# Sound Production and Modeling

**Perry R. Cook**
*Princeton University*

**This tutorial gives a brief overview of sound, describing sound as a physical phenomenon, computer representations of sound, and perception of sound by humans.**

Sound in multimedia, movies, games, virtual reality, and human–computer interfaces is a growing field that encompasses the disciplines of analog and digital signal processing, physics, speech, music, perception, and computer systems architecture. This overview of sound production and modeling techniques surveys the state of the art in sound technology.

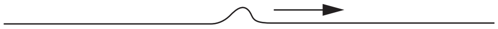
Sound has become a critical component in modern multimedia systems, especially multispeaker surround-sound entertainment systems. Many of the components and techniques in these systems are also applicable in virtual and augmented reality systems. With basic sound hardware now available for most (but not all, as in some palmtops) computer systems, and increasing processor speeds allowing direct real-time sound manipulation, enhancing multimodal computer–human interfaces by using the sonic channel is becoming commonplace. Sound in games and other graphics-intensive real-time applications is achieving higher levels of sophistication and promises to undergo even further advancements in realism and responsiveness.

## Sound waves in air and materials

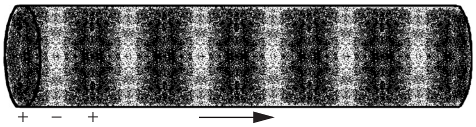
Sound is a wave phenomenon created by vibrations of physical objects or fluids. For any given medium, sound travels at a constant rate determined by the properties of that medium such as the density and bending/compression modulus. In stiff media (such as rigid bars and plates), the speed of sound is sometimes a function of the frequency of oscillation, with greater speed for increasing frequency. Some propagating sound waves are *transverse*, where the disturbance is orthogonal to the direction of propagation. This is evident in water waves, which displace the surface of the water up and down, yet travel perpendicular to the disturbance motion. Figure 1 shows a transverse wave in the wave traveling on a string. The transverse wave also occurs in many other solids such as membranes and plates. Other types of propagating sound waves are *longitudinal*, in which the disturbance is in the same direction as the propagation. This happens in air, as Figure 2 shows.

Waves propagate in physical objects, but the vibration can sometimes appear much differently to observers. For example, if we lift up a string under tension at the center and release it, it appears to flop up and down, rather than the appearance of waves traveling down and back along the string. Similarly, if we strike a drumhead in the center, it appears to flop up and down (like the string, but in 2D). If we were to pick up a string near the end, and if we could observe the vibration in slow motion, we might see the disturbance traveling down the string and back. It turns out that we can decompose any oscillation of a plucked string into two components—one going left and the other right on the string—and the observed standing-wave displacement is the sum of these two traveling-wave components (see Figure 3, next page). This is also true in most acoustical systems, such as a vibrating membrane, or the longitudinal oscillations in a tube filled with air.

The speed of sound in air is about 340 meters (1,100 feet) per second. This is an important number to remember when thinking about sound, especially when comparing the human auditory and visual systems. Because of the slow propagation speed, time and time delay are often the critical aspects of sound. For example, sound waves traveling down and back inside a trombone take about 0.005 seconds to complete their round-trip, causing the instrument (with the player's lip) to oscillate nat-

