[SQUEAKING]

[RUSTLING]

[CLICKING]

JOSH MCDERMOTT: Today, we're going to talk about pitch perception, and then get into auditory scene analysis, all right? So one of the really remarkable things about the sense of hearing is that most natural sounds contain many different frequencies. And yet, they sound like one thing.

So in particular, voices and instruments, among other sounds, tend to produce frequencies that are harmonics of a fundamental frequency. And so what that means is that the frequencies in the sound are integer multiples of a fundamental frequency. So here's just an example.

This is a schematic power spectrum. So we've got amplitude here. And this is frequency in hertz. But the harmonics here are labeled by their harmonic number. And that's because the frequencies that are in the sound have this very regular pattern to them. So we've got the fundamental frequency down here and 2 times that, 3 times that, 4 times that, all the way up. So on a linear frequency scale, these things are regularly spaced. So these are harmonics.

And so typically, when we talk about the sounds having a pitch, it usually refers to the fact that they're harmonic. And the pitch is thought to be the perceptual correlate of that fundamental frequency. So if you're playing a musical instrument and you're varying the pitch of the notes, what you're doing is varying the fundamental frequency of the notes. And that can be because you're varying the string length, for instance, or something else inside the instrument.

All right. So I was just playing you an excerpt of a tune called "Summer Madness" by Kool and the Gang. And I always like showing this when we talk about pitch because this has these very prominent octave jumps. So an octave is a ratio of 2. And so when you have two notes that are separated by an octave, that means that the fundamental frequency differs by a factor of 2. And then you can see all of the harmonics kind of stretch. And so we can just look at that one more time. So here's our live spectrogram. And this is Kool and the Gang.

[KOOL AND THE GANG, "SUMMER MADNESS"]

And so what you hopefully saw is that, after each one of those intervals, the harmonics all doubled. So the fundamental frequency doubled. And as a consequence of that, the spacing between the harmonics doubled. So that's the nature of harmonic sounds.

And so sounds can have the same fundamental frequency. And you might describe them as having the same pitch, despite them differing in other ways. So these are, again, schematic power spectra of six different sounds that all have the same fundamental frequency, despite having very different harmonic amplitudes. So we've got a piano, a trombone, tenor sax and then three spoken vowels.

And so you can see that the lowest frequency is the same across all of these sounds. And then the spacing between the frequencies they contain is the same. So they've got the same harmonic series, but they've been passed through different filters. And at the end of the lecture, we're going to talk about timbre. So these sounds we would all describe as having different timbres. But they've got the same fundamental frequency, or f0.

So another kind of important concept is that the fundamental frequency, it determines the period of the sound, even when it's not physically present. So this is a sound that has a missing fundamental. And it consists of the second, the third, and the fourth harmonic of 250 hertz. So the second harmonic of 250 hertz is 500 hertz, so twice that. Third harmonic is 750. And the fourth is 1,000.

And so you can see that, within this period of the sound, the 500 hertz component has two cycles. The 750 component has three cycles. And the 1,000 hertz component has four cycles. Now, if you add them all together to get the sound that this would correspond to, this is the waveform. So this is the sum of these three sinusoids.

And it's got some funny shape. But you can see that the period of the sound-- so the time interval over which the sound exerts one cycle after which it repeats-- is 4 milliseconds, which is 1 over the fundamental frequency of 250 hertz. So the fundamental frequency of this sound is 250 hertz, even though the sound does not actually contain any power at 250 hertz. So that 250 hertz component is missing physically from the sound. So the fundamental frequency is a more abstract notion. It describes the repetition rate of the sound.

And so the classical notion of pitch is that it is the perceptual analog of the fundamental frequency, or f0. And we typically think that sounds that have the same periodicity, which equivalently means they have the same f0, usually share the pitch. So here are some examples. So here's a pure tone.

[PURE TONE]

Here's a harmonic complex tone.

[HARMONIC COMPLEX TONE]

So this harmonic complex tone has the fundamental frequency that's equal to the frequency of the pure tone. But then it's got all these other harmonics on top of it. And so the waveforms, which are shown here, look really different. But when you listen to them, you might think there's something they have in common.

Now, in the bottom it's that same harmonic complex, but without the fundamental component. So the fundamental frequency of the sound is still that same f0, but the F0 is now physically missing from the sound.

[TWO NOTES]

All right. So you can tell that those sounds are not exactly the same, but there's some quality that they have in common. And we would often call that quality the pitch. So why does pitch matter? Well, it's a diagnostic property of some sounds. And for some familiar sounds, you know what their pitch is. And voices are a good example of this.

So this is a demo, where I have a recording of a person whose voice probably all of you are fairly familiar with. And then I have pitch shifted it artificially. So I'm going to play you the pitch-shifted versions. And then we'll play you the correct version. Raise your hand-- don't say who you think this is, but raise your hand if you think you can tell who this is.

[AUDIO PLAYBACK]

- There are a lot of points I want to make tonight, but the most important one--

[END PLAYBACK]

JOSH MCDERMOTT: OK, so a couple of you know. All right, let's play the really low one.

[AUDIO PLAYBACK]

- There are a lot of points I want to make tonight. But the most important one--

[END PLAYBACK]

JOSH MCDERMOTT: OK, let's go a little closer to the true pitch.

[AUDIO PLAYBACK]

- There are a lot of points I want to make tonight, but the most important one--

[END PLAYBACK]

[AUDIO PLAYBACK]

- There are a lot of points I want to make tonight, but the most important one--

[END PLAYBACK]

JOSH MCDERMOTT: All right, how many people know who that is? Raise your hand. All right, a few of you. But then here's the correct one.

[AUDIO PLAYBACK]

- There are a lot of points I want to make tonight, but the most important one is that, 20 years ago, I became the luckiest man on Earth because Michelle Obama agreed to marry me.

[END PLAYBACK]

JOSH MCDERMOTT: All right, so that's our former president. And I think probably you would agree that it's a little more recognizable when it's played at the correct pitch. And even those of you who could recognize that that was Barack Obama could probably tell that the pitch was not correct. So you know what the pitch is of familiar voices.

This is kind of a cool experiment that actually shows that you rely on pitch to recognize voices. So this was an experiment conducted by Malinda McPherson, a former graduate student here, who measured the ability to recognize celebrity voices. So she took all these recordings of famous people, did an experiment where participants heard excerpts of these people talking. And those excerpts could either be played at the original pitch or could be pitch-shifted up or down by up to an octave in either direction.

People would hear the excerpt, and then they would type in who they thought it was. And that was coded. And so this graph plots percent correct recognition as a function of the pitch shift. And so you can see that, when it's not shifted at all, people are recognizing a little over 50% of the celebrities. And that's just because lots of celebrities only some people know because not everybody watches the same TV shows and stuff like that. But as you shift the pitch, the recognition drops fairly precipitously. And even a pitch shift of three semitones-- that's a quarter of an octave-- is enough to give you a deficit.

