obstacle, the result is exactly the superposition of what the wave would have looked like without the obstacle and the *negative* of the transmitted wave through a hole with the same shape as the obstacle. That is, the obstacle subtracts from the total wave exactly what a hole would have made.

For instance, an object much smaller than a wavelength has very little impact on a passing wave. Just as a transmitted wave through a small hole diffracts over a wide range of angles, the impact of the object also spreads very widely, and is thus diluted. If you have observed a water wave pass by a vertical stick or post stuck in the bottom of a lake, you may have seen that the wave passes around the post and continues mostly undisturbed. We can understand this as a consequence of the way the water flows. Water moves pretty quickly to level out its own surface. In the one-half period that it takes for the water wave crest to pass the

stick, there is plenty of time for the water to flow around to the back of the stick, so that behind the stick the water moves up and down just as it would have from the wave alone. That way of understanding is not different from diffraction; it is just a specific, concrete realization of the more abstract diffraction idea. Both ways of thinking are also applicable to sound waves passing by small objects.

On the other hand, we saw that for holes larger than about five wavelengths the wave passes through the center undisturbed, at least for a short distance. Inverting that, behind an object with all cross-sectional dimensions much larger than a wavelength the disturbance will be (the original wave) − (an undisturbed wave) = nothing! The sound does not succeed in bending around the edges of the object that much, and there is a **sound shadow** quite analogous to shadows of light. Because of the comparison to wavelength, this particularly applies to high pitched sounds. Even then, the sound shadow may not extend very far behind the object. Light shadows extend far beyond moderately sized objects because light wavelengths are so tiny.

Between these two extremes is an obstacle of moderate size, compared to the wavelength. It turns out that the Fraunhofer pattern that appears (at an appropriate distance) behind a moderately sized hole is in phase with the incident wave. As a result, when the subtraction of Babinet's principle occurs, we find that in the corresponding location behind an obstacle the wave amplitude is reduced. This is a second sort of **sound shadow,** although in this case the intensity is only somewhat reduced. Moving further away behind the obstacle, the effect of the obstacle gets smaller and the sound intensity returns toward its unaffected value.

In real-world situations, listeners in a sound shadow are likely to hear the sound via other paths, such as reflections off other nearby objects. Understanding sound shadows is only the beginning of understanding the acoustic environment surrounding objects.

Chapter 167. Directional Hearing

A single ear can determine many things about a sound, but since an eardrum only measures an oscillation at one position in space, the eardrum cannot, of itself, determine the direction that a sound came from. Many animals have an external ear, the **pinna,** which receives sound preferentially in one direction. By combining that with the ability to aim the pinna in different directions, they can determine the source direction. But the human pinna, or **auricle,** cannot be aimed independent of moving the head.

The primary and most direct way that humans obtain **directional hearing,** also called **sound localization,** is through having two ears. Using two ears is called **binaural hearing,** and directional hearing is perhaps its primary, but not its only, benefit. By combining information from two positions in space, it is possible to determine the angle of the source relative to the axis through those two points. That still leaves ambiguity, as there is an entire cone of directions that are any given angle from our ear-to-ear axis. But combine that with the fact that most sounds that we care about come roughly from a horizontal plane, and the localization is pretty effective, leaving only ambiguity about whether the source is in front or behind us.

Additional information does come from the fact that human auricles collect sounds differently depending on the source direction. These differences depend on frequency, and influence both the efficiency with which sound is collected and minute time delays in doing so. These effects can be subconsciously compared to remembered sounds, to get clues about the source direction. But these clues are subtle and not terribly effective on their own (as the author can attest, since he is totally deaf in one ear, and thus relies solely on these secondary clues).

Human accuracy in determining direction is fairly impressive, with typical errors of only 3° when the source is nearly straight ahead.[56] Sounds to the side are more poorly localized, with typical errors of 10°, while sounds coming from behind are localized with intermediate accuracy.

[56] Jens Blauert, *Spatial Hearing* (Cambridge, MA: The MIT Press, 1997), 41.

