Acoustics for Musicians and Artists

Music 170/ICAM 103 course notes

DRAFT: November 24, 2014

Miller Puckette

These notes take the place of a textbook for the Fall 2013 offering of Musical Acoustics at UCSD (cross-listed as Music 170 and ICAM 103). The overall objective of the class is to provide both background and hands-on experience with sound as it appears in musical and artistic settings. Since most things to do with sound are now mediated by computers, computer audio techniques are foregrounded throughout these notes.

Examples in the book are realized using the Pure Data (Pd) software environment. Because it would take longer than the duration of this class to learn Pd well, a "patch library" of Pd functions ("objects") is provided here.

These notes will be growing as the course progresses. Eight chapters are planned, each one ending with an assignment with five or so pen-and-paper exercises and a project that can be carried out using Pd and the patch library.

Chapter 1

Sounds, Signals, and Recordings

Acoustics is the study of sounds, and for an artist or media researcher, the important things about acoustics might include: how to store and transmit records of sounds; how to use sound to sense things about the environment; how to generate synthetic sounds; or how to achieve a desired sound quality in an environment.

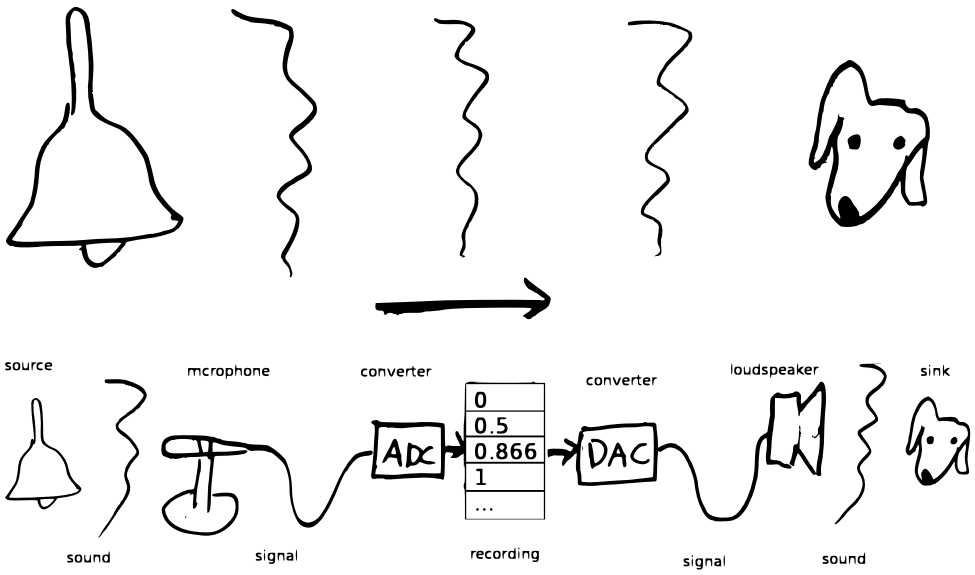
To be able to do things like this you'll need some understanding of how sounds behave in the real world. You'll also need to know something about how human hearing works, and a good bit about how to manipulate representations of sounds using computers.

1.1 Using Signals and Recordings to Mediate Sounds

Physically speaking, a sound is time-varying motion of air (or some other medium) with an accompanying change in pressure. Both the motion and the pressure depend on physical location. Knowing the pressure and motion at one point in the air does not inform you what the pressure and motion might be at any other point.

You can visualize sound this way:

Sounds can be mediated as signals, which in turn can be mediated as recordings. A *signal*, or, to be more explicit, an *analog* signal, is a voltage or current that goes up and down in time analogously to the changing pres-



sure at a fixed point in space. If we ignore for the moment any real-world limitations of accuracy, the analog signal provides an exact description of the time-varying pressure. (Usually an analog signal doesn't reflect the true pressure, but its deviation from the average atmospheric pressure over time, so that it can take both positive and negative values.) Mathematically, a signal may be represented as a real-valued function of time.

Analog signals can be *digitized* and *recorded* using a computer or other digital circuitry. A digital recording is just a series of numbers encoded in some digital representation. A single such number is commonly called a *sample*, although that term is often also used to describe a digital recording (such as you would play using a "sampler"), so here I'll try to remember to use the more precise term *sample point*.

Computer audio workflow usually goes along part or all of the chain of transmissions shown below:

The picture starts with a source emitting a real sound in the air. A microphone translates, in real time, the pressure deviation at a single point into an analog signal, encoded as a time-varying voltage. An *analog-to-digital converter* (ADC) converts the voltages to a digital recording (a series of numbers). These are no longer time-dependent; they may be stored in a file and accessed at a later time.

The rest of the diagram is the first part in reverse: first, a *digital-to-analog*

converter reads the stored numbers and regenerates a time-varying voltage; then a loudspeaker converts the voltages into a sound in the air.

This is the setup for computer-mediated sound manipulation today, used in recording, broadcasting, telephony, music synthesis, sound art, and many other applications. There might be more than one microphone and speaker, and while in digital form the recordings may be stored, combined with other recordings, moved from one place to another, or whatnot. In some situations, only the first or second part of the chain is needed; for instance, a digital keyboard instrument is essentially a computer that generates a recording and converts it into an analog signal.

In no situation is this setup capable of actually reproducing the sound that the bell emits. The microphone only measures the pressure at a single point in space, and the loudspeaker makes a new sound whose pressure variations at points close to the speaker are approximate reconstructions of the measured pressure at the microphone. But (in one way of thinking about it) the microphone does not distinguish among the infinitude of possible directions the sound it picked up was traveling. In theory we would need an infinitude of microphones to allow us to resolve that infinite number of possibly independently time-varying signals.

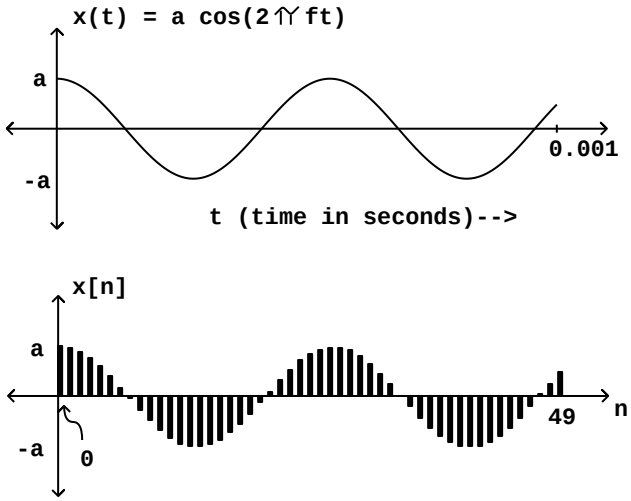
One other remark: although the recording in the middle of the diagram has no dependence on time, it is still possible to make the whole chain appear as if it is operating in real time, by quickly passing each arriving sample point (after processing it as desired) on to the DAC – perhaps 1/100 of a second or so after it is received from the ADC. That is how real-time audio processing software works.

1.2 Frequently Used Signals: Sinusoids and Noise

A *sinusoid* is a signal that changes sinusoidally in time, or its recording. As a signal (an analog function of time) it takes the form:

$$x(t) = a\cos(2\pi ft + \phi_0)$$

Here, the variable a is the sinusoid's *peak amplitude*, in other words, the amplitude ("bigness") of the signal at its peak. The variable f is the *frequency* in cycles per unit time. (If time is measured in seconds, then the frequency f is measured in cycles per second, also known as *Hertz*). The variable ϕ_0 (Greek letter phi) is the *initial phase*; the subscript 0 is there to indicate



taht we're talking about the phase at time zero, because we also use the word phase to mean the time-varying phase, equal to $2\pi ft + \phi_0$.

A sinusoid may be graphed like this (here the initial phase is zero and so isn't shown in the equation):

This signal cycles about 2.2 times in the 0.001 seconds shown; from this we can estimate that its frequency is about 2.2 cycles per 0.001 second, or, equivalently, 2200 Hertz.

The signal is *periodic*, i.e., it repeats the same thing over and over, potentially forever. The period is the number of seconds per cycle, and so it is the reciprocal of the frequency.

As a digital recording, the sinusoid we're looking at might be graphed like this:

Instead of a continuous function of time, we see a bar graph with 50 elements. (Alternatively, we could have printed out a list of 50 numerical values). There are 50 sample points per millisecond, or, equivalently, 50,000 sample points per second. We say that the recording has a *sample rate* of 50,000 samples per second, or to abuse language slightly, 50,000 Hz.

For either a recording or an analog signal, the frequency f and the period τ are related by: $f = 1/\tau$ or, equivalently, $\tau = 1/f$. The period is in time units and the frequency in cycles per time unit. In the case of a recording, one might specify time in either seconds or samples. If the sample rate is R samples per second, we may convert frequency or period from one to the

other. FOr example, the sinusoid above has a period of about 23 samples (you can count them). To learn the period in seconds, write

$$\tau = 23\text{samples}$$

$$= 23\text{samples} \cdot \frac{1\text{second}}{50000\text{samples}}$$

$$= \frac{23}{50000}\text{seconds}$$

and similarly the frequency is

$$f = 1/\tau = \frac{1}{23}\text{samples}^{-1} = \frac{50000}{23}\text{seconds}^{-1}$$

Here we can read a unit like seconds^{-1} equivalently as "per second" or, to be more explicit, "cycles per second" as we have been doing.

Here, for the record, is what a siuusoid might soind like:

SOUND EXAMPLE: a sinusoid, amplitude 0.1; frequency 1000 cycles per second (Hertz); 5 second duration.

The tone has an audible *pitch*, which is determined by its frequency. So the parameters a (the amplitude) and f (the frequency) correspond to audible characteristics of the sinusoid. Under normal conditions you won't hear the initial phase; indeed, if you tune into the sinusoid at some later point in time there will be a different initial phase but it's the same sinusoid with the same sound.

One other elementary signal type recurs throughout any study of acoustics, called *white noise*. As a recorded signal this is easy to describe: every sample point is a random number between $-a$ and a, where, again, a denotes the amplitude. (To be more pedantic, this is called uniform white noise to distinguish it from white noise whose sample points are chosen according to a Gaussian or other probability distribution; we won't worry about that here.)

SOUND EXAMPLE: uniform white noise, amplitude 0.1, 5 second duration.

Most people would not say that white noise has an audible pitch, and indeed it has no periodicity. White noise is also different from a sinusoid in that it is not deterministic; it is the result of a random process and if someone else generates a recording of white noise it most likely won't be equal to yours, although it should sound the same. So for instance if I added a sinusoid to

a recording of Beethoven's fifth symphony, and if you know its frequency, amplitude, and initial phase, you could subtract the sinusoid back out and recover the original recording; but if I added white noise you wouldn't be able to subtract it out unless I somehow sent you the particular recording of noise I had used.

1.3 Units of Pitch and Amplitude

Pitch is often described using logarithmic units (called *octaves*), for an exceedingly good reason: over the entire range of audible pitches, changing the pitch of a sound by an octave has a very uniform effect on the perceived pitch. Amplitude is also very often described in logarithmic units, called *decibels*, not because they are the best unit of loudness (that would be *sones*, to be discussed later) but rather because in sound engineering signals are often put through a series of operations that act multiplicatively on their amplitudes, and in such a situation it is convenient to deal in the logarithms of the amplitude changes so that we can add them instead of having to multiply. (Also, the range of "reasonable" amplitudes between just-audible and dangerously loud can reach a ratio of 100,000; one is immediately tempted to talk in logarithms just to be able to make reasonable graphs.)

Before we go any further I'll risk insulting your intelligence by reviewing logarithms, with musical pitch as the driving example. Choose, for the moment, a reference frequency equal to 440 Hz. We can raise it by octaves by successively doubling it, and lower it by halving it:

FREQUENCY	...	55	110	220	440	880	1760	3520	...
RATIO to 440		1/8	1/4	1/2	1	2	4	8	
OCTAVES		-3	-2	-1	0	1	2	3	

So if R is the ratio and I is the *interval* in octaves between 440 Hz. and our frequency f, the three are related as

$$R = f/440 = 2^I$$

$$f = 440 \cdot 2^I$$

or, solving for the interval I,

$$I = \log_2\left(\frac{f}{440}\right)$$

As we'll see in Chapter 4 (pitch and musical scales), it is customary in Western musical practice to use a different scale of pitches, measuring them in so-called *half steps*, defined as one twelfth of an octave:

$$H = 12 \cdot I$$

The choice of the reference pitch, 440 Hz, was arbitrary (although that particular frequency is often used as a reference.) If, for instance, we decided to use 220 Hz. as a reference our scale would then look like this:

FREQUENCY	...	55	110	220	440	880	1760	3520	...
RATIO to 440		1/4	1/2	1	2	4	8	16	
OCTAVES		-2	-1	0	1	2	3	4	
HALF STEPS		-24	-12	0	12	24	36	48	

The logarithmic scale of amplitude works similarly. We start by choosing a reference in the appropriate units, which could be, for example, one (for a sound recording), or one volt (for an analog electrical signal) or 0.00002 newtons per square meter (for a pressure deviation in air). Then we can measure the amplitude of any other signal, compared to the reference one, in decibels by an artificial construction that parallels the more natural way we dealt with pitch above. Taking the reference to be one with no units, the decibel scale is set up as shown:

AMPLITUDE	0.01	0.1	1	10	100	1000
DECADES	-2	-1	0	1	2	3
DECIBELS	-40	-20	0	20	40	60

In equations, the relative *level* L is related to the amplitude a by:

$$L = 20 \cdot \log_{10}\left(\frac{a}{1}\right)$$

Here we're explicitly dividing by one, the reference amplitude; if you use a different reference amplitude you should replace the "1" by that reference, the same way you did for the reference frequency, 440, in the calculation of pitch. To make the definitions as general as possible, it is customary to give the reference frequency and amplitudes names—f_{ref} and a_{ref}—to generalize the definitions of relative pitch and level as:

$$H = 12 \cdot \log_2\left(\frac{f}{f_{\text{ref}}}\right)$$

$$L = 20 \cdot \log_{10}\left(\frac{a}{a_{\text{ref}}}\right)$$

As with pitch, there are conventional choices for reference amplitudes, particularly for describing physical sounds, which we'll get to in chapter 6.

1.4 Word Size and Sample Rate of Recordings

Since the process of recording is essentially transcribing a continuous, real-valued function of time into a finite-sized array of digits, there is naturally a question of how much precision we will need in order to faithfully reproduce the analog signal we are recording. This has two aspects: first, what should be the *precision*, the number of binary digits we use to represent each individual sample point? And second, since we can only store a finite number of sample points per second, how many will be enough?

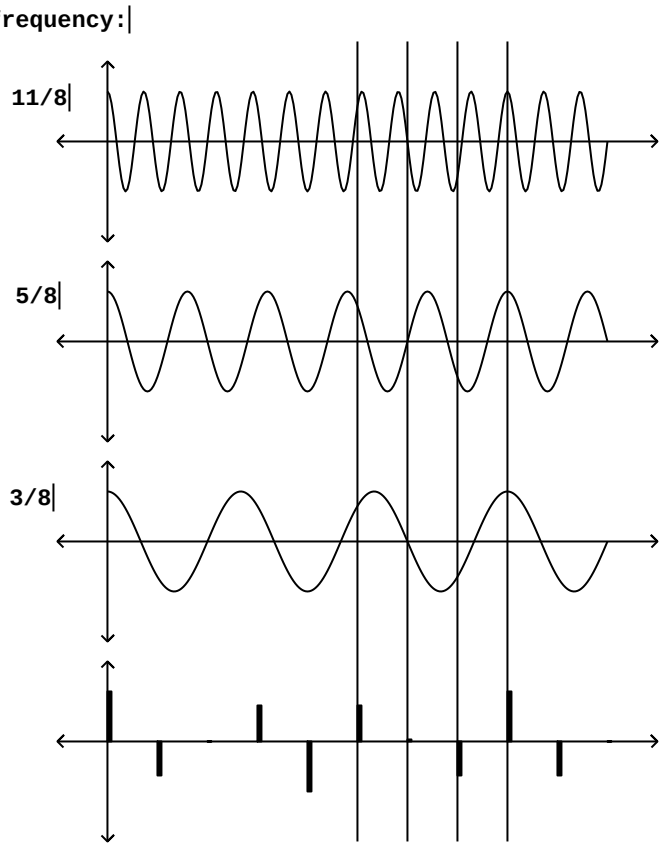
Precision is easily enough understood and decided upon. At stake here is a real number (with a range, for instance, of one volt) being transcribed as a binary number. The average error of the transcription is on the order of the least significant bit. If there are N bits, this is 2^N times smaller than the range of possible values the N-bit number can take. So the error, expressed in decibels with one volt as the reference amplitude, is:

$$L = 20\log_{10}\left(2^{-N}\right) = -N \cdot 20\log_{10}(2) \approx -6N$$

In other words, we get 6 decibels of precision for each additional bit we use to encode the sample points. Put another way, the *signal-to-noise ratio* (often abbreviated as SNR) is $6N$.

How much is enough? Well, for day-to-day work, 16 bits (for a SNR of 96 dB) should do it. For exacting situations or those in which you might have some a priori uncertainty as to the level of signal you are dealing with in the first place, it is often desirable to increase the precision further. In professional recording situations it is customary to use 24 bits for an SNR of 144 dB.

The question of sample rate is somewhat trickier. At first consideration one might suppose that a sample rate of 20 kHz should be adequate to represent any signal whose frequencies are limited to 20 kHz (usually considered the upper limit of human hearing). But this doesn't work. Suppose for simplicity that our sample rate is one (that is, one sample for each unit of time),

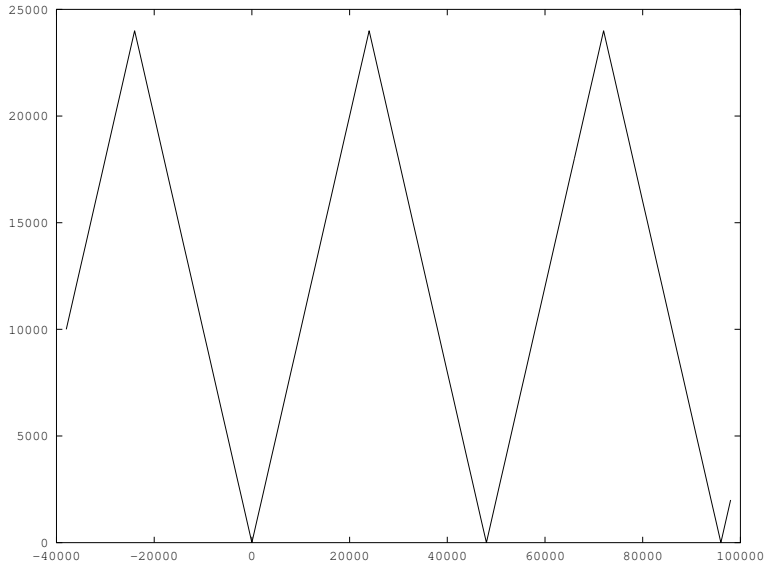


and suppose we want to record a sinusoid whose frequency is also one. Well, we'll sample the sinusoid at time points 0, 1, 2, etc., and... the instantaneous voltages at those time points will all turn out to be equal! We won't get any meaningful recording at all.

The picture below shows this effect in a slightly more general way: here again we set the sample rate to one, and consider the effects of sampling a sinusoid at frequencies of 3/8, 5/8, and 11/8:

Ouch: we get exactly the same recording from the three different sinusoids. This is an example of the phenomenon known as foldover. In general, any sinusoid whose frequency exceeds 1/2 the sample rate (that's called the *Nyquist frequency*) is exactly equal to another, lower frequency one.

So to represent sinusoids up to a frequency of 20 kHz, we need a sample rate of at least twice that. Because of various engineering considerations we will need an additional margin. The "standard" sample rates in widest use



in digital audio are 44100 Hz. (44.1 kHz), called the "consumer" or "CD" sample rate, or 48 kHz, called the "professional" one (although for various reasons people often record at higher rates still).

It is easy to come by a signal that is out of the range of human hearing, either electronically or physically; and it is also easy to write computer algorithms that generate frequencies above the Nyquist frequency as digital recordings. But once such a signal is recorded, standard playback hardware (DACs) will re-create them as the equivalent sinusoid, if any, that is within the range of human hearing, according to the following chart (which uses 48 kHz as the sample rate):

The phenomenon of synthesizing or recording one frequency and hearing another because of this ambiguity of frequencies in digital recordings is called *foldover*.

Recording sound at high quality can require a fair amount of memory; a six-channel, 48 kHz, 24-bit, one-hour recording would require over 3 gigabytes of storage, which might be inconvenient to store on one's computer and worse than inconvenient to share on a website, for example. For this reason much research has been done on data compression for audio recordings, and formats are now available that do an excellent job of reducing the size of a digital recording without changing the audible contents very much. These techniques would take many of pages of equations to describe, and anyway

most people don't seem to want to know how they work.

1.5 Fundamental Operations: Amplification, Mixing, and Delay

Once a sound is in the form of an analog signals or a digital recording, we can perform a variety of operations on it, three of which con be considered the most fundamental.

Amplification is the process of multiplying the signal or recording by a constant k. If $x(t)$ is a signal, the result is another function of time, $k \cdot x(t)$. If k is nonzero, and if (by any measure) the level of the original signal is L, the result will have a level equal to:

$$L + 20 \log_{10}(|k|)$$

If k is negative, the signal will also have been *inverted*. The constant k is called the *gain*. The change in level, $20 \log_{10}(|k|)$, is called the "gain in dB".

Mixing is the process of adding two or more signals. (The term is often enlarged to mean "amplifying them by various gains and then summing".) The result is usually louder than any of the signals to be summed, but not necessarily.

Delay refers to the process whereby a signal is replaced with an earlier copy of itself. Again using $x(t)$ to denote the original signal, choose a positive time value τ (Greek tau). The delayed signal is then $x(t - \tau)$. Note that we caon't apply a negative delay to a signal in time; that would amount to predicting the future (just tune into tomorrow's news).

The first two of these operations are applied to digital recordings in the same way as to real-time analog signals, but the last one, delay, is slighty different. If we have stored a recording of a sound, we can delay it by simply copying the numbers in the recording to the right or left (or up or down if you're imagining them vertically). The "delay" can be negative or positive. But note that, if you are trying to create the impression of real-time processing by continuously outputting the sample points soon after they arrive, you won't be able to shift them forward in time for the same reason you can't do that with an analog signal.

Exercises and Project

1. A recorded sinusoid has a sample rate of 48 kHz and a frequency of 440 Hz. What is its period in samples?

2. If a 1 volt amplitude signal is raised by 6 decibels, what's the resulting voltage?

3. What frequency is 1/2 octave above 440 Hz.?

4. If you record a signal with a word length of 8 bits, what is the theoretical signal-to-noise ratio?

5. If you generate a sinusoid of frequency 40 Khz, but only sample your sinusoid at a rate of 44.1 kHz, what frequency will you hear when you play it?

6. How many octaves are there in the human hearing range (between 20 and 20,000 Hz.)?

Project: *Why you shouldn't trust your computer's speaker.* In this project, you will determine your threshold of hearing as a function of frequency: that is, for each frequency, the minimum relative level at which you can hear whether a sinusoid is present or not. This is a generalization of you hearing range: outside your hearing range the threshold is infinite (no matter how loud you play the sound you won't hear it), but you would expect your ears to be somewhat less sensitive to the extremes than the middle as well.

This has been measured for "typical" young humans with increasing reliability and accuracy ever since a set of pioneering experiments in the 1930s by Fletcher and Munson; here is a good up-to-date article. The bottom curve in the graph is the "normal" threshold of hearing.

To do this yourself, get Pd and this patch library, following directions until you have verified that you can make "sinusoid" and "output" objects in a Pd document. All the patch you have to make is to connect a "sinusoid" to an "output".

Then, setting the "sinusoid" frequency to 1000, try one level (in dB) after another in the "output" object until you find a level at which you don't think you hear the difference when you toggle the sound on and off (using the toggle in the "output" object). This will be a crude process and you are unlikely to be sure to plus or minus 5 dB or so where the actual threshold lies. (It might also change with practice, or if you change you sitting posture, etc.) Use your computer speaker (if it has one), or headphones.

If, at 1000 Hz, you get a value above about 50, you might not be able to follow the curve as it rises at other frequencies; if so, try to turn your computer volume up so that you can turn the "output" control down lower.

Once you have this working at 1000, try other frequencies at octaves from it (going up: 2000, 4000, 8000, 16000; and going down: 500. 250. 125. 63. 31. 16) finding anew the threshold at each frequency and plotting it. If you can't hear it at all the answer is "infinite".

Graph this as best you can, and also write down whether you used the speaker on your laptop, or something else (headphones, stereo speakers, wires on the tongue, ...) How does the curve you got differ from the one in the link, and why might that be?

