[SQUEAKING]

JOSH MCDERMOTT: The problem that we've been talking about is the problem that there's this world out there with multiple sources, typically, that will make sound. And you receive a single sound waveform that enters your ears. You're trying to infer what happened in the world to cause the sound. So that's a specific example of the general problem of perceptual inference.

And so we've talked about how the typical formalization that we use to think about perceptual inference problems is given by Bayes' rule, where the idea is that we want to find the hypothesis, which is the potential state of the world that is most probable given the observation. So we're trying to find something that has high posterior probability. You can break that down in terms of the prior and the likelihood.

And so you look at Bayes' rule. And it kind of makes it seem like everything is kind of nice and simple. But in actuality, if you think deeply about what's required in order to really have a thorough understanding of one of these problems, there's a number of things that require specifying, which we've discussed before.

So we've got to say what the hypothesis space is. So what's the candidate set of states of the world that we're going to try to differentiate between? And that's often kind of non-obvious, or at least there's lots of different ways to specify the hypothesis space. So for instance, you could think of the hypothesis space as the actual waveforms that are produced by the sources. You could think of it as like parameters of those sources, for instance, like the pitch of the notes in some way of characterizing the timbre. There's lots of different ways to think about what exactly that is.

Then you have to establish the prior probability of the different hypotheses. You need a likelihood function, which involves a comparison of a given hypothesis to the sensory input. You need a means to find the hypothesis with the highest posterior probability. So just because you know how to calculate the posterior for one example hypothesis doesn't give you necessarily a way to find the one that is best. And then we ultimately want to describe all this stuff in terms of neurons.

OK. So where does all of the things we've been talking about leave us? So one of the basic questions that you could ask about auditory scene analysis is, what is the hypothesis space, and what's the nature of the prior over those hypotheses? And this involves implicitly making assumptions about sources and how they produce sounds.

And so we often think of the problem of perceptual inference as there being like a model of the world that defines a space of hypotheses. So in the case of auditory scene analysis, the world model would consist of sources in the world that can make sounds. They generate sounds, and they produce an observation, which is like a mixture of all the sounds that would be produced on their own.

OK, so Bayes' rule gives us a formalization for using a world model to do inference. And it looks really simple. You've got probability of the world state given the data being proportional to the likelihood times the prior. But there's a lot of other stuff that has to go into this.

Now so what's the contribution of all these illusions that we've been discussing? So I think one way to look at this is that these illusions suggest constraints on the brain's hypothesis space for sound sources. So we've been looking at all of these examples that seem to suggest that our auditory system is assuming the tendency of sources to make sounds that have abrupt onsets, that contain harmonic frequencies, that contain repeating elements, and that exhibits similar acoustic properties over time. Because all of those illusions, they're essentially sounds that when you listen to them get organized in your head in a certain way.

So for instance, we saw the harmonica's tuning effect. So you take a frequency, and you tune it and that as part of a harmonic complex. And then you hear it as a separate sound. So that's indicating that there's this assumption that's implicit in the way your auditory system works that sources in the world make harmonic frequencies. So that's one of the contributions of a lot of these illusions is revealing these constraints on sound sources and thus on the brain's hypothesis space.

And now, of course, these constraints, at some level, they come from the world. So we believe that these perceptual effects that relate to harmonic frequencies are abrupt onsets. They're there due to the fact that the sources in the world behave in a certain way due to physical laws. So there's sound sources in the world that are periodic in time that are harmonic in frequency or that tend to produce sounds in a way that creates this common onset across frequencies. So at some level, this stuff originates in the world. But then the illusions kind of both reveal those as being important things in the world but also things that are internalized by your auditory system.

We also looked at examples that speak to the tendency for environmental surfaces to produce reflections. Again, that's a statement about acoustics. But the illusion that is associated with that, the precedence effect, is demonstrating that that's been internalized by the auditory system. The illusions also kind of reveal that the auditory system is implicitly cognizant of the possibility that sounds can occur concurrently and that they can continue in the absence of overt evidence, all of these filling in illusions that we saw.