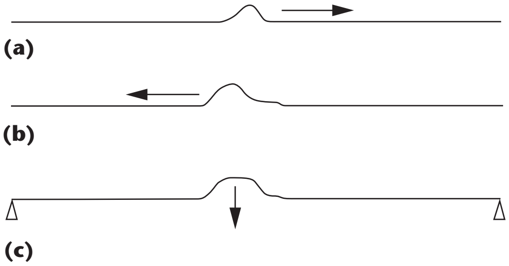


**1** Transverse wave on a string.

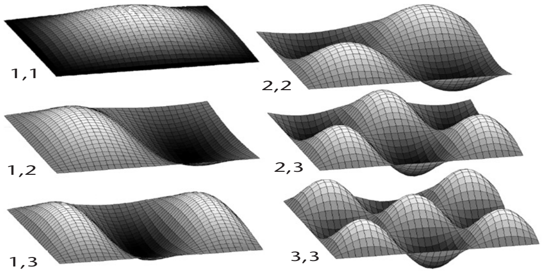


**2** Longitudinal waves in a pipe.

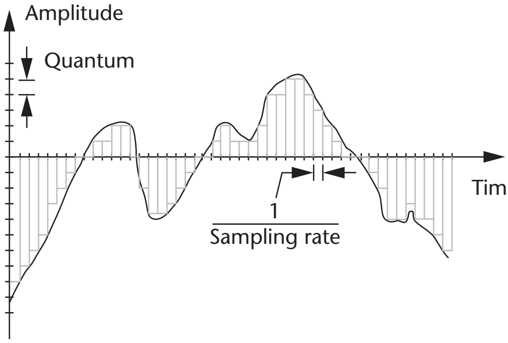
0272-1716/02/$17.00 © 2002 IEEE



3 (a) Left- and (b) right-going traveling-wave components sum to form (c) a composite displacement wave on string.

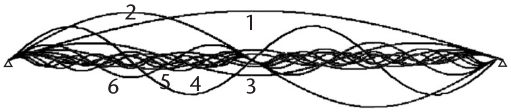


5 Modes of a vibrating rectangular membrane. The modes in the left column are labeled 1,1, 1,2 and 1,3 corresponding to the first, second, and third displacement modes in the horizontal (*x*) direction. Similar to the vibrating string in Figure 4, these modes are spatial harmonics, undergoing 0.5, 1.0, and 1.5 periods of a sinusoidal displacement function in the *x* direction, and all undergo 0.5 sinusoidal periods in the *y* direction. The right column shows the 2,2, 2,3, and 3,3 modes corresponding to the superimposed sinusoidal displacements in the *x* and *y* directions.



6 Sampling and quantization of a waveform.

urally at a low frequency of about 50 cycles per second. As another example, the reflections of our own sounds coming back from a wall 25 feet away are delayed by about 45 milliseconds. This might seem like an imperceptible delay, but as an exercise, go find a big isolated wall outdoors, step 25 feet away from it, and clap your hands a few times. Walk closer to the wall and clap, then farther away, and you'll be able to hear the difference in delay. To boggle your mind a little further, the last section will talk about delays on the order of microseconds that are perceptible to the human auditory system.



4 Modes of a vibrating string. Each mode is labeled with a harmonic number. The fundamental mode is number 1 and undergoes half of a sinusoidal excursion along the length of the string. The second mode undergoes one complete sinusoidal period along the string. The third  mode undergoes 1.5 periods of a spatial sine wave, and so forth.

## Modes of vibration

As I already mentioned, superimposed traveling waves can give the appearance and behavior of stationary standing waves. Another way to visualize the oscillation of systems such as strings, membranes, and enclosed tubes and spaces is to look at it as a superposition of standing sinusoidal modes of vibration. A simple definition for a system's modes is that they are that system's natural frequencies when it is excited and allowed to vibrate freely.

Figure 4 shows some of the sinusoidal modes of a vibrating string, and Figure 5 shows the first few vibrational modes of a struck rectangular membrane. Later, we'll talk about the Fourier transform, which lets us convert any shape or waveform into a sum of superimposed sinusoidal modes. Modes turn out to be an economical method to model and simulate some vibrating systems and are also important perceptually, as we'll discuss later.

## Digital sound

In computer systems, we capture, store, transmit, analyze, transform, synthesize, and play back audio in digital form. This means that to get an analog signal into the computer, we must sample it (measuring the instantaneous value) at regular intervals in time and then quantize it (rounding or truncating to the nearest digital number) to discrete values. The difference between quantization steps is called the *quantum*. The process of sampling a waveform, holding the value, and quantizing the value to the nearest number that can be represented in the system is an analog-to-digital (A to D, or A/D) conversion. Coding and representing waveforms in this manner is called pulse-code modulation (PCM). The device that does the conversion is an analog-to-digital converter (ADC, or A/D). The corresponding process of converting the sampled signal back into an analog signal is called digital-to-analog conversion, and the device that performs this is a DAC. Filtering is also necessary to reconstruct the sampled signal back into a smooth continuous time analog signal, and this filtering is usually contained in the DAC hardware. Figure 6 shows a waveform's sampling and quantization.