Chapters 168 and 169 detail the two different ways that binaural hearing contributes to directional hearing. They are both based on ways in which a single sound will be detected differently by the two ears.

Chapter 168. High Frequency Directional Hearing

One way a single sound can differ at the two ears is in loudness. As illustrated in Figure 168.1, the ear closer to the sound source receives the sound directly, but the sound must diffract around the head to reach the further ear. If the wavelength is short enough, this exposes the far ear to the reduced sound amplitude in the sound shadow of the head. The degree to which this happens depends on both the direction of the source and the wavelength. Nevertheless, our subconscious brain can learn that sound at a certain frequency (and wavelength) and with a certain difference in loudness between the two ears implies that the sound is arriving from a specific direction.

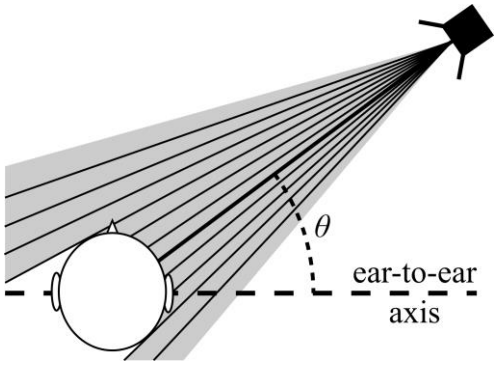

Figure 168.1
High frequency sound results in sound shadow behind head from source. Solid lines are rays.

To be effective, the wavelength of the sound must not be so long that the sound diffracts around the head with little reduction in amplitude. The disappearance of the sound shadow as the wavelength gets longer is a gradual change, so there is no exact prediction for this limit. Studies[57] have found that localization accuracy gets worse as the frequency drops from 3000 Hz to 2000 Hz. Thus, for this mechanism to work well, the wavelength must be shorter than roughly

$$\lambda_{\max} = \frac{s}{f_{\min}} = \frac{340\ \text{m/s}}{2500\ \text{Hz}} = 13.6\ \text{cm} \quad , \tag{168.1}$$

or about 80% of the diameter of a human head. This is consistent with the fact that the sound shadow is quite small when the wavelength roughly matches the obstacle size.

Localization accuracy also appears to get worse as the frequency rises above about 7000 Hz, for a wavelength shorter than about 5 cm. This may possibly be due to the sound shadow becoming complete at short wavelengths, so that sound is only heard in the near ear, and the auditory system can't compare two sounds. But it also may be due to a psychoacoustic difficulty in judging widely different loudness at the two ears.

Chapter 169. Low Frequency Directional Hearing

For a sound source that is to one side, the sound wave has to travel farther to get to one ear, so there is a time delay Δt between when any given feature of the sound wave reaches the two ears. The exact longer distance involves a path that curves around the head to the farther ear. However, for wavelengths much longer than one head diameter, diffraction of the sound leads the wave to be almost unaffected by the head, and it is adequate to measure the source-to-ear distances along straight rays, as is done in Figure 169.1.

A source direction directly to the side would give the maximum possible time delay, which for a typical size head would be

$$\Delta t_{\text{side}} = \frac{17\ \text{cm}}{340\ \text{m/s}} = 0.5\ \text{ms} \quad . \tag{169.1}$$

[57] Blauert, *Spatial Hearing,* 40.

Sources closer to the front-back axis would cause smaller time delays. Recall from Chapter 73 that the brain does receive information about when sound oscillation peaks arrive, which is the basis of the temporal theory of hearing. Our nervous system can use this binaural time delay to localize low frequency sounds.

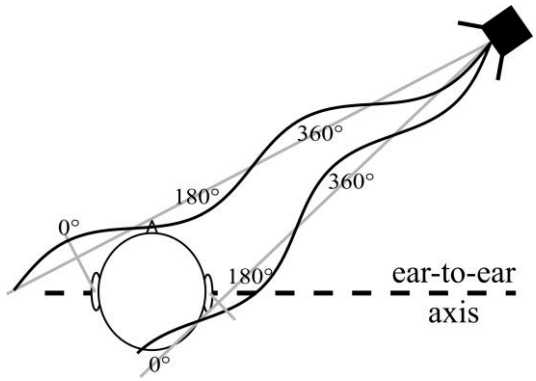

Figure 169.1
Snapshot of low frequency sound, causing phase difference between oscillations at the two ears.

