Introduction to sound

A few lectures from now, we will consider problems of spatial hearing which I loosely define as problems of using sounds to compute where objects are in 3D space. We’ll look at two kinds of spatial hearing problems. One involves sounds that are emitted by objects in the world. The other involves sounds that are reflected off objects in the world.

Emitted sounds are produced when forces are applied to an object that make the object vibrate (oscillate). When one moving object hits another object, the kinetic energy of the moving object is transformed into potential energy by an elastic compression. This elastic compression results in vibrations of the object(s) which dampen out over time depending on the material of the object. These object vibrations produce tiny pressure changes in the air surrounding the object, since as the object vibrates it bumps into the air molecules next to it. These air pressure changes then propagate as waves into the surrounding air.

Emitted sounds are important for hearing. They inform us about events occurring around us, such as footsteps, a person talking, an approaching vehicle, etc. Here we have a major difference between hearing and vision. Nearly all the visible surfaces reflect light rather than emit light. However, light sources themselves are generally not informative for vision, but rather their ‘role’ is to illuminate other objects, that is, shining light on other objects so that the visual system can use this reflected light to estimate 3D scene properties. Emitting sound sources are informative to us. They tell us about the location and material properties (or identities) of objects.

What about reflected sounds? To what extent are they useful? Blind people seem to use reflected sounds (echos) to navigate. As long as an environment is reasonably quiet, blind people can walk through an environment without bumping into walls. To do so, they use the echos of their footsteps and the echos of the tapping of their cane. They hear the reflections of these sounds off walls and other obstacles. What about people that have normal vision? (We probably also use echos, although not nearly as much as blind people. This hasn’t been studied much.)

Animals such as bats and dolphins actively use reflected sounds to decide both where objects are in space and what their material properties are. e.g. Bats can easily distinguish a flying moth from a blowing leaf. I will tell you about this a few lectures from now.

Pressure vs. intensity

Sound is a set of air pressure waves that are measured by the ear. Sound can travel through liquids and solids as well as air. We will restrict our discussion to air. Air pressure is always positive. It oscillates about some mean value $I_a$ which we call atmospheric pressure. The units of air pressure are atmospheres and the mean air pressure around us is approximately “one atmosphere”.

Sounds are small variations in air pressure about this mean value $I_a$. These pressure variations can be either positive (compression) or negative (rarefaction). We will treat the pressure at a point in 3D space as a function of time,

$$P(X,Y,Z,t) = I_a + I(X,Y,Z,t)$$

---

1. Another component of the kinetic energy is transformed into non-elastic mechanical energy, namely a permanent shape change. This happens when the object cracks, chips, breaks or is dented.
2. Typically only hot objects emit light. Electronic displays/lights e.g. LEDs are obvious exceptions to this statement. Not only are they non-hot light emitters, but the light patterns they emit are often meant to be informative – perhaps the light pattern you are seeing right now.
where $I(X,Y,Z,t)$ is small compared to atmospheric pressure $I_a$. Note that we are using “big” $X,Y,Z$ rather than little $x,y,z$, since we are talking about points in 3D space.

I emphasize that sounds are quite small perturbations of the atmospheric pressure. The quietest sound that we can hear is a perturbation of $10^{-9}$ atmospheres. The loudest sound that we can tolerate without pain is $10^{-3}$ atmospheres. Thus, we are sensitive to 6 orders of magnitude of pressure changes. (An order of magnitude is a factor of 10.)

**Decibels: units for measuring loudness**

To refer to the loudness of a sound, one generally refers to either pressure changes $I(X,Y,Z,t)$ or intensity changes. The two are not the same. Pressure change is a function that oscillates about 0. Intensity is the square of this function, and so is always positive. Intensity corresponds to energy per unit volume, namely, the work done to compress (positive $I$) or expand (negative $I$) the unit volume of air to produce the deviation from $I_a$.

Sounds can occur over such a large range of pressure changes. A common measure of “physical loudness” of a sound is a ratio of the pressure of the sound to that of some standard $I_0$ which is very soft sound (called the “threshold of hearing”). Moreover, loudness of sounds can occur over such a large range of levels, and so one uses measures loudness by the log of this ratio. One defines:

$$\text{Bels} = \log_{10} \frac{I^2}{I_0^2} = 2 \log_{10} \left| \frac{I}{I_0} \right|$$

Note that intensity is used rather than pressure.

It is common to use a slightly different unit, namely ten times Bels:

$$\text{Decibels} \equiv 10 \log_{10} \frac{I^2}{I_0^2} = 20 \log_{10} \left| \frac{I}{I_0} \right|$$

Multiplying by 10 is convenient because the human auditory system is limited in its ability to discriminate sounds of different loudnesses, such that we can just discriminate sounds that different from each other by about 1 dB. So, we can think of 1 dB as a just noticeable difference (JND).