So it's also often the case that information is conveyed by the way that pitch changes over time. And so that's certainly also very important in pitch and also in music. So prosody is the word that we use to refer to the pitch changes that occur in music. They often convey emotion and emphasis. So here's a particular example.

[AUDIO PLAYBACK]

- 903.

[END PLAYBACK]

JOSH MCDERMOTT: OK, so that's a person who is saying a number that doesn't have any particular emotional content. But you can tell they're really excited. Conversely--

[AUDIO PLAYBACK]

- December 4.

[END PLAYBACK]

JOSH MCDERMOTT: That person is really sad. They're just saying a date. But importantly, if we just shift the pitch up, so the pitch gets much higher, you can still kind of hear the emotion in their voice, because it's conveyed by the way the pitch is changing over time.

[AUDIO PLAYBACK]

- 903.

[END PLAYBACK]

[AUDIO PLAYBACK]

- December 4.

[END PLAYBACK]

JOSH MCDERMOTT:

OK, so that's relative pitch. All right. So classically, the problem of pitch perception has been construed as trying to understand how the brain estimates the f0 of a sound from its harmonics. And part of the idea of that is that a sound that has an f0 is harmonic. It consists of these different frequency components.

And the ear-- we often think of the ear as kind of doing a frequency analysis of a sound. So the sound comes into the ear, it's just a waveform, right? But it contains these discrete frequency components. And the ear is doing an analysis of the frequency content. And so, naively, you might think that the ear takes that sound, takes it apart into these constituent frequencies. And then you have to determine the properties of the sound, one of which is the fundamental frequency.

All right. And so one key idea in all of this is that the information that we have about the harmonics and the sound is limited by the frequency resolution of the cochlea. So what's shown here is a schematic power spectrum of a harmonic sound. This is a harmonic sound that's actually missing the fundamental. But you have all of the other harmonics. And this is a schematic of the filters that we think of as being in your cochlea.

So remember how, when we learned about the filters in your cochlea, we talked about how, in absolute terms, the filters at the low end of the cochlea-- so actually, at the apex, but tuned to low frequencies-- those tend to be narrower in hertz. And then as you move up the spectrum, the filters get wider. And that's what's shown here. So down here at the low end, the filters are narrow. And they get wider as you move up the spectrum.

So let's think about the representation of this harmonic sound when passed through these filters. So what's shown at the top is what's called an excitation pattern. Think of this as the magnitude of the response at each place along the cochlea. And you can see that you get these peaks and valleys here down at the apical end of the cochlea in response to some of the individual frequency components. But you can also see that those peaks and valleys peter out. So the higher numbered harmonics, you don't produce these discernible peaks and valleys.

So why is this? Well, let's think about the interactions of these harmonics with the filters. So this is harmonics 2 and 3. Harmonics 2 and 3 are actually quite widely spaced compared to the cochlear filter bank. And so the consequence is that they hit different filters. And so in particular, in between them, there's a filter that-- there's another filter there. That's the purple one. And so you get a peak and a trough there. And so that's true for a lot of these low-numbered harmonics. So we often call these harmonics resolved because they produce these discernible peaks in the excitation pattern of the cochlea.

By contrast, if you look up here at the high-numbered harmonics, you've got two filters here that are separated by the same number of hertz. And so they have the same separation relative to the harmonics. But they're much broader. And so now, the filters up there at the high are looking at a whole bunch of harmonics at once

And so the consequence is that the responses of different filters actually varies quite minimally. So even though you have a harmonic here and a harmonic here, there's very little variation in the response. And so this regime is classically referred to as being unresolved. So you're not able to resolve the individual components because of the coarseness of the tuning of the filter.

All right. So this is that same kind of concept depicted in a different picture. So now, we've got like the power spectrum up top, pass through the filter bank, you get the excitation pattern. The addition here is now, for different places along the cochlea, we're looking at the basilar membrane vibration over time

All right. And there's a few important things about this picture. So the first thing to note is that, down here at the apical end of the cochlea where it's tuned to low frequencies, if you look at a place that's being stimulated by a single harmonic, you can see that the basilar membrane vibration is dominated by that harmonic. It looks like a sinusoid.

So in this sense, the cochlea is really picking out that one low-numbered harmonic. Here's another place a little way up. It's the same kind of thing. But if you move a little ways up further, what you're seeing is a response that's no longer purely sinusoidal. In fact, there's amplitude modulation. Anybody remember what we call that?

AUDIENCE: Beating.

JOSH MCDERMOTT: Beating, yeah. So we're getting some beating. Remember, beating happens whenever there are two frequency components that are getting passed through the same-- so we hear beating. And it gets registered by the ear whenever they pass through the same filter. And that causes amplitude modulation. So that's what you're seeing. So this is beating in the response of the ear to the sound. And then as we move up the spectrum, the beating becomes more and more pronounced because there are more and more harmonics that are essentially making it through the same filter.

So up here at the basal end of the cochlea, when you're responding to high frequencies, the harmonics are unresolved, in the sense that the pattern of excitation along the cochlea is undifferentiated. But you do have this big temporal signal here. And in fact, the beating here happens at the fundamental frequency. So this has got the same period as the original sound.

So big picture is that there's these two cues that pop out of these kinds of diagrams as being plausible sources of information about the fundamental frequency of a sound. One is what are called place cues. So these are the peaks and valleys in the excitation pattern. And the other are timing cues here, which exist even for these high-numbered unresolved harmonics.

So now, I'm going to tell you a fact about pitch perception. And one very salient fact is that pitch discrimination turns out to be better in humans when resolved harmonics are present in the sound. And so you can do an experiment where you vary the harmonics that are present in a sound. And you can measure discrimination thresholds. So that would be like the just noticeable difference in fundamental frequency. It's how much you have to increase the pitch for somebody to tell that it went up or down.

So that's what's plotted here on the y-axis. This is the fundamental frequency difference that is the threshold for a bunch of humans. There are two different frequency ranges, which for our purposes doesn't really matter. And then the critical variable here is what's plotted on the x-axis, which is the lowest harmonic number in the stimulus.

And so if the lowest harmonic number is 3, that means that stimulus contains harmonics 3, 4, 5, 6, 7, up to some large number. If the lowest harmonic number is 10, that means that you start with the 10th harmonic. And you have 10, 11, 12, 13, 14, 15. And so what is evident from this results graph is that thresholds tend to be pretty good when the stimulus includes low-numbered harmonics. And then as you start to lose those low-numbered harmonics-- so when the lowest harmonic number is 8, or 10, or 12, or 15-- the thresholds increase.

All right. So here's a demo of that. So this is an example trial from an experiment. And so in this particular experiment, there are three stimuli. And you have to say which of them has a different pitch from the others. So let's start with something easy.