As illustrated in Figure 169.1, for periodic sounds this can equally well be considered a phase difference $\Delta\phi$ between the oscillations at the two ears. In fact, the picture is rather similar to two-source interference, except that the two waves, traveling in the opposite direction, are in phase where they meet instead of at the separated ends. The relationship between time delay and phase difference is the same as Eq. 21.2,

$$\frac{\Delta t}{T} = \frac{\Delta\phi}{360°} \quad . \tag{169.2}$$

Although our auditory system probably does not detect the phase difference per se, the physics of phase differences does impose a limitation on this mechanism. Rewriting Eq. 169.2 as

$$f = \frac{\Delta\phi}{360° \, \Delta t} \quad , \tag{169.3}$$

shows that as the frequency increases, a given time delay implies a larger phase difference. But a phase difference larger than 180° becomes ambiguous. For example, a phase difference $\Delta\phi = 190°$ is indistinguishable from a phase difference of $\Delta\phi = 190° - 360° = -170°$. Those two phase differences would imply sounds coming from completely different directions, on opposite sides of the head. So, this mechanism can only function for frequencies low enough that the phase difference is less than $\Delta\phi_{\text{max}} = 180°$.

For a sound coming from the side, where the limitation is most severe because Δt is largest, this leads to a maximum frequency

$$f_{\text{max}} = \frac{1}{T_{\text{min}}} = \frac{\Delta\phi_{\text{max}}}{360° \, \Delta t_{\text{side}}} = 1000 \text{ Hz} \quad . \tag{169.4}$$

Psychoacoustic studies indeed have found that localization accuracy does get worse as the frequency rises above this threshold.

Combining this information with Chapter 168 reveals that there is frequency gap, roughly 1 kHz to 2.5 kHz, where neither the time delay mechanism nor the sound shadow mechanism works very well. Our auditory system seems to make do, however, as typical localization errors only increase by a factor of two or three in that frequency range.

Localization based on time delay actually can occur for higher frequency sounds as well, based on the very beginning of sound (its **attack**) arriving at the ears with a time delay. This is very similar to the mechanism described above, but it could be thought of as a distinct localization mechanism. In fact, when there are conflicting indications of a sound's direction, the first one that we perceive usually dominates (the **precedence effect**) and our subconscious ignores the others. So even though this mechanism is not available throughout a sound, it is often very important in determining perception of direction.

Chapter 170. Stereophonics

Stereophonic sound refers to the reproduction of sound with two (or more) speakers in such a way that listeners perceive directionality, or more generally a complete sound space. Explorations into this idea started in the late 1800s, not long after the invention of electrical speakers.[58] It was not until around 1960 that stereo phonographs and stereo radio became available to the general public.

The most direct way to achieve this is to use earphones to deliver signals to the two ears independently. In theory, this could exactly mimic the experience of being present at the original sound production. Interestingly, when simple sounds are generated and modified in accordance with the mechanisms described in Chapters 168 and 169, listeners often describe the apparent source location as being *inside* their head. This highlights the importance of the secondary mechanisms described in Chapter 167, especially the effects of the auricle on the sound spectrum. To get fully convincing earphone stereo, the easiest method is to record the sound using a manikin head with complete auricles and microphones in the ear canal positions.

More commonly, stereo recordings are intended for playback on two stationary loudspeakers a few meters apart, which here we will approximate at point sources. It is quite a complex undertaking to understand the degree to which this setup can reproduce the experience of hearing the original source. We can start by considering a listener positioned symmetrically in front of two speakers emitting identical sounds, as in Figure 170.1. Because the setup is symmetric, we can be sure that both ears experience the same loudness and phase, and the apparent source direction will be straight ahead.

