Sound waves

Last lecture we considered sound to be a pressure function $I(X, Y, Z, t)$. However, sound is not just any function of those four variables. Rather, sound obeys the *wave equation*:

$$\frac{\partial^2 I(X,Y,Z,t)}{\partial X^2} + \frac{\partial^2 I(X,Y,Z,t)}{\partial Y^2} + \frac{\partial^2 I(X,Y,Z,t)}{\partial Z^2} = \frac{1}{v^2}\frac{\partial^2 I(X,Y,Z,t)}{\partial t^2}$$

where v is the speed of sound. This equation says that if you take a snapshot of the pressure function at any time t, then the spatial derivatives the pressure function at each point XYZ tell you how the pressure at the point will change as time varies. Note that this equation contains the consant v which is the speed of sound.

The speed of sound in air is about $v = 340$ meters per second, or 34 cm per millisecond. This is quite slow. (If you go to a baseball game and you sit behind the outfield fence over 100 m away, you can easily perceive the delay between when you *see* the ball hit the bat, and when you *hear* the ball hit the bat.) Amazingly, the speed of sound is so slow that our brains can detect differences in the arrival times of sounds at the left and right ear, and we use this difference to help us perceive where sound sources are. (We'll discuss this in the following few lectures.)

Also notice that the wave equation is linear in $I(X, Y, Z, t)$. If you have several sources of sound, then the pressure function I that results is identical to the sum of the pressure functions produced by the individual sources in isolation.

Today we will examine two types of sounds that are of great interest: music and speech. We will see how a frequency domain analysis is fundamental to both.

Musical sounds

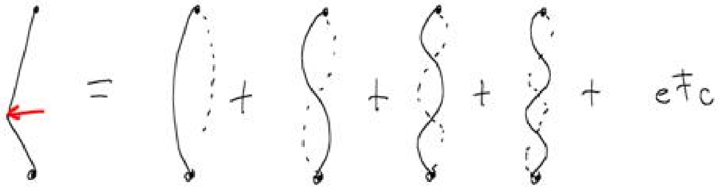
Let's begin by briefly considering string instruments such as guitars. First consider the vibrating string. When we pluck the guitar string, we are setting its initial shape to something different than its resting state. This initial shape and the subsequent shape as it vibrates always has fixed end points. The initial shape can be written as a sum of sine functions, specifically sine functions with value zero at the end points. This is summation is similar a Fourier transform, but here we only need sine functions (not sines and cosines), in particular,

$$\sin(\frac{\pi}{L}xm)$$

where $m \geq 0$ is an integer and L is the length of the guitar string. We have π rather than 2π in the numerator since the sine value is zero when $x = \frac{L}{m}$ for any m.

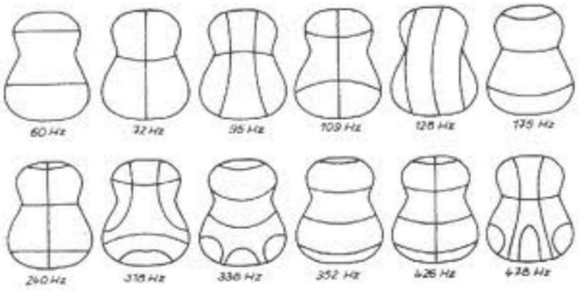
Physics tells us that if a string is of length L then its mode $\sin(\frac{\pi}{L}x)$ vibrates at a temporal frequency $\omega = \frac{c}{L}$ where c is a constant that depends on the properties of the string such as its material, thickness, tension. Think of each mode m of vibration as dividing the string into equal size parts of size $\frac{L}{m}$. For example, we would have four parts of length $\frac{L}{4}$. (See sketch in slide). You can think of each of these parts as being little strings with fixed endpoints.

Frequency m is called the *m-th harmonic*. The frequency $\omega_0 = \frac{c}{L}$ i.e. $m = 1$ is called the *fundamental* frequency. Frequencies for $m > 1$ are called *overtones*. Note harmonic frequencies have a linear progression $m\omega_0$. They are multiples of the fundamental.



Note that the *definition* of harmonic frequencies is that they are an integer multiple of a fundamental frequency. It just happens to be the case that vibrating strings naturally produce a set of harmonic frequencies. There are other ways to get harmonic frequencies as well, for example, voiced sounds as we will see later.