[THREE PITCHES]

1, 2, or 3?

AUDIENCE: 3.

JOSH MCDERMOTT: 3, good. So this was a stimulus that contained resolved harmonics. So I'm not going to go through the details of exactly how these stimuli were made. But they're passed through a filter and manipulated so that you've either got low-numbered harmonics that are resolved, thought to be resolved, or high-numbered harmonics that we call unresolved. So that's a 10% difference in fundamental frequency.

So just for comparison, a semitone, which is like the difference between adjacent white and black keys on a piano, that's about 6%. So 10%, in musical terms, is pretty substantial. So we expect you to be able to hear that difference because, if you couldn't, you wouldn't really be able to recognize a lot of melodies So that's for the resolved harmonics. Here's the same thing, but when the harmonics are unresolved. And it'll probably be harder for you to tell which one is different.

[THREE PITCHES]

Which one? Yeah, I think that's right. But it's a little harder. Now, let's go down to 2%.

[THREE PITCHES]

Which one is different? But that was obviously a little bit harder. But now, let's try it with the unresolved.

[THREE PITCHES]

Yeah, so that's pretty hard. So you're all normal, which is to say your pitch discrimination is better when you have low-numbered harmonics. So that's just this fact about human hearing. So the standard inference from that result is that it is suggestive that pitch perception involves some form of spectral pattern recognition.

So the fact that you're better when the stimulus contains these low-numbered harmonics suggests that these peaks and valleys are part of what is being recognized when you're detecting the pitch. And in this regime up here, where the harmonics are unresolved and all you have is that temporal response, that's where pitch perception is worse. You don't have that spectral pattern information. And so that seems to be important.

All right. So we often see these phenomena in perception. And it's natural to wonder why things are the way that they are. And so some of the new tools that we have at our disposal, due to the advances of machine learning, give us some new ways to investigate some of these questions about why things are the way they are.

And so one of the things that we can do to try to understand the nature of human perception is to optimize machine systems to solve what we think are some of the same problems that human perceptual systems are designed to solve. And then we can look at whether they do what people do and under what conditions.

So this is a model that was optimized to estimate the fundamental frequency of excerpts of speech or music superimposed on noise. So there's a model of the cochlea, and then a convolutional neural network. And so the model is trained to do this. And then you can simulate experiments on the model and ask whether the model exhibits the same kind of characteristics that are seen in humans.

So this is a results graph from human pitch perception that just summarizes what we saw earlier. So what's being plotted here is the f0 discrimination threshold. So that's your pitch discrimination threshold-- so down is better, up is worse-- as a function of the lowest harmonic number.

And so what this shows is that when the lowest harmonic number is low, the discrimination thresholds are pretty good, i.e. small. And then as the lowest harmonic number increases, they get worse. There's also this phase manipulation here that is not terribly relevant for our purposes.

And this is the model when you run it on a similar experiment. You can see that, qualitatively, you get a similar kind of result, that the model has low discrimination thresholds, i.e. good discrimination thresholds, in this regime where you have low-numbered harmonics, and then they get worse.

So this suggests that this particular performance characteristic emerges as a consequence of optimizing a system to estimate fundamental frequency. Now, of course, it doesn't tell you why that happens. But with the model, you can do experiments that would be difficult to do in biology. And in particular, you can ask whether the perceptual characteristics that are observed depend on the auditory diet that the system is optimized for.

So the results that I just showed you came from a model that was trained on excerpts of speech or music, so natural or naturalistic sounds. One of the most salient features of natural sounds is that they tend to be low pass. They tend to have more power at low frequencies than high frequencies. And that's evident here.

So this is a graph that shows a power spectrum. So the y-axis is power. The x-axis is frequency. The black line is the average spectrum of speech or music stimuli. And so you can see they have a fair amount of power at low frequencies. And then it drops off. And so what was done in this study is to create new sounds, either by filtering speech or music, or by generating synthetic tones that have particular spectral properties so that they were either low pass or high pass. So high pass sounds instead have more power at higher frequencies. So you can take these new auditory diets, and then optimize systems to estimate fundamental frequency under these different conditions.

And what was found is that you get very different psychophysical results on these models that are optimized on different kinds of sounds. And so this is that same experiment. So again, we have f0 discrimination thresholds as a function of the lowest harmonic number. So the black curve is the one that you get if you train the model on natural sounds.

And then if you filter the sounds to either be even more low pass or high pass, you get very different results curves. And here, if you take synthetic sounds, if you match the spectrum of the synthetic sounds to these natural sounds, you get the black curve out, which, again, looks qualitatively human-like. And then if you do these anti-match sounds, where they're instead high pass, you get kind of the opposite sort of pattern.

So this is suggesting that the dependence of human pitch perception on these low-numbered harmonics is because that's where the information is in natural sounds about fundamental frequency, in the sense that if you take a machine system, you train it to estimate fundamental frequency, that's the information that it ends up learning to use. And if you train it on different kinds of sounds, it develops a different strategy. What questions do you have about that?

OK. So big picture is we can take machine learning models and optimize them for biological problems, and then look at what they do. And what you can do with these machine learning models that you can't do with actual biological systems is you can optimize them for different kinds of worlds.

Think of this as like simulating evolution and development in alternative worlds where the nature of sounds is very different. So it can give you a glimpse of what might we be like if we had evolved and grown up in a very different kind of world. And in this particular case, it suggests that our pitch perception would be different. So it gives us one way to understand why we have these particular traits that we do.

OK. So let's talk very briefly about where all this might be happening in the brain. So this is a cross-section of the head showing you the key components of the auditory system. So we've got these different subcortical weigh stations, and then auditory cortex here in the temporal lobe.

So this is a view of the macaque brain from the side. And auditory cortex is mostly buried inside a sulcus. And so to view it, you typically cut out a piece of the brain. And it's very common to distinguish in the cortex between three regions-- the core, the belt that goes around it, and then the parabelt.

And these different regions are distinguished in part by different response properties, but also by their connections. So this is a crude diagram that denotes different areas and their connections. So this thing in the middle here is the core. And that's conventionally thought to consist of three regions, and then surrounded by the belt, and then the parabelt.

Now, one of the key things that you see when you look at the auditory cortex is what's known as tonotopy. And that refers to the fact that there is a map of frequency selectivity. So this is a picture that is the result of an fMRI experiment, where people are in an fMRI scanner and they're listening to pure tones. And the pure tones vary in their frequency.

So what an fMRI scanner does is it measures a signal from the brain that is believed to be related to neural activity. So you can then analyze the results. And for every little piece of the brain, you can ask which frequency caused it to respond the most. And so that's turned into a map here, where the voxels, or the little pieces of brain, are color-coded based on the frequency that caused the biggest response.