So, as we said, Bayes' rule gives us a formalization for using a model of the world to do inference. But it doesn't tell us how to do the inference. And this actually in practice turns out to be like often one of the hardest parts of making models of perception work.

And so what I mean by telling us how to do the inference is determining how to find the most probable hypothesis for a given observation. And the reason that this is in practice often pretty hard is indicated by this picture. And this is just like-- it's a very simplistic attempt to convey some of the issues. So this picture is plotting a hypothetical posterior distribution over a hypothesis space that has only two parameters.

So let's suppose that the hypothesis space that you're considering would be determined by these two parameters, say the frequency and the duration of a tone. So that would be like an incredibly simple hypothesis space. But as you can see in this simple example, the posterior probability has these multiple peaks throughout this space.

Now in actual problems, there are way more than two parameters. If you think of all of the parameters that would be necessary to specify even a moderately complicated auditory scene, it's quite a lot. So we've got this probability distribution defined over a pretty high-dimensional space. It's typically multimodal.

And so like in this particular example, we've plotted the full distribution. And so you can look at it with your visual system and say, oh yeah, this is the best point. But normally, you can't plot it, and you can't look at it with your visual system. And so you have to have some other means of finding the hypothesis that is the best.

Now the other added complexity is that in a lot of cases-- this picture is actually a little bit misleading because the number of parameters that defines your hypothesis space is not even fixed. So let's suppose, for instance, in the domain of auditory scene analysis, you often don't know how many sources are out there in the world. And so one of the questions is, well, is there two sources? Are there three? Are there four? And for those different numbers of sources, that would create different numbers of parameters. So you also have to be able to consider those different possibilities.

But the crux of the issue is that you can think of perceptual inference as a search problem. So we've got this huge space of possible hypotheses. You have an observation. And you somehow have to search through that space of hypotheses to find one that has good posterior probability given the observation. And these are hard search problems because the optimization landscape is complicated, and it's very high dimensional.

So one kind of algorithmic approach, which you can think of as mapping onto some of the things that we know about sensory systems, is to use a feed-forward algorithm to take clues that are in the sensory input and get you to the right part of the hypothesis space. And this is actually like a pretty common computational technique that is used in this neck of the woods.

So the idea is that you could take a feed-forward neural network that takes the sensory input and make some initial guesses as to what the parameters are. Now in practice, those guesses are not perfect. But they are often useful to get you to the right part of the space. And then you could do some kind of optimization procedure to try to find your way to the peak of the local distribution.

But that's the key thing to take away from this is that when we think about perception as Bayesian inference, instantiating that for the real problems that we solve is pretty difficult. And doing that right now is like on the bleeding edge of what is kind of possible in our field. And the crux of the issue is that you have to solve these very challenging search problems. So initially, you can try to define a hypothesis space, but then you still have this problem of finding which one is best given an observed sensory input. Any questions about that way of thinking about things? Yeah, Christa.

STUDENT: Did you say it's accurate to describe illusions as like finding a local but not local optimum? Or should you say that the analogy kind of breaks down?

JOSH MCDERMOTT: Well, OK. So illusions are interesting. They relate to this kind of picture in an interesting way sometimes because some illusions are bistable in the sense that you can actually hear them in one of two or sometimes three ways. And that kind of suggests that that's a situation where maybe there are two modes of the posterior that are both pretty good, and your perceptual system kind of switches between them. And that typically doesn't happen very much in the real world. One of the remarkable things about perception in the real world is you pretty much just always see or hear this one thing, and it usually seems to be pretty close to ground truth.

And illusions are interesting just because they're often not like that. So I would say that's-- I mean, that's one respect in which you can think of illusions, as being related to this, is that sometimes you experience bistability. And you could think of that as there being two good solutions or two approximately equally good explanations of a stimulus. Yeah.