Here are a few examples of sound levels:

<table>
<thead>
<tr>
<th>Sound</th>
<th>dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>jet plane taking off (60 m)</td>
<td>120</td>
</tr>
<tr>
<td>noisy traffic</td>
<td>90</td>
</tr>
<tr>
<td>conversation (1 m)</td>
<td>60</td>
</tr>
<tr>
<td>middle of night quiet</td>
<td>30</td>
</tr>
<tr>
<td>recording studio</td>
<td>10</td>
</tr>
<tr>
<td>threshold of hearing</td>
<td>0</td>
</tr>
</tbody>
</table>

**Speed of sound**

Sound travels through air at speed about $v = 340$ meters per second. 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.
A scale that is more relevant to the human is the time it takes a sound to go the distance of say 17 cm which is roughly the distance between the ears (340 m/s = 34 cm/s. So, a sound that arrives at the head from the left will reach the left ear 1/2 millisecond before the right ear. The brain is quite sensitive to this time difference. (You are not consciously aware of this difference but your brain does use it.)

The phenomenon is analogous to stereo vision where you are not consciously aware of the small spatial disparities between the left and right eye. In stereovision, your brain fuses the left and right eyes images and converts the disparities to depths. In binaural hearing, your brain fuses the left and right ear’s images and converts the temporal disparities to a direction in 3D space. More on this later...

**Wave equation**

We are considering sound to be a pressure function $I(X, Y, Z, t)$. However, sound is not just any function of 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. Notice that this equation is linear. 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. (It is difficult enough to perceptually disentangle multiple sounds, even with the property that the signal is linear!)

<table>
<thead>
<tr>
<th>frequency (Hz)</th>
<th>wavelength (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>343</td>
</tr>
<tr>
<td>10</td>
<td>34.3</td>
</tr>
<tr>
<td>100</td>
<td>3.43</td>
</tr>
<tr>
<td>1000</td>
<td>34.3</td>
</tr>
<tr>
<td>10,000</td>
<td>3.43</td>
</tr>
</tbody>
</table>

**Sound impulse**

Consider an isolated perturbation (a pulse) of air pressure at 3D point $(X_o, Y_o, Z_o)$ and at time $t = t_o$, e.g. due to some impact. You can imagine a digital sound generation system with a speaker that generates this impulse. Intuitively, we suppose that pressure is constant in the 3D world (complete silence) and then suddenly there is an instantaneous impulse in pressure at some particular spatial location. What happens?

Mathematically, we could model this sound source as an impulse function $\delta(X - X_o, Y - Y_o, Z - Z_o, t - t_o)$. How does this point impulse change over time? Think of a stone dropped in the water. After the impact, you get an expanding circle. Because sound also is obeys a wave equation, the same phenomenon occurs, except now you are in 3D and so you get a wave of expanding spheres. The speed of the wavefronts is the speed of sound. After one millisecond, the sphere is of radius 34 cm. After two milliseconds, the sphere is of radius 68 cm, etc.
In general, the area of a sphere is $4\pi r^2$, where

$$r = \sqrt{(X - X_0)^2 + (Y - Y_0)^2 + (Z - Z_0)^2}$$

and

$$r = v \cdot (t - t_0)$$

where $v$ is the speed of sound. At any $t$, the intensity of the sound is approximately constant along the sphere. (We ignore loss of energy due to friction/attenuation in the air, which in fact can be substantial for high frequencies). Thus, the intensity $I^2$ (which is energy per unit volume) on a sphere of radius $r$ must fall off as $1/r^2$. So, the pressure $I$ falls off as $1/r$. If the point $(X_0, Y_0, Z_0)$ is far from the origin, then the impulse will reach the origin at time $t = t_0 + \frac{r}{v}$ and the impulse will be roughly a plane in the neighborhood of the origin.