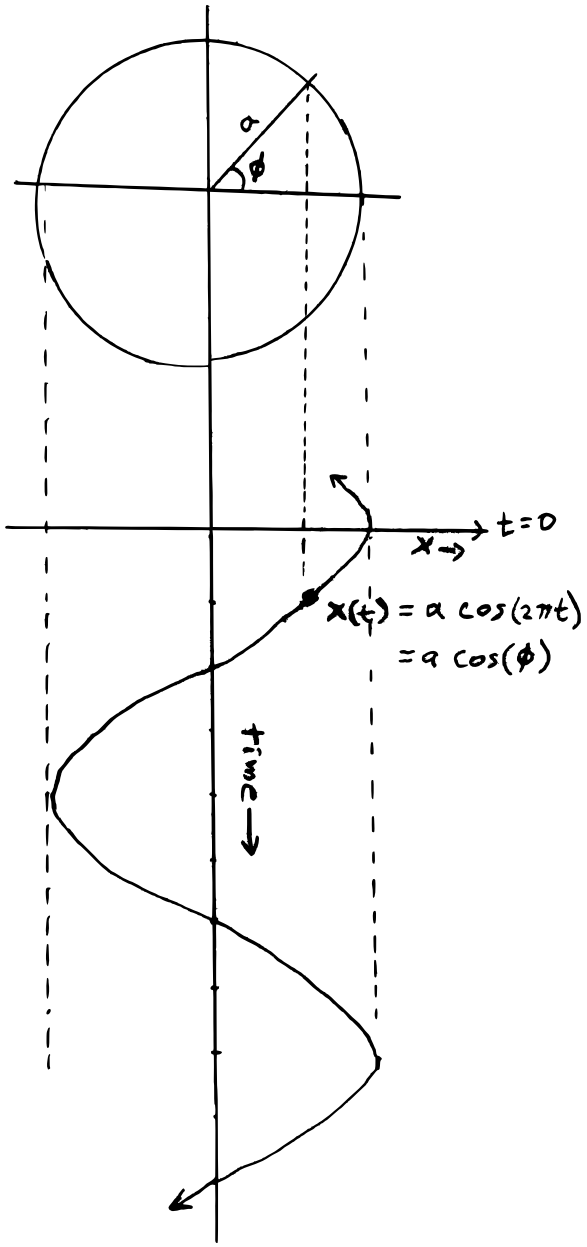
Chapter 2

Sinusoids

For several different reasons, sinusoids pop up ubiquitously in both theoretical and practical situations having to do with sound. For one thing, sinusoids occur naturally in a variety of ways, and if one happens to couple physically with the air and is of audible frequency and amplitude, we'll hear it. Second, sinusoids behave in simple and predictable ways when the elementary operations (amplification. mixing, delay; section 1.5) are applied to them. Third, one can add up sinusoids to make arbitrary signals or digital recordings (with some provisos having to do with convergence); this ability is extraordinarily useful for analyzing and synthesizing sounds.

2.1 Elementary Operations on Sinusoids

Here is a picture that might help visualize the mathematics of sinusoids. Imagine a point on the rim of a spinning bicycle wheel:



The progress in space of the point has horizontal (x) and vertical (y) components. If we forget the vertical component and graph just the horizontal component over time we get a sinusoid. If the point is initially at an angle

ϕ_0 from the x axis, we get the familiar formula:

$$x(t) = a \cdot \cos(2\pi f t + \phi_0)$$

where f, the frequency, is the number of revolutions per unit time, and the amplitude a is the radius of the wheel.

Now for the three elementary operations. First, amplification, say by a linear gain g, replaces $x(t)$ above with

$$g \cdot x(t) = ga \cdot \cos(2\pi f t + \phi_0)$$

If the gain is specified in decibels (say, g_{dB}), then we convert from decibels to a linear gain by applying the definition of decibels backward:

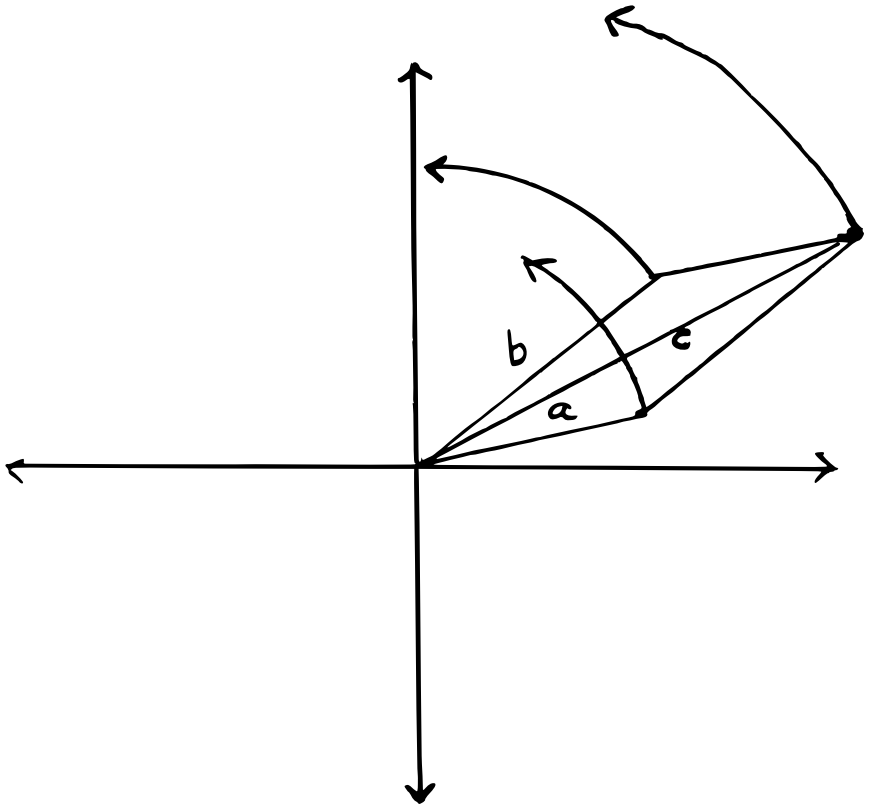
$$g = 10^{g_{dB}/20}$$

Applying a delay to a sinusoid equal to τ (or, if a recording, a time shift forward or backward by a positive or negative number equivalent to a time τ) has the effect of replacing t with $t - \tau$ in the formula:

$$x(t - \tau) = a \cdot \cos(2\pi f(t - \tau) + \phi_0) = a \cdot \cos(2\pi f t + (\phi_0 - 2\pi\tau f))$$

This leaves the amplitude a and the frequency f unchanged, but subtracts an offset $2\pi\tau f$ from the initial phase.

The effect of mixing two sinusoids (the third elementary operation) is more complicated. We'll start by supposing the two have equal frequencies (but not necessarily the same amplitudes or initial phases). Here is a picture:



The parallelogram represents the initial situation at time zero; the entire thing rotates about the origin as indicated by the arrows, without changing size or shape. If the initial phases of the two are ϕ_1 and ϕ_2, the angle between them is either plus or minus $\phi_2 - \phi_1$ and, by the law of cosines, we get

$$c^2 = a^2 + b^2 + 2ab \cdot \cos(\phi_2 - \phi_1)$$

(it doesn't matter which order ϕ_1 and ϕ_2 appear in the formula, since the cosine of the difference is the same either way). Depending on the phase difference, c may lie anywhere between $|b - a|$ (if $\phi_2 - \phi_1 = \pi$ so that the two sinusoids are exactly out of phase) and $a + b$ (if $\phi_2 - \phi_1 = 0$ so that they are perfectly in phase.)

The resulting initial phase depends in a complicated way on all of a, b, ϕ_1, and ϕ_2—the easiest way to compute it would be to convert everything to rectangular coordinates and back, but we will put that off for another day.

If the two frequencies are not equal—call them f and g—we can still apply the same reasoning, at least qualitatively. At time $t = 0$ we still get a

parallelogram, but now the two summands are rotating about the origin at different rates, so that the difference between the two phases, initially $\phi_2 - \phi_1$, is itself increasing or decreasing by a rate equal to the difference of the two component frequencies, that is, $g-f$. As a result, exactly $g-f$ times every unit of time, the parallelogram runs through its entire range of shapes and the resultant amplitude runs back and forth between its minimum and maximum possible values, $|b-a|$ and $a+b$.

If f and g differ by less than about 30 Hz., you can hear these changes in amplitude. This effect is called *beating*. At greater frequency separations you are likely to hear two separate tones, unless indeed they act as we'll describe in the next section:

2.2 Periodic and aperiodic tones

So far we have tacitly assumed that our ears can actually hear sinusoids as separate sounds, and that, presented with two or more sinusoids, we would be likely to perceive them as separate sounds. The truth is somewhat stranger: under the right conditions, our ears appear to have evolved to be able to distinguish periodic signals from each other, even if several of them with different periodicities are mixed together. (This is a good adaptation because it allows us to perceive the voices of other humans, which are approximately periodic most of the time, but rarely if ever sinusoidal.)

A signal is called *periodic* when, for some nonzero time duration τ, we have

$$f(t) = f(t+\tau)$$

for all t. We can apply this equation repeatedly to get:

$$\ldots = f(t-\tau) = f(t) = f(t+\tau) = f(t+2\tau) = \ldots$$

In other words, the signal repeats forever. Knowing the value of the function for one period, for example from $t = 0$ to $t = \tau$, determines the function for all other values of t.

If a function repeats after τ time units, it also repeats after 2τ, 3τ, ..., time units.. The smallest value of τ at which the signal repeats is called the signal's *period*.

A sinusoid whose frequency is f has period $1/f$. But an infinitude of other sinusoids repeat after $1/f$ time units. A sinusoid of frequency $2f$ has period

$1/2f$, and so repeats twice in a time interval lasting $1/f$. In general a sinusoid whose frequency is any integer multiple of f repeats (perhaps for the nth time) after an elapsed time of $1/f$. More generally, any signal obtained by amplifying and mixing sinusoids of frequencies that are all multiples of f will repeat after $1/f$ units of time, and therefore have a period of $1/f$ (if not some smaller submultiple of f).

Under reasonable conditions (f at least about 30; sinusoids at lower multiples of f having enough relative amplitude compared to the whole; no signal frequency other than f having an amplitude greater than the sum of all the others; at least some energy in odd-numbered multiples of f; etc.) we would hear such a mixture as a single tone whose pitch corresponds to f, which is then called the *fundamental frequency* of the mixture. The mixture will have the general form:

$$\begin{aligned} x(t) = a_1 \cos(2\pi f t + \phi_1) \\ + a_2 \cos(4\pi f t + \phi_2) \\ + a_3 \cos(6\pi f t + \phi_3) + \cdots \end{aligned}$$

only stopping, for a digital recording, at the Nyquist frequency, and possibly continuing forever for an analog signal.

Such a sum of harmonically sinusoids is known as a *Fourier series*, and although we won't prove it here, it's known that any "reasonable" periodic signal, (having a certain continuity property in time that any real signal should have) can be expressed as a Fourier series. Its digital recording can as well. This means that, in principle at least, you can synthesize any periodic signal if you can synthesize sinusoids.

The whole mixture is sometimes called a *complex periodic tone*, and the individual sinusoids that make it up are called *harmonics*. If all goes well, the perceived pitch of a complex periodic tone is that of its first harmonic, corresponding to the frequency f, which you can compute as in section 1.3.

It sometimes happens that a mixture of sinusoids that aren't collectively periodic somehow are perceived by the ear as a single entity (a tone) anyhow. Such a mixture could be written as:

$$\begin{aligned} x(t) = a_1 \cos(2\pi f_1 t + \phi_1) \\ + a_2 \cos(2\pi f_2 t + \phi_2) \\ + a_3 \cos(2\pi f_3 t + \phi_3) + \cdots \end{aligned}$$

and is called a *complex inharmonic tone*. The individual sinusoids that make it up are then called *partials* or *components*, and not harmonics - that term is reserved for the periodic case described earlier.

2.3 Special case: combining two equal-amplitude sinusoids

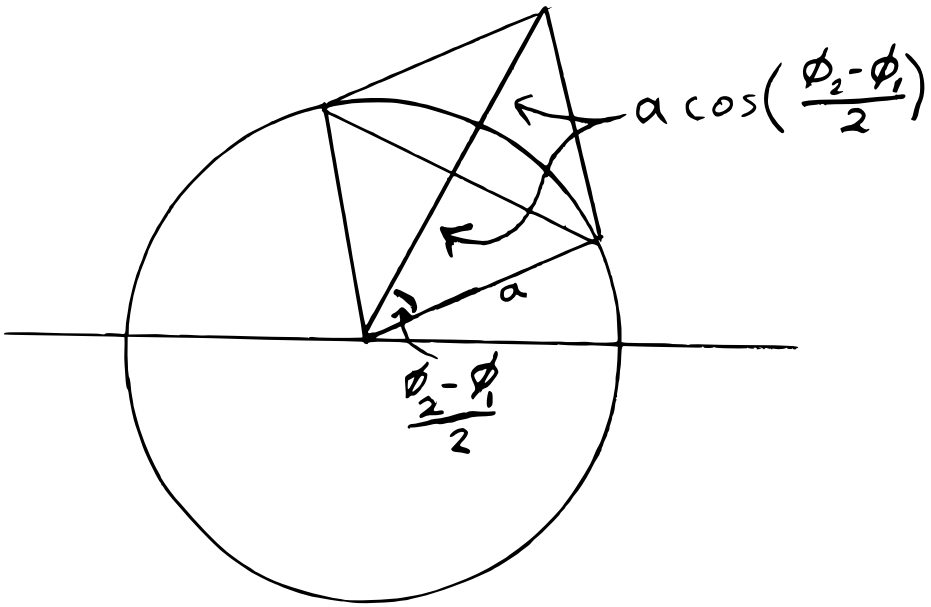
Suppose two sinusoids have the same amplitude a and frequency f, but different initial phases, ϕ_1 and ϕ_2. Our formula for the amplitude of the sum (from section 2.1) reduces to:

$$a_{\text{sum}} = a\sqrt{2 + 2\cos(\phi_2 - \phi_1)}$$

We can apply a standard trigonometric identity to get:

$$a_{\text{sum}} = 2a \cdot \cos(\frac{\phi_2 - \phi_1}{2})$$

This outcome is clear even if we don't remember that sort of identity; we can look at what the previous figure becomes when the two amplitudes are equal:



So not only is the amplitude increased (or decreased) by twice the cosine of half the phase difference; we also see that the initial phase of the resulting

sinusoid (which would have been complicated to calculate in general) is the average of the two initial phases ϕ_1 and ϕ_2.

As long as the amplitudes of the two sinusoids are the same, we can use the same picture to find the result of adding (mixing) two sinusoids of different frequencies f and g. To reduce clutter we'll leave out the initial phases to get the following formula:

$$a \cdot \cos(2\pi f t) + a \cdot \cos(2\pi g t) = 2a \cdot \cos(2\pi \frac{f-g}{2} t) \cos(2\pi \frac{f+g}{2} t)$$

This formula will recur often. I call it the Fundamental Law of Electronic Music, although perhaps that's overstating things a bit.

2.4 Power

Although the nominal (peak) amplitude of a sinusoid is a perfectly good measure of its overall strength, most signals in real life aren't sinusoids, and their peak amplitudes don't necessarily give a realistic measure of their strength. Also, you could wish to have a measure of strength that was additive, in the sense that, at least in good conditions, when you add two signals their measured strengths are added as well. The nearest thing we have to such a measure is the average *power*, which we will first motivate from physical considerations, then define, then show that it (at least sometimes) works the way we would wish.

The simplest way to motivate the definition of power is by considering a real-world analog, electrical signal. The amplitude (a function of time) is in this instance the time-varying voltage, customarily given the variable name V. We now suppose the signal is connected to a load of some sort, which has an electrical resistance R, measured in ohms. Power is voltage times current. To find the current I we apply Ohm's law to get:

$$I = V/R$$

and finally

$$\text{power} = V^2/R$$

We conclude that power, like amplitude, is a function of time; it is proportional to the square of amplitude. It is always either zero or positive.

Although we aren't ready to discuss real sounds in the air yet (we will be able to put that off until chapter 5 or perhaps even 6), the same reasoning will

apply. The amplitude is the (space-dependent) pressure. One can measure the power flowing through a specified area as follows: the pressure exerts a force on the area; as a result some air flows through the area, and the force times the velocity gives energy per second, which is the physical definition of power. The speed at which the air flows is proportional to the pressure (it's pressure divided by *impedance*)—a concept that generalizes resistance to describe "reluctance to move" in whatever medium we're talking about.) Power is then amplitude squared divided by impedance.

For digital recordings, we don't have a notion of physical impedance and so we just arbitrarily set it to one, giving

$$\text{power}(t) = [x(t)]^2$$

where $x(t)$ denotes the amplitude of the recorded signal. (Note that we're abusing notation here; recordings aren't functions of time, so t really stands for the time at which we mean to play the sample, or else the time at which we recorded it. The only true way to describe the variation of a recording is by talking about the memory addresses, or indices, of the sample points.)

So far we've only described *instantaneous power*, which is a time-varying function. The measure we're interested in is a signal's or recording's *average power*, which is simply the average, over some suitable period of time or range of samples, of the instantaneous power.

What is the average power of a sinusoid? Well, its square is

$$a^2 \cdot \cos^2(2\pi f t)$$

(we're dropping the initial phase which won't affect out calculation). Now use my Fundamental Law of Electronic Music with $g = 0$ (and omitting its own, arbitrary value of a):

$$a \cdot \cos(4\pi f t) + a = 2a \cdot \cos^2(2\pi f t)$$

(We omitted the $\cos(2\pi g t)$ term because $g = 0$ and the cosine of zero is one.) If we multiplied the right hand size by $a/2$ we would get the desired instantaneous power, so we multiply through by $a/2$ and swap sides, giving

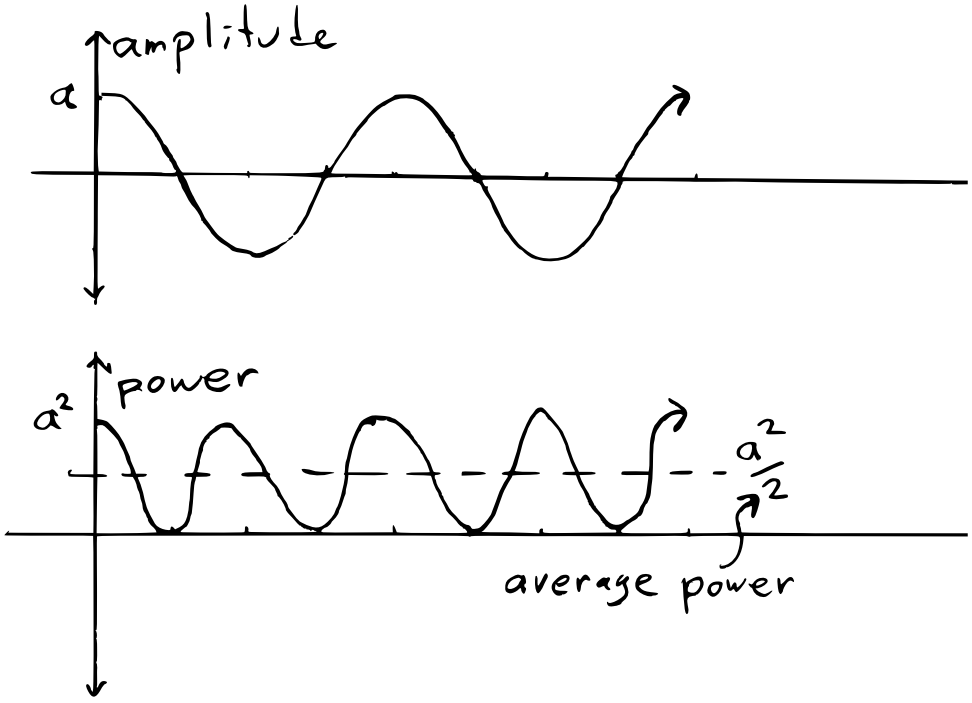
$$\text{power}(t) = \frac{a^2}{2}\cos(2\pi f t) + \frac{a^2}{2}$$

We want to know the average power. When we average the right-hand side of the equation, the cosine term averages out to zero, and so the average

power of the original sinusoid is given by:

$$(\text{average power}) = \frac{a^2}{2}$$

Here's what it looks like:



What happens when we add two sinusoids? Well, case one, they have the same frequency, and their amplitudes are a and b. Let c denote the amplitude of the resulting sinusoid (which will also have the same frequency). As we saw above, the three are related by the law of cosines:

$$c^2 = a^2 + b^2 + 2ab\cos(\phi_2 - \phi_1)$$

The three sinusoids have average power

$$P_a = \frac{a^2}{2}\ , \quad \text{etc}$$

so

$$P_c = P_a + P_b + ab\cos(\phi_2 - \phi_1)$$

About this we can at least say that, if we don't know what the relative phases of the two are, "on average" we expect the power to be additive because the cosine term is just as likely to be negative as positive.

Once again, we can deal with sinusoids of differing frequencies f and g by just letting the phase difference $\phi_2 - \phi_1$ precess in time at a frequency $|f - g|$. In this case the cosine term really does average out to zero no matter what the initial phases were. The power of the sum of the two sinusoids is the sum of the powers of the two summands.

In fact, the cosine term can be considered as the two sinusoids beating. If we want to measure the power accurately we must wait at least a few beats—the closer the two sinusoids are in frequency, the longer it will take our measurement to converge on the correct answer.

To calculate the average power of uniform white noise of amplitude a we have to do a quick integral; we get

$$P_{\text{noise}} = \frac{a^2}{\sqrt{3}}$$

Noise also has the property that it contributes power additively to a signal (as long as you don't add it to itself; see the next paragraph.)

It might seem that it is almost always true that adding two signals, with average power P_a and P_b, respectively, gives a signal of average power $P_a + P_b$; but beware the following counterexample: if you add a signal to itself you will double all its values and so the average power will be multiplied by 4, not 2. If you add a signal to its additive inverse (which has the same power as the original), the power of the sum will be zero. Also, if two sinusoids have the same frequency the average power of their sum will depend on the phase difference. There is a term for the situation in which you can simply add the average power of two signals to get the average power of the sum: such signals are said to be *uncorrelated*.

In general, scaling a signal (that is, multiplying all its values) by a factor of k scales the average power by a factor k^2, whereas accumulating k unrelated signals should be expected only to multiply the power by k on average.

2.4.1 Expressing Power In Decibels

In the previous chapter, we developed the notion of decibels for comparing the amplitudes of sinusoids. At that point we had no precise way to describe the amplitudes of signals in general, but now we do: by measuring their average power. If two signals have average power P_1 and P_2, their level

difference in decibels is:

$$L = 10 \log_{10}(\frac{P_1}{P_2})$$

We can quickly check that this is compatible with our earlier formula in terms of amplitude: two sinusoids of amplitude a_1 and a_2 would have average power a_1^2, a_2^2 and the above formula reduces to:

$$L = 10 \log_{10} \left((\frac{a_1}{a_2})^2 \right)$$

$$= 20 \log_{10}(\frac{a_1}{a_2})$$

so this new definition agrees with th eearlier one in section 1.3.

Exercises and Project

1. Two sinusoids with the the same frequency (440 Hz., say), and with peak amplitudes 2 and 3 are added (or mixed, in other words). What are the minimum and maximum possible peak amplitude of the resulting sinusoid?

2. Two sinusoids with different frequencies, whose average powers are 3 and 4 respectively, are added. What is the average power of the resulting signal?

3. Two sinusoids, of period 4 and 6 milliseconds, respectively, are added. What is the period of the resulting waveform?

4. Two sinusoids are added (once again)... One has a frequency of 1 kHz . The resulting signal "beats" 5 times per second. What are the possible frequencies of the other sinusoid?

5. A signal - any signal - is amplified, multiplying it by three. By how many decibels is the level raised?

6. What is the pitch, in octaves, of the second harmonic of a complex harmonic tone, relative to the first harmonic?

Project: comb filtering. In this project you will use the phase-dependent effect of combining two sinusoids to build the simplest type of digital filter, called a *comb filter*.

To start with, make a single sinusoid of frequency 100 Hz (using the sinusoid object in the course library for Pd). You can check the level of its output using the "meter" object; it should be about 97 dB.

Now put the sinusoid into a "vdelay" (variable delay) object, and connect the delay output as well as the original sinusoid output to the meter. When the delay is zero you should see something 6 decibels higher, about 103.

Now measure and graph the amplitudes you measure, changing the delay in ten steps from 0 to 0.005 seconds. (Hint: to make the graph readable, don't make the vertical axis linear in decibels; instead, perhaps make equal spaces for 0, 94, 97, 100, and 103). But if you really want a nice-looking graph and don't mind 5 extra minutes of effort, convert from decibels to power.

Now do the same thing (on the same graph with a different color or line style) with the sinusoid at 200 Hz. instead of 100 Hz. Do you see a relationship between the two?

Now put six sinusoids at 100, 200, 300, 400, 500, 600 Hz. into a "switch" object (that's primarily for convenience; connecting the six to the switch will add them.) Connect the switch output to both the delay and directly to the output as before. As you change the delay between 0 and 10 milliseconds, what do you hear? What special thing happens when you choose a 5 millisecond delay?

Chapter 3

Spectra

As we saw in Section 2.1, the three fundamental operations on signals have the property that if their inputs are sinusoidal (and if there is more than one input, if they share the same frequency), then the output is also a sinusoid at the same frequency. The only things that might change are the sinusoid's amplitude and/or initial phase.

It's also true of many physical objects or systems that, if you apply a sinusoidal force at one point and measure the resulting motion at another, you'll also get a sinusoid of the same frequency, and that the output amplitude is proportional to the input amplitude.