STUDENT: How good or bad are current networks at doing auditory scene analysis?

JOSH MCDERMOTT: The problem is certainly unsolved. So I mean, I would say the-- I mean, as with a lot of problems that involve signals, there's been enormous progress in the last 10 years. But in terms of having models that solve the problems in the way that people do, we're still fairly far from that. So one domain in which there's been a lot of progress is what's called source separation. So that's really the problem of taking a mixture of source signals and then estimating the constituent signals.

So imagine we have a recording of two people talking. You add that together. And then the task is to take that mixture and estimate the two source signals. So that's a problem that neural networks actually do a pretty good job of solving. You might ask, OK, well, can we take one of those systems and ask whether it explains all of the phenomena that we've been looking at in human auditory perception?

And we've actually done some of this in our lab. And typically, those systems, they really don't reproduce almost any of the phenomena that we've seen here. They seem to be doing something that's a bit different and a little bit more specific.

So I would say yeah, we're a little ways off from solving this in the way that people do. And part of it is related to some of these sorts of issues, which just turn out to be very challenging. Yeah. Any other questions?

So I want to turn briefly to thinking about what happens in the brain when we segregate sounds. And we're going to make a brief detour to talk in a little bit more detail about some issues in auditory neuroscience and in particular about receptive fields in the auditory system and how neurons represent sound.

So this slide is depicting something called the spike-triggered average. How many people have encountered a spike-triggered average? Yeah, maybe just a few. So they teach this in 940, which I think probably most of you haven't taken yet.

But so it's a method for characterizing the sensitivity of neurons in sensory systems. And the basic idea is actually pretty simple and pretty cool. So the notion is let's suppose you do an experiment where you're recording from a neuron in the brain of some organism. And you expose the organism to a continuous stimulus. So in this particular case, it's a visual-- it's a visual stimulus, which is a sequence of noise patterns-- pattern one, pattern two, pattern three, rapidly one after another.

So you're recording from the neuron, and you keep track of when it fires an action potential. And so the notion is that every time there is an action potential, you look back in the stimulus history, and you grab the stimulus that preceded the action potential by some amount of time, say 30 milliseconds, 40 milliseconds. And you collect that set of stimuli. So there's this long sequence of frames, and these are just the five ones that preceded these spikes.

So you collect all these stimuli that preceded the spikes, and then you average them. It's called the spike-triggered average. So every time there's a spike, you average the stimulus that preceded the spike. And so what that gives you is the average stimulus that precedes a spike.

So it's kind of like you could think of that as representing the feature in the input that's the neurons responding to it. And it turns out that if under certain assumptions, you can think of the spike-triggered average as an estimate of the filter that the neuron is instantiating. So if the neuron is a filter followed by a simple non-linearity, the spike-triggered average will give you the filter.

So it's a fairly common method for characterizing the tuning properties of neurons in sensory systems. So this is an example from the visual domain. And we'll talk more about this kind of stuff when we get to vision. But you can also do this in the auditory domain. So imagine instead of showing you these sequence of frames, we play you this.

[RANDOM NOISES]

OK, you get the idea. So it's a very extended kind of random noise stimulus. And we record from a neuron in your brain. And then every time there's a spike, we average the little section of the stimulus that preceded the spike. And so if we did this in a stage of your brain called the inferior colliculus-- so this is a subcortical region of the auditory pathway like a few synapses up from the cochlea.

This is the spectrogram of the stimulus that would precede a spike. And so what you can see is that there is a little burst of energy-- that's the red-- and then a little dip in the energy. And at a particular frequency, it looks like it's around 8 kilohertz.

So this particular neuron is being driven by amplitude modulation at a particular rate. So you can see that the cycle time here is 4 or 5 milliseconds, something like that. So it's pretty rapid. It's a very rapid modulation at a particular audio frequency.

