How Digital Audio Works

A thorough explanation of how digital audio works is well beyond the scope of this manual. What follows is a very brief explanation that will give you the minimum understanding necessary to use MSP successfully.

For a more complete explanation of how digital audio works, we recommend *The Computer Music Tutorial* by Curtis Roads, published in 1996 by the MIT Press. It also includes an extensive bibliography on the subject.

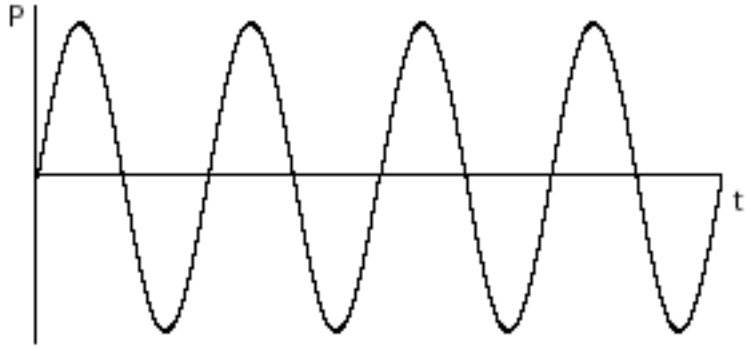
Sound

Simple harmonic motion

The sounds we hear are fluctuations in air pressure—tiny variations from normal atmospheric pressure—caused by vibrating objects. (Well, technically it could be water pressure if you're listening underwater, but please keep your computer out of the swimming pool.)

As an object moves, it displaces air molecules next to it, which in turn displace air molecules next to them, and so on, resulting in a momentary "high pressure front" that travels away from the moving object (toward your ears). So, if we cause an object to vibrate—we strike a tuning fork, for example—and then measure the air pressure at some nearby point with a microphone, the microphone will detect a slight rise in air pressure as the "high pressure front" moves by. Since the tine of the tuning fork is fairly rigid and is fixed at one end, there is a restoring force pulling it back to its normal position, and because this restoring force gives it momentum it overshoots its normal position, moves to the opposite extreme position, and continues vibrating back and forth in this manner until it eventually loses momentum and comes to rest in its normal position. As a result, our microphone detects a rise in pressure, followed by a drop in pressure, followed by a rise in pressure, and so on, corresponding to the back and forth vibrations of the tine of the tuning fork.

If we were to draw a graph of the change in air pressure detected by the microphone over time, we would see a sinusoidal shape (a *sine wave*) rising and falling, corresponding to the back and forth vibrations of the tuning fork.



Sinusoidal change in air pressure caused by a simple vibration back and forth

This continuous rise and fall in pressure creates a wave of sound. The amount of change in air pressure, with respect to normal atmospheric pressure, is called the wave's *amplitude* (literally, its "bigness"). We most commonly use the term "amplitude" to refer to the *peak amplitude*, the greatest change in pressure achieved by the wave.

This type of simple back and forth motion (seen also in the swing of a pendulum) is called *simple harmonic motion*. It's considered the simplest form of vibration because the object completes one full back-and-forth cycle at a constant rate. Even though its velocity changes when it slows down to change direction and then gains speed in the other direction—as shown by the curve of the sine wave—its average velocity from one cycle to the next is the same. Each complete vibratory cycle therefore occurs in an equal interval of time (in a given *period* of time), so the wave is said to be *periodic*. The number of cycles that occur in one second is referred to as the frequency of the vibration. For example, if the tine of the tuning fork goes back and forth 440 times per second, its *frequency* is 440 cycles per second, and its *period* is $^{1}/_{440}$ second per cycle.

In order for us to hear such fluctuations of pressure:

- The fluctuations must be substantial enough to affect our tympanic membrane (eardrum), yet not so substantial as to hurt us. In practice, the intensity of the changes in air pressure must be greater than about 10^{-9} times atmospheric pressure, but not greater than about 10^{-3} times atmospheric pressure. You'll never actually need that information, but there it is. It means that the softest sound we can hear has about one millionth the intensity of the loudest sound we can bear. That's quite a wide range of possibilities.