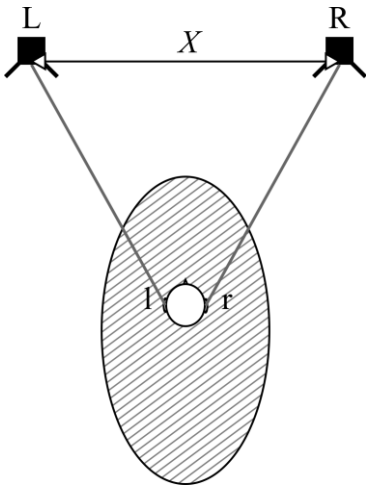

Figure 170.1
Listener located symmetrically with stereo speakers. Gray lines l & r are suggestive of earphone approximation. Hashed area has optimal stereo listening.

Let's additionally make the **earphone approximation,** which is that each ear is primarily influenced by the corresponding speaker (L to l and R to r). In this simplified model, the effects from Chapters 168 and 169 can be applied directly to adjust the apparent direction, but adjusting the relative loudness or phase from the two speakers. The listener's auricles provide the secondary cues that the sound source is towards the front.

What if the listener moves? If the head in Figure 170.1 rotates clockwise, for instance, ear r moves away from speaker R, and ear l moves closer to speaker L. Qualitatively, those are the same changes as if a single source were between the speakers, so the illusion is somewhat maintained. If the head slides to the right, however, problems arise. Ear r moves towards speaker R and ear l moves away from speaker L. Both the time-delay and loudness mechanisms would then predict that the apparent source direction would shift to the right of the listener.

In addition, the earphone approximation has relatively weak justifications. Sounds of roughly 100 Hz to 500 Hz, which are important for music, have wavelengths somewhat less than typical speaker separations. We know from Chapter 156 that they will combine to form interference patterns. This makes an analysis of the relative phases experienced by the two ears quite difficult. (Although at these frequencies, loudness differences do not contribute to sound localization, the interference can cause unwanted variations in loudness, both overall and differentially at the two ears.)

In the end, this is a question for which it is easier to test experience than to predict results based on theory. It is found that two-speaker stereophonic systems can be quite effective, as long as the listeners are in a

[58] "The Telephone at the Paris Opera," *Scientific American,* December 31, 1881, 422–423.

favorable region. This region is roughly an oval along the line bisecting the speakers, shown as a hashed area in Figure 170.1. If the distance between the speakers is x, then this oval runs between $x/2$ and $3x/2$ from the line connecting the speakers, and is about $x/2$ wide. Moving out of this area, the stereophonic effect degrades.

Chapter 171. The Back Wave Problem

For many types of speaker diaphragms, when the front side moves forward to compress the air just in front of the diaphragm, the back side (by moving in the same direction) rarifies the air just behind the diaphragm. As the diaphragm vibrates, the two effects on the air have exactly the same frequency, but they are out of phase.

A wave can spread from each of these sources, which this book will call the **front wave** and the **back wave.** If the two were allowed to combine while they are out of phase, that could result in destructive interference, thus reducing the amplitude of the sound leaving the speaker. There are a number of ways that speakers are designed to avoid this problem, or even benefit from the back wave.

Chapter 172. Avoiding Diffraction

When a speaker diaphragm moves forward in its oscillation, its purpose is to compress the air in front of it. If it moves too slowly, then instead of compressing, the air will simply flow around the edge. In fact, for many diaphragm types the backside of the diaphragm is moving in the same direction as the front side, and thus trying to rarify the air. The air can flow from the front around the edge to the back, and no sound is created. The question is, how slowly is too slowly?

It often happens in physics that there are several very different ways to approach a question. Each of them may be equally accurate, and yet one may be better suited to answer a particular question than another is. This is an example where a shift of perspective is helpful. According to Huygens' principle (detailed in Chapter 157), the creation of a wave at the surface of a flat diaphragm is equivalent to the transmission of a wave through a hole of the same size. This means that the sound wave emanating from the front of the speaker will show the same diffraction patterns as reviewed in Chapter 165.