A fundamental mathematical law of digital signal processing states that if an analog signal is bandlimited with bandwidth *B*, we can periodically sample the signal at sample rate 2*B* and exactly reconstruct it from the samples. If components are present in a signal at frequencies greater than half the sampling rate, these components will not be represented properly and will alias as frequencies different from their true original values. So if we properly bandlimit signals before sampling them, we can exactly get them back, but not really. Because of quantization when the signal values are rounded or truncated, the small rounded differences between the original signal and the quantized signal are lost forever. A rule of thumb for estimating the noise introduced by quantization is 6*N* dB, where *N* is the number of bits we use to represent the signal. This means that a system using 16-bit linear quantization will exhibit a signal-to-quantization noise ratio of approximately 96 dB.

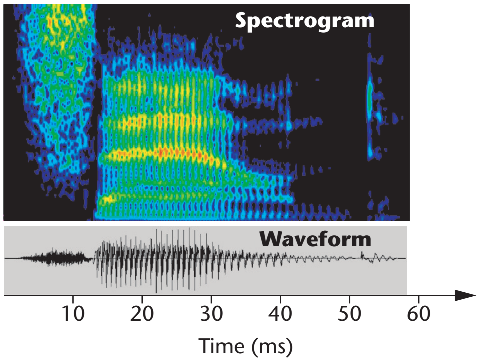
Young humans with normal hearing can perceive frequencies from roughly 20 Hz to 20 kHz, thus requiring a minimum sampling rate of at least 40 kHz. Speech signals are often sampled at 8 kHz or 11.025 kHz, while music is usually sampled at 22.05 kHz, 44.1 kHz (the sampling rate used on audio compact discs), or 48 kHz. Other popular sampling rates include 16 kHz (often used in speech-recognition systems) and 32 kHz. The maximum sampling rate available in most multimedia systems is 48 kHz, but new systems and standards are proposing or using sampling rates of 96 or 192 kHz. Most PC-based multimedia audio systems provide two or three basic sizes of audio words. Sixteen-bit data is common because this is the data format compact disc systems use. Eight-bit data is equally common and is usually used to store speech data. Twenty-four-bit data has recently become more popular.

## Perception of sound and the FFT

The importance of modes in modeling some vibrating systems isn't just a mathematical/geometric trick or convenience. In fact, a mechanism in our inner ear performs the function of turning the eardrum's time oscillations into frequency-dependent nerve firings in the brain. So in general, we create sound in the time domain but perceive many aspects of these objects and processes in the frequency domain.

Just as we can pass light through a prism to break it into the individual light frequencies from which it's composed, we can separate sound into individual simple frequencies (sine waves). A mathematical technique, called a *frequency transform*, uniquely converts a time domain waveform into a set of frequency components. The set of individual amplitudes and phases of the sines that make up a sound are called a *frequency spectrum*. Using the frequency spectrum to inspect aspects of a sound is called *spectral analysis*. The process of solving for the sine and cosine components of a signal or waveform is called Fourier analysis, or the *Fourier transform*. The digital version of the Fourier transform is the discrete Fourier transform (DFT), and a fast algorithm for computing it (if the length of the signal being transformed is a power of 2) is the fast Fourier transform (FFT).

A *spectrogram* (or sonogram) is a time-varying FFT plot showing time on the abcissa, frequency on the height axis, and intensity at each frequency in time shown as a brightness of color. Figure 7 shows a spectrogram of an utterance of the word sound. Figure 7 shows many things about sound perception as a function of the spectral properties. One is the noisy "sss" section from 0.0 to 0.01 seconds. This is characterized by the fuzzy energy at high frequencies and relatively less energy at low frequencies. Another thing to note is that we can see the individual harmonic (integer multiples in frequency) modes of the pitched voice source in the "ahh" and "ooo" and "nnn" sections from 0.012 to 0.05 seconds. We can also clearly see differences in the relative energy at different frequencies in these different vowel and liquid-consonant (phonetic name for the "nnn") sounds. Finally, we can see the stop consonant "d," followed by a noisy burst of air sound as the "d" is released.