- The fluctuations must repeat at a regular rate fast enough for us to perceive them as a sound (rather than as individual events), yet not so fast that it exceeds our ability to hear it. Textbooks usually present this range of audible frequencies as 20 to 20,000 cycles per second (*cps*, also known as *hertz*, abbreviated *Hz*). Your own mileage may vary. If you are approaching middle age or have listened to too much loud music, you may top out at about 17,000 Hz or even lower.

Complex tones

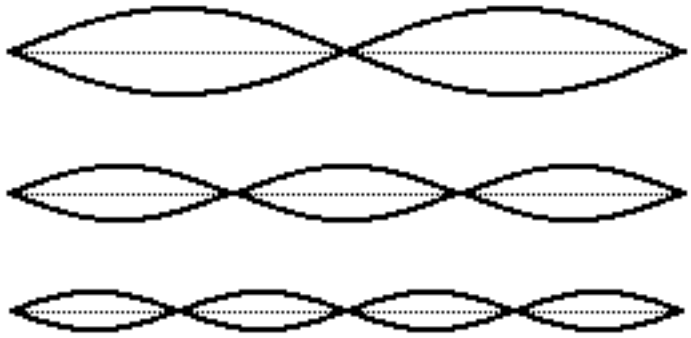
An object that vibrates in simple harmonic motion is said to have a resonant mode of vibration— a frequency at which it will naturally tend to vibrate when set in motion. However, most real- world objects have *several* resonant modes of vibration, and thus vibrate at many frequencies at once. Any sound that contains more than a single frequency (that is, any sound that is not a simple sine wave) is called a *complex tone*. Let's take a stretched guitar string as an example.

A guitar string has a uniform mass across its entire length, has a known length since it is fixed at both ends (at the "nut" and at the "bridge"), and has a given tension depending on how tightly it is tuned with the tuning peg. Because the string is fixed at both ends, it must always be stationary at those points, so it naturally vibrates most widely at its center.

A plucked string vibrating in its fundamental resonant mode

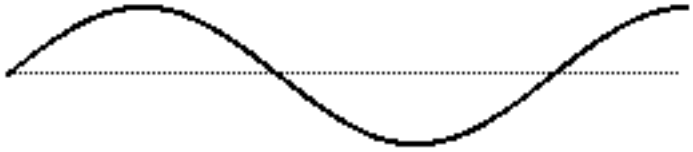
The frequency at which it vibrates depends on its mass, its tension, and its length. These traits stay fairly constant over the course of a note, so it has one fundamental frequency at which it vibrates.

However, other modes of vibration are still possible.



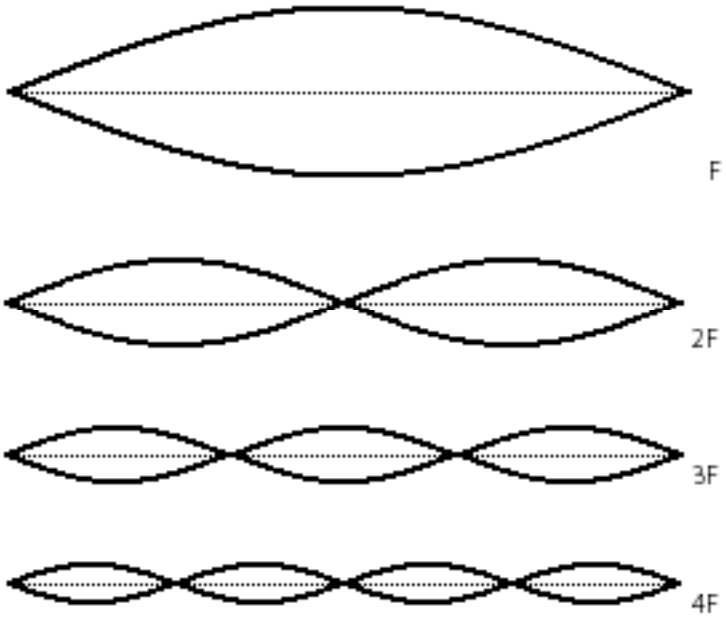
Some other resonant modes of a stretched string

The possible modes of vibration are constrained by the fact that the string must remain stationary at each end. This limits its modes of resonance to integer divisions of its length.