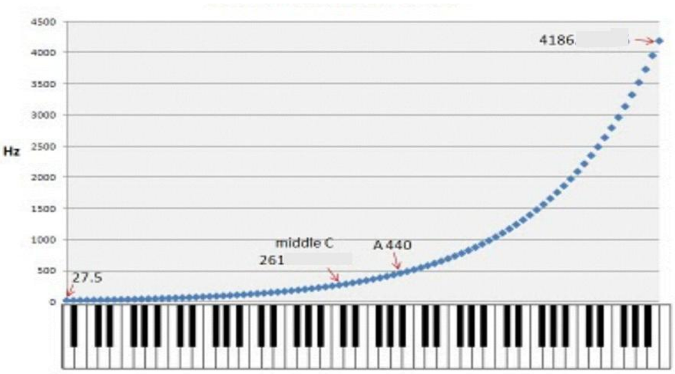
For stringed instruments such as a guitar, most of the sound that you hear comes not from the vibrating strings, but rather the sound comes from the vibrations of the instrument body (neck, front and back plates) in response to the vibrating strings. The body has its own vibration modes as shown below. The curved lines in the figure are the nodal points which do not move. Unlike the string, the body modes do not define an arithmetic progression.



For another example, see
`http://www.acs.psu.edu/drussell/guitars/hummingbird.html`

In western music, notes have letter names and are periodic: A, B, C, D, E, F, G, A, B, C, D, E, F, G, A, B, C, D, E, F, G, etc. Each of these notes defines a fundamental frequency. The consecutive fundamental frequencies of the notes for any letter (say C) are separated by one octave. e.g. A, B, C, D, E, F, G, A covers one octave. Recall from the linear systems lecture that a difference of one octave is a doubling of the frequency, and in general two frequencies ω_1 and ω_2 are separated by $\log_2 \frac{\omega_2}{\omega_1}$ octaves.

An octave is partitioned into 12 intervals called *semitones*. The intervals are each $\frac{1}{12}$ of an octave, i.e. equal intervals on a log scale. A to B, C to D, D to E, F to G, and G to A are all two semitones, whereas B to C and E to F are each one semitone. (No, I don't know the history of that.) It follows that the number of semitones between a note with fundamental ω_1 and a note with fundamental ω_2 is $12 \log_2 \frac{\omega_2}{\omega_1}$. To put it another way, the frequency that is n semitones above ω_1 is $\omega_1 2^{\frac{n}{12}}$. The notes on a piano keyboard are shown below, along with a plot of their fundamental frequencies.



Notice that the frequencies of consecutive semitones define a geometric progression, whereas consecutive harmonics of a string define an arithmetic progression. When you play a note on a piano keyboard, the sound that results contains the fundamental as well as all the overtones - which form an arithmetic progression. When you play multiple notes, the sound contains the fundamentals of each note as well as the overtones of each. [ASIDE: The reason why some chords (multiple notes played together) sound better than other has to do – in part – with the distribution of the overtones of the notes, namely how well they align. Details omitted.]

Speech sounds

Human speech sounds have very particular properties. They obey certain physical constraints, namely our anatomy. Speech sounds depend on several variables. One is the shape of the *oral cavity*, which is the space inside your mouth. This shape is defined by the tongue, lips, and jaw position which are known as *articulators*. The sound wave that you hear has passed from the lungs, past the vocal cords, and through the long cavity (pharynx + oral and nasal cavity) before it exits the body. The shape of the oral cavity is determined by the position of the tongue, the jaw, the lips.

Consider the different vowel sounds in normal spoken English "aaaaaa", "eeeeeee", "iiiiiii", "oooooo", "uuuuuuu". Make these sounds to yourself and notice how you need to move your tongue, lips, and jaw around. These variations are determined by the positioning of the articulators. Think of the vocal tract (the volume between the vocal cords and the mouth and nose) as a resonant tube, like a bottle. Changing the shape of the tube by varying the articulators causes different sound frequencies that are emitted from you to be amplified and others to be attenuated.

Voiced Sounds

Certain sounds require that your vocal cords vibrate while other sounds require that they do not vibrate. When vocal cords are tensed, the sounds that result are called *voiced*. An example is a tone produced by a singing voice. When the vocal cords are relaxed, the sounds are called *unvoiced*. An example is whispering. Normal human speech is a combination of voiced and unvoiced sounds.