Let's make the connection to diaphragm speed. A more slowly moving diaphragm means a lower frequency sound, which means a longer sound wavelength, which for a constant diaphragm size (equivalent to hole size) means moving further down in Table 165.1. For the upper rows of the table, with $a > 2\lambda$, the wave projects forward, so there is no interaction with the edge or backside of the diaphragm, and we are safe from the originally posed problem.

At the bottom of Table 165.1, where $a = \lambda$, the diffracted wave spreads over a full hemisphere to the diaphragm edge. We must now address two ways in which the speaker diaphragm is different from a hole. First, while a wave diffracting through a hole can't spread more than 90° because the wall is in the way, a wave diffracting from an isolated diaphragm can spread by more than 90°, wrapping around in the backwards direction. Second, we return to the fact that the back side of the diaphragm is simultaneously creating a second wave aimed in the backwards direction. At the position of the diaphragm itself, the two waves are exactly out of phase, compression of one happening with rarefaction of the other.

Putting this all together, for wavelengths that are longer than the size of the speaker (the diameter for circular diaphragms), the back wave diffracts around to the front and tends to cancel out the sound creation, because it is out of phase with the front wave. The longer the wavelength, the worse this effect will be. This is the diffraction way to understand the problem of air flowing from one side to the other. It is a more complex way to understand what is happening, but it allows us to answer the question of what frequencies are affected.

The final result is that small speakers can exhibit poor efficiency for sounds at frequencies low enough to have wavelengths longer than the speaker size. This is the reason that small speakers can sound "tinny." But there are ways to address this problem other than making the speaker larger. One way is to provide a **baffle,** a rigid panel or surface with the speaker set in the center. The baffle prevents the back wave from diffracting around to the front. Thus, it becomes the size of the baffle, rather than the size of the speaker, which determines the wavelength and frequency at which sound production starts to become a problem. If a speaker is placed in a wall, or in some other way that the back and front sides are completely separated, the wall can be considered an **infinite baffle.**

Along the way, the diffraction picture and Table 165.1 allow us to identify another potential problem. For higher frequencies with shorter wavelengths, the sound wave does not project to the side very well. What if you want to listen to the speaker without being directly in front of it? This is why speakers intended for high frequencies are often dome shaped instead of flat. This mimics a wave crest from a point course, helping to push the air in a wider range of angles. Yet, doing this can make the diffraction problem worse for lower frequencies.

Chapter 173. Eliminating the Back Wave

The back wave can be removed from the picture entirely by placing the diaphragm on the surface of a sealed box. Then the back wave cannot get out of the box to cause a problem. This arrangement is called an **acoustic suspension** (also called **air suspension**) speaker enclosure. The name comes from the fact that when the diaphragm vibrates, it must compress and rarify the air inside the speaker relative to its equilibrium volume. As described in Section 132b, the air thus provides much of the restoring force that pushes the diaphragm back towards its equilibrium position.

There still must be something solid that holds the moving parts of the speaker in the right position. But with the internal air providing much of the restoring force, the solid support can have a relatively low spring constant. This makes it possible to design these speakers to vibrate with large amplitudes, which as described in Chapter 126 is important for generating strong bass frequencies. The large amplitude design tends to result in a low efficiency for converting electrical power into sound power. But acoustic suspension speakers are quite popular because of their ability to realize a relatively flat response function in a compact size.

Chapter 174. Using the Back Wave

The wave coming from the back of the diaphragm could be put to use if it were made to be in phase with the front wave. The basic idea of the **acoustical labyrinth** speaker enclosure is to achieve this by sending the back wave through a tube that is close to one-half wavelength long, illustrated in Figure 174.1. By the time the wave has traveled that half-wavelength, it exits the enclosure in phase with the front wave and reinforces it.