So very low, best frequencies would be dark blue and very high would be red. And so what you see when you do an experiment like this is typically this kind of gradient, where you see a stripe here where the best frequency is high. And then it transitions to the best frequency being low. And then it switches back again to the best frequency being high. So it's a very prominent, functional, anatomical thing that you see in the auditory cortex. Now, it's not totally clear why this is there and why it's important. I mean, it's essentially recapitulating the tonotopy that you see as early as the cochlea. But it's there. Yeah?

AUDIENCE: Is this the fMRI scan of the macaque brain? Or--

JOSH MCDERMOTT: No, this is a human. Yeah. Yeah, and so I would say, although it's not totally clear why this is particularly interesting or important, it's there. And it's there in everybody. And it pretty much always looks pretty much the same. So everybody's got a map like this. It's more or less always in the same place. Yeah?

AUDIENCE: Is there a pattern to it that I'm not seeing?

JOSH MCDERMOTT: High, low, high. So high is red. Low is blue. High is red. So that's the stereotyped pattern that you see when you do the experiment. And then the high regions come close to connecting at the top. Yeah. So one of the reasons why this is worth knowing about is because it's there and it's because it's prominent. It's also worth knowing about because, scientifically, it turns out to be a fairly useful landmark. So oftentimes, we will measure tonotopy as a reference point. And it's useful to compare other responses that you might observe in the auditory cortex to that tonotopic map.

And so one of the places where this turns out to be useful is when we think about the neural basis of pitch. So one approach that has been used for a long time to try to isolate parts of the brain that process pitch is to look for regions that respond more to tones than to noise. And that's what's shown in the top plots.

So we have the right hemisphere and the left hemisphere. And these are probabilistic maps of the likelihood that a voxel will significantly respond more to tones, so harmonic tones, than to noise. And so you see these regions that have that property. And so the bottom part, the borders of that region are compared to that tonotopic map that we just looked at. So again, we've got high, low, high.

And you can see that this region that is responding more to tones than to noise, that a lot of people have hypothesized is involved in pitch perception, overlaps with the low frequency part of the tonotopic map, and then extends beyond it. So again, the reasons why this is where it is are not well understood. But it's a regularity that is there and is kind of interesting.

All right. So what is the evidence that this thing is actually involved in pitch perception? Well, I just told you about how pitch perception seems to be dominated by low-numbered harmonics, in the sense that your discrimination thresholds are best, they're lowest, when the stimulus contains low-numbered harmonics.

And so it turns out that, if you measure the response of this brain region as a function of the harmonic composition of tones that people are listening to, the response is largest when the tones contain low-numbered harmonics. So what these graphs are plotting is the fMRI response measured as percent signal change-- that's on the y-axis-- for different stimulus conditions that vary in the lowest harmonic number. And again, there's two different frequency ranges. And it's not critical for our purposes.

And what you can see is that, when the stimulus contains low numbered harmonics up to maybe the eighth, you get a higher response. And then after that, it kind of drops off. And so these two graphs, they look like the flip of these two graphs. So the response of the brain is high whenever the discrimination thresholds are low-- that is, they're good. And the response gets lower whenever discrimination is worse. And so the fact that there's this relationship between this stimulus parameter that affects perception is consistent with the idea that this brain region is involved in some way in mediating the perception of pitch. Yeah?

AUDIENCE: Why is it higher for the [INAUDIBLE]?

JOSH MCDERMOTT: Say that louder?

AUDIENCE: Why is it higher for the high frequency ranges?

JOSH MCDERMOTT: You mean, why does this go up a little bit higher? I don't know. And if you squint, maybe it looks like there's a similar effect here perceptually. This might just be a fluke. I mean, real data. So we don't expect them to be exactly the same. I'm not sure if that's worth writing home about or not. Yeah, it's a good question. What other questions you got?

All right. We're going to talk about timbre now. So the last thing we're going to talk about before we get into auditory scene analysis is timbre. So timbre is-- it's frustrating because, on the one hand, it's maybe the most interesting thing about sound. On the other hand, it's defined as what a whole bunch of things leave out.

So it's classically defined as the other dimension, in addition to loudness and pitch, along which our perception of sounds can vary. And it's pretty clearly not just one dimension. But the number of dimensions is unclear. It's not even really even clear how to think about it. But it's nonetheless a thing that people talk about a lot. And it references lots of things that are important.

So classically, timbre is the thing that differentiates different instruments. So this is an example. We've got a piano note, a trombone note, and a tenor sax note. And the fundamental frequencies are all the same. So classically and musically, we would say that they have the same pitch. They're the same note. But they sound really different.

So what is it that determines timbre? And one factor that determines timbre is the relative amplitude of overtones. And so this is a kind of a cool demo that shows how different overtones contribute to timbre. So listen to this.

[AUDIO PLAYBACK]

- The effect of spectrum on timbre. You will hear the sounds of two instruments built up by adding partials one at a time.

[SINGLE NOTE]

[TWO NOTES]

[THREE NOTES]

[MULTIPLE NOTES, INSTRUMENTS]

[MULTIPLE NOTES, INSTRUMENTS]

[MULTIPLE NOTES, INSTRUMENTS]

[MULTIPLE NOTES, INSTRUMENTS]

[MULTIPLE NOTES, INSTRUMENTS]

[END PLAYBACK]

JOSH MCDERMOTT: So the point of that is that thing initially didn't sound like what it sounded like at the end. And then we add in those frequency components. And you build up this sense of what it is. And here's the same thing here.

[SINGLE RESONATING NOTE]

[SINGLE RESONATING NOTE]

[TWO RESONATING NOTES]

[MULTIPLE RESONATING NOTES]

[MULTIPLE RESONATING NOTES]

[MULTIPLE RESONATING NOTES]

[MULTIPLE RESONATING NOTES]

[MULTIPLE RESONATING NOTES]

So hopefully, you had the experience that that gradually became more and more of a realistic rendition of a guitar being plucked. So the timbre, the sense of what the instrument is, it's some operation that's being performed on the entire sound. And the relative amplitudes of the different overtones is part of the story.

But another interesting fact, just at least about musical instruments, is that the high and the low tones from a musical instrument, they generally have different amplitude spectra. And what I mean by that is that the shape of the spectrum is distinct. It's not the case that you can take the amplitude spectrum of a very low note and just translate that up and end up with a high note that sounds like that instrument.

So for instance, on the piano, the low notes will typically have a lot more energy in the high overtones, whereas the high notes have a lot more energy at the fundamental. And so the consequence of this, like I said, is if you take one note played on an instrument and simply scale all the frequencies up or down, it doesn't sound like the same instrument. So this is a classic demonstration of this effect. So check this out.

[AUDIO PLAYBACK]

- Change in timbre with transposition-- a three-octave scale on a bassoon is presented, followed by a three octave scale of notes that are simple transpositions of the instrument's highest tone. This is how the bassoon would sound if all its tones had the same relative spectrum.