This is at least one reason (perhaps the primary reason) why sinusoids are important: if you can describe a signal in terms of sinusoids, and if you know that some operation or other acts on sinusoids only by changing their amplitudes and phases, then you might be able to find what the system will do to your overall signal.

In general, a description of the makeup of a signal as sum of sinusoids is called a *spectrum*, and is the subject of this chapter. In the next section we'll try to make a more precise description of two possible definitions for the spectrum of a signal. In section 3.2 we'll consider how systems that preserve sinusoids—specifically, *filters*—affect the spectra of signals they are applied to.

The spectrum of a signal can be related to what we hear as the signal's *timbre* (a catch-all term that just means what a thing sounds like), and so operations (such as filters) that have predictable effects on signals' spectra

can be very useful in synthesizing and processing sounds. In section 3.3 we'll take up one example of this, called *subtractive synthesis*.

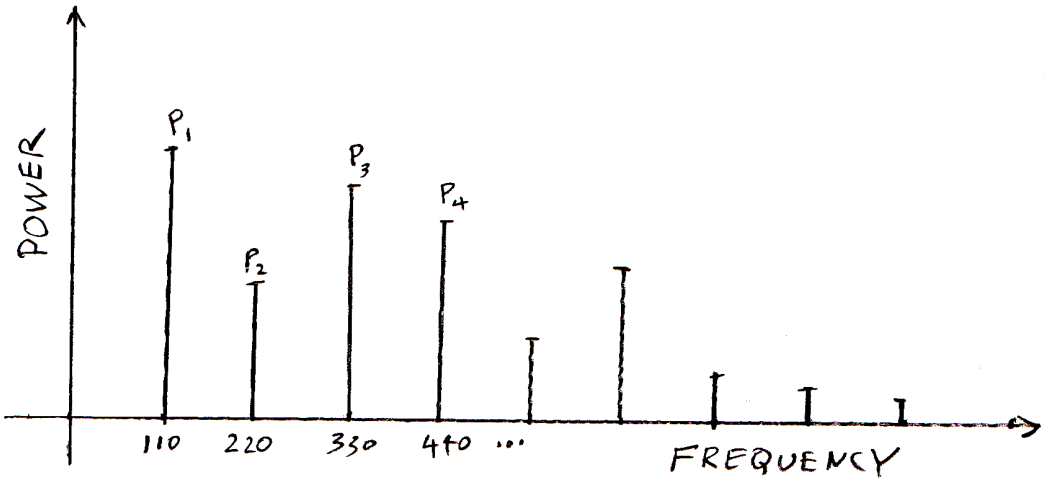
3.1 Definitions and Examples

In the field of acoustics, one sees at least two important types of spectra: real ones (usually obtained by making measurements on a digital recording) and idealized ones (that might arise, for example, in theoretical analyses of various systems or be specified out of thin air for compositional or other reasons). In either case, a spectrum can be thought of as a graph whose horizontal axis shows frequency and whose vertical axis shows the relative strength of a signal (or other thing) at each frequency.

Real, measured spectra are usually (perhaps always) represented as continuous functions of frequency. Idealized spectra are often protrayed as using only a discrete set of frequencies; we will look at this situation first.

3.1.1 Discrete Spectra

Suppose for example we either have, or want to generate, a periodic signal whose fundamental frequency is 110 Hz. In section 2.2 we claimed (without proof and with some waving of hands about continuity requirements) that such a signal could be written as a sum of sinusoids with frequencies 110, 220, 330, and so on. Each of these components has its own amplitude and initial phase. In situations where we don't care about the phase, we can represent the signal's Fourier series graphically, for example like this:



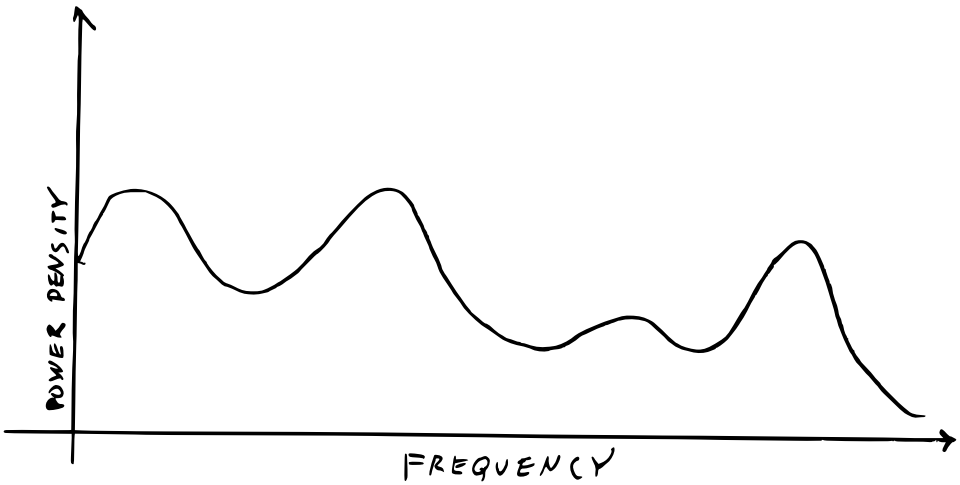
Here the numbers P_1, ..., represent the average power of the sinusoidal components. (Alternatively we could specify their peak amplitudes since the two are related by $P = a^2/2$. I chose power instead of amplitude to make clear the parallel between this and the following picture.)

Such a spectrum is called *discrete* because all the power is concentrated on a discrete set, that is, a set containing finite number of points per unit of frequency. The example here is furthermore a *harmonic* spectrum, meaning that the frequencies where there is power are all multiples of a fundamental frequency that is within the audible frequency range (in this case, 110 Hz). This is the spectrum of a complex periodic tone (section 2.2).

A discrete spectrum could also describe a complex inharmonic tone, in which case we say that the spectrum, too, is inharmonic.

3.1.2 Continuous Spectra

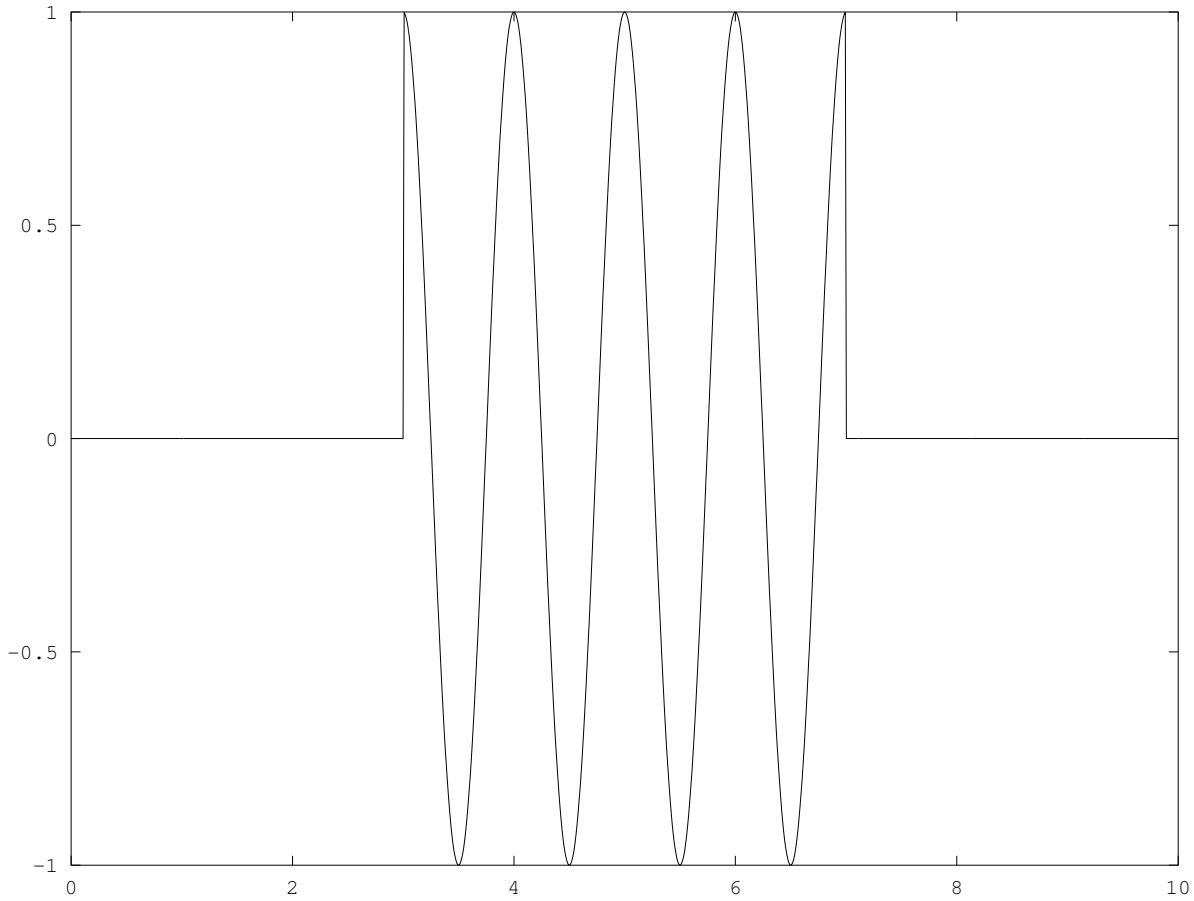
Signals or recordings that occur in nature never have discrete spectra; their spectra are *continuous* functions of frequency. A signal's continuous power spectrum might look as shown:



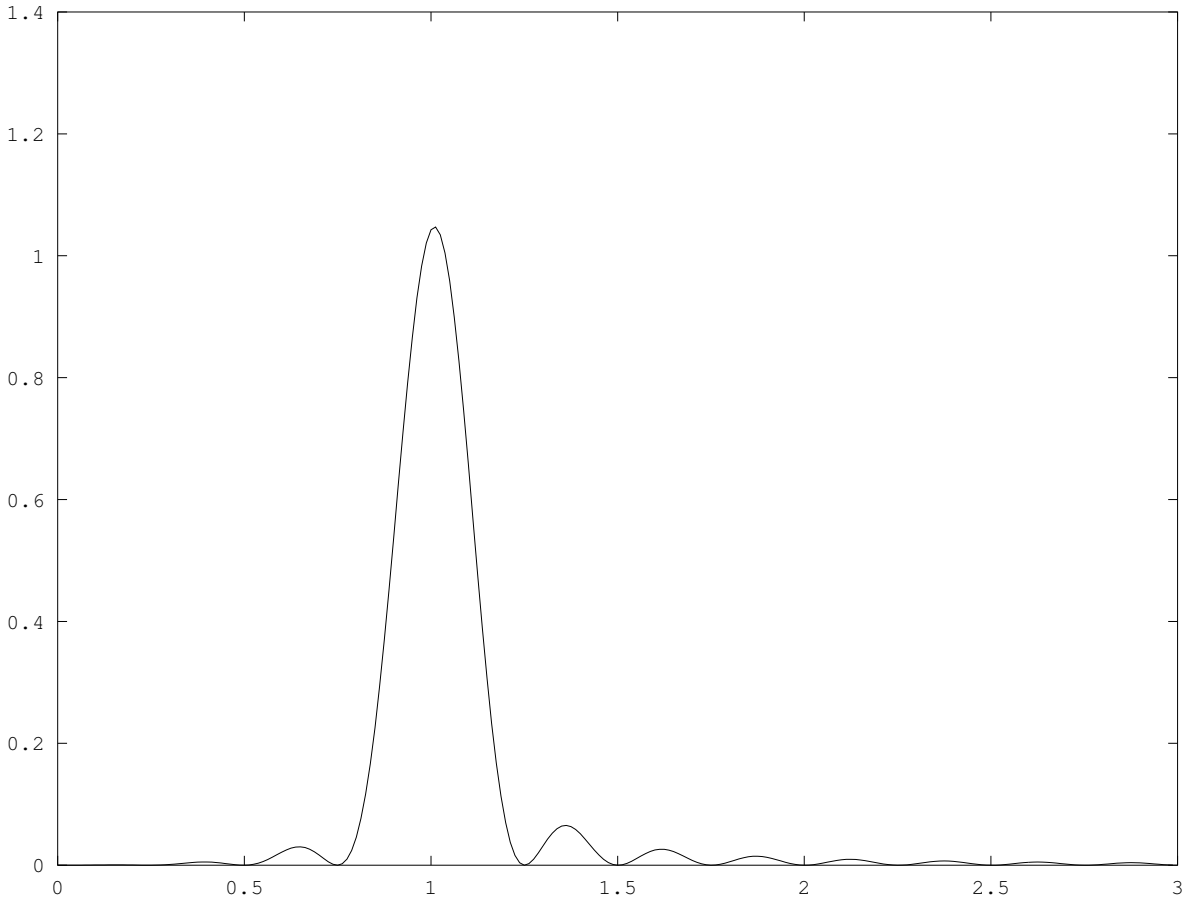
Continuous power spectra can be (and perhaps usually are) measurements of a real signal over a finite period of time (for a signal) or a finite number of sample points (for a recording). A continuous power spectrum has a physical meaning: the area under the curve over a range of frequencies (say from f_1 to f_2 is the total average power of the signal between those two frequencies. The area under the entire curve (from zero frequency to the highest possible one) is the total average power of the signal.

To put this another way: the continuous power spectrum describes how the average power of the signal is distributed over frequencies. Its units are power per frequency (for instance, watts per Hz.).

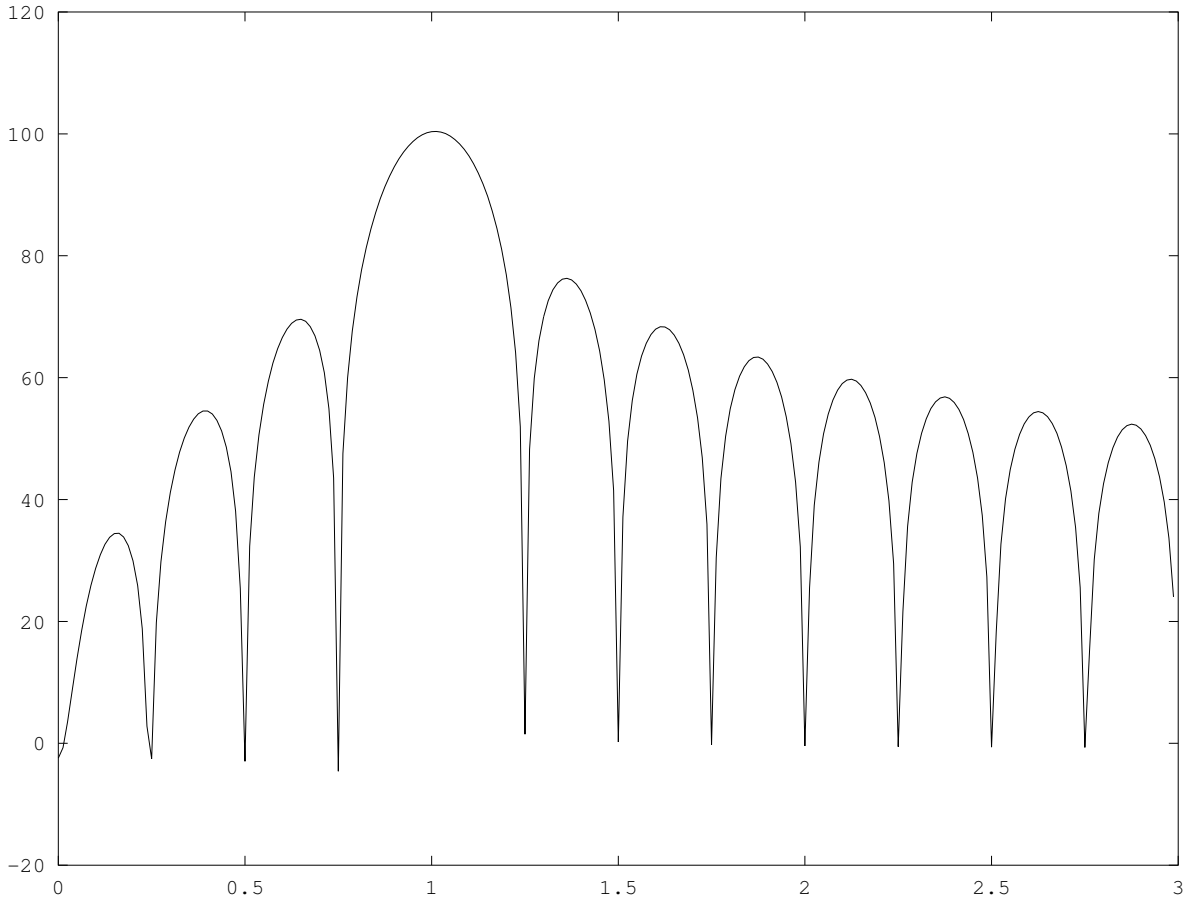
As an example, here is a segment of a sinusoid, that is, a sinusoid that is only computed for a finite amount of time. Although the true signal is a digital recording, it is labeled as if it were a signal depending on time; it has four cycles, at a frequency of one cycle per unit of time:



Here is the signal's measured power spectrum; the horizontal axis is frequency in cycles per unit time and the vertical axis is power pre frequency, normalized so that the peak is one:



There is a peak centered about a frequency of one, with width 1/2. There are other visible peaks, called *sidelobes*, which look a good bit smaller than the main, "real" one. To see better what happened to our measurement, we relabel the vertical axis to show relative power in decibels, with the peak normalized to 100:



We see that, if we don't have the luxury of waiting forever, our sinusoids can look very impure indeed. In general, if we only have access to a short segment of a signal (in this example, we had only four periods of a sinusoid), so that we have nothing better to suppose than that the signal is zero outside the time window we're looking at, our attempts to measure the signal's spectrum will give us only a blurry, out-of-focus result.

Why, then, don't we just collect a very long sample of the signal? Perhaps there is a practical reason we couldn't do that, but there's a deeper consideration: real signals that might arise in music, speech, or communications often change rapidly over time and in order to try to resolve how they are behaving in time we're obliged to examine them on short time scales. And the shorter the time scale we look on, more blurred in frequenct the spectrum will become. This limitation is well known in physics— it's caller the Heisenberg Uncertainty Principle.

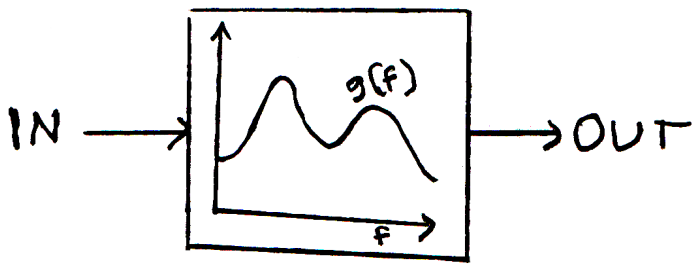
3.1.3 Short-time Spectra

Conceptually, the spectrum of a signal is an average over all of time. However, it is often desirable to find out what the spectrum of a signal is over a specific duration of time, or even to split a signal up into a sequence of short recordings and measure the spectra of each of these recordings separately. In this way, for a single input sound or recording, you would get a time-dependent spectrum.

To do this, for any time t you would consider a small interval of time (say, from t to $t+\tau$ for some fixed interval of time τ), and make up a new recording that consists only of those samples lying within the interval. (You can consider this extracted recording as being equal to zero outside the interval, exactly like the short sinusoidal burst we analyzed above.) and take the spectrum of that. The spectrum of this extracted signal is called a *short-time spectrum*. It depends on the choice of τ; the larger the segment you analyze the more sharply you can resolve frequencies but the less precisely you can resolve features in time.

3.2 Filters

Not only is a signal's spectrum a useful descriptive device, but it is one that we have some power to modify. Probably the most important tools we have for doing this are *filters*. A filter, for our purposes, is a process through which we can send any signal or recording, that multiplies the spectrum of the signal by a function of frequency known as the filter's *frequency response*. As a block diagram it might look like this:



If $g(f)$ is the filter's frequency response, and if you put in a sinusoid with

amplitude a_{in} and frequency f, the output of the filter will be a sinusoid with the same frequency but an amplitude of

$$a_{out} = g(f) \cdot a_{in}$$

The output might have a larger amplitude than the input at some frequencies and a smaller one at others, depending on whether g is larger or smaller than one. The way we are defining gain implies that it is always zero or positive.

That implies that the power changes like this:

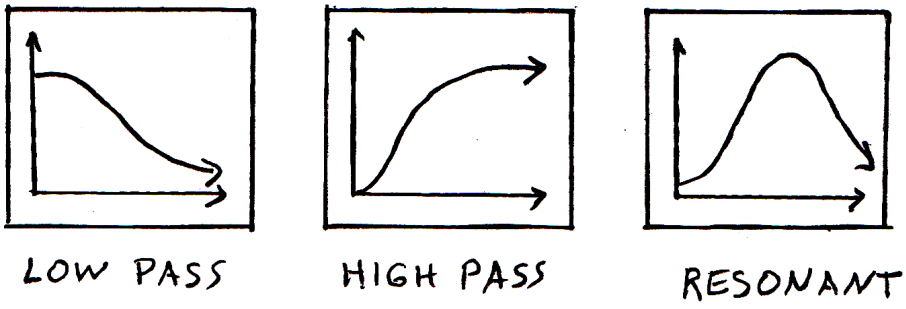
$$P_{out} = [g(f)]^2 \cdot P_{in}$$

and the level in decibels, like this:

$$L_{out} = L_{in} + 20 \log_{10}(g(f))$$

In a widely agreed-upon confusion of terminology, the gain in decibels, equal to $20 \log_{10}(g(f))$, is often called the *frequency response in decibels*.

Of all the sorts of filters one could design, three specific types, *low-pass*, *high-pass*, and *resonant* filters, appear often. They have frequency responses as suggested here:

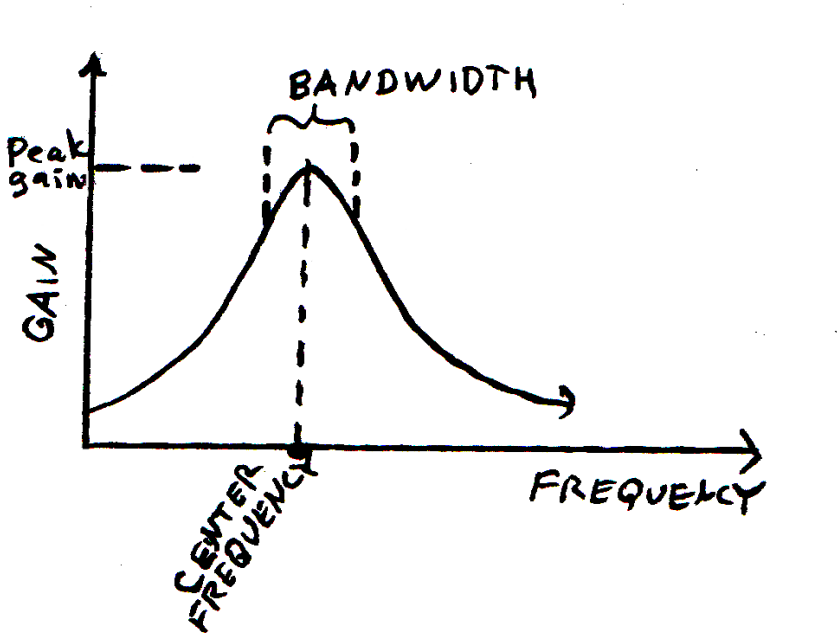


Low-pass and high-pass filters are often used to get rid of, or at least decrease, the amplitude of high or low frequencies, respectively.

Resonant filters have at least two interesting functions. First, they can be used to imitate physical systems, such as cavities filled with air (your mouth or your ear canal, for example). A wah-wah guitar pedal is nothing but a foot-controlled resonant filter. Second, they permit us to (approximately) pick out a portion of a signal's spectrum in order to measure it (that's

how one measures spectra in the first place) or in order to be able to treat different frequencies, that might be simultaneously present in a complex sound, in independently controllable ways.

A resonant filter typically has three parameters: a peak gain, a center frequency, and a bandwidth, as shown below:



The peak gain and center frequency have precise definitions (they are the two coordinates of the point at the apex of the curve). The bandwidth is a looser notion. One often-used measure is the distance between two points on the frequency axis where the curve is a specific relative amount (often 3 decibels) lower than the peak.

3.3 Application: Subtractive Synthesis

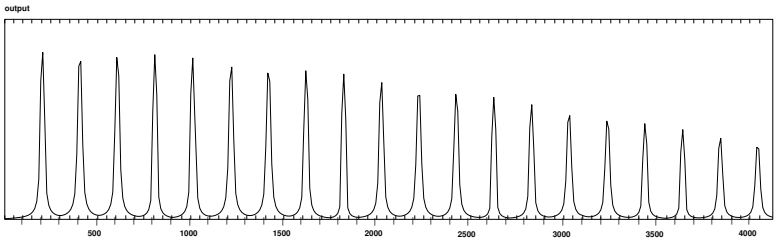
Real-world sounds frequently have time-varying timbres. The most prominent example of this is the human voice, in which (as we'll see in Chapter 5) the formation of vowels and consonants is reflected in variations in the voice's short-time spectrum, which can also be heard as changes in timbre. Other musical instruments behave in the same way; for instance, brass in-

struments tend to sound brighter when they are played louder, and it would be desirable to be able to make electronic instruments whose sounds can change in similar (or perhaps opposing) ways.