So this is known as a spectro-temporal receptive field. It is usually abbreviated STRF. And people who work in the field will just call this a STRF. So if you go to a neuroscience conference and you hear people talking about STRFs, that's what they're talking to. That's what they're talking about. So this is an example where this neuron is responding to changes in amplitude in a particular frequency range.

So remember how when we were talking about the cochlea and auditory nerve fibers, we talked about how they are tuned in frequency. And when we were discussing that, we were talking about audio frequency.

So you can take a waveform. You can do a Fourier transform. And you decompose that into sinusoids. Those are audio frequencies.

But there's a lot of information in sound that is conveyed by amplitude modulation. And remember how we talked about how a sound can have an envelope?

So here is an example sound. The waveform is in blue. It's actually wiggling up and down. It's just so dense that you don't see it. But there's the instantaneous amplitude, which is plotted in red, that kind of goes up and down.

And if we zoom in on a particular frequency channel, you can see that the blue waveform here is oscillating up and down, and then the red envelope is kind of changing at a much slower rate So this is what we call a cochleagram. So what the way that you generate this picture is you take a filter bank that's kind of replicating the frequency tuning of the ear. And you apply that filter bank to a sound. You then measure the envelope of every frequency channel's output, and you plot that in grayscale as a row.

So every row in this picture is the instantaneous amplitude that varies over time of a particular filter. You can think of that as a particular place on the cochlea. And so this looks a lot like a spectrogram. The only difference being that we made it with a model of the ear. And so it's got the tuning properties that you see in here. So remember how in the ear, the filters get broader when they're tuned to high frequencies. So that happens here-- stuff like that.

Now what's shown here is the power spectrum of the response of one particular filter. In this case, it's one that's centered at looks like 350 hertz-- so kind of down here And there's two. So remember, a power spectrum plots power as a function of frequency. And the blue graph here shows you the power spectrum of the output of the filter.

So this tells you the audio frequencies that are present in that output. And this is a bandpass filter. And so you can see that there's power between around 280 and 400 hertz. So there's positive stuff here, and then it goes to 0. So that's a bandpass filter.

Now the red plot here is the power spectrum of the envelope of the filter of the red curve up here And what does this show us?

Well, it shows us that the frequencies that are in the envelope are low frequencies. The envelope is low pass. So that says that the instantaneous amplitude of the output of that filter is varying relatively slowly. It's defined by low frequencies. And these are modulation frequencies.

So the key takeaway message from this slide is that we have audio frequencies. That's kind of what you get if you do a Fourier transform of sound. We also have modulation frequencies. That's what you get if you take the envelope of part of a sound, and you look at the frequency decomposition of that envelope. And the envelope tends to vary more slowly.

And so it contains low frequencies. And so whereas the auditory nerve is tuned to audio frequency. If you look later in the auditory system, you'll find tuning to modulation frequency. And so we just saw an example of that with the spectro-temporal receptive field.

This is an example of a bunch of different neurons in the midbrain. And each one of these curves is a tuning curve that plots the response of the neuron as a function of the modulation frequency. So we have a stimulus, and we are varying how rapidly the amplitude is modulated. [MODULATION SOUNDS]

So that's modulation frequency. And what you can see-- so this particular neuron here responds best at looks like around 70 hertz. This particular neuron here responds best at around 40 hertz. This particular neuron here is down around 25 hertz. And this looks like a filter bank.

So it's a set of bandpass filters that are tuned to different frequencies, but they're modulation frequencies. So it's kind of conceptually similar to the filter bank that we use to model the cochlea. But instead of being tuned to audio frequency, it's tuned to modulation frequency.

So this is a picture that kind of captures that idea. So this is a model of auditory signal processing that extends from the cochlea to the midbrain and thalamus. And so what happens here? Well, we start out with a sound waveform. That gets processed by a bank of filters.