This mode of resonance would be impossible because the string is fixed at each end

Because the tension and mass are set, integer divisions of the string's length result in integer multiples of the fundamental frequency.



Each resonant mode results in a different frequency

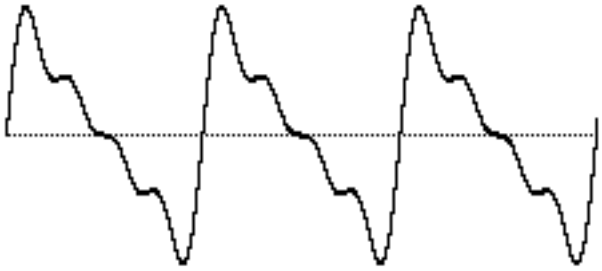
In fact, a plucked string will vibrate in all of these possible resonant modes simultaneously, creating energy at all of the corresponding frequencies. Of course, each mode of vibration (and thus each frequency) will have a different amplitude. (In the example of the guitar string, the longer segments of string have more freedom to vibrate.) The resulting tone will be the sum of all of these frequencies, each with its own amplitude.

As the string's vibrations die away due to the damping force of the fixture at each end, each frequency may die away at a different rate. In fact, in many sounds the amplitudes of the different component frequencies may vary quite separately and differently from each other. This variety seems to be one of the fundamental factors in our perception of sounds as having different *tone color* (i.e., *timbre*), and the timbre of even a single note may change drastically over the course of the note.

Harmonic tones

The combination of frequencies—and their amplitudes—that are present in a sound is called its *spectrum* (just as different frequencies and intensities of light constitute a color spectrum). Each individual frequency that goes into the makeup of a complex tone is called a *partial*. (It's one part of the whole tone.)

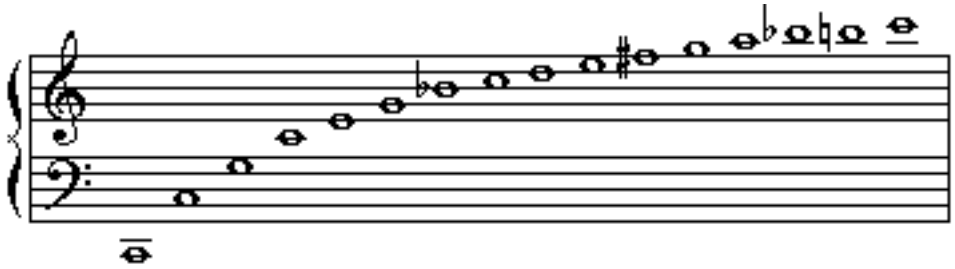
When the partials (component frequencies) in a complex tone are all integer multiples of the same fundamental frequency, as in our example of a guitar string, the sound is said to have a *harmonic spectrum*. Each component of a harmonic spectrum is called a *harmonic partial*, or simply a *harmonic*. The sum of all those harmonically related frequencies still results in a periodic wave having the fundamental frequency. The integer multiple frequencies thus fuse "harmoniously" into a single tone.



The sum of harmonically related frequencies still repeats at the fundamental frequency

This fusion is supported by the famous mathematical theorem of Jean-Baptiste Joseph Fourier, which states that any periodic wave, no matter how complex, can be demonstrated to be the sum of different harmonically related frequencies (sinusoidal waves), each having its own amplitude and phase. (*Phase* is an offset in time by some fraction of a cycle.)

Harmonically related frequencies outline a particular set of related pitches in our musical perception.



Harmonic partials of a fundamental frequency f, where f = 65.4 Hz = the pitch low C

Each time the fundamental frequency is multiplied by a power of 2—2, 4, 8, 16, etc.—the perceived musical pitch increases by one octave. All cultures seem to share the perception that there is a certain "sameness" of pitch class between such octave-related frequencies. The other integer multiples of the fundamental yield new musical pitches. Whenever you're hearing a harmonic complex tone, you're actually hearing a chord! As we've seen,

though, the combined result repeats at the fundamental frequency, so we tend to fuse these frequencies together such that we perceive a single pitch.