[BASSOON TUNE]

[ELECTRONIC BASSOON TUNE]

[END PLAYBACK]

JOSH MCDERMOTT: So you can tell, in that version, really, it's only the high notes that actually sound like the bassoon. So the low notes have that same spectrum as the high note, but just transposed down. And your brain has learned that that's not what a bass sounds soon like-- that's not what a bassoon sounds like.

So the basis for that kind of effect is still not known. It's probably learned from exposure to instruments, but we don't really know. And really, I should say that the reason why the bassoon has that acoustic property, I think, is fairly straightforward, which is that a lot of instruments have a relatively fixed filter. So the body of the instrument is this thing that's constant.

And then it's excited with a sound that can vary in fundamental frequency. But because the filter is fixed, the shape of the sound ends up being competent, in some sense. But then you're varying the fundamental frequency. And that causes the relative amplitudes of the different harmonics to change, depending on the pitch of the note. But how you actually learn that that's how instruments do their thing is unclear.

So another stimulus property that influences timbre is the envelope of a sound. So that refers to the way that the intensity changes over time. So this can be really different for something that's plucked versus struck or blown. So these are waveforms of two notes played on a violin. In the first one, it's plucked. So when you pluck the string, you get this very rapid increase in energy that slowly decays, whereas when you bow a string, there's a slow buildup. And then it's more sustained as the bow travels across the string.

And so the importance of the envelope can be seen by playing things backwards, because the amplitude spectrum is unchanged when you play things backwards. And so if you've ever played around with playing sounds backwards, you'll know that they sound kind of funny. And here's an example.

[AUDIO PLAYBACK]

- The effect of tone envelope on timbre. You will hear a recording of a Bach chorale played on a piano.

[BACH CHORALE]

[END PLAYBACK]

JOSH MCDERMOTT: OK. Now, what they did-- this is a really old demo. And so they got a very talented musician to play the Bach chorale backwards. They recorded that. And then they time-reversed the tape. So now, you have the Bach chorale where all the notes are in the correct order, but each note is actually the time reversal of what it's supposed to be. And so it sounds pretty funny.

[AUDIO PLAYBACK]

- Now, the tape of the last recording is played backwards so that the chorale is heard forward again, but with an interesting difference.

[BACH CHORALE, STRETCHED]

[END PLAYBACK]

JOSH MCDERMOTT: All right, you get the idea. So that's the piano played backwards. It doesn't sound at all like a piano anymore. And so obviously, music producers use effects like this now in modern times all the time to create psychedelic effects and things like that. So the amplitude envelope matters a lot.

There's all kinds of really interesting timbral effects with voices. And in general, a lot of people will often say that musical instruments are intended to mimic voices. And in some contexts, maybe that's true. But voices do a lot of things that musical instruments often don't. And in particular, typically, the fundamental frequency of the voice will change fairly rapidly over time. Certainly, that happens in speech and often, when people sing, if there's any kind of vibrato. And so that turns out to be this very distinctive thing that makes something sound like a voice. Let's look at the spectrogram.

[AUDIO PLAYBACK]

- Demonstration 24, role of frequency modulation on voice perception. This demonstration is quite complex. Please refer to the booklet for a complete explanation. First, we hear a pure tone. Then it is joined by other partials that could make up a vowel. Our perception is that the pure tone continues on, accompanied by a rich one. Finally, when an identical vibrato is added to all the harmonics, the sounds fuse and a sung vowel emerges. This sequence is heard for three different vowels.

[VIBRATING TONES]

[END PLAYBACK]

JOSH MCDERMOTT: OK, I think that's enough. You get the idea. But it's quite striking. So you get this thing that goes from sounding like a synthetic tone to something that really sounds voice-like, because you've learned to associate vibrato with that, presumably. What questions have you got about timbre? Yeah?

AUDIENCE: I'm kind of curious how the effect of-- I'm not sure what to call it, reverb? But the effect of the space that the sound exists in has on timbre. I'm thinking like with the saxophone. Sometimes, if you play with a different mouthpiece, it will be described as having a brassier timbre or a non-brassy timbre.

JOSH MCDERMOTT: Yeah. Well, that I mean-- would you describe that as-- that's not about the space. That's just you're changing the instrument, right?

AUDIENCE: Well, I guess. Yeah, the place where the sound first comes in.

JOSH MCDERMOTT: Yeah. I mean, again, you can think about this in a lot of different ways. But I mean, you are changing the sound. It's going to probably filter the sound in different ways. And the question is, why does that make it feel different? Why do we associate those sounds with-- I don't know-- different emotional emphasis or something? And it's not totally clear.

Some people would say that the emotional associations that you have with timbre maybe derive from experience with voices, that you hear people vocalizing under different emotional conditions. And then some instrument effects are trying to mimic that. That's possible and probably true to some extent. But yeah, it's always a little hard to say. Yeah, so timbre, it's very-- yeah, it's a lot of fun to listen to. It's hard to talk about. And yeah, often hard to think about. Yeah?

AUDIENCE: I guess you preempted my question a bit. But has anyone tried to subdivide timbre into things aside from what we just discussed about, like envelope and stuff?

JOSH MCDERMOTT: Not really. So the main thing that's been done is there's been a fair bit of work on instrument sounds, where you can do things like ask people to rate the similarity of different instrument sounds. And then there's analyses that you can do to try to figure out what are the main dimensions that determine the variation and similarity. And the main thing that comes out of that are really two main dimensions.

One is the speed of the attack. So it's how abrupt the amplitude envelope is. So that's one of the things that we talked about. And then the other is the spectral centroid-- so how bright or dull the sound is. So some sounds have a lot more energy at high frequencies. Some have a lot more energy at low frequencies. So that's what comes out of that.

And it's always a little underwhelming, because you listen to instruments. And they're so much fun to listen to. And it feels very rich. And obviously, when you listen to a musician, you can hear all kinds of stuff about how they're playing. So there's a lot more to it. But yeah, that's the main thing that's been done. Yeah?

AUDIENCE: Besides bright and dull, how do people describe timbre, especially in the science fields?

JOSH MCDERMOTT: Yeah, they don't. So we talk about brightness and sharpness. I mean, there's a very small number of words. So yeah, there's sharpness. There's roughness. That's how amplitude-modulated something is. So if a sound has got a lot of amplitude modulation in a certain range, we'll call it rough because it kind of sounds rough. And yeah, I mean, people just make stuff up. And different people will use different words. And they're inconsistent in how they use them and stuff. It's hard to talk about these things. Yeah, so it's interesting that it's not very well-connected to language.

Yeah, OK. Last thing we'll talk about here is intensity and loudness. So loudness is like this other very obvious property of sound. Some sounds are quiet. Some sounds are loud. It's a very important thing to actually understand because it's one of the main things that people complain about.