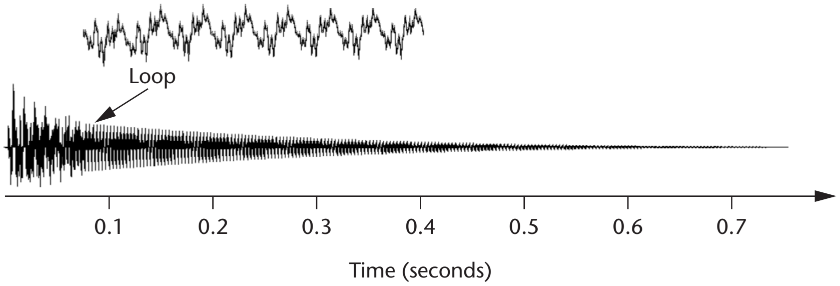


7 Spectrogram and waveform of utterance of the word sound.

## Synthesis: Waveform, spectrum, and physics

The majority of computer audio comes from the playback of stored PCM waveforms. Single-shot playback of entire segments of stored sounds is common for sound effects, narrations, prompts, musical segments, and so on. For many musical sounds, it's common to store just one loop, or table, of the periodic component of a recorded sound waveform and play that loop back repeatedly. This is called *wavetable synthesis*. For more realism, the attack or beginning portion of the recorded sound is stored in addition to the periodic steady-state part. Originally called sampling synthesis in the music industry, all synthesis involving stored PCM waveforms has more commonly become known as wavetable synthesis. Filters and envelopes (time-varying gain functions) are usually added to wavetable synthesis to control spectral brightness as a function of intensity and to get more variety of sounds out of a given set of samples. We can only pitch shift a given sample so far in either direction before it begins to sound unnatural. We deal with this by storing multiple recordings of the sound at different pitches, and switching or interpolating between these upon resynthesis—a process called *multisampling*. Figure 8 (next page) shows the initial attack of a plucked string, followed by the repetition of a short loop with time-varying decay to model the remainder of the sound.

Synthesis of signals by adding fundamental waveform components is called *additive synthesis*. Because we can uniquely represent any function as a linear combination of sinusoidal components, the powerful tool of Fourier analysis gives rise to a completely generic method of sound



**8** **Wavetable synthesis of a plucked guitar sound.**

analysis and resynthesis. When only a few sinusoidal components (a few natural modes of a system) exist, additive synthesis can be efficient and flexible. In this case, called *modal synthesis*, we can use sinusoidal oscillators or resonant filters to model the individual modes. However, many sounds have a significant number of components that vary rapidly in magnitude and frequency.

Many sounds yield well to subtractive synthesis, where we filter a complex source signal to yield a sound close to the desired result. The human voice is a subtractive synthesizer because the complex waveform produced by the vocal folds is filtered and shaped by the vocal tract tube's resonances. Thus, the voice is well modeled using subtractive-synthesis techniques.

Other synthesis techniques, such as frequency modulation (FM) and wave shaping, use nonlinear function transformations of simple waveforms (often sine waves) to create complex spectra. Other synthesis techniques exploit the statistical properties of some sounds (especially noisy sounds). Many systems support mixtures of techniques, and with the increase of software-based sound synthesis, we should expect to see more flexible systems in the future that use hybrid techniques.

Physical-modeling synthesis endeavors to model and solve the acoustical physics of sound-producing systems to synthesize sound. Unlike additive synthesis, which can use one powerful generic model for any sound, physical modeling requires a different model for each separate family of musical instruments or sound producing objects. One of the many potential benefits to physical modeling is the natural expressive control available when using a true physical simulation.

## Hearing sound in spaces

The human auditory system is remarkably adept at using only the (essentially 1D vibration) information at the eardrums to determine the locations of sound-producing objects. The strongest perceptual cue for left–right sound location is the time delay between our two ears. Remembering our earlier discussion of sound propagation speed, and assuming that the distance between our ears is about 9 inches, we could perform some simple experiments to convince ourselves that humans can discern interaural time delay cues of only a few microseconds. The next strongest cue for sound location is amplitude difference between the two ears. Other cues include filtering related to the shadowing effects of the head and shoulders as well as complex filtering functions related to the twists and turns of the pinnae (outer ears).