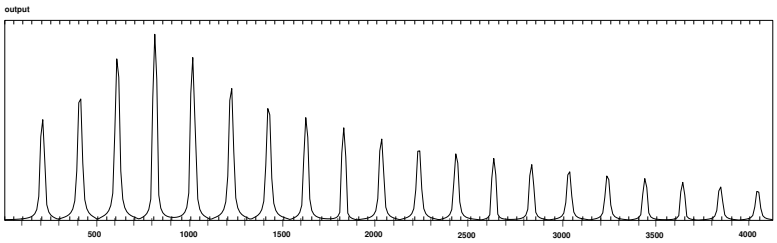
One excellent way to do this is to start with either noise or a complex sum of sinusoids (there are many ways to come by one of these, for instance simply by playing a short recording in a loop many times per second) and apply a filter whose properties vary appropriately with time. Here, for instance, is a simple sound to begin with:

SOUND EXAMPLE: Recording of a pulse, repeated 110 times per second; 5 second duration.

Here is its measured spectrum:



If we send that recording through a resonant filter, with center frequency 880 and bandwidth about 300, we get a spectrum like this:



The center frequency or bandwidth could be varied in time. In the following sound example we vary the center frequency from 110 to about 3000 and back:

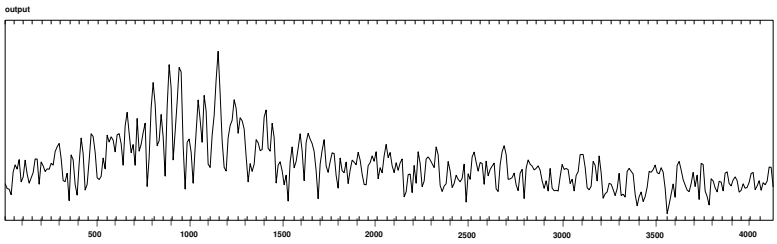
SOUND EXAMPLE: subtractive synthesis in which the sound described above is put through a sweeping filter.

One powerful aspect of this particular technique is that one isn't limited to using very simple recordings as sources. To make just one example, white noise (which we've mentioned before but haven't been able to do very much with) is an excellent starting point for subtractive synthesis. Applying

exactly the same sweeping filter before to a recording of white noise gives us this:

SOUND EXAMPLE: subtractive synthesis; same filter, white noise as input.

Here is the spectrum corresponding to the one above (880 Hz. center frequency):



3.4 The Inner Ear as Filterbank

Some limited insight about how timbre (which is a subjective quality of a sound) might relate to spectrum (a measurable property of a signal or recording) can be gained by studying how hearing works. Hearing is in general a ferociously complicated thing that humans are unlikely ever to understand, except in essentially trivial aspects. However, the things we do pretend to understand can often suggest interesting things to try in working with computer-mediated sound.

The active element of human hearing is the part of the human body that translates mechanical motion into nerve activation. This is a tiny, coiled, worm-shaped device in the inner ear called the *cochlea*. Vibrations that travel down its length get stronger and weaker as they travel in such a way that different frequencies are more prominent at different locations. It is a great simplification but not a complete misrepresentation to regard this as an array of resonant filters wired in parallel. Such an array is called a *filterbank*. Among other things, the ear seems to estimate the short-time spectrum of incoming sounds by measuring the average power, over short intervals of time, that shows up at various points along the cochlea.

Something is known about the frequency response of the "filters" that predict cochlear vibrations up and down its length. As a very rough indication, the bandwidths are about 100 Hz. up to a center frequency of about 550 Hz (or, equivalently, up until the lower edge of the band is 500 Hz.) For center frequencies above 550 Hz. the bandwidth is about 20 percent of the

frequency of the lower side of the band (or, equivalently, 18 percent of the center frequency). If you lay filters out side to side obeying these proportions, it takes about 24 of them to fill the frequency range of human hearing. This set of frequency ranges (often specified as the 24 intervals between the 25 frequencies 0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500) are known as the *bark scale*. So 100 Hz. is one bark, 1080 Hz. is 9 barks, and so on.

A range of frequencies corresponding to one location along the cochlea (from 100 to 200 Hz, say) is called a *critical band*. Critical bands overlap; for instance, 110 to 210 Hz. would also be considered one. The bandwidth of a critical band is always one bark.

Measuring the short-time power spectrum of a signal arranged in Barks can offer a rough visual idea of the overall loudness and timbre of a sound. In particular, the perceived strength of a signal within a critical band appears to depend on the total power within the band. This total power then has to be converted to perceived loudness (using a conversion unit called the *sone*) and then all the loudnesses are added (in sones) to get the resulting overall loudness.

Exercises and Project

1. In a complex, periodic tone, how many harmonics lie between two and three octaves above the fundamental (not including the lower and upper limit)?

2. What is the interval, in half tones (twelfths of an octave), between the second and third harmonic of a complex harmonic tone?

3. A low-pass filter has a frequency-dependent gain of

$$g(f) = \frac{1}{\sqrt{1 + f/(1000Hz)}}$$

What is the gain, in decibels, at 1000Hz?

4. If you send a sinusoid at frequency 100 Hz. and average power one, through the filter of exercise 3, what is the average power of the output?

5. What is the lowest-frequency pair of partials of a 1000-Hz.complex harmonic tone that lies within a critical band?

6. If two frequencies above 550 Hz. are separated by one bark, how many half-tones are they apart?

Project: Critical bands and loudness. This project tries to investigate how loudnesses of clusters of sinusoids are perceived differently when they are spaced withing a critical band than otherwise. For this experinent you should try to set yourself up with a reasonable listening environment, either using headphones or playing through a stereo (but not your laptop speaker).

Start by connecting a single "sinusoid" object with frequency 1000 Hz. to an "switch" object (these objects are both in the Music 170 library).

Now make another version (in the same patch) with four sinusoids tuned to 960, 980, 1000, anad 1020 Hz.. Connect all four to th input of a second "output" so that you can turn them on and off as a group, independently of the first one.

Make a third group of objects in the same way (or just duplicate the second group) but now set the frequencies to 500. 1000, 2000, and 4000.

Now, by turning them on and off (using the onoff control on the three output objects) equalize the outputs until all three are at a comfortable (reasonably soft) listening level. (If you have to push any of the output gains past about 90 dB, you should turn up your speaker instead. On my system I'm using gain values between 50 and 70.)

Now adjust the three output gains so that, as you turn them on one at a time, you judge them to have roughly equal loudnesses. Write down the three gain values you had to use to equalize them.

Since the four frequencies are roughly at the same level on the equal-loudness contour chart (Wikipedia is your friend), the different frequencies should be less a factor than the spacing. Is it in fact nearly true (or totally false) that in the close spacing example, you ended up adjusting the comples tone so that its power was roughly equal to the power of the single 1000 Hz. tone? Is that still true when the four frequencies are spread widely (500-4000)?

Chapter 4

Pitches, Intervals, and Scales

This chapter is about how Western musical tradition treats pitch, and why. Since pitch is primarily heard (by most people) in terms of ratios of frequencies, it is natural to use a logarithmic scale to assign pitches (which are subjective) to (objective) frequencies. But one has to pick a scale, that is, a ratio that corresponds to one unit of interval. This ratio in the West is the twelfth root of two, approximately equal to 1.059. That particular number turns out to be such a good choice of interval to measure pitches by, that it came to rule over a millennium of Western art music. Although we won't be concerned with all the historical details, this chapter will try to explain what's so special about the twelfth root of two as a unit of pitch.

The punch line is that this particular logarithmic scale turns out to have a surprisingly high number of sweet-sounding intervals in it. To develop this idea we first have to figure out what makes some intervals sound sweeter than others; this is pretty well explained by what is known as the Helmholz theory of consonance and dissonance (section **??**). Then we will investigate the actual intervals that arise in the Western scale (section **??**). Finally we'll consider some of the consequences of the way pitch is organized in Western music and consider some alternative ways to organize pitches.

4.1 The Helmholz Theory of Consonance and Dissonance

It's a commonplace that some intervals sound sweet and some sound sour, like this:

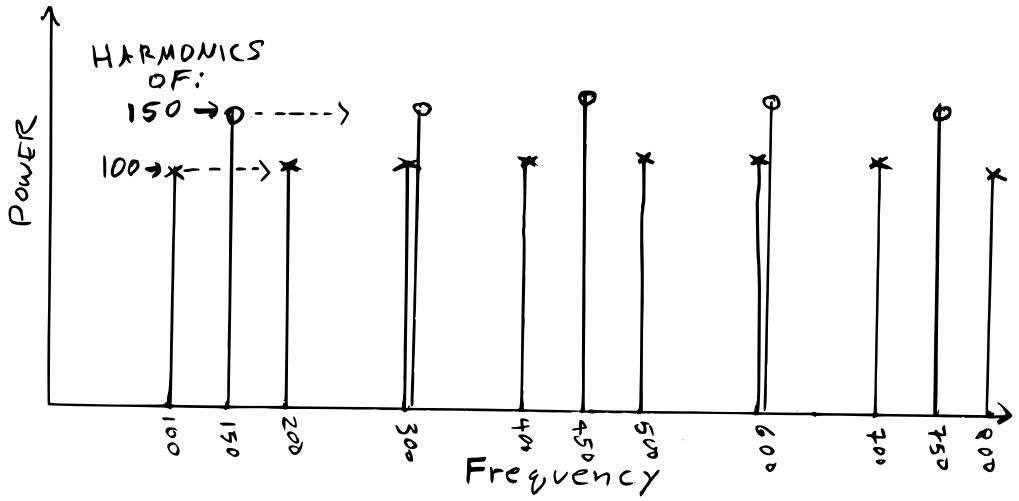
SOUND EXAMPLE 1: a musical fifth (usually considered sweet sounding)

SOUND EXAMPLE 2: a tritone (sour by comparison).

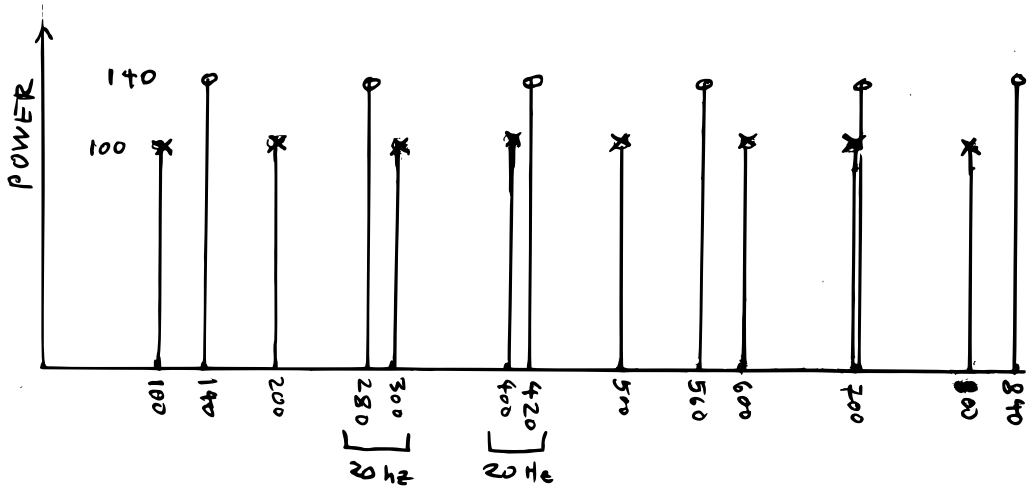
To call these sweet and sour is a rather clumsy metaphor. In musical language, we refer to a sweet-sounding interval as *consonant* and a sour-sounding on as *dissonant*—terms that can be taken to mean "going together" and "not going together". (Even this more neutral-sounding terminology carries an implicit value judgment that should not be accepted unquestioningly.) It turns out that the two intervals above have a physical difference that correlates with people's judgment of consonance and dissonance (as they are measured by psychoacousticians in experiments), that fits into what we know today as the theory of consonance and dissonance.

Although the theory of consonance and dissonance is usually associated with Hermann von Helmholz (1821 - 1894), many of its ideas and concepts date back further, even to ancient Greece; and the theory was much elaborated upon (and argued with) over the century since Helmholz published his contributions. The theory seems to have finally been brought to a definitive form in Plomp and Levelt's very readable and persuasive 1965 paper on the subject.

In the theory, we consider two complex periodic tones, that is, tones that may be written as a sum of sinusoids with frequencies in the ratios 1:2:3:..; in other words, tones all of whose partials are tuned to multiples of a fundamental frequency. Here is what happens when the two fundamental frequencies are chosen, for instance, as 100 and 150 Hz:



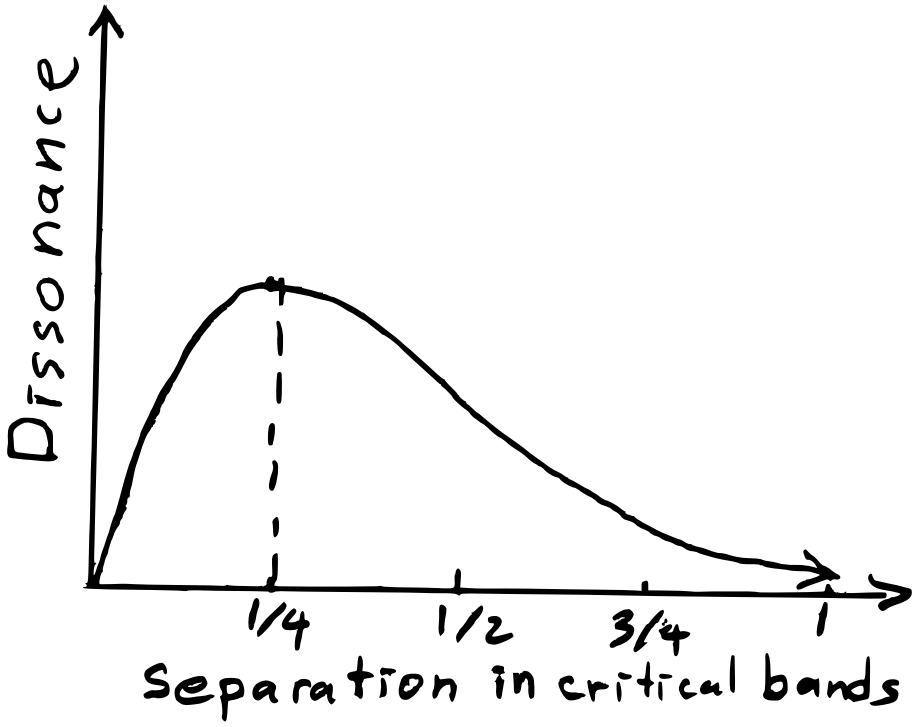
(To make the picture easy to see, all the harmonics of the 100 Hz. tone are given the same power, and so are all the multiples of the 150-Hz. tone; but the theory doesn't rely on that fact. Also, the double peaks at 300 and 600 Hz. are in fact single sinusoids; they're drawn this way for clarity). Here, on the other hand, is the situation when the fundamental frequencies are 100 and 140:



These pictures roughly correspond to the two sound examples above. The first one is consonant and the second one, dissonant. The Helmholz theory explains the consonance of the first example and the dissonance of the second one, by the absence or presence of awkward pairs of sinusoids (in this example there are two: 280 and 300 Hz, and 400 and 420 Hz.) These pairs

are far enough apart to be perceived separately but close enough to interfere with each other by vibrating in heavily overlapping regions of the cochlea (Section 3.4).

Plomp and Levelt go so far as to posit the consonance and dissonance of two sinusoids as a function of their separation in critical bands, thus:



Under this rule, the two pairs of sinusoids in the dissonant example above are almost as dissonant as possible (20 Hz. being close to 1/4 of a 100-Hz. critical band). The wider separations in the first example are about 1/2 of a critical band and contribute much less dissonance.

It's unavoidable that multiples of two fundamental frequencies would give rise to close neighbors here and there. The special reason the closely placed harmonics that occur in the first example didn't contribute to dissonance is that they landed right on top of each other. For this to happen, the ratio between the two pitches must be an integer ratio. For instance, for the third harmonic of one tone to coincide with the second harmonic of another, the fundamentals must be in a 2:3 ratio.

Here are the definitions of some intervals given by integer ratios between one and two (that is, within an octave), arranged from the most consonant to the most dissonant. The names are what they are for music-theoretical reasons too abstruse to explain here:

```
RATIO NAME
1:1   unison
2:1   octave
3:2   fifth
4:3   fourth
5:4   major third
5:3   major sixth
8:5   minor sixth
6:5   minor third
```

4.2 The Western Musical Scale

In many situations it's a good, practical move to choose, out of the set of all possible musical pitches, a reasonably small set of pitches, called a *scale*, to which you would restrict yourself when writing music. One reason for this might be that instruments, such as pianos or fretted guitars, are often designed to play a discrete set of pitches out of the whole continuum. (But if we consider that vocal music predates the development of keyboard and fretted instruments, this may cease to seem a compelling reason). Another consideration might be that you would want to be able to write music down. It would be impractical to write all the pitches as numerical frequencies, so in practice (in the West as well as elsewhere) musical traditions have settled on sets of pitches, typically between 5 and 21 in an octave, out of which a working musical context might use 5 to 7 at a time. For example, the Western scale has 12 pitches per octave, and one often chooses a musical *key* which implies a choice of 7 out of the 12.

Now suppose we wanted to divide the octave into n equal intervals to make up a musical scale. (Using equal sized intervals sounds like a good choice; it's like using a ruler whose marks are spaced regularly along it. You could reasonably request, for instance, that the interval you hear when you play the first and third notes on the scale should be the same interval you get from the second to the forth, or the third to the fifth, and so on.) If we call the interval between two successive pitches in the scale h, then the interval

between the first and third is h^2, and so on; the whole octave is a ratio of h^n. Since we know an octave is a ratio of 2:1, we get

$$h^n = 2$$

or, solving for h,

$$h = \sqrt[n]{2} = 2^{1/n}$$

Would our scale (with n equal steps per octave) have an interval that approximates a fifth (ratio of 3:2)? To answer this it's best to go back to computing things in octaves. One step is $1/n$ octave. How many octaves is the interval 3:2? From section 1.3 we get the interval in octaves is:

$$I = \log_2(3/2) \approx 0.58496$$

Now for instance if we decided to put five steps in an octave the available intervals would be 0.2, 0.4, 0.6, or 0.8 octaves; we see that three steps gives us an approximate fifth with an error of $0.6 - 0.58496 \approx 0.015$ octave. If, instead, we divide the octave in six parts, the closest interval to the fifth we can find is $4/6 \approx 0.6667$ octave for a much larger error, 0.081 octaves.

How bad are those errors? If a critical band is taken as an 18% increment in frequency, it is then

$$\log_2(1.18) \approx 0.24$$

and the worst possible mis-tuning is then 1/4 of that (according to the Helmholz theory) or 0.06 octave. So our 6-tone scale, with its error of 0.081 octaves, has about the sourest fifth you could ask for.

To gauge how well a scale is doing at hosting consonant intervals, we can study how its fifths and major thirds come out (the other intervals listed above can be formed from octaves, fifths, and major thirds.) Here is a table of the results. The first row gives the number of steps we divide the octave into; the two other rows give the error, in thousandths of an octave, of the fifth and major third:

```
division      3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
fifth error  82 85 15 82 14 40 29 15 40  2 30 14 15 22  3
third error  11 72 78 11 36 53 11 22 42 11 14 35 11  9 28

division     18 19 20 21 22 23 24
fifth error  26  6 15 14  6 20  2
third error  11  6 22 11  4 18 11
```

The first column that gives even half-decent approximations for the fifth and the third has 12 steps per octave. (Column 19 looks comparable and 22 even slightly better, but it would be hard to argue that the improvement is so great as to merit adding all those extra notes. Think what a piano would have looked like.)

How good or bad is the 12-steps-per-octave scale at reproducing a major third and a fifth? Well, the approximations turn out to be four and seven half-steps:

$$2^{4/12} \approx 1.2599$$

$$2^{7/12} \approx 1.4983$$

Starting with the better one (the fifth), we can now consider what happens when we play two pitches separated by seven steps on the 12-note-per-octave scale. Choosing 220 Hz. for the bottom pitch (A below middle C), we now see how closely the third harmonic of that tone meets up with the second harmonic of the tone 7 steps higher:

$$3 \cdot 220 = 660$$

$$2 \cdot 220 \cdot 2^{7/12} \approx 659.25$$

So we are a mere 3/4 Hz off - the two partials will beat slightly less than once per second. Now for the major third, for which we compare the fifth harmonic of the base tone with the fourth harmonic of the tone four steps up:

$$5 \cdot 220 = 1100$$

$$4 \cdot 220 \cdot 2^{4/12} \approx 1108.7$$

Whether this is a reasonable result (i.e., whether this interval should be regarded as consonant) is more a matter of taste than of measurable scientific fact. Some people complain about it, and some instruments (brass and voice, for instance) have a tradition of slightly altering a pitch here and there to make thirds sound more consonant when they appear in the music.

The other consonant intervals (minor third, fourth, major and minor sixths) may be built out of combinations of the octave, fifth, and major third. For example, to go up a fourth (4:3) ratio), we can go up an octave (2:1) and then down a fifth (2:3). This is approximated in the Western scale by going up 12 steps (the octave) and down 7 (the fifth), or in other words, by going up 5 steps. Similarly, the minor third (6:5) is up a fifth (3:2) and down a major third (4:5); this gives 7-4=3 steps. Here is a summary of how the intervals

fare in the Western scale (with the error now reported in steps—twelfths of an octave—to make them more readable):

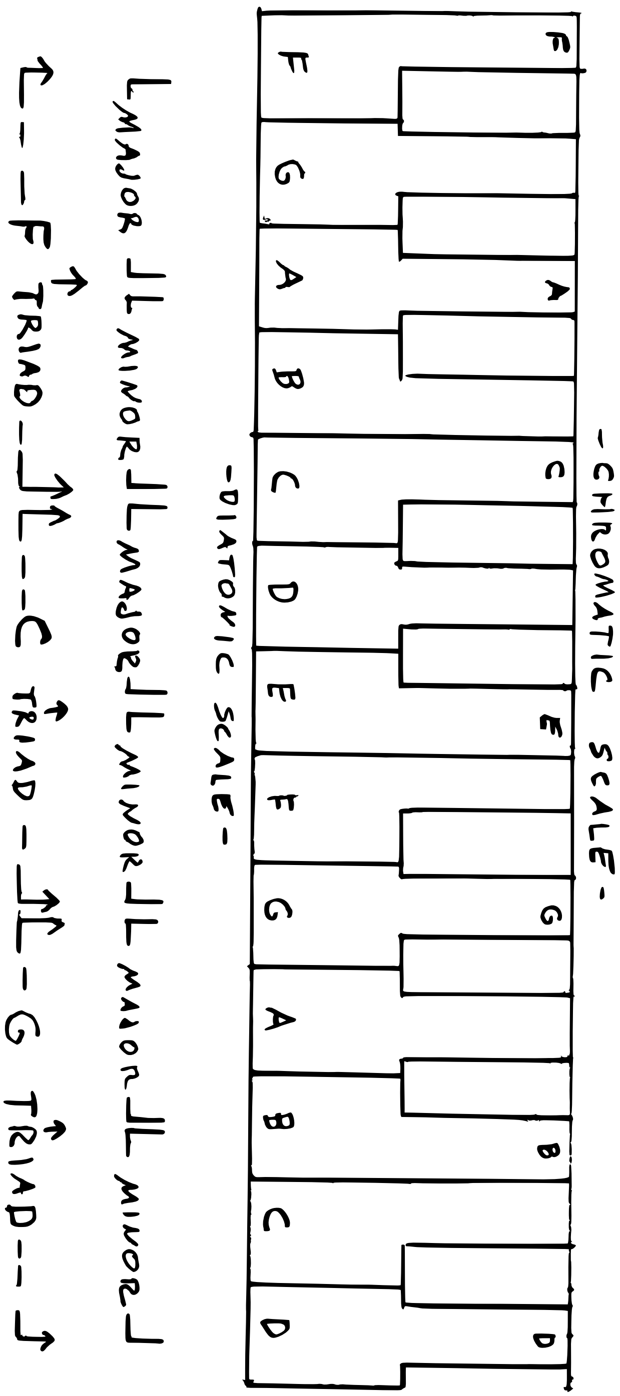
interval	ratio	steps	error (in steps)
unison	1:1	0	0
minor third	6:5	3	-0.156
major third	5:4	4	0.137
fourth	4:3	5	0.020
fifth	3:2	7	-0.020
minor sixth	5:8	8	-0.137
major sixth	5:3	9	0.156

4.3 The Diatonic Scale

A quick look at a piano keyboard will suggest (and a few minutes spent studying the way pitches are represented in Western musical notation will amply confirm) that the twelve steps of the octave are not considered equal; instead, they are organized in a highly non-uniform way that at first seems highly non-intuitive.

If someone had asked me to design the piano keyboard I would have simply put the even-numbered keys on the bottom and the odd-numbered ones on top; that way, once you learned to play a piece you could quickly transpose it to another key just by moving your hands right or left. But there is deep wisdom—learned over a thousand years or so with many spilled tears and even some blood—in the layout that we now use. With the benefit of hindsight, we can see why things are as they are in a fairly simple way, and although we shouldn't forget that there are rich historical resonances here, we can cheerfully leave them for a course in music history and confine our own study to the acoustics of the situation.