So you're living in a city. There's too much noise. And so you complain to the authorities. And then they're supposed to do something. And so in order to do that, they need to be able to measure or estimate the loudness. Or if you're a car manufacturer, people will complain if the car is too loud. And so you want to be able to model and predict how loud the car noise is.

And so this is something that's been the subject of quite a bit of research. In the very early days of psychophysics, there was a lot of work done on what's called magnitude estimation, where people would do these experiments where they'd play people, in this case a sound. And they'd ask them to rate how loud it is, just using a number.

And so the person who most is associated with this approach is S. S. Stevens, who is a psychophysicist who was working at Harvard like in the '40s. And the conclusion of these experiments is that loudness can be described as a power law. So the loudness that someone will ascribe to a sound is proportional to the physical intensity of the sound raised to a power of approximately 0.3. It's an empirical fact.

So if you have sound intensity on this axis and the magnitude estimation on this axis, you get a lawful relationship. So what's significant about this? Well, again, this tends to be fairly consistent across people. And you can use this to predict the extent to which a change in sound level, that you can measure in decibels, would translate to a change in loudness.

And so what that formula tells us is that a 10-decibel increase in the level of a sound will give approximately a doubling in loudness. So if you go from something that's 60 dB to something that's 70 dB, it should sound about twice as loud. And that's because, when you have a 10 dB increase in level, that's a 10-fold increase in intensity. And so 10 to the power of 0.3 is about 2. All right. So that's just a useful rule of thumb to know about.

So that's one fact that's worth knowing, I think, about loudness and intensity. Another fact that is worth knowing about is something that we talked a little bit about earlier in the class. And that is that it's kind of interesting that people can discriminate intensities quite well over this enormous dynamic range. And so this is something that, again, has been measured quite a lot. This is one example.

So this is a graph that is plotting the intensity discrimination threshold, in decibels, as a function of the intensity of the sounds. So 0 here, this is in SPL, dB SPL. So remember, that's relative to this reference that's thought to be close to the threshold of hearing. So 0 dB SPL basically is just what you can just barely hear. So when you get down to these really low stimulus intensities, the thresholds are not so good. But then once you're above about dB SPL, all the way up to 85, the thresholds are less than 1 decibel.

All right. Again, it's just a fact about what people can discriminate. Intensity resolution is pretty good. And so one of the reasons that people have thought this was interesting is that, if you look in the nervous system, in particular in the auditory nerve, most auditory nerve fibers have narrow dynamic ranges.

So what that means is that if you measure the response of a nerve fiber-- in this case, this is spikes per second-- as a function of the intensity of a stimulus-- in this case, this is the tone at the characteristic frequency-- what happens is that, well, when the intensity is too low, you just get the spontaneous rate of the fiber. And then at some point, you get a threshold. And that means that the response starts to increase. But then at some point, you saturate. It's like that's the highest intensity over which the response will change.

And so what you can see here is that, here, the threshold of the tone is maybe about decibels. And then it's saturating here at maybe-- I don't know-- 50 or 55. So that's a very fairly narrow dynamic range of maybe 25-30 dB. It gets a little more complicated because, if you look at the auditory nerve, there tend to be these different types of fibers.

So some of the fibers are like this one, where the dynamic range is fairly narrow and the spontaneous rate is high. So on its own, without any sound stimulation, the neuron is spiking around 30 spikes per second. You also have nerve fibers that have lower spontaneous rates and that have a wider dynamic range, and a fact about the auditory nerve.

So the puzzle here is that we seem to be able to register and discriminate intensities over this enormous range, from all the way down here to all the way up like above here. But the nerve fibers only change their response over a fairly narrow range. So the question is, well, once the nerve fiber is saturated, the response isn't changing, if you increase the stimulus beyond that. And so how could you tell that the intensity was continuing to go up?

All right. So that's what's called the dynamic range problem. It's a well-known thing people have puzzled about. And there's a handful of solutions that have been proposed over the years. One is what is called off-frequency listening. And so the idea there is that-- so let's suppose you're discriminating the intensity of a tone at 1 kilohertz. When the intensity gets really high, the nerve fibers that have a characteristic frequency near a kilohertz may be saturated. But the ones that are further away won't be. And so you could somehow listen to those. So that's off-frequency listening.

There's also the notion that you might use temporal information, like phase locking. So you see increases in synchrony with levels. Of course, you can only use this when you have phase locking, which is for frequencies that are below 4 kilohertz. And then the other proposal is that these low spontaneous rate fibers might be really important. So those have higher thresholds, but greater dynamic range. And people have wondered about, well, if they're so important, why are they a relatively small proportion of the auditory nerve? All right. So that's the dynamic range problem. And I would say the solution to this is still something that is not-- there's not total consensus about how this works.

So we've talked about, over the past couple lectures, the properties of individual sounds. We describe individual sounds in terms of their loudness, their location, their pitch, their timbre. So all of these are perceptual variables that tell us things about the physical events in the world that produce the sounds, which is presumably the point of hearing them in the first place. So timbre is the richest, but it's the least defined and the least understood of these dimensions.

And what we're going to talk about now is what happens when more than one sound occurs at once, when we have an auditory scene. Before I do that, any questions about loudness or anything else? Yeah?

AUDIENCE: I have a question about the tonotopy that we saw in the fMRI scan.

JOSH MCDERMOTT: Yeah?

AUDIENCE: [INAUDIBLE] similar attributions in the auditory cortex in deaf individuals?

JOSH MCDERMOTT: So you're asking, is there a way to activate this bit of cortex in someone who's deaf?

AUDIENCE: Yeah.

JOSH MCDERMOTT: Yeah. So I think that the standard thinking on someone who lacks a sense-- so deaf or blind-- is that the cortex that's normally allocated to that sense tends to get repurposed.

AUDIENCE: So what kind of stimuli would activate-- produce similar activations, I guess?

JOSH MCDERMOTT: Well, I don't know if what would produce similar activations. But I think it seems-- I'm not sure if this is actually known. But it seems very plausible that, in someone who is deaf, that you might get visual stimuli actually activating the auditory cortex. So you often see the reverse thing. So people who are blind, when they're listening to sounds, you often get a lot of visual cortex activation that you wouldn't necessarily see in someone who's sighted. So the idea is that the cortex gets repurposed. And there's lots of debate over like, well, exactly what can the cortex be repurposed for. And I think that's unresolved. Yeah. What other questions you got? Yeah?

AUDIENCE: Does the loudness estimate depend on the frequency of the sound? Are there some frequencies that may appear louder or quieter?

JOSH MCDERMOTT: Yes. There are probably small effects of frequency. I mean, obviously, your detection threshold varies quite a lot depending on frequency. And so you'd want to compensate for that. And once you've compensated for that, there's probably some effects. I don't actually know what they are. The thing that has a bigger effect on loudness is bandwidth, which is on your problem set. Cool. OK, let's start talking about auditory scene analysis.