So each one of these pictures here represents the frequency tuning of one filter in your cochlea. So this is frequency, and that's the response.

So this is a high-frequency filter. That's a low-frequency filter. This is the output of that filter when it's applied to that sound signal. So this is a low-frequency filter, and you can see the blue curve is kind of wiggling at a pretty slow rate. This is a higher-frequency filter, and you can see the blue curve is wiggling at a faster rate.

But the output of each of those filters is characterized by an envelope. And you can extract that with a nonlinear operation. And that's just what these blue curves are.

And then these envelopes can then be passed through a second set of filters. These are modulation filters. So they are also tuned in frequency. But this modulation frequency. So again, this one is tuned to low frequencies. And so you can see that the output varies rather slowly. This one's tuned to high frequencies, and the output varies more quickly.

So we often think of sensory systems as kind of having these cascades of filtering operations followed by non-linear operations followed by more filtering operations. And this is just one example in the auditory system.

So I showed you an example of a STRF, an STRF, in the inferior colliculus where things look nice and simple. If you make the same kind of measurement in the auditory cortex, things are a little more complicated. So these are four examples of spectro-temporal receptive fields measured in the auditory cortex. So again, you can think of this as the stimulus that's most likely to elicit a spike in a particular neuron. And the general takeaway is they're more complicated.

You can see that some of the neurons are tuned for it looks like kind of ripples in the spectrum, maybe with some kind of difference in frequency content. There's still some tuning to amplitude modulation, but it just gets more complicated.

And so there is a standard model of the auditory cortex that's inspired by these kinds of operations. And it looks like this, where you've got the output of the early auditory system, which we think of as a time-frequency representation. That then gets passed through a set of filters that are now tuned in both time and frequency. And so they're kind of sensitive to these features in the earlier representation. So these are again modulation filters, but they're now spectro-temporal modulation filters that are tuned in both time and frequency.

But the important thing to take away from this is that we get of second set of filters that's kind of tuned to higher-order structure that's operating on the output of an earlier set of filters. Any questions about audio frequency or modulation frequency or these cascades of filters? Yeah.

STUDENT: I'm curious. These all have very nice waveforms. But like when you're actually looking at the brain, can you directly measure this and see this kind signal transformation, or is it more abstract?

JOSH MCDERMOTT: I mean, OK, the difference is that normally, if you're making a measurement from neurons, you're measuring action potentials. And so like the notion is that typically-- when we talk about these filters kind of existing in the brain, normally, what we mean is that the neuron is implementing that filter plus a stage of generating action potentials. So there's this spike generation phase that's kind of missing from the picture.

And so that adds a layer of complexity. But modulo that, yeah, I mean, you do see traces of stuff like this. I mean, these pictures-- like this picture. I mean, this is derived from measurements that are made from one neuron in the inferior colliculus. I mean, these pictures are derived from measurements made from individual neurons in the cortex. So yeah, I mean, this is what you get. Yeah.

So what is all this stuff good for? So one kind of interesting application of this is to understand the representation of sounds in the human brain. And one kind of opportunity to do this kind comes from epilepsy patients.

So in general, with human neuroscience, because typically the methods are noninvasive, they're very coarse both spatially and temporally. So we have fMRI. We have MEG. We have EEG. And they're making measurements of signals from the brain. But typically, those measurements are kind of pooling over time and space to a pretty good degree. And they're pretty indirect.

If you have somebody who has bad epilepsy, sometimes the epilepsy interferes with quality of life to the extent that they decide that they want to have an operation to try to remove the part of the brain that's causing their seizures. So the person will go to the hospital, and oftentimes there will be a grid of electrodes that is placed on their cortex. And they'll be monitored for some period of time. Often, this is because the surgeon is trying to determine exactly where the seizure is coming from.