Let $I_{\text{src}}(t_0)$ be a constant that indicates the strength of the impulse. Then at a distance $r$ away from the source, for example at the origin, the impulse is a function,

$$I(t) = \frac{I_{\text{src}}(t_0)}{r} \delta(r - v(t - t_0)).$$

Of course, a real sound source won’t be just a single impulse, but rather will have a finite time duration. Think of a person talking or shaking keys, etc. Even an impact which seems to have quite a short duration will in fact have a duration over tens of milliseconds, as the vibrations of the object dampen to zero. We can model a more general sound source as a sum of impulses

$$I(t) = \sum_{t_0} \frac{I_{\text{src}}(t_0)}{r} \delta(r - v(t - t_0)).$$

Keep in mind our earlier assumption that the source is at a point $(X_0, Y_0, Z_0)$ that is a large distance $r$ away. The pressure signal $I(t)$ measured at the origin in fact represents the pressure signal in a spatial neighborhood of the origin. That is, although this space-time neighborhood is defined by four variables $(X, Y, Z, t)$, the pressure in fact only depends on one variable namely $r - v(t - t_0)$ where, for a small neighborhood, $r$ is roughly constant along planes whose normal is $(X_0, Y_0, Z_0)$ i.e. the direction from the origin to the source. We define the normal of this plane by two angles $(\theta, \phi)$ which are spherical coordinates. So, we can write $I(t; \theta, \phi)$. We can define the spherical coordinates such that the north and south pole are directly above and directly below (up/down, in head coordinates), respectively. Then $\theta$ is the azimuth (longitude) and $\phi$ is the elevation (latitude). “Straight ahead” marks zero azimuth ($\theta = 0$). We will return to these terms later.

**Interaural Timing Differences**

To compare the arrival time difference for the two ears, we begin with a simple model that relates the pressure signals measured by the left and right ears:

$$I_l(t) = I_r(t - \tau)$$
where $\tau$ is the time delay. The auditory system is not given $\tau$ explicitly, of course. Rather it has to estimate it. We can formulate this estimation problem as follows. Find $\tau$ that minimizes:

$$
\sum_{t=1}^{T} (I_l(t) - I_r(t - \tau))^2.
$$

Intuitively, we wish to shift the right ear’s signal by $\tau$ so that it lines up in time with the left ear’s signal. If the signals could be lined up perfectly, then the sum of square differences would be zero. Note that, to find the minimum over $\tau$, the auditory system only needs to consider $\tau$ in the range $[-\frac{1}{2}, \frac{1}{2}]$ ms, which is the time it takes sound to go the distance between the ears.

**ADDED: (not mentioned in lecture)**

Minimizing (1) is equivalent to minimizing

$$
\sum_{t=1}^{T} I_l(t)^2 + \sum_{t=1}^{T} I_r(t - \tau)^2 - 2 \sum_{t=1}^{T} I_l(t)I_r(t - \tau).
$$

The summations in the first two terms are slightly different because of the shift $\tau$ in the second term. However, if $\tau$ is small relative to the time duration of $T$ samples (if $T$ samples cover several milliseconds), then the vast majority of sample points are shared by the two summations. Moreover, if the mean square intensity of $I_r(t)$ and $I_l(t)$ are roughly the constant over an interval of $T$ samples $\pm \frac{1}{2}$ms, then we can ignore the explicit dependence on $\tau$ in the second term. One typically makes this assumption, and finds the $\tau$ that maximizes

$$
\sum_{t=1}^{T} I_l(t)I_r(t - \tau).
$$

which is the cross-correlation of $I_l(t)$ and $I_r(t)$. Thus, the auditory system would try to find the $\tau$ that maximizes the cross-correlation.

**Interaural Intensity Differences**

One limitation of the above model is that it does not consider the shadowing that is caused by the head. Because of this shadowing, the intensity of the sound at the left ear may be different from that at the right ear. To account for the above intensity difference, the above model could be generalized to find the $\tau$ that minimizes:

$$
\sum_{t=1}^{T} (I_l(t) - \alpha I_r(t - \tau))^2.
$$

where $\alpha$ is an unknown scaling factor. Verify for yourself that posing the problem this way doesn’t fundamentally change the minimization discussed above: expanding the above expression we would see that the minimum is achieved again by finding the $\tau$ that maximizes the cross correlation above.
ADDED: (not mentioned in lecture)

As I will discuss in upcoming lecture, the shadowing by the head turns out to be useful as a cue since the relative loudness in the two ears is (like timing difference) a cue to position. To compare the loudnesses (in dB) of the sound over some short time interval, we could compute the decibel (dB) difference,

$$\Delta I = 10 \log_{10} \frac{\sum_{t=1}^{T} I_l(t)^2}{\sum_{t=1}^{T} I_r(t)^2}$$

The time range and sampling rate of the sum is not specified here, but since there may be interaural delays of up to $1/2$ ms, the time range should be at least a few milliseconds so the ears are measuring the same sound sources over this time interval.

Note that by using a log scale, we factor out the absolute level of the sound. If we make the sound source twice as loud, then it is twice as loud in the two ears which tells us nothing about location. Taking the ratio automatically cancels out this scale factor.