So the problem that we're going to be talking about is that many acoustic events happen in the world at once. But your ear only receives one sound wave. And so the sound wave that is received by your ear is the sum of those that would have been produced by the individual events if they happened on their own. Somehow, the brain has to distinguish the different events that happened in the world from the single signal that it receives.

So for instance, if I'm talking and somebody walks in and you hear the door close, you need to be able to distinguish that door closing from my voice, because you might want to know whether somebody came in. Similarly, if you're at a restaurant, and you're trying to talk to somebody, and there's all this noise, you need to be able to distinguish the different sound sources. So this happens all the time. And people are really pretty good at it. So here's the classic cocktail party version of the problem. So you're talking to somebody. You want to understand this.

[AUDIO PLAYBACK]

- She argues with her sister.

JOSH MCDERMOTT: But maybe what comes into your ear is this.

- She argues with her sister.

JOSH MCDERMOTT: Or this.

[INTERPOSING VOICES]

JOSH MCDERMOTT: Or this.

[INTERPOSING VOICES]

[END PLAYBACK]

JOSH MCDERMOTT: So the presence of the other talkers obscures a lot of the structure. And the target utterance creates this very complicated signal. But the speech remains largely intelligible. And people are still substantially better at this than present day speech recognition algorithms. Although, that may not be true for very much longer.

All right. So one interesting perspective on this problem, which is what we talked about in the very first lecture to the class, is that you can think of the problem that the auditory system is solving as being akin to me giving you this equation and asking you to solve for x. So that's a ridiculous thing for me to ask you to do. But that's like exactly what's happening at some level, because you're measuring this one signal that is the sum of multiple signals that were generated in the world. And then you're trying to estimate one or more of those one signal, the thing that you're trying to understand.

All right. So this is a classic example of an ill-posed problem. It's ill-posed because there are many combinations of source signals that could add up to be equal to the observed mixture signal. So for every sound wave that the ear receives, there are really an infinite set of sound sources that could have generated it. So it's ill-posed.

So as with any ill-posed problem, the brain has to make its best guess as to how many sources there are and what they are based on what it knows about the world. So another kind of fun way to think about this problem is this pictorial description that's normally attributed to Bregman.

So imagine the situation here, where we've got a lake. There's a bunch of boats on the lake. There are these two little inlets that we've carved out of the edge of the lake. And we lay down two handkerchiefs so that the handkerchiefs float on the surface of the water. And we are allowed to measure the motion of these two handkerchiefs. And from the motion of the pieces of cloth, the handkerchiefs, we have to determine how many boats are in the lake, where they are in the lake, how big they are, or maybe where they're going, stuff like that.

All right. So the analogy here-- we've got two ears. The ears are just measuring vibrations in the air. Those vibrations are originating with sources in the world. So it's kind of amazing. So how can we do this? Well, I think the crux of the issue is that the sounds that we have to hear, the sounds that occur in the real world, are not random. So this is a brief demonstration of the fact that sounds in the world are not random. So here are some real-world sounds.

[ENGINE ROARING]

[DOORBELL RINGING]

[HAWK CALLING]

[SWOOSHING]

OK, so those are just sounds that you might hear in your everyday life. Let's compare those to sounds that are random. So I made these sounds to be fully random. So what does that mean? Well, a sound is just a big vector of numbers, where the number at every point in time tells you the pressure at that point in time.

So you can make a random sound by randomly sampling each one of those numbers from an independent distribution. I used a Gaussian, I think. All right. So you could do this yourself in Python. So these are fully random sounds. Each sample is an independent draw. Here's what they sound like.

[STATIC SOUND]

All right. Let's listen to another one.

[STATIC SOUND]

All right. Let's do it again.

[STATIC SOUND]

So it's pretty clear that we would have to sit here generating these random sounds for a very, very, very long time before you got a doorbell or a hawk or the sound of toothbrushing. And so what that tells us is that the sounds that we hear in life are a tiny, tiny, tiny fraction of the space of all possible sounds.

And that's really the key to us being able to solve auditory scene analysis, is that, when we describe the problem like this-- well, we're giving you this equation with two unknowns. The fact is that x and y can't take on every possible value. They live in this very, very specific subset of the space of all possible numbers. And so that's what makes the problem solvable. Real world sounds are a very small portion of all possible sounds.

So we talked, at the very start of the class, of how we often think of perception as unconscious probabilistic inference. It's an inference process because the problems are ill-posed. The solution is not uniquely determined. And you have to make your best guess as to what happened in the world to cause a stimulus, given what you know about the world. So that's probabilistic inference, making your best guess. It's unconscious, because you're not actually aware of that happening. You just end up hearing something or seeing something. This is just happening under the Hood.

And the general framework that we use to think about probabilistic inference is Bayes' rule. So here's the setting. When we're perceiving something, there is an observation, which is the sensory data that we receive. In this case, this is an image. It could be a sound. So that's the image.

There is a set of possible hypotheses for what could have happened in the world to cause the image. So hypothesis A is that there are these two objects in the world. One is light green. And it's got that funny pattern of stripes. The other is darker green. And it's got that curved thing on it. Hypothesis B is that there are these two objects in the world. There's one surface that is like a two-tone surface. And then there's the letter B. Hypothesis C is that you have that same surface, but now there's a different letter there. So that's the observation.

Now, when we are doing probabilistic inference in this setting, we are trying to find the hypothesis that is most likely given the observation. You get sensory data. And you want to figure out what possible state of the world is most likely to have caused that sensory data. And the hypotheses here are different possible states of the world that might have caused that sensory data.

So how do we evaluate that probability? Well, this is an expression for that probability. That's what's called the posterior probability. It's the probability of the hypothesis, given the observation. That vertical line means given. And Bayes' rule states that the posterior probability, the probability of A, in this case, given O, is equal to the product of the probability of A on its own-- that's called the prior-- and the likelihood, the probability of the observation given the hypothesis.

And then it's normalized by the probability of the observation. But typically, we're working in a setting where the observation is known. And so that's going to be the same across all possible hypotheses that we consider. So really, the thing that will cause the posterior to vary across hypotheses are the prior and the likelihood.

So let's think through, in this simple setting, what that would correspond to. So the prior is a reflection of the fact that some hypotheses are a priori much more likely than others, just based on the way the world works. And so the idea in this simple example is that this hypothesis is a priori not very likely because objects like this tend not to occur in the world, whereas these two hypotheses have fairly high prior probability. This is just a reasonable looking surface. And then the world has lots of letters in it. So the idea in this contrived example is that the prior is high for these two and low for that one.

Now, the likelihood is the probability of the observation given the hypothesis. That's going to tell us to what extent could a hypothesis actually have caused the observation. And so in this contrived example, the idea is that hypothesis A and B both do a very good job of explaining the observation, because if you put these two objects next to each other, or you put the B on top of this, you're going to reproduce pretty much exactly the observed image.