So the person will be hanging out in the hospital with a grid of electrodes on their brain, essentially waiting for a seizure to happen so that they can make measurements and figure out where they need to perform the operation. And during this period of time, they're often willing to participate in experiments. And so this is of a branch of human neuroscience where people make intracranial recordings in this kind of setting. And so you can, for instance, show the person movies or play them sounds, or ask them to do math problems and measure responses in their brain.

And the key difference over a lot of the methods that we normally use in human neuroscience is that the resolution is substantially better. So you place a grid of electrodes on the brain, and the temporal resolution is typically much better than you would get with fMRI. And the spatial resolution is way better than you would get with MEG or EEG and usually better than what you would get with fMRI as well.

So I'm going to tell you about one application of this type of method to understanding auditory scene analysis. And there's two things that I have to tell you about. And you're going to have to mostly accept them on faith. It's really one thing.

So we've talked about how you can make measurements from the brain and infer spectro-temporal receptive fields. And what you're going to have to accept is that if you have characterized the receptive fields of the neurons and if you measure their responses to a sound, you can reconstruct an estimate of the stimulus that the person is hearing. So given you can measure the responses to a sound. And if you have some prior measurements of their responses to sound and you've characterized like what the neurons respond to, you can then estimate what the person's listening to.

And so the study that I'm going to tell you about is an attempt to answer the question of, what stimulus is represented by the brain when you are listening to a mixture of sounds and trying to pay attention to one of them? So it's the classic cocktail party problem. You have two people who are talking. You're trying to listen to person A. The question is, what is being represented in your brain when that is happening?

And so the way that works is you do an experiment where you play the person a mixture of talkers. You cue them to attend to talker one. And then you do the reconstruction, and you compare that to the spectrogram of the mixture. And as well as of the individual talkers. And the question is, will the representation in your brain actually be a lot more similar to the voice that you're attending than to the voice that you're trying to ignore?

So here's an example of where these-- of these recordings are made. So this is in the auditory cortex, typically over the superior-temporal gyrus. This is an example of the spike-triggered average or the receptive field that you would get from one electrode. So this is an electrode that happens to be responsive to very high frequencies. And this is a depiction of one example stimulus of a mixture of two people talking.

So this is frequency. This is time. And now every point in the spectrogram has been color-coded depending on whether speaker 1 or speaker 2 is kind of dominating the mixture at that particular point.

So usually, the energy in a mixture is predominantly coming from one speaker or the other. And so at all the places where it's blue, that's where speaker 1 has the most energy. All the places where it's red, that's where speaker 2 has the most energy.

And so this is an example here where they're plotting the response of this one particular electrode, which is tuned to these high frequencies. And you can see in that high-frequency band, there are these two places that are kind of circled. So one of the circles is a place where speaker 1 has this high-frequency energy, and another place is where speaker 2 has this high-frequency energy.

So first, look at the dashed lines here. Those show the responses that we've measured in the brain when either speaker 1 or speaker 2 is being presented alone. And so you can see that the dashed line here has a peak which corresponds to that burst of high-frequency energy that's in speaker 2. The dashed line, when it's read, has a peak that's later in the trial that corresponds to that later burst of high-frequency energy. So the point is that the electrode is just responding to the content of the stimulus here.

Now the really critical case is where the person is being presented with a mixture of these talkers and either is attending to speaker 1 or attending to speaker 2. And that's the solid lines, the solid red and the solid blue. So this is a case where the physical stimulus is the same, but the listener is directing their attention to one talker or the other.

And what you're supposed to take away from this is that the brain responds during that condition. When you're attending to speaker 1 is more like the response to speaker 1 alone, and when you're attending to speaker 2, it's more like the response to speaker 2 alone. So the inference here is that the brain is preferentially representing the talker that you're paying attention to.

So this is just one kind of example from one electrode. This is what happens when you reconstruct the full stimulus. So here, we have speaker 1 and speaker 2. Speaker one is saying the sentence that's up at the top-- ready tiger go to green five now. Speaker 2 saying ready ringo go to red two now.