Here is a diagram showing a span of 22 steps of the Western scale, as they would appear on a piano keyboard:


-CHROMATIC SCALE-
-DIATONIC SCALE-
F G A B C D E F G A B C D
MAJOR
MINOR
MAJOR
MINOR
MAJOR
MINOR
F TRIAD
C TRIAD
G TRIAD

All the 22 pitches shown appear as evenly spaced stripes on the right-hand side of the diagram. These are the pitches of the 12-note-per-octave Western scale. For historical reasons, and somewhat over-poetically, this is called the *chromatic scale*. On the left-hand side of the keyboard you see only seven out of every twelve pitches; these are labeled A through G. (The labels repeat because pitches that are separated by an octave are given the same label). They are the piano's white keys, and they comprise the *diatonic* scale, so named because they are (mostly) spaced two steps apart..

The seven pitches per octave that make up the diatonic scale are called *naturals* (as in "G natural") to distinguish them from the other five, which are called *accidentals*. Accidentals are named for an adjacent natural, as in "D sharp" (the pitch between D and E) or "D flat" (between C and D).

A simple description of the diatonic scale is that it consists of three *major triads*. These are chords (collections of pitches) separated by a major and a minor third in turn. The diagram shows the three triads: an F triad (F, A, and C), a C triad (C, E, and G) and a G triad (G, B, and D); the notes C and G are in common between them. This is a natural way to choose 7 of the 12 pitches that have the maximum possible number of consonant intervals between them; in addition to the 3 major and 3 minor thirds, there are five pairs of fifths. (The other intervals are also available by reducing an octave by each of these three.)

These seven pitches (F, A, C, E, G, B, and D) can be re-ordered to get the pitches (A, B, C, D, E, F, G) that we know as the diatonic scale. By convention this scale is often arranged in the order (C, D, E, F, G, A, B, C) (repeating the C at either end of the scale); in this form it is called the *C major scale*.

This scale has many wonderful properties, but perhaps the most important is that, if we shift the entire thing by a fifth, we get back almost all the same notes. To see this we'll go back to its arrangement as three major triads, as in the diagram, and shift upward in pitch. Because we designed it as three major triads joined end to end, the first five pitches (forming the first two triads) land on other notes in the scale. Of the other two, the D shifts up to an A (you can check this by counting up 7 steps from the middle 'D', landing on 'A'. The B, however, lands on F sharp, the pitch between F and G. The new, shifted scale has the pitches (C, E, G, B, D, F sharp, and A), or, in letter order, (A, B, C, D, E, F sharp, G). Going further, we can shift the scale up or down a fifth, an arbitrary number of times, by changing one pitch in the scale for each shift. Shifting a musical scale (or a chord, or an

entire piece of music) by a fixed interval is called *transposition*.

Starting again with the pitch F, we now consider what happens if we repeatedly transpose it (but just F now, not the whole scale) by a fifth. We get the pitches (F, C, G, D, A, E, B, F sharp, C sharp, G sharp, D sharp, A sharp, F)—after which, being back at F, the sequence repeats. We ended up hitting each of the twelve pitches exactly once: first all the naturals, then all the accidentals. This arrangement is called the *circle of fifths*, and it sends music theorists into paroxysms of joy.

4.4 The Just-Intoned Scale

The Western chromatic scale is not without its discontents, who often complain about the poor accuracy of approximating major thirds as four twelfths of an octave. We can fix that if we are willing to relax the requirement that our scale have equal steps. (In fact, we could then do anything we wanted).

Returning to the keyboard diagram above, we could simply assign frequencies to the diatonic scale so that all the marked intervals are exactly correct. An interval that is an exact ratio of integers, such as 3:2 or 5:4, is called a *just interval*, and the scale we then get is called a *just-intoned scale*. (The word *intonation* is music jargon for "tuning".)

To construct the just-intoned scale we figure out the frequency for each pitch as an interval from C. So for instance, the pitch A is a minor third below C, for a relative frequency of 5/6. We then raise or lower it by octaves until it resides withing an octave above the original C; in this case we have to go up an octave, giving 5/3. Continuing in this way we get the just-intoned scale in C as shown:

C	D	E	F	G	A	B
1	9/8	5/4	4/3	3/2	5/3	15/8

This is all excellent except for two things: first, there has to be a different scale starting at each key. (Even if you put in extra pitches for the accidentals, you can't get all the intervals the same and so transposing such a scale gives you a whole new set of pitches.)

Second, certain of the intervals aren't what you'd wish for. In particular, the interval from D to A, ideally 3:2, is a thorny 40/27; and the interval

from E to A, ideally 4:3, is instead 25/12. If we were to move A and E over to fix those intervals, E would land at

$$\frac{9}{8} \cdot \frac{3}{2} \cdot \frac{3}{4} = \frac{81}{64}$$

which is off by a factor of:

$$\frac{81/64}{5/4} = \frac{81}{80}$$

This interval is called the *syntonic comma*, and it plays an amusing role in early, faltering attempts to find a musical scale in which all the intervals work. For instance, one approach was to nudge each of G, D, A, and E, successively, by 1/4 of the syntonic comma in order to get the E to be just - this was called *mean-tone temperament*. (In general, the process of slightly adjusting pitches up and down to compromise between incompatible interval constraints is called *tempering*. For this historical reason, the modern scale, whose intervals are all equal, is often said to have *equal temperament*.)

Although it's easiest to relax and let the modern Western scale rule over your music, the investigation of alternative pitch scales has been, and continues to be, an exceedingly fruitful avenue for composers including Harry Partch, Alvin Lucier, Charles Dodge, John Chowning, and Rand Steiger (and certainly many others), who have found highly individual ways of organizing sets of pitches.

Exercises and Project

1. In the Western tempered scale, if A is tuned to 440 Hz., what is the frequency of the C below it?

2. What is the frequency of the same C, under the same conditions, using the just-intoned scale in C instead of the tempered one?

3. How many half-tones, in the Western tempered scale, are there between the fundamental and the seventh partial? If the fundamental is tuned to a note on the Western Scale, how far is the nearest note on the scale to the seventh partial?

4. How many distinct major thirds can be formed using the 7-note diatonic scale? (Count two of them as being 'the same' if they differ by an octave).

5. What is the frequency ratio (as an exact number) between B and the next F above it in the Western tempered scale?

6. How many half-tones is the syntonic comma (as defined in Section 4.4)?

Project: How much detuning makes an interval sound sour? This project is a test of the Helmholz theory of consonance and dissonance. The interval we'll work on is the fourth below 440 Hz. (and later, 220 Hz.)

First, using "sinusoid" objects, make a perfect fourth using the frequencies 440 and 330. You can connect them to the same "output" object so that they have the same amplitude as each other. Now drag the 330 Hz. tone down in frequency until, to your ears, the result starts to sound "sour". How many Hz. did you have to decrease the 330-Hz. tone to make it sour? (If it never sounds sour to you at all, just report that.)

Now do the same things with pulse trains. You'll need the "pulse" object which is in version 2 of the Music 170 library (a folder named m170-function-library-v2, uploaded Oct. 15) - if you have version 1 get the new one (and change Pd's path or your working directory accordingly). When you've got it updated you can type "pulse" into a box to make a pulse generator.

Make two of them, frequencies 440 and 330, with "BW" (bandwidth) set to 2000, and connect them to an "output" object as you did with the sinusoids. Now reduce the 330-Hz. one to 329. What do you hear?

Now reduce it further until it sounds sour. How many Hz. less than 330 did you have to go? Was it further away than the tempered fourth (329.628)?

One could think that the number of Hz. you have to mis-tune an interval to get sourness might be a constant or else that it might be a constant proportion (i.e., interval). To find out, repeat the experiment for 220 Hz. and 165 Hz. Again, decrease the lower frequency (165) until you think it sounds sour. How many Hz. did it take and is it more nearly the same frequency difference or the same proportion?

Chapter 5

The Voice

From the point of view of humans, the human voice is arguably the most important kind of sound, and the ability to make vocal sounds, and to perceive them, is an ability whose importance is at least on a par with walking and seeing.

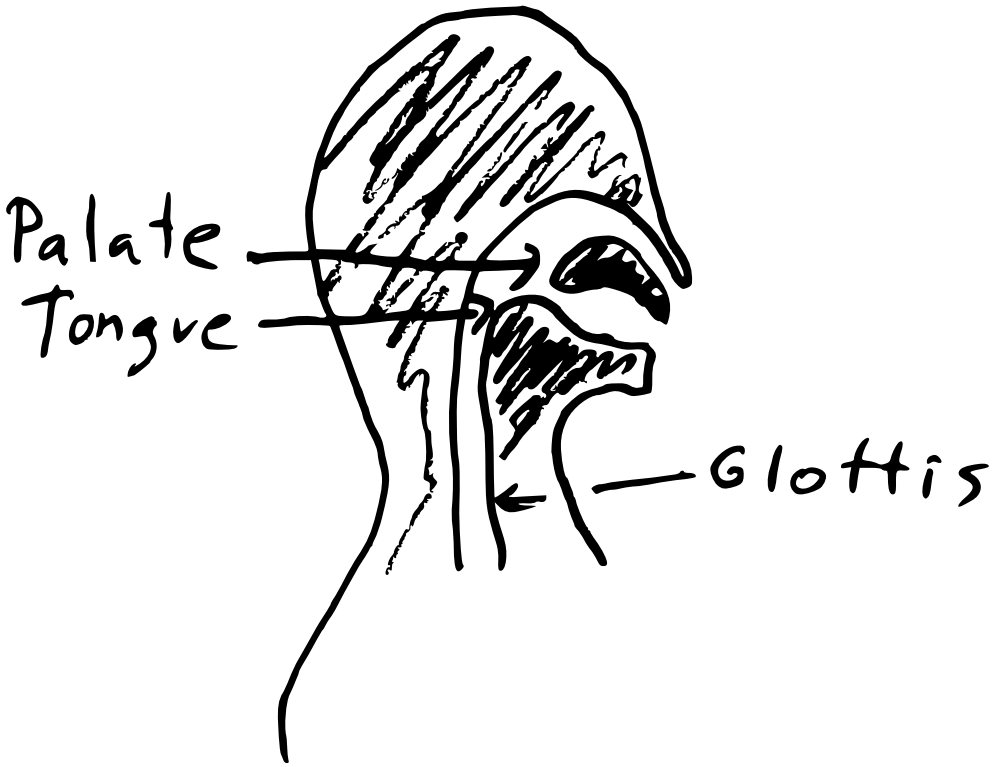
Until very recently in human history, most people could neither read nor write, and although the visual arts were often employed to impart messages (for instance, through story-telling mosaics or paintings), spoken language was the only practical way to warn someone that there might be a snake in a nearby bush. Perhaps as a result, human hearing and speech have evolved together to the point that we have an extraordinarily well-developed ability to perceive even subtle nuances in the human voice, and to vocalize in ways that can exploit our hearing abilities.

In addition to speech as communication, the voice is almost certainly the original medium for making and transmitting music. The function of music in human life is a deep mystery, but the fact that no human culture is without music suggests that it somehow plays a fundamental role. It seems certain that the ways we make and listen to music are in some deep way (or ways) connected to the use of the voice as the primary carrier of human language, even if we don't (and perhaps never will) know how either music *or* language really works.

In the following sections we'll look at natural production of vocal sounds (both speech and singing), and what can be said about the nature of the sounds that are prduced. We'll also describe a formal model of the human voice, at the basis of many speech analysis tools and synthesizers.

5.1 Vocal Sound Production

Here is a very simplified (and somewhat mis-proportioned) drawing of the parts of the body used for making vocal sounds:



To make a pitched sound, you push air out of your lungs (not shown) while closing a space within your *glottis*, which is a kind of encumbrance in your throat that forces the air to pass through a narrow slit between two *vocal folds*. These vibrate against each other, somewhat the way a trumpet player's two lips do. When they open, a short burst of air called a *glottal pulse* emerges. When things are running smoothly, these pulses emerge regularly, between about 50 and 600 times per second, and result in a periodic *pulse train* with an audible pitch.

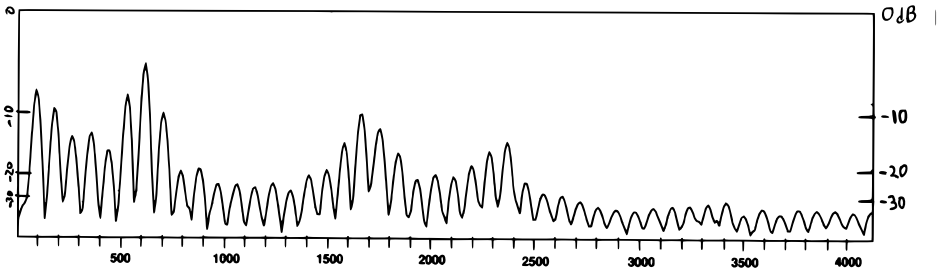
This pitched sound then travels through the mouth and/or nose to become sound in the air. (There are also vibrations in the flesh, but for now we'll only worry about the vibrations in the air). The sound passes through larger

or smaller cavities depending on the placement of the tongue, palate, lower jaw, and lips. The net effect is that the air passages filter the sound from the glottis on its way out to the air.

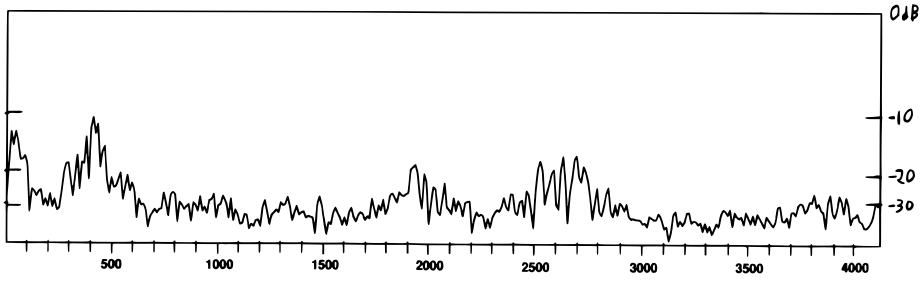
Depending on the placement of the tongue, etc., the filter's frequency response can change very quickly. This changing frequency response affects the timbre of the sound of the glottis as it appears in the outside air.

Unpitched sounds are made by relaxing the glottis (so that it no longer vibrates) but constricting one or another area to cause turbulence in the passing air: anywhere from the throat (for the English 'h' sound) to the teeth and lips (for 'f'). Depending on where these sounds originate, they are either fully, partly, or barely at all filtered by the air passages through the throat, mouth, and nose.

Here is the measured spectrum of a vowel (the 'a' in 'cafeteria'):



and here is the measured spectrum of a consonant (the 't' in the same word):



It is important to remember that these are in no sense canonical measurements of the particular vowel and consonant - the spectra are constantly changing in time and if the same word were uttered again (even by the same speaker) the result would almost certainly be different.

The vowel example shows the characteristic form of a periodic signal (even though it is only approximately so). The fundamental frequency is about 85 Hz, (the marker at 1000 Hz. is between the 11th and 12th peak). Certain

peaks are higher than their neighbors—they are peaks among the peaks. These are the 1st, 7th, 19th, and 27th peaks, at frequencies of about 85, 600, 1700, and 2350 Hz. These may be thought of as corresponding to resonances (peaks in the frequency response) of the filter that is made by the throat, nose, and mouth.

Although it isn't accurate, a simple model for the situation would be that the glottis is putting out a signal in which all the harmonics have equal height. (Actually, we think they drop off gradually with frequency, but we can pretend they're all equal and that the differences are all because of the filtering.)

The peaks in the frequency response of this filter (or, almost equivalently, the peaks in the spectrum that are higher than their neighbors) are called *formants*. The ear appears to be very sensitive to them. Their placement (their frequencies, bandwidths, and heights relative to each other) are properties of the throat-mouth-nose filter, and they roughly characterize the vowel that is being spoken. For instance, looking up the phoneme corresponding to the 'a' in 'cafeteria' on Wikipedia, we find that we should expect formants at 820 and 1530 Hz—a very poor match to what the picture above shows.

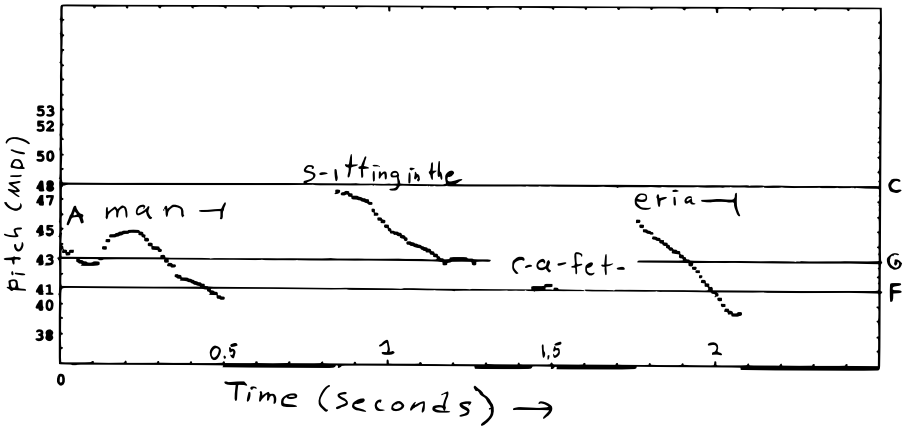
Consonants are more complicated. One class, the *fricatives*, are static in character (like vowels, they can last as long as breath permits), but are noisy and not periodic. Examples are 's', 'f', and 'sh'. Others are equally noisy but are generated by making short explosions that can't be sustained over time. These include 't' (shown above) and 'p'; they are called *plosives*. Others are essentially parasites on vowels; they can't be made except at the start or beginning of a vowel because they consist of rapid changes in filtering as one passage or another closes or opens. These are called *voiced* consonants. Examples are 'b', 'd', and 'g'.

Together the vowels and consonants are called *phonemes*, and they can be considered the basic sonic building blocks of language.

Speech and singing can be thought of as centered on the production of vowels. Vowels tend to have much longer durations than consonants, and the consonants can be thought of as decorations at the beginnings and ends of, and in between, vowels.

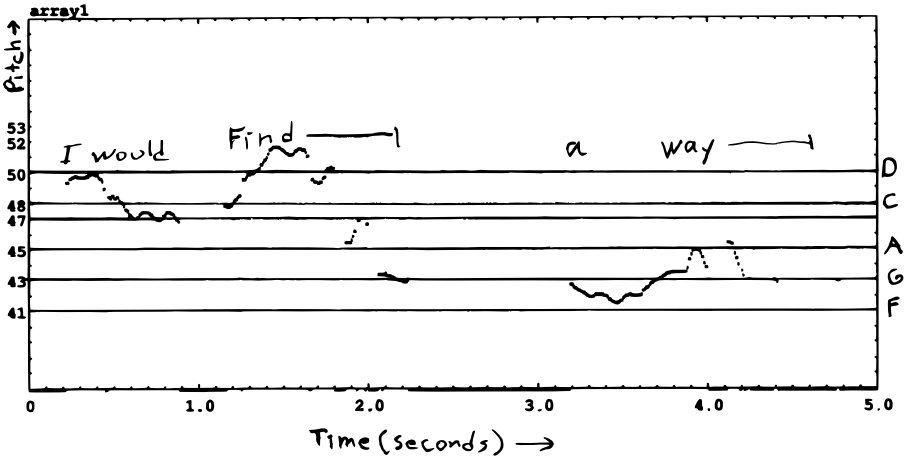
5.2 Pitch

The pitch of the voice is usually set by the frequency of glottal pulses during vowels or voiced consonants. Whether in speech or music, the pitch changes continuously (unlike a piano string, for instance, whose pitch is approximately constant.) Here is an example from a spoken phrase ("a man, sitting in the cafeteria"):



SOUND EXAMPLE 1: a spoken phrase ("A man, sitting in the cafeteria").

And here is Johnny Cash singing the last line of "Hurt":



SOUND EXAMPLE 2: a phrase sung by Johnny Cash at age 70.

The pitch is sometimes clearly defined and sometimes not. The voiced portions of speech are frequently nearly periodic and may be assigned a pitch

experimentally. Unvoiced consonants, generated partly (and often entirely) by air turbulence, don't have a readily assignable pitch (and neither does silence.)

What is the difference between singing and speech? A first attempt at an answer might be, "in singing the pitch sticks to a scale, and individual notes are characterized by having steady pitch in time". But the above example doesn't seem to support that statement at all.

It is often claimed that singing has a systematically different timbre—and correspondingly differing vowel spectra—from voice. On looking more closely, that seems in fact to be an artifact of Western art music conservatory training, and although their singing styles might have their own spectral ideosyncracies, these seem to be more reflective of style (or stylization) than of the essence of singing itself.

There are some general trends (but there are exceptions to all of them!): singing is often slower than speech (in particular, vowels are elongated more than consonants); it often has a wider range of pitch variation; its pitch patterns are often repeated from one performance to another in a way we wouldn't expect of speech; and if there is an accompaniment with discernible rhythms and pitches, singing is more likely to follow them than is speech. But the most general answer is probably that nobody knows the real difference, even though it is usually immediately clear to the listener.

One phenomenon that's present in some (but not all) singing, and nearly absent from speech, is *vibrato*. Traces of vibrato can be seen in the singing example above, at the end of the words "find" and "way", where the pitch wavers up and down. Physiologically, vibrato is produced by (roughly cyclicly) increasing and decreasing the air pressure underneath the glottis, which makes the pitch and power both vary upward and downward, and also changes the timbre. Typically, vibrato in singing cycles between 4.5 and 7 Hz, and the variation in pitch may be on the order of a half tone. (This is in agreement with the pitch trace of the singing example; in both areas the vibrato is about 5 Hz,) and, while it's unclear what the depth might be in the first instance ("find"—because the pitch is simultaneously sliding downward 9 half tones!), at the end of "way" (which is heard as the pitch A), the pitch trace varies between about G and A.

In this example the pressure variation is enough that the glottis apparently stops vibrating altogether during part of the vibrato cycle; in less extreme situations it will often be possible to trace the pitch all the way through the

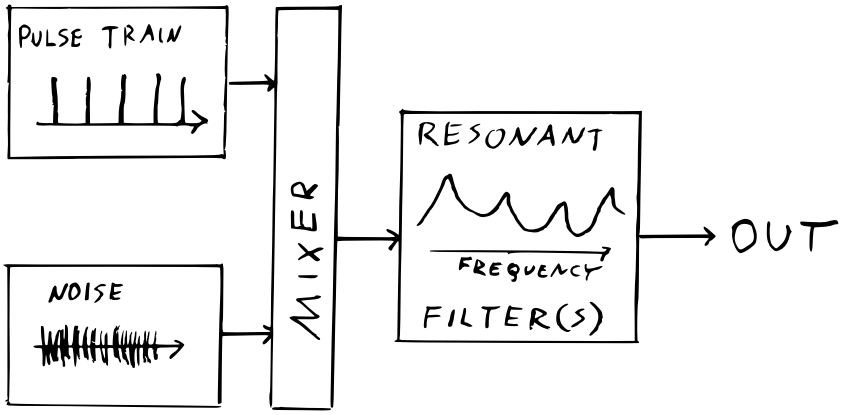
vibrato cycle.

In both areas of the trace the vibrato increases over time; this is quite common in the West, in both popular and classical idioms, and both in the voice and in other instruments.

5.3 Modeling the Voice

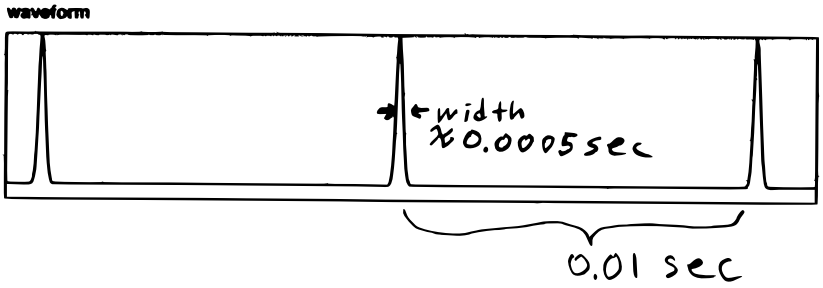
Over years of voice research, a sort of de facto model has emerged that people frequently refer to. This model is not only useful for synthesizing vocal sounds, but also for understanding the voice in its natural habitat (by, for instance, recording the voice's output and trying to make the model fit it). For example, voice recognition is often done by applying this model to a real voice and seeing what parameters one would have to supply the model with to get the observed output.

The model follows this block diagram:



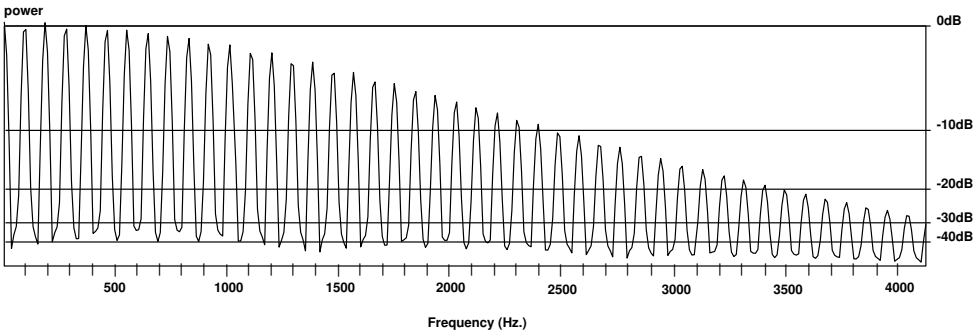
The pulse train models the glottis and a noise generator models turbulent noise. The mixer selects one or both of these sources depending on which source is active. A filter (or filters) models the vocal tract as it enhances some frequencies more than others.