Three-dimensional audio processing and modeling endeavors to use headphones or speakers to place sources at arbitrary perceived locations around the listener's head. If we use headphones and synthesize the appropriate cues into the binaural signals presented to the two ears, we can directly manipulate virtual source locations. But if users move their head, the image of the virtual sonic world moves too. Head tracking is required to keep the virtual sonic world stable in the observer's real world. Using just a pair of stereo speakers requires more signal processing to cancel the effects of each speaker signal getting to both ears. VR systems using helmets with vision systems often have an integrated stereo audio capability, and we can use the head-tracking systems used to manipulate the visual displays in these systems to manipulate the audio displays as well.

Researchers have historically used multispeaker systems such as quadrophonic sound and Ambisonics in certain research and artistic settings, but other communities (such as entertainment and gaming) haven't broadly adopted them for various reasons, including economy, multiple competing standards, and lack of commercially available program material. Recently, various surround sound formats such as Dolby Pro-Logic, Dolby 5.1 Digital, DTS, AC3, and Sony Super-CD are entering the home-entertainment market, including dedicated game boxes. These new standards promise to give a multispeaker capability to many more multimedia systems owners, making immersive audio experiences in the home more pervasive. In addition to the movie and audiophile music titles that have initially driven the market for these new systems, we now see many new games and other multimedia content emerging that take advantage of multiple speaker systems.

## Conclusion

Sound pervades our lives, sometimes in invasive and irritating ways, but more often as an enhancement or complement to other sensory information. It's often said that sound guides the eyes, but we only need to close our eyes for a moment to experience the amazing variety of information that our ears provide and often more quickly and richly than any other sense. Active research areas in sound include

- efficient parametric algorithms for real-time synthesis of sound;
- exhaustive (but provably physical) synthesis methods;

- user interfaces for sound manipulation;
- sound at the user interface;
- modeling and rendering of sonic environments
- 3D sound;
- sound in virtual and augmented reality;
- immersive sound systems;
- sound in arts and entertainment; and
- verification and testing of synthesis, rendering, and interaction systems.

Most of these topics are essentially open-ended, with much work still remaining to be done in all areas of computational sound. Future hardware capabilities, algorithmic advances, and application areas remain to be researched and exploited. ■

***Perry Cook*** *is an associate professor in the Computer Science Department, with a joint appointment in the Music Department, at Princeton University. His main research areas are physics-based models for sound synthesis, human perception of sound, audio at the human–computer interface, and devices and systems for interactive sound control and artistic performance. He has a BA in music (studying voice and electronic music) from the University of Missouri at the Kansas City Conservatory of Music and a BSEE in engineering from the University of Missouri. He also has a PhD in electrical engineering from Stanford University, where he was technical director of the Center for Computer Research in Music and Acoustics.*

## Recommended Further Reading

Each of these works is what I consider the best, single book reference on its topic.

D. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, San Diego, 1994.

R. Bracewell, *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 1986.

P. Cook, *Music, Cognition, and Computerized Sound*, MIT Press, Cambridge, Mass., 1999.

P. Cook, *Real Sound Synthesis for Interactive Applications*, AK Peters, Natick, Mass., 2002.

K. Pohlmann, *Principles of Digital Audio*, McGraw-Hill, New York, 2000.

C. Roads, *A Computer Music Tutorial*, MIT Press, Cambridge, Mass., 1996.

C. Roads and J. Strawn, eds., *Foundations of Computer Music*, MIT Press, Cambridge, Mass., 1985.

J.O. Smith, *Digital Waveguide Modeling of Musical Instruments*, book in progress at http://www-ccrma.stanford.edu/~jos/waveguide/.

K. Steiglitz, *A Digital Signal Processing Primer, With Applications to Digital Audio and Computer Music*, Addison Wesley, Reading, Mass., 1995.

*Readers may contact Perry Cook at the Dept. of Computer Science, Princeton Univ., 35 Olden St., Princeton, NJ 08544-2087, email prc@cs.princeton.edu.*

For further information on this or any other computing topic, please visit our Digital Library at http://computer.org/publications/dlib.