The best labyrinth tube length L depends on what wavelength you want to assist. Speaker cones typically have poor response in the bass, so the labyrinth is chosen to assist at the lowest frequencies. Chapter 140 noted that superposition always increases amplitude if the phase difference is between $-90°$ and $+90°$. Therefore, the labyrinth tube length is chosen to be one-quarter of the longest wavelength to be boosted,

$$\lambda_{\max} = 4L \quad . \tag{174.1}$$

This back wave takes one-quarter of a period to traverse the labyrinth. Consider the back of the diaphragm to have a $+180°$ phase difference relative to the front. By the time the back wave comes out, it has fallen a quarter cycle behind, to be only $+90°$ ahead of the front wave.

The time required for sound from the diaphragm back to travel the labyrinth is always the same. But for slightly shorter wavelengths, the period of a cycle is slightly shorter, so that the travel time becomes more than one-quarter of a period. The back wave comes out less than $+90°$ ahead of the front wave. For sound with a wavelength half of the longest,

$$\lambda = \frac{1}{2}\lambda_{\text{max}} = 2L \quad , \qquad (174.2)$$

the back wave will exit the labyrinth in phase with the front wave. Constructive interference continues for wavelengths as short as

$$\lambda = \frac{1}{3}\lambda_{\text{max}} = \frac{4}{3}L \quad , \qquad (174.3)$$

for which the sound is delayed enough to exit $-90°$ behind the front wave.

And then the problems with this clever system show up. For even shorter wavelengths (with higher frequencies), how do we prevent the back wave from exiting out of phase with the front wave? In fact, there is a whole sequence of wavelengths,

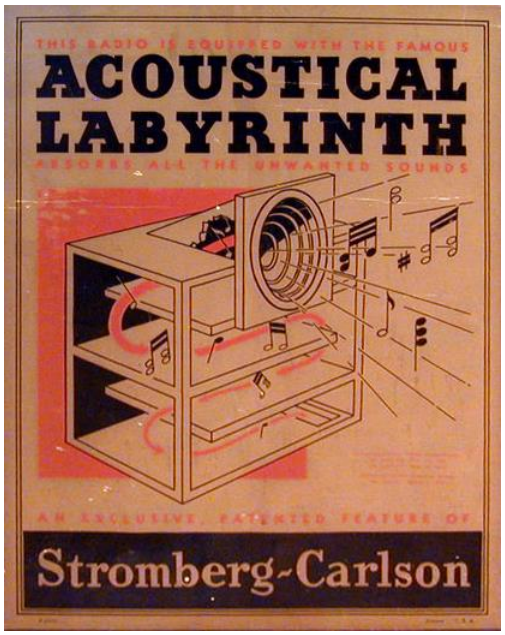

Figure 174.1
Label from 1939 radio illustrates its acoustical labyrinth technology.

$$\lambda = \frac{L}{n} \quad , \quad n = \text{positive integer}, \qquad (174.4)$$

for which a whole number of cycles fits in the labyrinth, so that the back wave exits in phase with the back of the diaphragm and out of phase with the front wave. To get around this, the material and shape of the tube walls are chosen so that shorter wavelengths tend to be absorbed, so that they exit the tube with much reduced amplitude. Part of the reason for the folded shape of the labyrinth tube is so that it fits in a reasonably sized box, but another part is that the twists and turns assist in this absorption of shorter wavelengths. Beyond that, the absorption of shorter wavelengths will not be described in this book.

Chapter 175. Bass Reflex Speakers

175a. Phase in a Driven Oscillator

Chapter 68 describes driven vibrations, where an object experiencing a restoring force is also shaken by an external force. The chapter described the resulting motion's frequency (the same as the external force's frequency) and amplitude (especially large if resonance occurs), but it skipped the third characteristic that describes any sinusoidal motion, the initial phase.

The driving force could have any initial phase. So instead of the initial phase of the object's oscillation, it is easier to describe the phase difference $\Delta\phi$ between the driving force and the oscillation.

$$\Delta\phi = \phi_{\text{object}} - \phi_{\text{drive}} \qquad (175.1)$$

Since the two have the same period, this phase difference will be the same for all time. As you can see in Figure 175.1, $\Delta\phi$ will always be less than zero, meaning that the object motion will always be at least a little behind the driving force. That's pretty reasonable, since the object is responding to the driving force. The standard terminology is that the object oscillation **lags** the driving force.