Voiced sounds are formed by regular pulses of air from the vocal cords. There is an opening in the vocal cords called the *glottis*. When the vocal cords are tensed, the glottis opens and closes at a regular rate. A typical rate for glottal "pulses" for adult males and females are around 100 and 200 Hz *i.e.* about a 10 ms or 5 ms period, although this can vary a lot depending on whether one has a deep versus average versus high voice. Moreover, each person can change their glottal frequency by varying the tension. That is what happens when you sing different notes.

Suppose you have $n_{glottal}$ glottal pulses which occur with period $T_{glottal}$ (time between pulses). The total duration would be $T = n_{glottal} T_{glottal}$ time samples. We can write the sound source pressure signal that is due to the glottal pulse train as:

$$I(t) = \sum_{j=0}^{n_{\text{glottal}}-1} g(t - jT_{glottal})$$

where $g()$ is the sound pressure due to each glottal pulse. We can write this equivalently as

$$I(t) \ = \ g(t) * \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{glottal}).$$

Each glottal pulse gets further shaped by the oral and nasal cavities. The oral cavity in particular depends on the positions of the articulators. If the articulators are fixed in place over some time interval, each glottal pulse will undergo the same waveform change in that interval. Some people speak very quickly but not so quickly that the position of the tongue, jaw and mouth changes over time scales of the order of say 10 ms. Indeed, if you could move your articulators that quickly, then your speech would not be comprehensible.

One can model the transformed glottal pulse train as a convolution with a function $a(t)$, so the final emitted sound is:

$$I(t) \ = \ a(t) * g(t) * \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{glottal})$$

So you can think of $a(t) * g(t)$ as a single impulse response function. The reason for separating them is that there really are two different things happening here. The glottal pulse $g(t)$ is not an impulse function and it is different from the effect $a(t)$ of the articulators. Each glottal pulse produces its own $a(t) * g(t)$ pressure wave and these little waves follow one after the other.

Let's next briefly consider the frequency properties of voiced sounds. If we take the Fourier transform of $I(t)$ over T time samples – and we assume the articulators are fixed in position so that we can define $a(t)$ and we assume $T_{glottal}$ is fixed over that time also – we get

$$\hat{I}(\omega) \ = \ \hat{a}(\omega) \ \hat{g}(\omega) \ \mathbf{F} \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{glottal}).$$

You can show (in Assignment 3) that

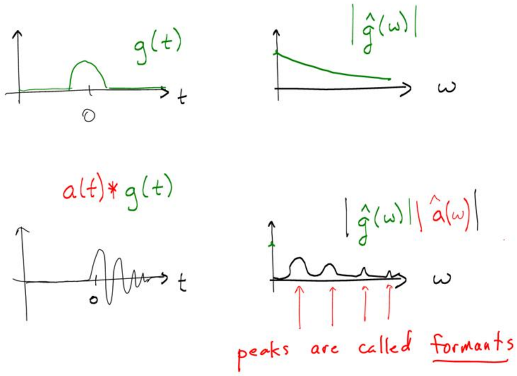
$$\mathbf{F} \sum_{j=0}^{n_{\text{glottal}}} \delta(t - jT_{glottal}) = n_{glottal} \sum_{j=0}^{T_{glottal}-1} \delta(\omega - jn_{glottal})$$

So,

$$\hat{I}(\omega) \ = \ \hat{a}(\omega) \ \hat{g}(\omega) \ n_{glottal} \sum_{j=0}^{T_{glottal}-1} \delta(\omega - jn_{glottal})$$

This means that the glottal pulses cancel out all frequencies except other than those that are a multiple of $n_{glottal} \ = \frac{T}{T_{glottal}}$, that is, the number glottal pulses per T samples. I emphasize here that this clean mathematical result requires that the sequence of glottal pulses spans the T samples, and the period is regular and the articulators are fixed during that interval.

Measurements show that the glottal pulse $g(t)$ is a low pass function. You can think of it as having a smooth amplitude spectrum, somewhere between a Gaussian amplitude spectrum which falls off quickly and an impulse amplitude spectrum which is constant over ω.



The effect of the articulators is to modulate the amplitude spectrum that is produced by the glottal pulses, namely by multiplying by $\hat{a}(\omega)$. This amplifies some frequencies and attenuates others. (It also produces phase shifts which we will ignore in this analysis, but which are important if one considers the wave shape of each pulse.) The peaks of the amplitude spectrum $|\hat{g}(\omega) \ \hat{a}(\omega)|$ are called *formants*. As you change the shape of your mouth and you move your jaw, you change $a(t)$ which changes the frequencies of the formants. I will mention formants again later when I discuss spectrograms.