The glottal pulse generator should output a *pulse train* as shown here:



SOUND EXAMPLE 3: A synthetic pulse train.

In this example there are 100 pulses per second. The signal is periodic and so will have partials at 100, 200, ..., Hz. Their relative amplitudes are controlled by the width of the pulse: roughly speaking, the partials' amplitudes slope downward to zero over a frequency range that is inversely proportional to the width of each pulse. In this example the pulses are 1/4000 second in duration so we would expect about 40 partials. Here is the measured power spectrum of the pulse train shown above:

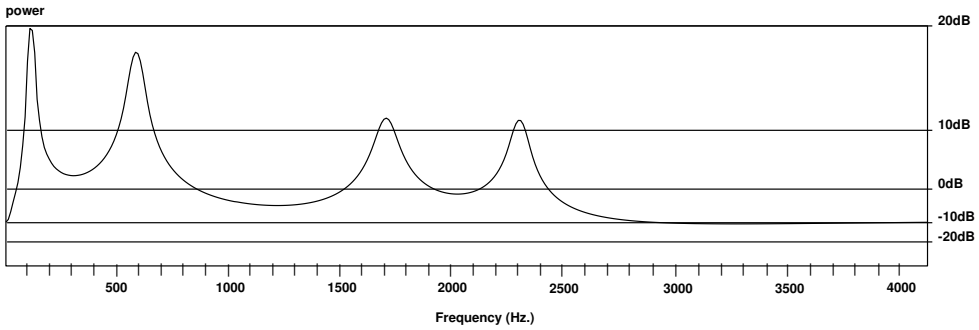


In real speech, the time duration of glottal pulses varies, roughly as a function of lung pressure (the higher the pressure, the narrower the pulse) so that louder or more forceful speech is brighter in timbre than soft speech.

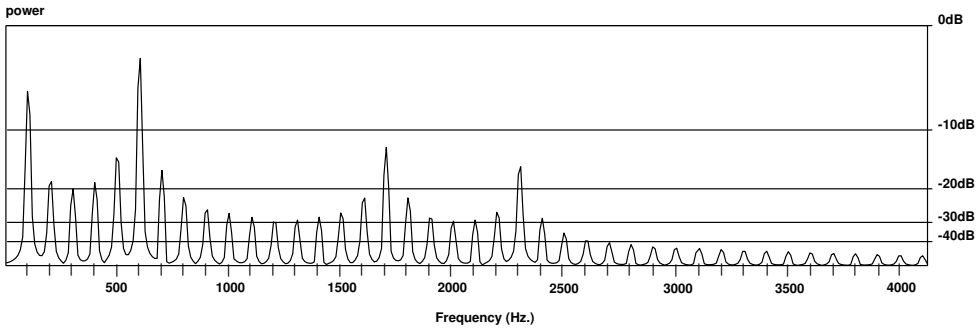
This glottal pulse train is mixed with noise. This is an oversimplification—it would capture the way the vocal tract filters turbulence around the glottis itself, such as the air leaks one hears clearly in the singing example above, but other kinds of turbulence aren't really filtered the same way as the glottal pulse train.

The vocal tract is modeled using a *filter*. Filters, which were introduced in Chapter 3, can be thought of as frequency-dependent amplifiers, in that passing a sinusoid through a filter outputs a sinusoid of the same frequency but a different amplitude, depending on the filter's frequency response. Here,

for example, is the frequency response of a filter intended to model the short 'a' vowel whose spectrum is shown above:



Here is the power spectrum of the result of filtering the pulse train with that filter:



SOUND EXAMPLE 4: The pulse train shown above, filtered.

This is far from a convincing human voice. The most glaring problem is that it is completely static; in natural voices, whether spoken or sung,the pitch, amplitude, and timbre are constantly varying with time. Arguably, it is the nature of the time-variations, rather than any given static snapshot, that gives the human voice its character.

By imitating these changes in time, speech synthesizers, which are usually constructed more or less along the lines described here, can often be made intelligible. But to make one sound natural, so that a listener would mistake one for a real voice, seems to be far from our present capabilities.

Speech recognition is done essentially by carrying this process out in reverse. Given a recorded sound as input, we measure its time-varying spectrum and try to fit a glottal pulse train (at some frequency and pulse width), amplitudes for the pitch and noise components, and a filter frequency response that best fit the measured spectrum. By looking for peaks in the filter

frequency response we would find the frequencies, relative amplitudes, and bandwidths of any formants. Those (especially the formant frequencies, and how they are changing in time) can be matched to the known behaviors of phonemes in the language being spoken. This is compared to a huge dictionary of phonetic pronunciations of words and phrases, and the best fit becomes the output of the speech recognizer.

5.4 Failures of the Model

Real voices don't really follow this model terribly well. All sort of imperfections (raspiness, scratchiness, breathiness, etc.) result from variations in the way a real glottis vibrates compared to our theoretical one. There are at least two separate modes of vibration of the glottis in most healthy speakers (in singing they are called the low and high registers), which result in differently shaped pulses with different spectra.

Vowels aren't really characterized by formant location; peoples' differently shaped mouths and throats unavoidably introduce variations in formant structure.

In English, pitch variations in speech aren't considered part of the phonetic structure of the language, but in many other languages (called *tonal languages*), the pitch and/or the way the pitch changes in time may make the difference between one phoneme and another.

Moreover, phonemes aren't produced independently of each other; the way any phoneme is joined to the ones before and after it affects how it is produced. In practice, it often isn't even possible to say definitively at what point one phoneme ends and the following one starts. It frequently happens that, when trying to extract a phoneme in a sound editor, one hears a different phoneme entirely.

Exercises and Project

It's hard to see how to assign exercises to this chapter which is mostly descriptive. Instead, these exercises serve as a cumulative review of Chapters 1-5.

1. Suppose a signal is a sum of three sinusoids, each with peak amplitude 1, at 300, 400, and 500 Hz. What is the signal's average power?

2. What is the period of the signal of exercise 1?

3. How many barks wide is the spectrum of the signal of exercise 1?

4. What is the name of the interval from the lowest to the highest of the component sinusoids?

5. How many half-steps is the lowest of the component sinusoids above middle C?

6. By how many dB does this signal's power exceed the power of its lowest component (i.e., of a 300 Hz. sinusoid of peak amplitude 1)?

Project: multiplying sinusoids. This project is a demonstration of my Fundamental Law of Computer Music (last formula of Section 2.3).

First, make a "sinusoid" object and connect it both to an "output" (so you can hear it) and to a "spectrum" (so you can see it). To start with, tune the sinusoid to 500 Hz. and turn the spectrum object on (turn on the "repeat" toggle and optionally increase the rate, say to 5 or 10 Hz). With the scale control at 100 you should see a peak at 500 Hz. on the horizontal scale and just reaching the top of the graph (0 dB.) (Really, I should be reporting the RMS power as -3 dB; I'll fix this next release, but no matter for now.)

Now bring out a second oscillator at 100 Hz. and multiply it by the first one you have (using a "multiply" object). Disconnect the first oscillator from the spectrum object and connect the output of the multiply object instead; do the same with the output object so you hear the product of the two sinusoids. What are the frequencies and amplitudes of the resultant sinusoids? (To measure the amplitude I used the "scale" control to scale them back up to 0dB; that told me the relative mplitude in dB compared to the original sinusoid.

Now repeat the process to make a product of *three* sinusoids: disconnect the output of the multiplier, and introduce a third sinusoid and a second multiply object; multiply the new sinusoid by the product you already made of the first two. Set the third sinusoid to 50 Hz. and connect the output of the multiplier (the product of the three sinusoids) to the output and spectrum objects. What frequencies and amplitudes do you see now?

Now that you've done this you're welcome to change the frequencies of the three oscillators and see how the result behaves. If you find that interesting (I do!) try multiplying by a fourth sinusoid in the same way and enjoy.

Chapter 6

How Sound Moves in the Air

Up to now we have dealt with signals and recordings, and ignored the phenomenon of sound itself. But unless you are using headphones to listen to a purely electronic signal, there will be one or more stages in any chain of audio production that is mediated by real, acoustic, in-the-air sound. There is much to know about how sound is propagated through the air and how objects absorb, reflect, and emit sounds. In the context of this course we will only have room and time for the very basics.

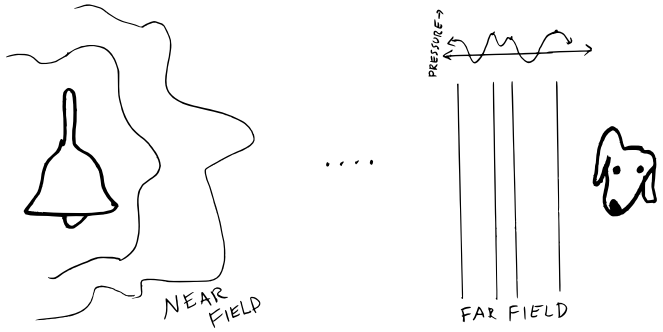
6.1 Modeling Sound Transmission

Air has mass, and it also acts as a spring—it can be compressed by applying force. Like any mass-and-spring system, it is able to store and transmit energy in the form of vibrations. In the particular case of air, these vibrations move in waves from one place to another. There are two kinds of variables to be concerned with: variations in air pressure from place to place and over time, and physical displacement of the air itself. At a given point in space and time, the pressure is a number in units of force per area (for instance, Newtons per square meter). One almost always speaks of the pressure as being positive if it is greater than atmospheric pressure, and negative if it is less – in other words, we tacitly subtract one atmosphere from the variables we use for pressure.

The displacement is a vector; it has magnitude and direction. When the air is at rest the displacement is zero. When sound is present, the displacement changes with time and as a result the air also has a non-zero velocity that

can be calculated from the displacement as a function of time.

One can idealize sound as having a source, and as traveling through the air to the point at which we hear or measure it. If a listener or microphone is distant enough from the source, the chain of transmission through the air can be thought of as shown:



The sound first passes through an area called the *near field*, in which the strength and even the direction the sound is traveling in might vary from point to point. If, however, we go out to a distance several times the size of the source, we can say approximately that the sound is moving *unidirectionally*. The pressure, displacement, and velocity of the air depend (approximately) only on one dimension, in the direction from the source to the listener. If the listener moves in a perpendicular direction to that axis, the sound is (approximately) unchanged.

Sound travels at some 343 meters per second (in the air, at sea level; this varies slightly with temperature and altitude). In English units this is about 767 miles per hour, or 1167 feet per second, or, in very round numbers, about one foot per millisecond. This is the *velocity of sound* which we'll call c.

If the source and listener are separated by a distance r, the sound arrives at the listener with a delay $\tau = r/c$. If the delay τ is known, the distance can be found using $r = \tau c$.

6.1.1 The Doppler Effect

If the source and/or the listener is moving in space so that the distance between the two is changing, then the delay between the source and the

listener changes as well. This results in a speeding up or slowing down of the sound.

To see this in detail, suppose the distance is changing at a speed s, measured in units of distance per time as is proper for a velocity. Then over a period of time τ the distance changes by τs and the time of transmission by $\tau s/c$. So the receiver receives the emitted sound not in time τ, but in an amount of time that is changed by $-\tau s/c$. The time overall over which the sound of length τ arrives is thus:

$$\tau - \tau s/c = \tau \cdot (1 - s/c)$$

This implies that the sound gets a speed change by a factor of

$$\frac{\tau}{\tau \cdot (1 - s/c)} = \frac{1}{1 - s/c}$$

If the distance is changing at speeds much smaller than that of sound (usually the case in day-to-day experience) the relative speed change is minus the speed that the distance is changing, divided by the speed of sound. So, for instance, to get a transposition of one half tone (6 percent), one needs for the distance to change by 6% of the speed of sound, or about 46 miles per hour. If the distance is decreasing with time, the pitch goes up, and if it is increasing, the pitch goes down.

6.2 Power, Intensity, and Sound Pressure

The pressure is a simpler variable to deal with than the velocity, and the simplest measure of the strength of a sound is a measure of power derived from the pressure. If, at a point in space, the pressure is a function of time $p(t)$, the *effective sound pressure* at that point is the root mean square of $p(t)$, that is, the square root of the average of $p^2(t)$. This is usually put in units of decibels and called the *sound pressure level* or SPL for short. To do so we need a reference value which is conventionally set at 0.00002 (20 millionths) newtons per square meter, which is roughly the quietest sound that would be audible under ideal conditions (the threshold of human hearing at 1 kHz). Thus if the RMS sound pressure is p, the SPL is

$$\mathrm{SPL} = 20 \log_{10} \left(\frac{p}{p_0} \right)$$

with

$$p_0 = 0.00002 \frac{\text{newton}}{\text{meter}^2}$$

It is often useful to know how the SPL of a sound relates to the power of the sound at its source and the distance from the source to the listener. To do this we need a measure of how much power a unidirectional sound carries as a function of its SPL. This power flow (sometimes called the *intensity*, which we'll denote by I) is in units of power per unit area, for instance, watts per square meter. (Caveat: there appear to be conflicting definitions of intensity; see for instance Wikipedia). To find I requires a more complicated calculation than can reasonably be presented here, but the upshot (for an ideal gas at least) is this:

$$I = \frac{p^2}{\rho c}$$

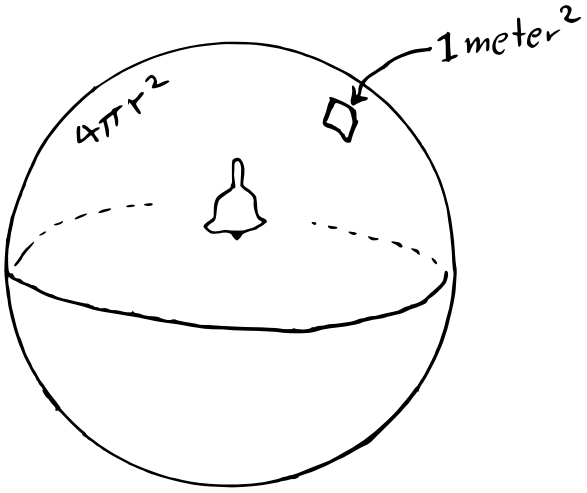
where ρ is the density of air, which in San Diego is about 1.225 kilograms per cubic meter. Solving for p gives:

$$p = \sqrt{\rho c I}$$

so that the SPL as a function of I is:

$$SPL = 10 \log_{10} \left(\frac{\rho c I}{{p_0}^2} \right)$$

Now if we know the total power w of the source, and the distance r, we can compute the power per unit area at a distance r, assuming the total power is evenly distributed over a sphere of radius r (and, hence, surface area $4\pi r^2$):



So the intensity (power flow) is

$$I = \frac{w}{4\pi r^2}$$

and plugging this into the formula for SPL gives:

$$SPL = 10 \log_{10} \left(\frac{\rho c w}{4\pi {p_0}^2 r^2} \right)$$

For example, suppose a one-watt source is one meter away from you. The stuff inside the logarithm is approximately:

$$\frac{1 \cdot 1.225 \cdot 343}{\cdot 4\pi \cdot 0.00002^2} \approx 8.36 \cdot 10^{10}$$

(That is, it's about 10^{11} times more powerful than the threshold of hearing), and so we get:

$$SPL \approx 10 \log_{10}(8.36 \cdot 10^{10}) \approx 109$$

Now increasing the distance to two meters reduces the power flow by a factor of four, and so the SPL goes down by 6 dB. To get the same SPL as before you'd need a four-watt speaker.

6.3 Plane Waves

Since, at least when we're not too close to the source (and still not considering issues such as deflections by other objects), the sound is roughly moving unidirectionally, it is useful to study unidirectional waves before studying more complicated ones. It is natural to start with sinusoids. A mathematical construct called the *plane wave* (but perhaps better described more explicitly as a *sinusoidal plane wave*), is a sound whose pressure changes sinusoidally at each point in space (so that a microphone would pick up a sinusoid), and that has a fixed direction. Three parameters (an amplitude, a frequency, and a direction in space) determine a plane wave completely. For simplicity, we'll choose a fixed direction, along the x axis to the right. The plane wave's pressure, as a function of time and space, can be written as:

$$p(x, y, z, t) = a \cos(2\pi f(t - x/c) + \phi)$$

Here a is the peak amplitude, in units of pressure. The frequency f is in cycles per unit time (Hz., for example), and ϕ is the initial phase at the origin where $x = y = z = 0$.

The independent variables y and z don't appear in the equation; the pressure isn't changed if we move in the y or z direction.

6.3.1 ... As a Function of Space

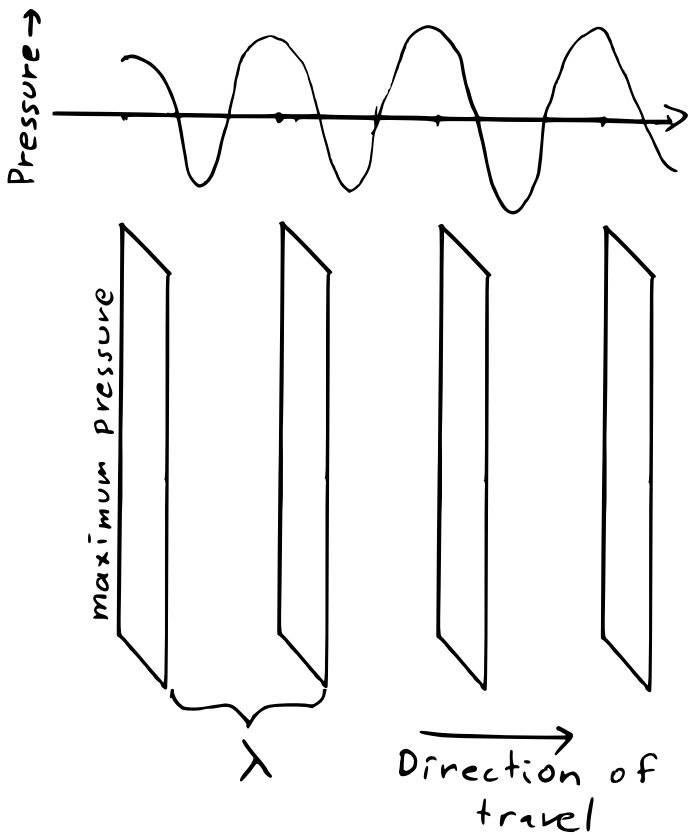
We now look at the plane wave's shape at an instant in time, say $t = 0$. At that moment the plane wave is a function of three spatial dimensions (but only depends on one of them):

$$p(x, y, z, 0) = a\cos(2\pi(f/cx + \phi))$$

This has the form of a sinusoid but instead of depending on time it depends on space. One cycle occupies a length equal to c/f (since increasing x by that much would add 2π to the phase). This distance is called the *wavelength* and is customarily denoted by λ (Greek lambda):

$$\lambda = \frac{c}{f}$$

If, for instance, we set the initial phase ϕ to zero, the locations at which the plane wave has zero phase are at $x = \ldots, -\lambda, 0, \lambda, 2\lambda, \ldots$. These points are a series of planes in space as shown:



We can re-write the plane wave using λ in the place of c/f:

$$p(x, y, z, t) = a\cos(2\pi(ft - x/\lambda) + \phi)$$

6.3.2 ... As a Function of Time

At a fixed point in space, (but now allowing time to vary), the pressure varies sinusoidally. We can re-group the original equation for pressure as:

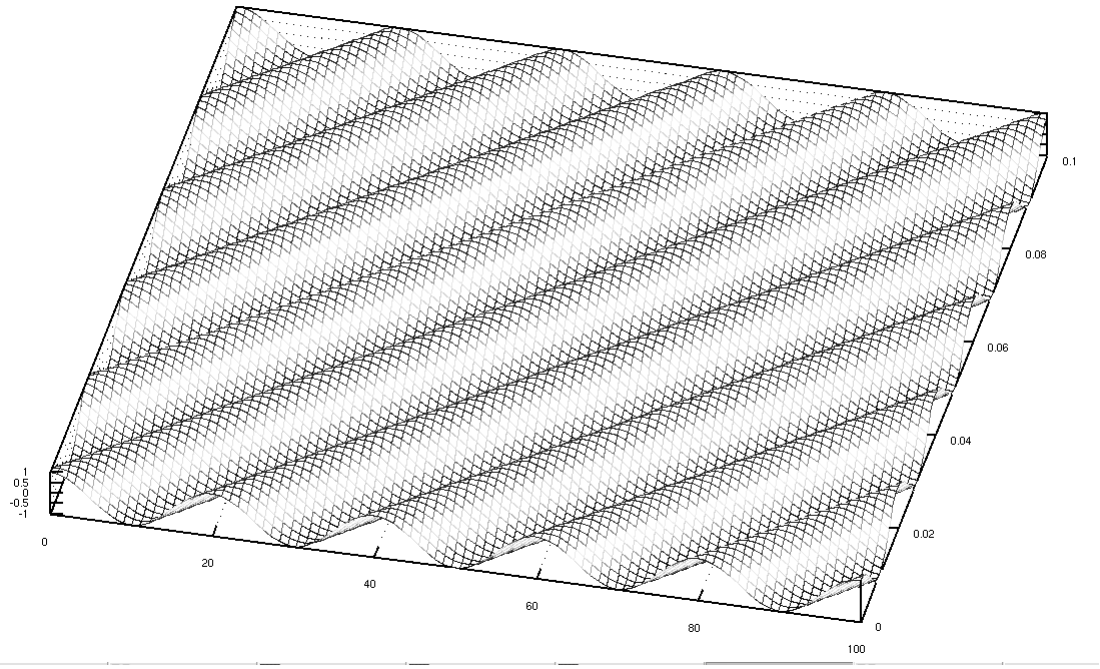
$$p(0, 0, 0, t) = a\cos(2\pi ft + (2\pi x/\lambda + \phi))$$

showing that at the point (x, y, z) the initial phase of the sinusoid is $2\pi x/\lambda + \phi$. So anywhere in space we get the same amplitude and frequency, but a space-dependent initial phase.

Placing two microphones (say, at points (x_1, y_1, z_1) and (x_2, y_2, z_2) will pick up sinusoids whose initial phases differ by $2\pi(x_2 - x_1)/\lambda$.

6.3.3 ... as a Function of x and t

For a third point of view, here is a graph of the plane wave as a function of x and t, as a three-dimensional surface:

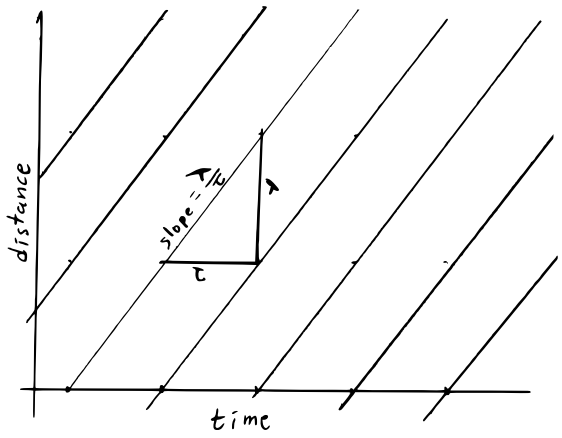


Here the horizontal axis in front shown distance in feet from 0 to 100, and the other horizontal axis (running from front to back in perspective) shows time from 0 to 0.1 second. To keep the numbers round, the speed of sound was rounded off to 1000 feet per second. The wavelength was chosen as 20 feet, corresponding to a frequency of 50 Hz. There are crests separated in space by λ =20 feet), and in time by $1/f = \tau = 1/50$ of a second.

The locations of the crests form a series of parallel lines defined by setting the phase to any multiple of 2π:

$$2\pi ft - \frac{2\pi x}{\lambda} + \phi = 2n\pi$$

for integer values of n:



Over each period τ, the crests of the plane wave march forward λ units of distance, so that each one is replaced by the one in back of it. The slopes of the lines are all equal to λ/τ, which equals c.

It is an interesting property of sound in the air that (for practical purposes) plane waves of different frequencies all move a the same speed c. (This isn't necessarily true in other media, and is quite noticeably not the case in solids).

To give an idea of the scale of reasonable wavelengths for sound: at 20 Hz. the wavelength is about 50 feet and at 20 kHz, it's about 0.6 inches.

6.3.4 Directional Waves as Superpositions of Plane Waves

If an arbitrary sound is traveling unidirectionally (say, in the x direction) it can, theoretically speaking and with some provisos,, be expressed as a s mixture of a (possibly infinite) number of sinusoidal plane waves, which correspond to the way its value at one point (a signal) could be written as a sum of sinusoids. In considering how a unidirectional sound behaves in space, it suffices to consider how its component sinusoids would behave separately.