Recall the specific example of an object hanging from a spring which is held by a vertically moving hand. The amplitude of the hand's motion is to be kept at a constant x_{m0}. For very slow motion, meaning the frequency is nearly zero, the spring stays at its equilibrium length. That is why the object's amplitude matched the hand's amplitude x_{m0}. That also means that the two are moving in phase, $\Delta\phi = 0$.

If the driving frequency is increased towards the object-and-spring's natural frequency, the object's amplitude increases. A measure of that increase is the quality factor Q. As that happens, the object motion starts to lag the hand's motion, until at the resonance frequency the phase lag is one quarter cycle, $\Delta\phi = -90°$. This means that in resonance, at the moment when the hand is at its highest position, the spring is at its equilibrium position, about to be compressed by the upward-moving object.

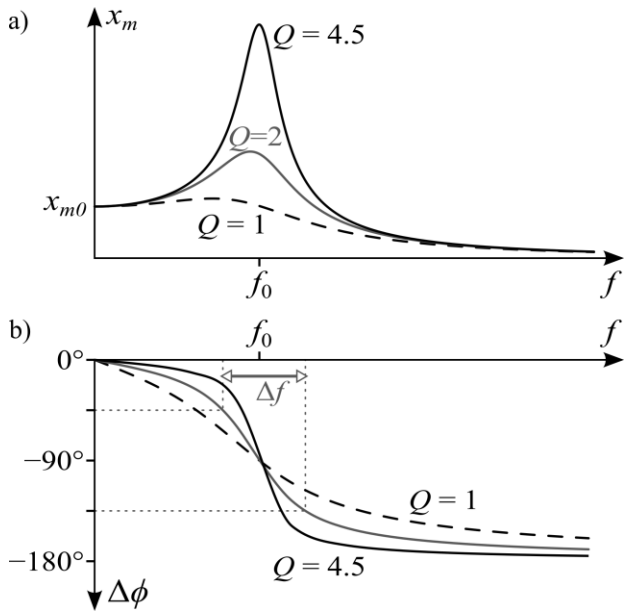

Figure 175.1

Amplitude and relative phase of driven oscillator versus driving frequency. Corresponding curves use the same line style.

If the driving force is sped up to an even higher frequency, then the object's amplitude decreases. In a sense, the object is unable to keep up with the rapidly vibrating hand. The phase lag continues to increase in magnitude, until at very high driving frequencies the object's small oscillations are nearly out of phase with the driving force. The object moves up when the hand moves down, and vice versa.

As shown by the various curves in Figure 175.1(b), how rapidly the phase difference changes with frequency depends on the quality factor. To characterize that, you can find the frequencies which cause a phase difference of $-45°$ and $-135°$, shown for the $Q = 2$ curve in Figure 175.1(b) by Δf. The separation of those two frequencies works out to be

$$\Delta f = \frac{2\pi}{Q} f_0 \quad . \tag{175.2}$$

The stronger the resonance, the more quickly the system switches between mostly in phase and mostly out of phase. The Δf interval is nearly centered on f_0 for systems with $Q > 5$.

175b. Bass Reflex Speaker Enclosure

The vibration of the back of the diaphragm could be put to use if it were made to be in phase with the front wave. Figure 175.2 shows a schematic of a **bass reflex speaker enclosure,** which uses the phase behavior of a resonance to achieve this.

The enclosure acts as a driven Helmholtz resonator, operating in the reverse sense of the one in Chapter 96. The parts in Figure 175.2 have been labeled to match those in Figure 28.1. The speaker diaphragm (shown as a cone) is mounted in a box which is sealed except for a cylindrical **port** or **vent,** which is a short tube open at both ends. That port, colored gray in the figure, is the neck of the Helmholtz resonator. When the backside of the diaphragm pushes and pulls on the enclosure volume, it tries to drive the Helmholtz resonator into oscillation.