As mentioned above, the sum of delta functions nulls out frequencies except those that happen to be part of an arithmetic progression of fundamental frequency $\omega_0 = n_{glottal} = \frac{T}{T_{glottal}}$, that is, $n_{glottal}$ samples per T time steps. However, we often want to express our frequencies in cycles per second rather than cycles per T samples. The typical sampling rate used in high quality digital audio is 44,100 samples per second, or about 44 samples per ms.[1] To obtain the frequency in cycles per second, convert $n_{glottal}$ pulses in T samples to pulses per sample by $n_{glottal}/T$ and the convert to cycles per second by multiplying by 44,100 samples per second.

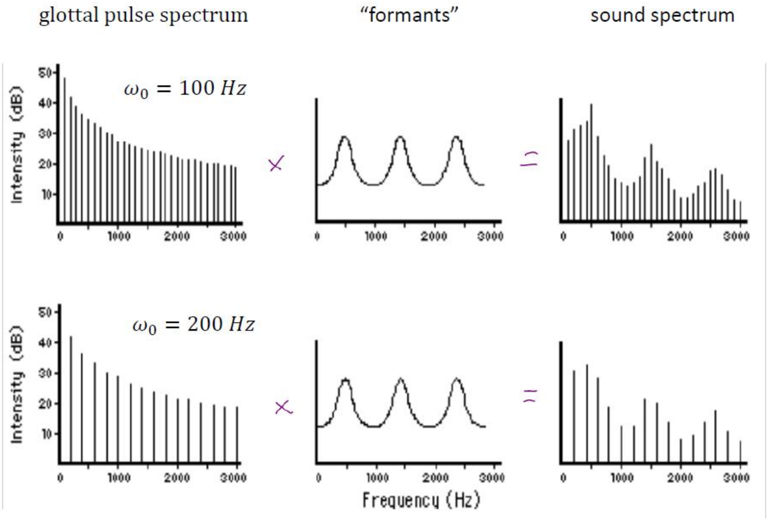
This sampling rate is not the only one that is used, though. Telephone uses a lower sampling rate, for example, since quality is less important.

The frequency 44,100 * $n_{glottal}/T$ is the fundamental frequency in cycles per second, which correponds the glottal pulse train. As mentioned earlier, in adult males this is typically around 100 Hz for normal spoken voice. In adult females, it is typically around 200 Hz. In children, it is often higher than 250 Hz.

The two rows in the figure below illustrate a voiced sound with fundamental 100 and 200 Hz. The left panels shows just amplitude spectrum of the glottal pulse train. The center panels illustrate the

[1] One often uses 16 bits for each of two channels (two speakers or two headphones).

amplitude spectrum of the articulators for several formants. The right panel shows the amplitude spectrum of the resulting sound.



Unvoiced sounds (whispering)

When the vocal cords are relaxed, the resulting sounds are called *unvoiced*. There are no glottal pulses. Instead, the sound wave that enters the oral cavity can be described better as noise. The changes that are produced by the articulators, etc are roughly the same in voiced versus unvoiced speech, but the sounds that are produced are quite different. You can still recognize speech when someone whispers. That's because there is still the same shaping of the different frequencies into the formants, and so the vowels are still defined. But now it is the noise that gets shaped rather than glottal pulses.

I mentioned in the lecture that the noise $n(t)$ produced by expelling air from the lungs has a flat amplitude spectrum, hat is, prior to the reshaping of the spectrum by the articulators. The sound that comes out the mouth is $n(t) * a(t)$ and that sound is shaped by the articulators.

Consonants

Another important speech sound occurs when one restricts the flow of air, and force it through a small opening. For example, consider the sound produced when the upper front teeth contact the lower lip. Compare this to when the lower front teeth are put in contact with the upper lip. (The latter is not part of English. I suggest you amuse yourself by experimenting with the sounds you can make in this way.) Compare these to when the tongue is put in contact with the front part of the palate vs. the back part of the palate.

Most *consonants* are defined this way, namely by a partial or complete blockage of air flow. There are several classes of consonants. Let's consider a few of them. For each, you should consider what is causing the blockage (lips, tongue, palate).