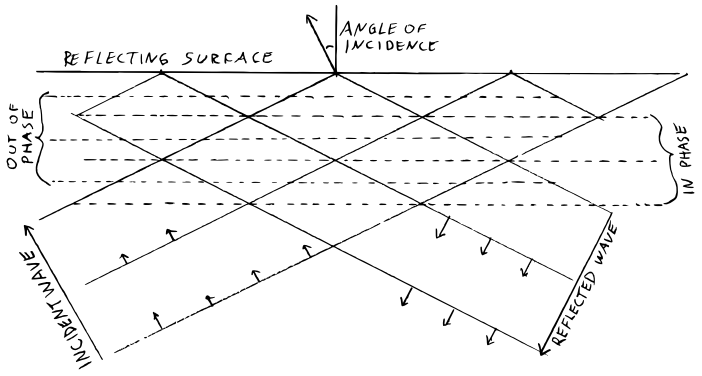
6.4 Reflection

When a unidirectional sound encounters a flat, non-moving surface such as the wall of a room, it is reflected. If the wall is many times larger in both

dimensions than one wavelength of the sound, its reflection is also approximately a plane wave. (So, if a sound has many frequency components, it might be observed that the higher frequencies in the sound are reflected roughly as light off a mirror, but that lower frequencies are dispersed in many directions. We'll return to this later; the effect is called *diffraction*.

Focusing for now on the situation where the reflection may be considered a plane wave (as well as the incident wave), we can describe the whole sound field as a superposition (mixture, or sum) of two plane waves, with equal frequencies and amplitudes but different directions.

We can see how these will interact by considering, at various points in space, how the phases of the two plane waves compare. Where they're equal, the sound pressure level will be 6 db higher than that of the incident wave. Wherever they differ by π, the two will cancel out and you will get silence. This combination of regions in space having higher and lower amplitudes of sound present is called an *interference pattern*. In the example we're considering, the situation looks like this:



There are evenly spaced planes (shown as horizontal lines parallel to the plane of reflection) where the two plane waves are in phase, and others where they are exactly out of phase and cancel.

Looking along one of the planes where the phases are aligned (say, right at the wall itself) we see that the waves are apparently elongated; but as we watch the waves "move" along the wall we'll also be distracted to notice they appear to be moving faster than the speed of sound. (Still, at any location where there is sound present, if we place a microphone we'll pick up the original frequency.)

6.5 Standing Waves

An important particular case is that n which the incident sound is perpendicular to the reflecting plane, so that the sound heads directly toward it and is reflected in exactly the opposite direction it came from. Suppose for simplicity that the incident plane wave is traveling in the x direction and it is reflected off the perpendicular plane $x = 0$. The incident and reflected plane waves are then:

$$P_{\text{incident}}(x, y, z, t) = a\cos(2\pi(ft - x/\lambda))$$

$$P_{\text{reflected}}(x, y, z, t) = a\cos(2\pi(ft + x/\lambda))$$

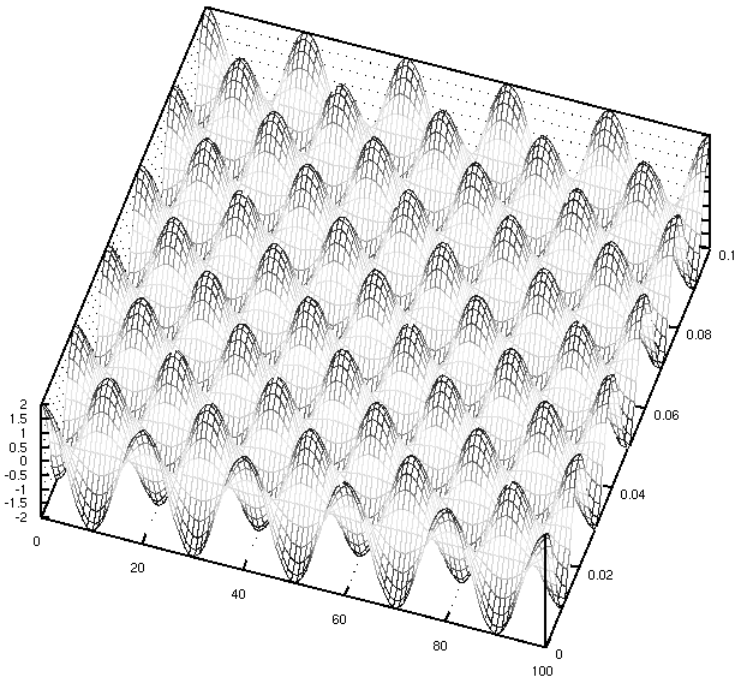
and if we sum them we can use my "Fundamental formula of computer music"

$$\cos(a) + \cos(b) = 2\cos(\frac{a+b}{2})\cos(\frac{a-b}{2})$$

to get:

$$P_{\text{incident}}(x, y, z, t) + P_{\text{reflected}}(x, y, z, t) = 2a\cos(2\pi ft)\cos(2\pi(x/\lambda))$$

The way this formula depends on t and x is special: one term depends only on x and the other on t. So there is one, unchanging waveform in space, given by $\cos(2\pi x/\lambda)$, whose amplitude is changed globally by the factor depending on t, but which no longer appears to move in space (or to change its shape at all). This is called a *standing wave*. Here is a graph of a standing wave as a function of one spatial dimension x and time:



Exercises and Project

1. A car is moving toward you at 60 miles per hour. Its horn is blowing at 440 Hz. By how many half tones does the sound rise above 440 Hz because of the car's motion?

2. Suppose the same car has the same horn blowing, but now you hear a pitch of 400 Hz. Assuming the car is moving directly away from you, how fast is it moving (in miles per hour)?

3. A sound has a wavelength of 4 feet. What is its frequency?

4. If two sounds are a perfect fifth apart, what is the ratio of their wavelengths?

5. Suppose that ten feet away from a loudspeaker the SPL is 60 dB. (and that the speaker is the only object making sound nearby). What is the SPL 20 feet away from the loudspeaker?

6. A 440-Hz. sinusoid is traveling in a plane wave in the x direction. Two

microphones are placed at $x = 0$ and $x = 1$ foot, respectively. What is the phase difference, in radians, between the signals picked up by the two microphones?

Project: *frequency response of a bandpass filter*

This project shows how to measure the frequency response of a filter, whether it's a designed one (as in this case) or it's something that acts as an unintentional filter (such as a loudspeaker that doesn't have a flat frequency response—and, in fact, none of them do.)

The filter we'll measure is the bandpass filter supplied in the music 170 library (called "bandpass"). It's a classical filter design that appears often in digital audio applications.

To measure it, make a "sinusoid" object and pass it through a "bandpass" object. Make two "meter" objects, and connect one to the output of the oscillator (so that you see what you're inputting to the filter) and one to the filter output so you can see how the two levels differ.

If you want to save time later, you can slightly complicate the patch by inserting a multiplier between the oscillator and its two connections (with a constant to multiply it by) so that you can adjust the oscillator's output to a round number in dB; but this isn't necessary to finish the project.

We're interested in two settings of the filter: the center frequency should be 1000 Hz, and the value of "Q" set to 10 and to 20. For each of these two filter settings do the following:

Set the oscillator's frequency to a series of values separated from 1000 by half octaves:

```
31, 44, 62, 88, 125, 176, 250, 353, 500, 707, 1000,
1414, 2000, 2828, 4000, 5656, 8000
```

With these numbers evenly spaced on the horizontal axis (a logarithmic scale), plot on the vertical axis the gain in decibels (the output level of the filter minus the input level). These numbers will all be negative. (Suggestion: find all the 34 values first—each filter's gain at each of the 17 frequencies shown—so that you will know what the bounds of the graph should be.) Draw two traces, one for each of the two filters. Enjoy the fact that at high frequencies you get two nearly parallel lines. How many decibels per octave do the filters' frequency responses drop off by at frequencies above about 2000?

Chapter 7

Sound Radiation and Measurement

Because sounds in real air are so different from signals, the things that convert between the two—microphones and speakers—aren't neutral elements that we can ignore but, instead, often have a huge impact on a sound production chain,. Both speakers and microphones come in a bewilderingly complicated array of types, shapes, and sizes, and what is well suited to one task might be ill suited to another.

Of the two, speakers are the more complicated because their greater size means that the spatial anomalies in their output are much greater. Microphones, on the other hand, can often be thought of as sampling sounds at a single point in space. So that's where we will start this chapter.

7.1 Microphones

At a fixed point in space, sound requires four numbers to determine it: the time-varying pressure and a three-component vector that gives the velocity of air at that point. All four are functions of time. (There are other variables such as displacement and acceleration, but these are determined by the velocity.)

Suppose now that there is a sinusoidal plane wave, with peak pressure a, traveling in any direction at all. It still passes through the origin since it is

defined everywhere. Its pressure at the origin is still the function of time:

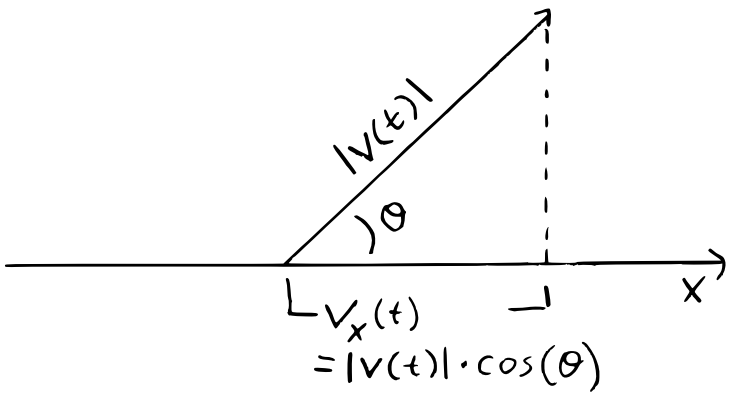
$$p(t) = a\cos(2\pi f t)$$

(Once more we're ignoring the initial phase term). The velocity is a vector $v(t)$, whose magnitude we denote by $|v(t)|$. In a suitably well-behaved gas, this magnitude is equal to:

$$|v(t)| = \frac{c}{p_a} p(t)$$

(Here, c is the speed of sound, as defined in Section 6.3, and p_a is atmospheric pressure, about 101,000 newtons per square meter.)

Now suppose the plane wave is traveling at an angle θ from the x axis, and that we want to know the velocity component in the x direction. (Knowing this will allow us also to get the velocity components in any other specific direction because we can just orient the coordinate system with x pointing in whatever direction we want to look in.) The situation looks like this:



The desired x component of the velocity is equal to the projection of the vector $v(t)$ onto the x axis which is the magnitude of $v(t)$ multiplied by the cosine of the angle θ:

$$v_x(t) = \cos(\theta)\frac{c}{p_a} p(t)$$

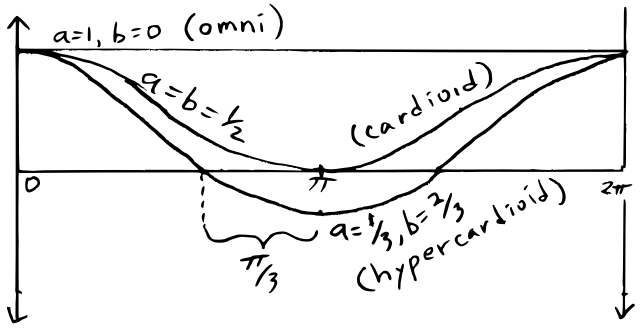
Ideally a microphone measures some linear combination of the pressure $p(t)$ and the three velocity components $v_x(t), v_y(t), v_z(t)$. Whatever combination of the three velocity components we're talking about, we can take to be only a multiple of the x component by suitably choosing a coordinate system so

that the microphone is pointing in the negative x direction (so that it picks up the velocity in the positive x direction—microphones always get pointed in the direction the wind is presumed to be coming from). Then the only combinations we can get are of the form

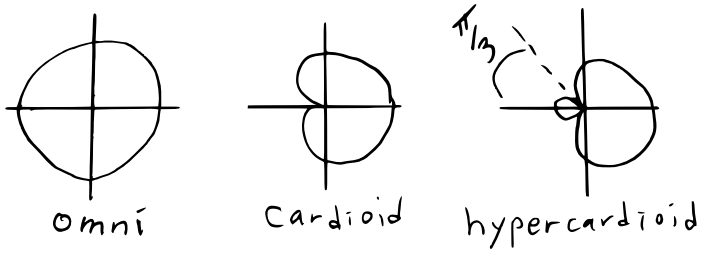
$$ap(t) + b\frac{p_a}{c}v_x(t)$$

where we can choose the parameters a and b freely. (We took a slight liberty and inserted a term p_a/c so that a and b have the same normalization.)

Now we consider what happens when a plane wave comes in at various choices of angle θ. If the plane wave is coming in in the positive x direction (directly into the microphone) the $\cos(\theta)$ term equals one, and we plug in the formula for v_x to find that the microphone picks up a signal of strength $a+b$. If, on the other hand, the plane wave comes from the opposite direction, the strength is $|a - b|$. Here are three choices of a and b, all arranged so that $a + b = 1$ so that they have the same gain in the direction $\theta = 0$:



Here are the same three examples, shown as polar plots, so that θ is shown as the angle from the positive x axis, and the absolute value of the magnitude is the distance from the origin—this is how microphone pickup patterns are most often graphed:

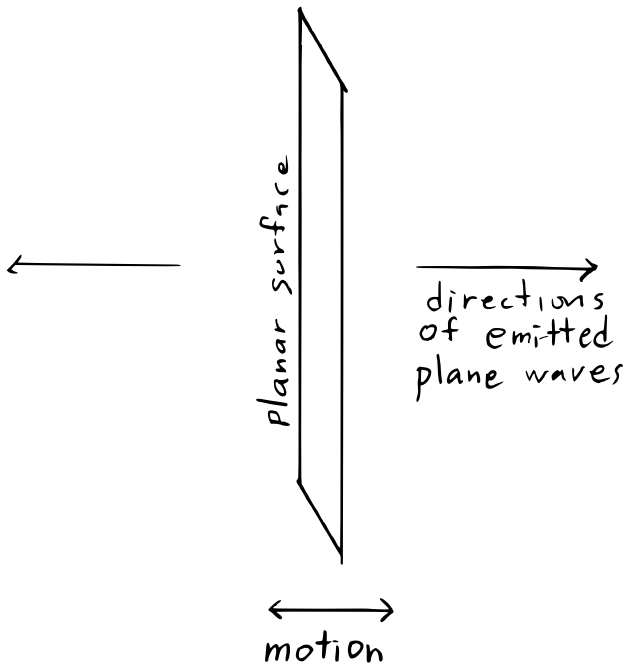


Note that the picture is now turned around so that the mic is now pointing in the positive x direction, in order to make it agree with the convention for graphing microphone pickup patterns. Also, for the example at right, for angles close to π the response is negative, but I showed the absolute value to avoid having the curve fold inside itself to avoid confusion.

In the special case where $a = 1$ and $b = 0$ (i.e. the microphone is sensitive to pressure only and not at all to velocity), there is no dependence on the direction at all. Such a microphone is called *omnidirectional*. If $a = b = 1/2$, so that there is no pickup in the opposite direction ($\theta = \pi$), the microphone is *cardioid* (named for the shape of the polar plot above). If $b > 1/2$ there's still an angle θ at which there's no pickup, and at greater angles than that the pickup is negative (out of phase). Such a microphone is called *hypercardioid*. In general, a microphone's pickup gain as a function of angle is called its *pickup pattern*.

7.2 Radiation From Large Planar Objects

In order to study how vibrating solid objects (such as loudspeakers) radiate sounds into the air, it's best to start with a simple, idealized situation, that of an infinitely large planar surface. The scenario is as pictured:



Supposing that the planar surface is vibrating, sinusoidally, in the direction shown, so that its motion is

$$x(t) = d\cos(2\pi f t)$$

with the amplitude d in units of distance. On each side of the plane this generates a plane wave moving away from the solid plane. (There's no choice but to emit a plane wave because of the symmetry of the situation; nothing in the setup depends on y or z.)

Paying attention to the right-hand side of the plane (the left-hand side acts in the same way but with the sign of x terms in the equations negated), the plane wave has pressure equal to

$$p(x, y, z, t) = a\cos(2\pi f(t - x/c))$$

where a is the peak pressure. As we saw in the previous section, the velocity is given by:

$$v(x, y, z, t) = \frac{ca}{p_a}\cos(2\pi f(t - x/c) + \phi)$$

From this we can deduce the physical displacement (by integrating the velocity, or, if you haven't studied calculus, by making arguments based on

spinning bicycle wheels):

$$x(x, y, z, t) = \frac{ca}{2\pi p_a f} \sin(2\pi f(t - x/c) + \phi)$$

This must equal the x position of the solid plane, and so to find the constants ϕ and a, we equate the two. After some rewriting we get:

$$\phi = \pi/2$$

$$d = \frac{ca}{2\pi p_a f}$$

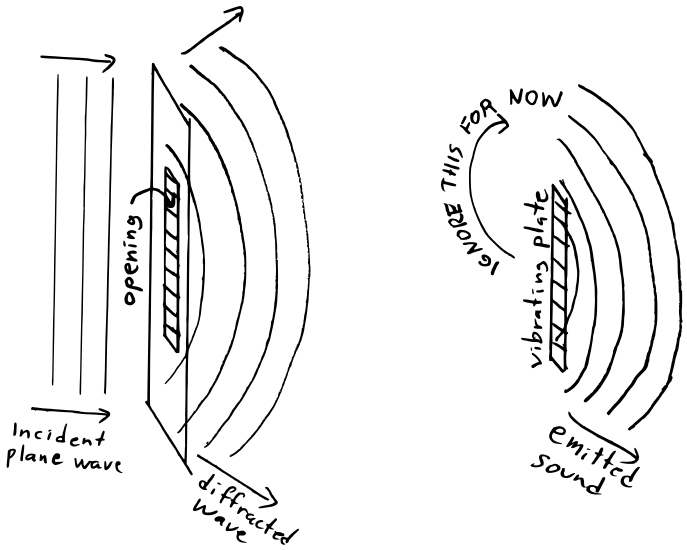
This shows that, to generate a plane wave of a desired amplitude a, we have to make the solid plane move physically at an amplitude that depends inversely in frequency. Generating low-frequency sounds requires more physical motion (and hence is harder and more expensive to do) than generating high-frequency ones at the same SPL. In fact, generating a 20 Hz. sound takes 100 times the physical motion that would be required to generate a 2 kHz, one (and consulting of the equal loudness contours, we see that to get an equal loudness the factor rises further yet!) This is why your computer speakers don't do well at low frequencies.

7.3 Radiation From Rectangular Objects: Diffraction

From the preceding section we can predict in general terms that a very large, flat vibrating object will emit a unidirectional beam of sound. This is theoretically true enough, but in our experience sound doesn't move in perfect beams; it's able to round corners. This phenomenon is called *diffraction*.

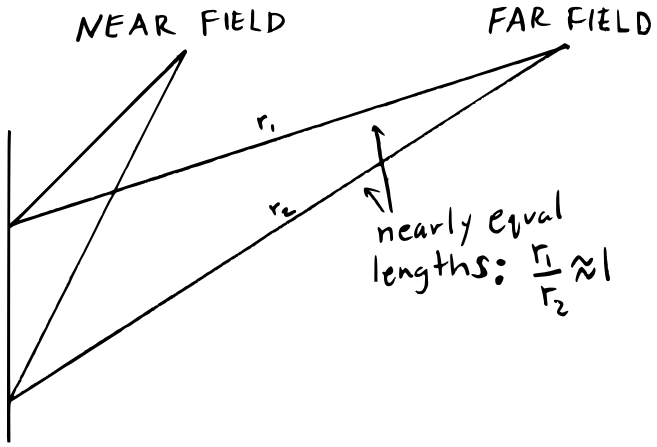
Although sound moving in empty space may be described as plane waves, this description breaks down when sound is absorbed or emitted by solid object, and also when it negotiates corners. Typically, higher-frequency sounds (whose wavelengths are shorter) tend to move more in beams, whereas lower-frequency, longer-wavelength ones, are more easily diffracted and can sometimes seem not to move directionally at all.

To study this behavior, we can set up a thought experiment, which comes in two slightly different forms shown below: on the left, sound passing through a rectangular window, and on the right, sound radiating from a rectangular solid plate:



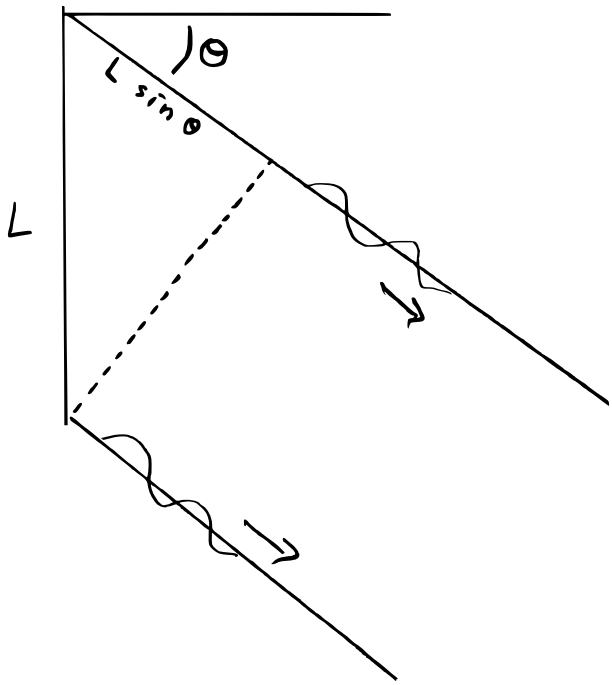
The two situations are similar in that, from the right-hand side of each solid object, we can approximately say that everything we hear was emitted from one point or other on either the hole (in the left-hand-side picture) or on the vibrating solid (on the right). In both situations we're ignoring what happens to the sound on the left-hand side of the object in question: in case of the opening, there will be sound reflected backward from the part of the barrier that isn't missing; and for the vibrating solid, not only is sound radiated on the other (left-hand) side, but in certain conditions, that sound will curve around (diffract!) to the side we're listening on. Later we'll see that we can ignore this if the object is many wavelengths long (at whatever frequency we're considering), but not otherwise.

In either case, we consider that the rectangular surface of interest radiates as if all of its infinitude of points radiated independently and additively. So, to consider what happens at a point in space, we simply trace all possible segments from a point on the surface to our listening point. The diagram below shows two such radiating points, and their effects on two listening points, one nearby, and the other further away:



At faraway listening points, such as the one at right, the distances from all the radiating points are roughly equal (that is, their ratios are nearly one). The RMS amplitude of the radiation is proportional to one over the distance (as we saw in Section 6.2). This region is called the *far field*. At closer distances (especially at distances smaller than the size of the radiating surface), the distances are much less equal, and nearby points increasingly dominate the sound. In particular, close to the edge of the radiating surface, it's the points at the edge that dominate everything.

In the far field, the amplitude depends on distance as $1/r$, and on direction in a way that we can analyze using this diagram:



For simplicity we've reduced the situation to one spatial dimension (the same thing will happen along the other spatial dimension too.) A segment of length L is radiating a sinusoid of wavelength λ. We assume that we're listening to the result a large distance away (compared to L), at an angle θ from horizontal (that is, off axis). If θ is nonzero, the various points along the segment arrive at different times. They're evenly distributed over a time that lasts $L\sin(\theta)/\lambda$ periods; we'll denote this number by k.

If $\theta = 0$, so that $k = 0$ as well, we are listening head on to the radiating surface, all contributing paths arrive with the same delay, and we get the maximum possible amplitude. On the other hand, if $k = 1$, we end up summing the signal at relative phases ranging from 0 to 2π, an entire cycle. A cycle of a sinusoid sums to zero and so there is no sound at all. Assuming L is at least one wavelenth so that $\lambda/L < 1$, the sound forms a beam that

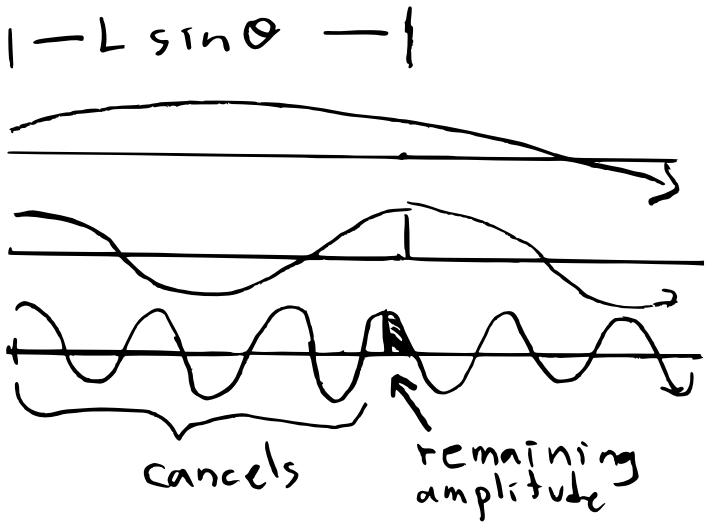
spreads at an angle

$$\theta = \sin^{-1}\left(\frac{\lambda}{L}\right)$$

For small values of λ/L (i.e., when L is many wavelengths long) the beam spreads at approximately:

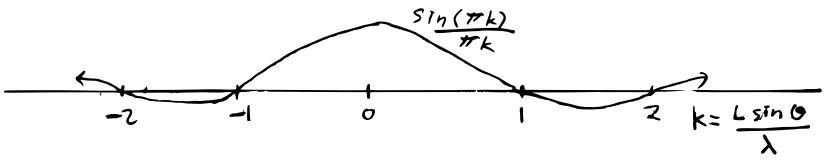
$$\theta \approx \frac{\lambda}{L}$$

In somewhat more detail, here is what happens when sinusoids f varying frequencies are summed over the fixed length $L\sin(\theta)$:



The only way to get a large sum is to have the wavelength be substantially larger than $L\sin(\theta)$. For smaller wavelengths, for which several cycles fit in the length, even if the number of cycles isn't an integer so that the whole sum doesn't cancel out, there is only a residual amount left after all the complete cycles are left to cancel out (the dark region in the bottom trace above, for example).