Inharmonic tones and noise

Some objects—such as a bell, for instance—vibrate in even more complex ways, with many different modes of vibrations which may not produce a harmonically related set of partials. If the frequencies present in a tone are not integer multiples of a single fundamental frequency, the wave does not repeat periodically. Therefore, an *inharmonic* set of partials does not fuse together so easily in our perception. We may be able to pick out the individual partials more readily, and—especially when the partials are many and are completely inharmonic—we may not perceive the tone as having a single discernible fundamental pitch.

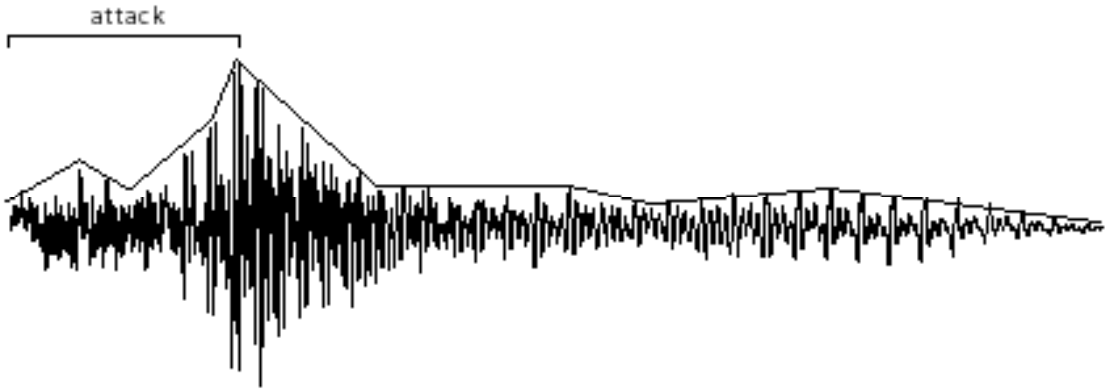
When a tone is so complex that it contains very many different frequencies with no apparent mathematical relationship, we perceive the sound as *noise*. A sound with many completely random frequencies and amplitudes—essentially all frequencies present in equal proportion—is the static-like sound known as *white noise* (analogous to white light which contains all frequencies of light).

So, it may be useful to think of sounds as existing on a continuum from total purity and predictability (a sine wave) to total randomness (white noise). Most sounds are between these two extremes. An harmonic tone—a trumpet or a guitar note, for example—is on the purer end of the continuum, while a cymbal crash is closer to the noisy end of the continuum. Timpani and bells may be just sufficiently suggestive of a harmonic spectrum that we can identify a fundamental pitch, yet they contain other inharmonic partials. Other drums produce more of a band-limited noise—randomly related frequencies, but restricted within a certain frequency range—giving a sense of pitch range, or non-specific pitch, rather than an identifiable fundamental. It is important to keep this continuum in mind when synthesizing sounds.

Amplitude envelope

Another important factor in the nearly infinite variety of sounds is the change in over-all amplitude of a sound over the course of its duration. The shape of this macroscopic over-all change in amplitude is termed the *amplitude envelope*. The initial portion of the sound, as the amplitude envelope increases from silence to audibility, rising to its peak amplitude, is known as the *attack* of the sound. The envelope, and especially the attack, of a sound are important factors in our ability to distinguish, recognize, and compare sounds. We have very little knowledge of how to read a graphic representation of a sound wave and hear the sound in our head the way a good sightreader can do with musical notation.

However, the amplitude envelope can at least tell us about the general evolution of the loudness of the sound over time.



The amplitude envelope is the evolution of a sound's amplitude over time

Amplitude and loudness

The relationship between the objectively measured amplitude of a sound and our subjective impression of its loudness is very complicated and depends on many factors. Without trying to explain all of those factors, we can at least point out that our sense of the relative loudness of two sounds is related to the ratio of their intensities, rather than the mathematical difference in their intensities. For example, on an arbitrary scale of measurement, the relationship between a sound of amplitude 1 and a sound of amplitude 0.5 is the same to us as the relationship between a sound of amplitude 0.25 and a sound of amplitude 0.125. The subtractive difference between amplitudes is 0.5 in the first case and 0.125 in the second case, but what concerns us perceptually is the ratio, which is 2:1 in both cases.