- fricatives (narrow constriction in vocal tract):
  - voiced: z, v, zh, th (as in *the*)
  - unvoiced: s, f, sh, th (as in θ)

- stops (temporary cessation of air flow):
  - voiced: b, d, g
  - unvoiced: p, t, k

  These are distinguished by where in the mouth the flow is cutoff. Stops are accompanied by a brief silence

- nasals (oral cavity is blocked, but nasal cavity is open)
  - voiced: m, n, ng

  You might not believe me when I tell you that nasal sounds actually come out of your nose. Try shutting your mouth, plugging your nose with your fingers, and saying "mmmmm". See what happens?

Spectrograms

When we considered voiced sounds, we took the Fourier transform over T samples and assumed that the voiced sound extended over those samples. One typically does not know in advance the duration of voiced sounds, so one has to arbitrary choose a time interval.

Often one analyzes the frequency content of a sound by partitioning $I(t)$ into blocks of B disjoint intervals each containing T samples – the total duration of the sound would be BT. For example, if $T = 512$ and the sampling rate is 44000 samples per second, then each interval would be about 12 milliseconds.

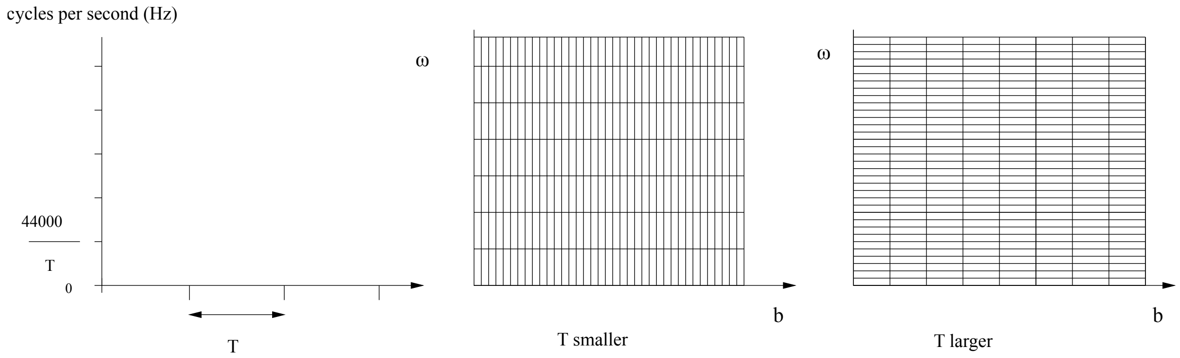
Let's compute the discrete Fourier transform on the T samples in each of these block. Let ω be the frequency variable, namely cycles per T samples, where $\omega = 0, 1, \ldots, T - 1$. Consider a 2D function which is the Fourier transform of block b:

$$\hat{I}(b, \omega) \ = \ \sum_{t=0}^{T-1} I(\ b\ T + t)\ e^{-i \frac{2\pi}{T} \omega t}.$$

Typically one ignores the phase of the Fourier transform here, and so one only plots the amplitude $\mid \hat{I}(b, \omega) \mid$. You can plot such a function as a 2D "image", which is called a *spectrogram.*

The sketch in the middle shows a spectrogram with a smaller T, and the sketch on the right shows one with a larger T. The one in the middle is called a "wideband" spectrogram because each 'pixel' of the spectrogram has a wide range of frequncies, and the one on the right is called a narrowband spectrogram because each 'pixel' has a smaller range of frequencies. For example, if $T = 512$ samples, each pixel would be about 12 ms wide and the steps in ω would be 86 Hz high, whereas if $T = 2048$ samples, then each pixel would be be 48 ms wide and the ω steps would be 21 Hz.

Notice that we cannot simultaneously localize the properties of the signal in time and in frequency. If you want good frequency resolution (small ω steps), then you need to estimate the frequency components over long time intervals. Similarly, if you want good temporal resolution (i.e. when exactly does something happen?), then you can only make coarse statements about which frequencies are present "when" that event happens. This inverse relationship is similar to what we observed earlier when we discussed the Gaussian and its Fourier transform.



Examples (see slides)

The slides show a few examples of spectrograms of speech sounds, in particular, vowels. The horizontal bands of frequencies are the formants which I mentioned earlier. Each vowel sound is characterized by the relative positions of the three formants. For an adult male, the first formant (called F1) is typically centered anywhere from 200 to 800 Hz. The second formant F2 from 800 to 2400 Hz, F3 from 2000 to 3000 Hz.