So the likelihood, the probability of this observation, given this hypothesis or this hypothesis, is high. By comparison, this hypothesis produces fairly low likelihood because it's got the wrong letter. So the image that would be created by putting the C on top of this surface is not going to be very similar to that observed image.

All right. So the idea is that we're trying to find the hypothesis that gives us a high posterior probability. And that's proportional to the product of the prior and the likelihood. So you want the prior to be high. And you also want the likelihood to be high. The prior being high tells you that the hypothesis is reasonable. The likelihood being high tells you that it can explain the sensory data.

And so in this contrived example, the idea is that really only hypothesis B has both high prior probability and a high likelihood. Hypothesis A has a high likelihood, but low prior probability. Hypothesis C has high prior probability, but low likelihood. So this is a mathematical framework for thinking about probabilistic inference, key idea being that there are these two ingredients here. There's the prior-- how likely different states of the world are, given what we know about the world-- and the likelihood, which is, can the hypothesis explain the data that we observe?

All right. So if we want to explain perception in this kind of way, as perceptual inference, there's a lot of things that actually have to be specified. So one of the things that you have to specify is the hypothesis space. So how are we actually going to define the states of the world that the perceptual system is actually trying to estimate?

If we define the hypothesis space, then we also need to know the prior probability of all the different hypotheses that live in that space. We need a likelihood function, which tells us how to compare a hypothesis to the sensory input. If we've got the prior and the likelihood, that gives us the posterior. But the whole idea behind the perceptual inference is you want to find the hypothesis that gives you a high value for the posterior.

And so that means you need some way to find the hypothesis that has, say, the highest posterior probability. And in practice, that actually is often highly non-trivial because the space of all the hypotheses may be enormous. And the posterior distribution may be multimodal. So finding the value that has the-- finding the hypothesis that has the highest posterior may be a really difficult search problem.

And then ultimately, if you want to understand how all this happens in the brain, we need a description of the underlying representations and of the inference in terms of neurons. All right. So there's lots of different things that we actually really need to nail down to really say that we understand this.

And these different ingredients that are here on this slide are related to this idea that there are different levels of analysis at which you can talk about an information processing problem. And the idea of there being levels of analysis was most famously articulated by somebody named David Marr, who was a vision researcher who worked here at MIT back in the '70s. He wrote a very influential book called *Vision,* which was his view for how to study vision. And the lasting contribution of that is the set of distinctions into levels of analysis that you still hear a lot about today. And you've probably heard about them in other classes.

So Marr distinguished between three levels of analysis. The first and the highest level is the computational level. And that involves asking what the problem is that is being solved by the system and what the constraints are that allow that problem to be solved. So defining the problem to be solved is akin to identifying the hypothesis space. So what is the thing that you're trying to estimate from the sensory data? So it could be the fundamental frequency. It could be like where something is coming from in the world. It could be the shape of an object.

And it also involves determining how you would assign probabilities to hypotheses-- so asserting that there's prior probability that varies over hypotheses, likelihood, and so forth. So you can also think of this computational level as nailing down the mapping between the input and the output. So a computational level description of a problem defines that mapping. So it says, what are the assumptions that are being made by the perceptual system to solve this problem? And those assumptions determine the mapping between the input, say the sound or the image, and the output, which is your estimate of the property of the world that you're interested in.

Now, then there's also the algorithmic level, which involves identifying the approach to finding the solution. And so one way to think about this, if we're thinking about probabilistic inference, is in describing an optimization procedure to find the best hypothesis, or the most probable hypothesis, given the sensory input. And as I said, that often involves very hard search problems because there's a huge hypothesis space. It's not obvious which one is the best. And then there's the level of the implementation, which is how you would instantiate a solution to a problem in hardware. And so this, in the context of neuroscience, would involve a description of neural circuitry.

So this is an excerpt of a figure from Marr's book, where-- these are, in his words, those three levels. So there's the computational theory, the representation and algorithm, and then hardware implementation. Any questions about levels of analysis? And I should say, we will return to these ideas repeatedly over the course of the class. And you'll think about them as being applied to lots of different aspects of perception.

All right. So we're going to talk about auditory scene analysis, which you can think of as this problem of inferring sources in the world from a sound signal that reaches your ears. So as we've discussed, if we want to think about this as perceptual inference, we need to specify the hypothesis space, the prior probability of hypotheses, a whole bunch of stuff.

And as a starting point, let's think about this about these priors. So priors reflect the regularities of real world sounds. So we've just seen this demonstration that the sounds that we hear in the real world are not random. So some sounds are a lot more likely than others. So the doorbell is the kind of thing that can occur in life. There's other things that wouldn't.

And these regularities that exist in real world sounds, they derive from the processes that generate them. So sounds, they're generated by lawful physical processes, for the most part. And those are governed by physical laws. And so you can't get any old thing. There's certain things that can result from that.

And so as with other areas of perception, illusions can tell us a lot. So illusions can reveal the constraints on sound generation that have been internalized by our brains, that make auditory scene analysis possible via priors that we have on what kinds of sounds are likely and what are not.

And so this is often talked about in the language of grouping cues. So we talked about cues when we were talking about sound localization. So the notion of a cue is a stimulus property that's informative about something in the world. So we talked about how there are binaural cues, like interaural time differences and level differences, and spectral cues from your pinna that tell you about elevation.

Similarly, or analogously, grouping cues are thought of as stimulus properties that are informative about the organization of sound into sources. We believe that these cues derive from statistical properties of natural sounds. And we'll talk about a number of examples. And really, two obvious examples are shown here-- onset and offset and harmonicity. And this is just a cochleagram, a cochlear representation of a speech signal.

And you can see that you get these patterns here, where the energy begins and ends, across a wide range of frequencies, more or less at the same time. So there's a common onset or a common offset. And that's because people open and close their mouth. And so you get a termination of energy or a burst of energy. And that common onset and offset is something that happens a lot in natural sounds. And the idea is that, when that is observed in the sound signal, that gives you a clue that all of those frequencies that have the same onset are caused by the same thing in the world.

Similarly, we just talked about how lots of sounds in the world are harmonic, in particular sounds that are produced by our voice. And that shows up as these horizontal stripes here for individual harmonics. And so the idea is that if you observe harmonic frequencies in the input to the ear, that's a pretty strong cue that those frequencies actually came from the same source in the world, because sources in the world tend to be harmonic.

And so what we're going to get into next time is a bunch of really remarkable demonstrations that give us evidence that your brain has internalized a whole bunch of these properties of sound. And it causes you to organize sound into distinct sources. These are illusions of perceptual organization. So we'll end there. Have a great weekend. Recitations tomorrow. Problem set is due tomorrow. I'll see you on Tuesday.