Does a sound with twice as great an amplitude sound twice as loud to us? In general, the answer is "no". First of all, our subjective sense of "loudness" is not directly proportional to amplitude. Experiments find that for most listeners, the (extremely subjective) sensation of a sound being "twice as loud" requires a much greater than twofold increase in amplitude. Furthermore, our sense of loudness varies considerably depending on the frequency of the sounds being considered. We're much more sensitive to frequencies in the range from about 300 Hz to 7,000 Hz than we are to frequencies outside that range. (This might possibly be due evolutionarily to the importance of hearing speech and many other important sounds which lie mostly in that frequency range.)

Nevertheless, there is a correlation—even if not perfectly linear—between amplitude and loudness, so it's certainly informative to know the relative amplitude of two sounds. As mentioned earlier, the softest sound we can hear has about one millionth the amplitude of the loudest sound we can bear. Rather than discuss amplitude using such a wide range of

numbers from 0 to 1,000,000, it is more common to compare amplitudes on a logarithmic scale.

The ratio between two amplitudes is commonly discussed in terms of *decibels* (abbreviated dB). A *level* expressed in terms of decibels is a statement of a ratio relationship between two values—not an absolute measurement. If we consider one amplitude as a reference which we call *A0*, then the relative amplitude of another sound in decibels can be calculated with the equation:

$$level\ in\ decibels = 20 \log10\ (A/A_0)$$

If we consider the maximum possible amplitude as a reference with a numerical value of 1, then a sound with amplitude 0.5 has $^1/_2$ the amplitude (equal to $10^{-0.3}$) so its level is

$$20 \log10\ (0.5/1) \_ 20\ (-0.3) = -6\ dB$$

Each halving of amplitude is a difference of about -6 dB; each doubling of amplitude is an increase of about 6 dB. So, if one amplitude is 48 dB greater than another, one can estimate that it's about 2^8 (256) times as great.

Summary

A theoretical understanding of sine waves, harmonic tones, inharmonic complex tones, and noise, as discussed here, is useful to understanding the nature of sound. However, most sounds are actually complicated combinations of these theoretical descriptions, changing from one instant to another. For example, a bowed string might include noise from the bow scraping against the string, variations in amplitude due to variations in bow pressure and speed, changes in the prominence of different frequencies due to bow position, changes in amplitude and in the fundamental frequency (and all its harmonics) due to vibrato movements in the left hand, etc. A drum note may be noisy but might evolve so as to have emphases in certain regions of its spectrum that imply a harmonic tone, thus giving an impression of fundamental pitch. Examination of existing sounds, and experimentation in synthesizing new sounds, can give insight into how sounds are composed. The computer provides that opportunity.

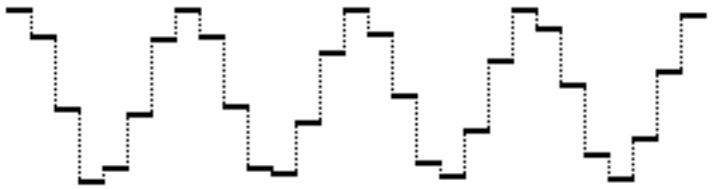
Digital representation of sound

Sampling and quantizing a sound wave

To understand how a computer represents sound, consider how a film represents motion. A movie is made by taking still photos in rapid sequence at a constant rate, usually twenty-four frames per second. When the photos are displayed in sequence at that same rate, it fools us into thinking we are seeing *continuous* motion, even though we are actually seeing twenty-four *discrete* images per second. Digital recording of sound works

on the same principle. We take many discrete samples of the sound wave's instantaneous amplitude, store that information, then later reproduce those amplitudes at the same rate to create the illusion of a continuous wave.

The job of a microphone is to transduce (convert one form of energy into another) the change in air pressure into an analogous change in electrical voltage. This continuously changing voltage can then be sampled periodically by a process known as *sample and hold*. At regularly spaced moments in time, the voltage at that instant is sampled and held constant until the next sample is taken. This reduces the total amount of information to a certain number of discrete voltages.