**Cone of confusion**

Note that timing differences do not uniquely specify direction. Consider the line passing through the two ears. This line and the center of the head, together with an angle $\theta \in [0, \pi]$, defines the so-called cone of confusion. If we treat the head as an isolated sphere floating in space, then all directions along a single cone of confusion produce the same intensity difference and the same timing difference.

Does this provide an ultimate limit on our ability to detect where sounds are coming from? No it doesn’t and the reason is that the head is not a sphere floating in space. The head is attached to the body (in particular the shoulders) which reflects sound in an asymmetric way, and the head has ears (the pinna) which shape the sound wave in a manner that depends on the direction from which the sound is coming. As we will see next lecture, there is an enormous amount of information available which breaks the cone of confusion.
Musical sounds

Let’s look briefly at string instruments such as guitars. First consider the vibrating string. Then endpoints of the string are held fixed. Physics tells us that the temporal frequency of vibration of a string of length $L$ is

$$\omega = \frac{c}{L}$$

where $c$ is a constant that depends on the properties of the string such as its material, thickness, tension, etc. Each mode of vibration divides the string into equal size parts of size $L, \frac{L}{2}, \frac{L}{3}, \frac{L}{4}, \ldots$ and so on. 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. The temporal frequencies of the vibration modes are $\frac{c}{L}j$ where $j = 1, 2, \ldots$. Frequency $\omega j$ is called the $j$th harmonic. The frequency $\omega_0 = \frac{c}{L}$ i.e. $j = 1$ is called the fundamental frequency. Frequencies for $j > 1$ are called overtones.

For stringed instruments, most of the sound is produced by vibrations of the instrument body (neck, front and back plates). The body has its own vibration modes. Unlike the string, the body modes do not define an arithmetic progression. See slide 7 for illustrations.

In western music, we have notes 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 any letter separated by an "octave". e.g. A, B, C, D, E, F, 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. More generally, two frequencies $\omega_1$ and $\omega_2$ are separated by $\log_2 \frac{\omega_2}{\omega_1}$ octaves.)

In western music, an octave is partitioned into 12 intervals called semitones. The intervals are each $\frac{1}{12}$ of an octave, i.e. equal size in log space. The intervals are also called semitones. 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. So, for example, the number of semitones between some fundamental $\omega_0$ and some other frequency $\omega$ is $12 \log_2 \frac{\omega}{\omega_0}$. To put it another way, the frequency that is $n$ semitones above $\omega_0$ is $\omega_0 2^{n/12}$. Note that frequencies of consecutive semitones define a geometric progression, whereas consecutive harmonics of a string defined an arithmetic progression.

Speech sounds

Human speech sounds have very particular properties. They obey certain physical constraints. I don’t mean that each human language has a relatively small number of words. Rather, I mean that the human anatomy itself constrains the kinds of sounds that can be produced when a person speaks. We all have roughly the same physical dimensions. Some people are smaller or bigger than average - children are smaller than adults - and there are differences in quality of the sounds of different people’s voices. But the similarities between voices are significant nonetheless. (I emphasize this has very little to do with the language a person is speaking.)

The sound that is emitted by a person depend on several factors. One key factor 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 you hear is a wave that 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

---

3No, I don’t know why. I have not read the history of this.
tongue, the jaw, the lips. Consider the different vowel sounds in normal spoken English “aaaaaa”, “eeeee”, “iiiiii”, “oooooo”, “uuuuuu”. 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 to be amplified and others to be attenuated.

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 not tensed, the sounds are called unvoiced. An example is whispering. Normal human speech is a combination of voiced and unvoiced sounds.

**Voiced 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 spoken speech has about a 10 ms period. You can change this frequency by providing different amounts of tension. That is what happens when you sing different notes.

Each glottal pulse get shaped into a wave which depends on the position of the articulators. As long as the articulators are fixed in place over some time interval, each glottal pulse will undergo the same waveform change in that interval. Some people talk very quickly but not so quickly that the position of the tongue, jaw and mouth changes over time scales of the order of 10 ms.