If we now consider what happens when we fix L and λ (that is, we consider only a fixed frequency), we can compute how the strength of transmission depends on the angle θ. We get the result graphed here:



We get a central "spot" of maximum amplitude, surrounded by fringes of much smaller amplitude, interleaved with zero crossings at regular intervals. This function recurs often in signal processing (since we often end up summing a signal over a fixed length of time) and it's called the "sinx" (pronounced "cinch") function:

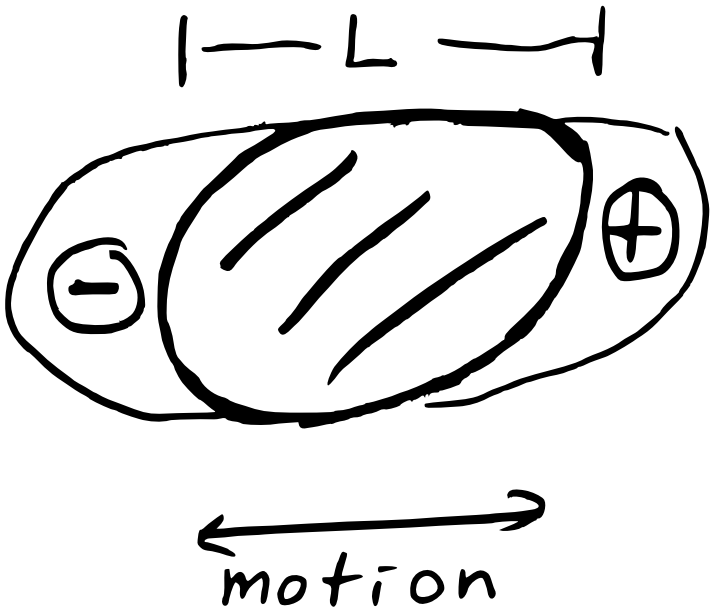
$$\text{sinx}(x) = \begin{cases} 1 & x = 0 \\ \sin(x)/x & \text{otherwise} \end{cases}$$

One thing that happens isn't very well explained by this analysis: if what we're talking about is a vibrating rectangle (instead of an opening), some of the radiation actually diffracts all the way around to the other direction. This is, roughly peaking, because the near field is itself limited in size, and so can be thought of as an opening whose radiation is in turn diffracted all over again. In situations where the wavelength of the sound is larger than the object itself, this is in fact the dominant mode of radiation, and everything pretty much spills out in all directions equally.

The dependence of the strength of transmission of sound emitted by an object, depending on angle, is called the object's *radiation pattern*. In general it depends on frequency, as in the simple example that was worked out here.

7.4 Radiation From Real Objects

Although a few objects in everyday experience radiate sound roughly spherically by injecting air directly into the atmosphere (the end of a sounding air column, for instance, or a firecracker), in practice most objects that make sound do so by vibrating from side to side (for instance, loudspeakers, or the sounding board on a string instrument). Assuming the object is approximately rigid (not a safe assumption at high frequencies but sometimes OK at low ones), the situation is as pictured:



As the object wobbles from side to side (supposing at the moment it is wobbling to the right) it will push air into the atmosphere on its right-hand side (labeled "+" in the diagram), and suck an equal amount of air back out on its left-hand side (marked "-").

Denoting the diameter of the object as L, we consider two cases. First, at frequencies high enough that the wavelength is much smaller than L, sound radiated from the two sides will be at least somewhat directional, so that, listening from the right-hand side of the object, for example, we will hear primarily the influence of the "+" area. The sound will radiate primarily to the left and right; if we listen from the top or bottom, the two will arrive out of phase and, although in practice they will never cancel each other out, they won't exactly add either. The sound radiated will be at least somewhat directional.

At low frequencies, the sound diffracts so that the radiation pattern is more uniform. At frequencies so low that the wavelength is several times the size L, the sound from the "+" and "-" regions arrive at phases that differ only by $2L/\lambda$ cycles, and if this is much smaller than one, they will nearly cancel each other out. We conclude that in the far field, *vibrating objects are lousy at radiating sounds at wavelengths much longer than themselves.*

If we stay in the near field, on the other hand, the "+" region might be proportionally much closer to us than the "-" region, so that the cancellation doesn't occur and the low frequencies aren't significantly attenuated. So a microphone placed within less than the diameter of an object tends to pick up low frequencies at higher amplitudes than one placed further away. This is called the *proximity effect*, although it would perhaps have been better terminology to consider this the natural sound of the object and to give a name to the canceling effect of low-frequency radiation at a distance instead.

Exercises and Project

1. A sinusoidal plane wave at 20 Hz. has an SPL of 80 decibels. What is the RMS displacement (in millimeters.) of air (in other words, how far does the air move)?

2. A rectangular vibrating surface one foot long is vibrating at 2000 Hz. Assuming the speed of sound is 1000 feet per second, at what angle off axis should the beam's amplitude drop to zero?

3. How many dB less does a cardioid microphone pick up from an incoming sound 90 degrees ($\pi/2$ radians) off-axis, compared to a signal coming in frontally (at the angle of highest gain)?

4. Suppose a sound's SPL is 0 dB (i.e., it's about the threshold of hearing at 1 kHz.) What is the total power that you ear receives? (Assume, a bit generously, that the opening is 1 square centimeter).

5. If light moves at $3 \cdot 10^8$ meters per second, and if a certain color has a wavelength of 500 nanometers (billionths of a meter - it would look green to human eyes), what is the frequency? Would it be audible if it were a sound? [NOTE: I gave the speed of light incorrectly (mixed up the units, ouch!) - we'll accept answers based on either the right speed or the wrong one I first gave.]

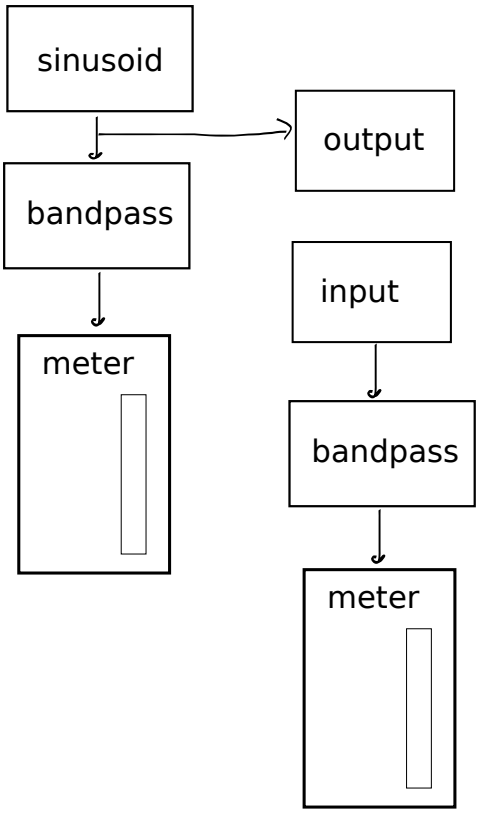
6. A speaker one meter from you is paying a tone at 440 Hz. If you move to a position 2 meters away (and then stop moving), at what pitch to you now hear the tone? (Hint: don't think too hard about this one).

Project: *why you really, really shouldn't trust your computer speaker.*

We know from project 1 that computer speakers perform badly at low frequencies, sometimes failing to do much of anything below 500 Hz. But

within the sweet spot of hearing, 1000 to 2000 Hz, say, are things starting to get normal? On my computer at least, it's impossible to believe anything at all about the audio system, as measured from speaker to microphone.

In this project, you'll measure the speaker-to-microphone gain of your laptop. but instead of looking at a wide range of frequencies, we're interested in a single octave, from 1000 to 2000, in steps of 100 Hz. The patch is somewhat complicated. You'll make a sinusoid to play out the speaker (no problem) but then you'll want to find the level, in dB, of what your microphone picks up. Since it will pick up a lot of other sound besides the sinusoid, you'll need to bandpass filter the input signal from the microphone to (at least approximately) isolate the sinusoid so you can measure it. Here's the block diagram I used:



To use it, set the Q for both bandpass filters to 10. Set the frequency of the sinusoid, and the center frequencies of both filters, to 1000. You should notice that the gain of the filter isn't one - this is why you have to monitor

the signal you're sending out the speaker through a matching filter, so that you're measuring the sinusoid's strength the same way on the output as on the input.

Now choosing a reasonable output level, and pushing the input level all the way to 100 (unity gain), verify that you're really measuring the input level in its meter (by turning the output off and on—you should see the level drop by at least about 20 dB when it's off. If not, you might have the filters set wrong, or perhaps Pd is mis-configured and looking for the wrong input.) Also, don't choose an output level so high it distorts the sinusoid - you should be able to hear if this is happening. Most likely an output level from 70 to 80 will be best.

The gain is the difference between the input and output levels (quite possibly negative; that's no problem). Now find the gain (from output to input, via the speaker and microphone) for frequencies 1000, 1100, 1200, ..., 2000 (eleven values). For each value, be sure you've set all three frequencies (the sinusoid and the two filters). Now graph the result, and if you've got less than 15 dB of variation your computer audio system is better than mine.

Chapter 8

Measurement, Control, and Interactivity

Historically, interactivity has played a central role in electronic music, since electronic musical instruments have usually used acoustical musical instruments as points of departure, and musical instruments are all designed to be played. In the design of electronic musical instruments, the way the instrument is played has often been at least as important as the method of sound generation, and in many important cases (the Theremin, for example), the means of playing the instrument is the instrument's defining characteristic.

In the electronic arts, too, the way sound is generated (if at all) may be less important than the way the art object responds to the changing situation, either by sensing peoples' voluntary or incidental actions, or by responding to environmental changes, which often arrive in the form of signals. Even if some signal (temperature, for example) doesn't have an acoustical origin, operations on signals that we've presented as acoustical can still be helpful. In any situation in which the quantities of interest are functions of time, we can act as if they were sounds.

To make an object that allows interactivity on the basis of signals we will need tools for measurement and for controlling aignal generators as functions of measured quantities.

8.1 Control in general

From a traditional viewpoint, the object of control is to find the inputs to a system that achieves a desired output. For instance, if you're heating a house, you have a desired ideal temperature and you can turn the heater on and off in order to try to keep the actual temperature as close as possible to the desired one. There is a whole theory about this, called, appropriately enough, control theory. This might be a reasonable way to think about the situation in which a violin player, for example, is trying to control the pitch the instrument is putting out. Assuming you can measure (hear) the violin's output, you then move your finger in one direction or the other to correct whatever the error might be. In other situations there isn't a specifically desired output; it might instead be desirable that the object behave in some complicated or unpredictable way.

Central to control theory is the possibility of feedback: a human (or a machine) measures what the output is at a moment in time and adjusts the input as a function of the measured output. If you can make an accurate measurement of the output, and if there aren't other external forces generating inputs to a system (noise, for instance, or variations in time of the system's characteristics), you should be able, by trial and error, to get to any desired output that is in the range of the system. But of course, it's often the possibility of other external forces (such as an antagonistic player) or time variation (the ice sculpture is gradually melting) that makes the whole thing interesting.

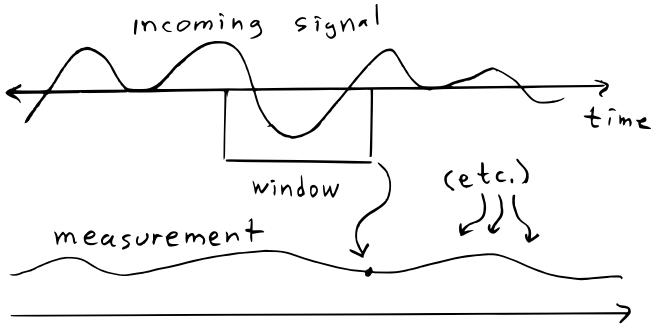
Even in the absence of such interference, it may be important not only to be able to reach a desired result but to be able to do so in a fixed amount of time, or make the output reach a succession of values at specified times. Or perhaps it's not important to arrive at specific pitches at specific times, but rather to make any of a huge number of possible pitch curves that will plant the desired perceived pitches in the listener's head. Such things are sometimes made easier through an understanding of control theory but are often only accessible through a long learning process.

8.2 Measurement

In some situations measurements are best made directly by hardware (photoresistors, accelerometers, etc), and appear to the computer as voltages

that can be converted to signals using either the computer's audio ADCs or other types of ADCs that are more appropriately adapted to "control" inputs. (ADCs that re optimized for audio input often have built-in high-pass filters that reject constant offsets, such as the DC output of a stationary accelerometer, that might be important in a non-audio context, whereas many kinds of physical measurements can be transmitted at much lower bit rates, and hence more efficiantly, than audio ADCs use.)

Other types of measurement can be done on audio (or other) signals to generate other signals. An example we're already mentioned is measuring the average power of a signal. A more sophisticated (and trickier) example is measuring an audio signal's pitch. In measurements either of power or pitch, one usually wants not a single answer but, instead, a series of estimates made at different times. Each estimate then depends on values of the signal over an interval of time called a *window*, as shown:



For each sample of the output, a separate analysis is run over the window consisting of the most recent N samples of input. The parameter N is called the window size. (In practice, it's often not necessary to recompute an analysis such as power or pitch for every sample of output but at longer intervals, since they don't tend to change as fast as the samples of the signal being analyzed.)

Most algorithms for making measurements on signals do some form of averaging to make aggregate estimates about the signal's behavior over the span of a window. The larger the window, the more accurate such a measurement may potentially be. On the other hand, it may be desirable to keep a window size small if a high time resolution is needed. There is often a trade-off between time resolution and accuracy.

As a matter of efficiency many sorts of measurements (wither directly har-

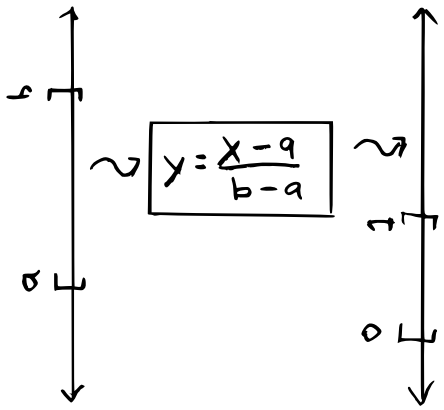
vested from the outside world or measured from other signals) can be represented as signals at a lower sample rate than the audio sample rate, but on the other hand some require higher ones than the audio sample rate. In practice audio software often maintains multiple sample rates to address these situations, but we won't worry about that for now.

8.3 Manipulating Signals as Controls

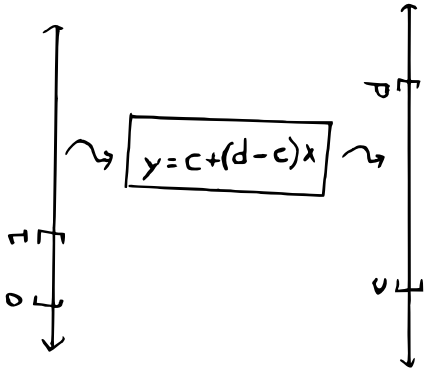
A signal may be used to control the generation or processing of another signal. The distinction between an "audio signal" and a "control signal" is almost a purely psychological one; for instance, one might control the amplitude of a signal a by multiplying it by a (more slowly changing) signal b. But we might instead regard a as the "control" signal and b and the "audio" one. But even though the difference has no substance, it is a useful one to maintain because certain operations are more likely to come up in "control" usage than "audio" usage. And signals that are thought of as "measurements" in the senses described above are likely to be used in "control" contexts.

The most frequent, and perhaps the most fundamental, issue that comes up in control is *scaling*. If you have a measurement whose natural range is from a to b (for instance, a thermometer outside in San Diego might give outputs ranging from 40 to 80), and if you want to control something whose range is from c to d (for instance, you might want the frequency of an oscillator to range from 110 to 440), you will at the very least want to be able to convert one range to another. In real situations there will typically be many different such ranges to convert between, and the most efficient way to manage this is often to standardize on a range (most conveniently, the *unit interval*, which reaches from 0 to 1) and to be able to convert signals with other ranges to and from that one.

To convert a signal ranging from a to b to one whose range is the unit interval, first subtract a (so that the range is now from 0 to $b-a$), and then divide by $b-a$ to re-scale the upper value to 1, as shown:

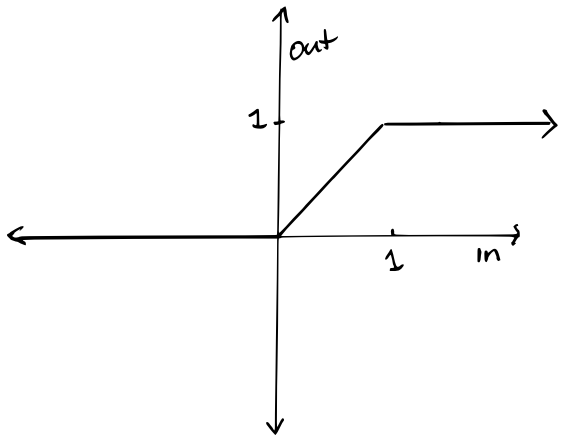


And supposing that we have a signal whose range is the unit interval and that we wish to make it range from c to d, we do the reverse: first re-scale it so that it ranges from 0 to $d - c$, and then add c to slide the range to where it should start:



In these drawings we have chosen $b > a$ and $d > c$ but we could have chosen the reverse, in which case the conversion would invert the direction of the signal.

Going back to the thermometer example, it might be a good thing to deal with cases where the thermometer strays outside of its intended range. The easiest way to do this is to *clip* the signal to its desired range. Assuming that we have scaled the signal so that its desired range is the unit interval, we can then clip the signal by replacing values outside the range with the appropriate endpoint (0 for negative numbers, and 1 for numbers greater than one). That is equivalent to applying this function to the signal:



Anohter class of operations on signals deals with their behavior in time. Some of these, such as delays, are already familiar as audio operations. In particular, filtering, which is useful in audio processing as a way to modify the spectrum of a signal, may be used on control signals for a different purpose: smoothing a control that changes too abruptly or noisily. For example, if we wish to control the amplitude of an audio signal using a switch (which appears to us as a signal that jumps between 0 and 1), we might wish to alter the switch's output so that it ramps between 0 and 1 over an interval of time on the order of 1/20 second. One conceptually simple way to do this is simply to low-pass filter the signal at 20 Hz. Doing this allows us to avoid the click that sounds when a signal changes discontinuously.

8.4 Event Detection

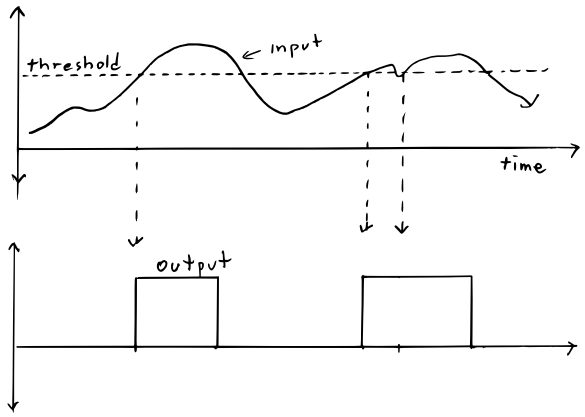
In computer audio applications (and, more generally, in the electronic arts), it is frequently desirable to detect a natural event or a human action and to make a causal response to it. An *event* can be loosely defined as the knowledge that a certain thing has happened at a certain time, often accompanied with some data to further describe it. For example, if you press a key on a musical keyboard, this can be made to generate an event in software, that might be accompanied with data specifying which key was pressed and how hard.

Other things treated as events in interactive computer software might include user input on a computer, arriving network packets, or the ringing of a virtual alarm clock. Here, since we're focusing on audio signals, we'll

only worry about a specific class of events that occur when we detect some feature in an audio signal. One might wish to be able to detect features such as the presence of speech, or the arrival of a specific pitch from an instrument, or whatnot, and the detection of the event might require sophisticated software.

Here, we'll just look at what might be the simplest example, which is threshold detection, in which we generate an event whenever a signal exceeds a fixed *threshold*, such as the temperature of a room rising above 70. Since the thing being measured could itself be the result of many different possible calculations, even though the notion of threshold detection is very simple, it can be very powerful.

Although in most software events are treated as a completely different type of data from signals. for our purposes, since we've only manipulated signals so far in these course notes, we'll offer a notion of threshold detection that results in a pulsed signal, as shown:



The output is a rectangular pulse. Unlike the pulses we've seen before that are for listening to, and that are rounded to control the audio bandwidth, a rectangular pulse changes discontinuously from one sample to the next to mark an event. (This is the way an event would be marked in an analog synthesizer; they are not much used in computer audio but we're using one here so that we can stay in the framework of audio signals.)

We could now design signal operations that are *triggered* by pulses, in the way an analog sequencer or envelope generator would; but for now we'll concent ourselves with just obtaining one.

Exercises and Project

[These are all review problems.]

1. A square plate is vibrating sinusoidally to create a 'beam' of sound. (Idealize this as in chapter 7 to a 1-foot line segment). At what frequency must it vibrate so that the beam spreads 30 degrees to either side (that is, so that the intensity drops to zero 30 degrees off axis)?

2. If you wish to form a beam with the same dispersion (spread), at a frequency two octaves lower, by what factor would you have to increase the dimensions of the square plate?

3. How many watts should you emit from a speaker (assuming the sound goes equally in all directions) to reach a sound level of 80 dB at a distance of 10 meters? (Assume you're away from any reflecting objects so that you only need consider the direct sound.)

4. What is the wavelength, in air, of the musical F above A440 (the musical A defined as 440 Hz.)? (You can answer in feet with c=1000 feet per second, or in meters at 343 M/sec.) Assume we're using the tempered scale.

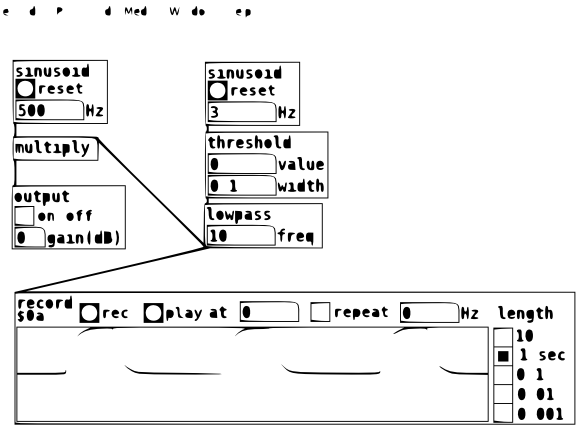
5. If a critical band is 300 Hz. wide, what, approximately, is its frequency range (given by its bottom and top frequencies)?

6. How fast must a sound source be moving away from you so that Doppler shift makes the pitch of the sound decrease by one octave?

Project: *low-pass filtering as a smoothing operation.*

Perhaps the most fundamental and important tool in dealing with sounds is controlling amplitudes by applying a gain to a signal. You do this any time you change the volume on your phone, for example. It's not as simple as it sounds. If you change the gain of an amplifier too quickly the sound will not just change amplitude but will often make an audible clicking sound as it does so. This is a major problem if the quality of the sound matters.

Here is a patch to demonstrate/test this idea:



The three objects on the left are a straightforward sinusoid with amplitude control via a "multiply" object. On the left, we're generating a signal to turn the sinusoid on and off, by thresholding another, slow sinusoid (the one at top left.) The threshold signal is a series rectangular pulses, three per second, each one 0.1 seconds long.

We're using a "lowpass" object to smooth the edges of the pulses. The cutoff frequency of the low-pass filter determines the sharpness of the edges. In the picture above, the cutoff frequency is set very low to exaggerate this effect so that you can see it. You might want to try values between 2 and 20 Hz. to see how they affect the picture you get from the "record" object at bottom.

The assignment is to find out how much smoothing you need to be able to turn the sinusoid on and off without hearing an audible "click" or "pop". This will turn out to depend on the frequency of the sinusoid.

First, set the frequency of the sinusoid (at upper left) to 2000. Adjust the low-pass filter's cutoff frequency to 20000 Hz. (essentially no filtering at all) and enjoy the clicks. Then drop the frequency to 5000, 2000, 1000, 500, etc., until you find the value at which you just hear a sinusoid turning on and off without artifacts. (Don't be a perfectionist... you can always convince yourself you hear a clock or pop, just get it so that it's not easily audible.)

Now do the same thing with the sinusoid set to 500 Hz, and finally repeat the experiment after replacing the sinusoid with a "noise" object. What are the three values you had to set the low-pass filter to to hear "clean" turn-ons and turn-offs for the three situations (2000, 500, noise)?

Index