Time-varying voltage sampled periodically

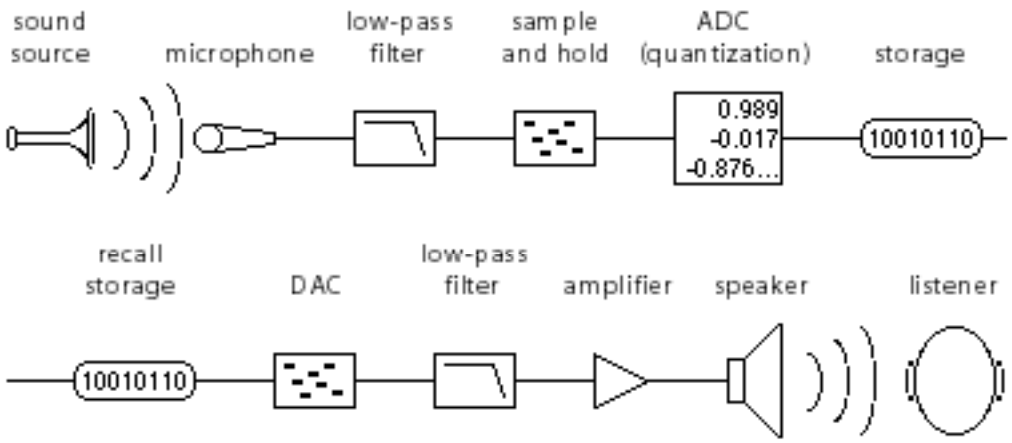
A device known as an *analog-to-digital converter* (ADC) receives the discrete voltages from the sample and hold device, and ascribes a numerical value to each amplitude. This process of converting voltages to numbers is known as *quantization*. Those numbers are expressed in the computer as a string of binary digits (1 or 0). The resulting binary numbers are stored in memory — usually on a digital audio tape, a hard disk, or a laser disc. To play the sound back, we read the numbers from memory, and deliver those numbers to a *digital-to-analog converter* (DAC) at the same rate at which they were recorded. The DAC converts each number to a voltage, and communicates those voltages to an amplifier to increase the amplitude of the voltage.

In order for a computer to represent sound accurately, many samples must be taken per second— many more than are necessary for filming a visual image. In fact, we need to take more than twice as many samples as the highest frequency we wish to record. (For an explanation of why this is so, see *Limitations of Digital Audio* on the next page.) If we want to record frequencies as high as 20,000 Hz, we need to sample the sound at least 40,000 times per second. The standard for compact disc recordings (and for "CD-quality" computer audio) is to take 44,100 samples per second for each channel of audio. The number of samples taken per second is known as the *sampling rate*.

This means the computer can only accurately represent frequencies up to half the sampling rate. Any frequencies in the sound that exceed half the sampling rate must be filtered out before the sampling process takes place. This is accomplished by sending the electrical signal through a *low-pass filter* which removes any frequencies above a certain

threshold. Also, when the digital signal (the stream of binary digits representing the quantized samples) is sent to the DAC to be re-converted into a continuous electrical signal, the sound coming out of the DAC will contain spurious high frequencies that were created by the sample and hold process itself. (These are due to the "sharp edges" created by the discrete samples, as seen in the above example.) Therefore, we need to send the output signal through a low-pass filter, as well.

The digital recording and playback process, then, is a chain of operations, as represented in the following diagram.



Digital recording and playback process

Limitations of digital audio

Sampling rate and Nyquist rate

We've noted that it's necessary to take at least twice as many samples as the highest frequency we wish to record. This was proven by Harold Nyquist, and is known as the *Nyquist theorem*. Stated another way, the computer can only accurately represent frequencies up to half the sampling rate. One half the sampling rate is often referred to as the *Nyquist frequency* or the *Nyquist rate*.

If we take, for example, 16,000 samples of an audio signal per second, we can only capture frequencies up to 8,000 Hz. Any frequencies higher than the Nyquist rate are perceptually "folded" back down into the range below the Nyquist frequency. So, if the sound we were trying to sample contained energy at 9,000 Hz, the sampling process would misrepresent that frequency as 7,000 Hz—a frequency that might not have been present at all in the original sound. This effect is known as *foldover* or *aliasing*. The main problem with