For a given positioning of the articulators, suppose you have \( n_{\text{pulse}} \) glottal pulses which occur with period \( T_g \), so \( T = n_{\text{pulse}}T_g \). We can write the sound source pressure signal that results as:

\[
I_{\text{src}}(t) = \sum_{m=0}^{n_{\text{pulse}}-1} g(t - mT_g)
\]

This signal is periodic with period \( T_g \), where \( T_g \) is about 10 ms. You can think of each \( g() \) as the result of a glottal pulse pressure wave that has been shaped into particular waveform by the oral and nasal cavities, i.e.

\[
I_{\text{src}}(t) = g(t) * \sum_{m=0}^{n_{\text{pulse}}} \delta(t - mT_g)
\]

The articulators then shape the source sound. They enhance some frequencies, attenuate some frequencies, and cause phase shifts to each frequency. We can model the changes simply as a convolution with a function \( a(t) \), so the final sound produced is:

\[
I(t) = a(t) * g(t) * \sum_{m=0}^{n_{\text{pulse}}} \delta(t - mT_g)
\]

Another way to think of it is that each glottal pulse produces a \( a(t) * g(t) \) 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 $I(t)$, we get

$$\hat{I}(\omega) = \hat{a}(\omega) \hat{g}(\omega) \mathbf{F} \sum_{m=0}^{n_{\text{pulse}}} \delta(t - mT_g)$$

You can show (see Exercises) that

$$\mathbf{F} \sum_{m=0}^{n_{\text{pulse}}} \delta(t - mT_g) = \frac{T}{T_g} \sum_{j=0}^{T_g-1} \delta(\omega - j \frac{T}{T_g})$$

So,

$$\hat{I}(\omega) = \hat{a}(\omega) \hat{g}(\omega) \frac{T}{T_g} \sum_{j=0}^{T_g-1} \delta(\omega - j \frac{T}{T_g})$$

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. See the slides for illustration.

Note that the sum of delta functions zeros out frequencies except those that happen to be part of an arithmetic progression of fundamental frequency $\omega_0 = n_{\text{pulse}} / T_g$, that is, $n_{\text{pulse}}$ samples per $T$ time steps. Or if you want to express it in cycles per second, it would be $n_{\text{pulse}} / T$ samples per second, or just $1 / T_g$ samples per second. For example, if the glottal pulses are every 10 ms, then the fundamental frequency would be 100 Hz. If the glottal pulses were every 5 ms, then the fundamental frequency would be 200 Hz.

The articulators modulate the amplitude spectrum that is produced by the glottal pulse train, by multiplying by $\hat{a}(\omega)$. This amplify some frequencies and attenuate others. (It also produces phase shifts which we will ignore.) The peaks of the amplitude spectrum of $\hat{a}(\omega)$ are called formants. As you change the shape of your mouth and you move your jaw, you vary the locations of the formants. See the slides for illustration. I will mention formants again when I discuss spectrograms.

Unvoiced sounds

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 is much closer to noise. The changes that are produced by the articulators, etc are roughly the same in whispered vs. non-whispered speech, but the sounds that are produced are quite very different, acoustically speaking. You can still recognize different unvoiced sounds of course. For example, the different vowels ”ooo”, ”eee”, etc. That’s because there is still a shaping of the noise.

Consonants

A related way to produce sounds is to restrict 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, front palatte, back palatte?).

- 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 ”mmmmmm”. See what happens?

Spectrograms

If we take the Fourier transform of a sound signal $I(t)$ of length $T$ samples, then we can get frequencies $\omega$ from 0 up to $T - 1$. The frequencies would be in units of cycles per $T$ samples, not cycles per second. (ASIDE: The typical sampling rate used in high quality digital audio is 44,000 samples per second, or equivalently, 44 samples per ms.)

The Fourier transform $\hat{I}(\omega)$ would be based on all the samples. This is not so useful since the sound source may change over the duration of the sound. A more useful representation for analyzing the frequency content of a sound is to partition $I(t)$ into $B$ disjoint intervals each containing $T_{\text{block}} = \frac{T}{B}$ samples. For example, if $T_{\text{block}} = 512$ and the sampling rate is 44000 samples per second, then each interval would be about 12 milliseconds (about ”4 meters of sound”) which is still a short interval. It is about the period pulses in a typical glottal pulse train for an adult male voice.

Let’s compute the discrete Fourier transform on the $T_{\text{block}}$ samples in each of these intervals. Let $\omega$ be the frequency variable, namely cycles per $T_{\text{block}}$ samples, where $\omega = 0, 1, \ldots, \frac{T_{\text{block}}}{2} - 1$. Consider a 2D function which is the Fourier transform of block $b$:

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

Typically one ignores the phase of the Fourier transform here, and so one only plots the amplitude $|\hat{I}(b, \omega)|$. 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 $T_{\text{block}}$, and the sketch on the right shows one with a larger $T_{\text{block}}$. For example, if $T_{\text{block}} = 512$, each pixel would be about 12 ms wide and the steps in $\omega$ would be 86 Hz high, whereas if $T_{\text{block}} = 2048$ then each ”pixel” would be be 48 ms wide and the $\omega$ 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 $\omega$ 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. We saw this uncertainty principle before when we discussed the Gaussian and its Fourier transform. Its the same idea here.

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.