Suppose that the speaker cone has an electromechanical response curve that is reasonably flat above some frequency f_c, but which has declining response below that frequency. In order to boost the response at

frequencies below f_c, the Helmholtz resonator is tuned to have a resonance frequency that is lower, $f_H < f_c$.

Let's consider the consequences, passing from high to low frequencies, from right to left in Figure 175.1. At frequencies well above the resonant frequency f_H (which is f_0 in Figure 175.1), the amplitude of the resonator is small, and the enclosure acts much like an air suspension enclosure. For decreasing frequencies, the phase difference between the air in the port (the mass of the Helmholtz resonator) and the back of the diaphragm shifts from $-180°$ towards $-90°$. But since the front and back of the diaphragm are exactly out of phase, this means that the phase difference between port and diaphragm front starts at 0° and increases towards $+90°$ at f_H. That is, sound from the port will constructively interfere with and reinforce the sound from the diaphragm with growing amplitude.

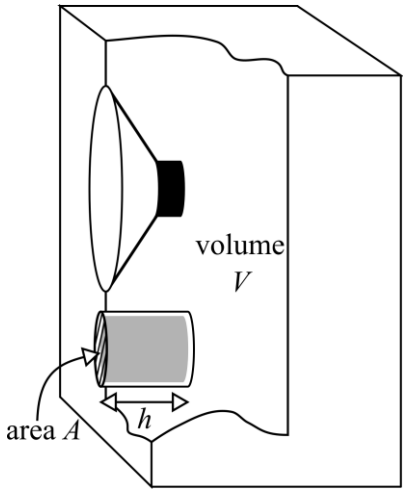

Figure 175.2

Cut-away picture of bass reflex speaker, combining speaker cone with open port, which together with cabinet form a Helmholtz resonator.

As the frequency drops to below f_c but above f_H, the resonance of the enclosure causes a larger amplitude port air vibration. Although the port and diaphragm front are not perfectly in phase, the phase difference increasing towards $+90°$ still results in reinforcement. This allows the bass reflex enclosure to extend good performance down to f_H.

Continuing to lower the frequency below the f_H, the left side of Figure 175.1(b) shows the phase of the port relative to the diaphragm changing from $-90°$ towards 0°. Relative to the front of the diaphragm, that puts the port at 90° increasing towards 180°, getting increasingly out of phase. The amplitude of the driven Helmholtz resonator is getting smaller, too. Not only is the electromechanical response is getting worse (because we are below f_c), destructive interference makes the response of the complete system even more worse. So, we pay for the better response between f_c and f_H by having really bad response below f_H.

Two factors limit how much this trick can help. First, although a strong enclosure resonance (high Q) would give a large boost to the response near f_H, it would do so over a narrow range of frequencies (see Eq. 175.2). Second, Chapter 99 explained that having a stronger resonance implies having longer ringing. Bass reflex speakers have a small time lag in the reproduced sound, which must not be too long. The Q must be kept moderate, not more than very roughly 5, both to boost a wider range of frequencies and to keep the time lag small. To do this, materials that absorb sound energy are intentionally included in the enclosure, which leads to rather low electroacoustic efficiencies.

Manageably sized bass reflex enclosures can be designed with f_H near the lower limit of human hearing. They do need to be a few cubic feet, so they aren't really portable, but they are smaller than acoustic labyrinth speakers. They have become a very popular choice for home sound systems since they were first invented in the 1960s.[59]

If you have read Chapter 174 about the acoustic labyrinth speaker enclosure, there is an interesting comparison between it and the bass reflex enclosure. A difficulty encountered with the acoustic labyrinth is that there is a set of wavelengths (and corresponding frequencies) for which the effect is the exact opposite of the intended goal. The bass reflex enclosure is easier to engineer because a Helmholtz resonator only has one special frequency, instead of many.

[59] A. N. Thiele, "Loudspeakers in Vented Boxes: Part I," *J. Audio Engineering Soc.* 19 (1971; reprint from *Proc. IRE Australia,* 1961): 382–392, and "Part II": 471–483.