You may wonder, why are the sentences so weird? And the reason is that this is a corpus of sentence that was specifically designed for this type of experiment. And the idea is that there is a call sign on, which in this case, would be tiger or Ringo. And ahead of the trial, you tell the person, you got to pay attention to the person who says tiger.

So then you hear the mixture. You hear somebody say tiger, and then you're supposed to be tracking the rest of what they say. And you have to report the color and the number. Like green 5 would be the correct answer here.

And so the idea is that this gives you a way to ask people what they're hearing, and they have this discrete set of things that they can choose from. And so the experiment is kind of simple to conduct and code.

So now we get a mixture of the two talkers-- again, color-coded in the way that we discussed. And then these four panels are showing you the stimulus reconstructions that's made from the recordings from the brain. And so these first two that's called single sp1 and sp2, those are the cases where the person is hearing only a single talker. And the argument is that the reconstruction of the stimulus is supposed to look like the original stimulus but kind of blurrier.

So there's definitely some information that gets lost because the measurements from the brain are still very coarse. But you capture the coarse spectro-temporal structure of the original talker. And these and these look different. And that one looks like that one. And that one looks like that one. That's what you're supposed to take away from this.

So now you get the reconstruction of the mixture where the person is attending to speaker 1 or attending to speaker 2. And so what you're supposed to take away from this is that this reconstruction kind of looks like the reconstruction of the individual talker on its own. And this reconstruction, where the persons attending to speaker 2 looks like this one.

And what's shown down here is kind of a superposition of all of these things so you can make that comparison. But I think just this visual comparison is supposed to be enough. So this is just one example.

And to quantify this across like a large set of trials, they did this analysis where they computed the correlation between the stimulus reconstructions from the mixtures and the individual talkers. And specifically, you've got the correlation between the mixture and the reconstruction of speaker 1 and the reconstruction of speaker 2 when you're either attending to speaker 2-- that's the red dots-- or attending to speaker 1-- that's the blue dots. And so what you're supposed to take away from this graph is that the red dots are mostly above the diagonal, and the blue dots are mostly below the diagonal. And so that means that when you're attending to speaker 2, the reconstruction is more like the reconstruction for speaker 2 than the reconstruction of speaker 1 and vice versa when the attention is switched. So what this is showing us again is that the representation in the auditory cortex seems to be predominantly that of the talker that you are paying attention to.

The other thing that they did that's kind of cool is they separately did this analysis for trials where people answered correctly. So again, they're cued to attend to one of the talkers, and they correctly report the words that talker said. When they do the analysis on trials where people made mistakes, you no longer see the difference between the red and the blue dots. And that's consistent with the idea that the mistakes that people make are kind of attentional selection errors.

They don't manage to correctly attend to the person they're trying to pay attention to. And so the brain representation doesn't isolate the representation of that talker. What questions do you have about this?

STUDENT: Yeah, so when the dots are on the wrong side for the correct trials, when the dots are on the wrong side of the graph, is that more like the participant-- like it was a fluke or something? They just guessed correctly?

JOSH MCDERMOTT: I mean, my guess is a lot of this is just measurement noise. I mean, like the brain measurements are not-- none of this is perfect. The brain measurements are not perfect. The reconstruction algorithm is not perfect.

Yeah, you're always going to get some of that. Yeah, because you just have this one set of electrodes that's on the surface of the cortex. And yeah.

So I think that's some of that is bound to happen. So I'm not sure that is necessarily meaningful. In principle, I mean, it could be the case that maybe some of those dots that are on the wrong side are cases where the person wasn't as attending as successfully, and maybe they were not as confident in their answer or something. It's possible, but hard to know. Yeah.

So what have we learned from these kinds of experiments? So I would say that these type of experiments, they give us some indication of where in the brain individual sources can be selected with attention. In this case, it's the superior temporal gyrus.

So what does this tell us about the problem of auditory scene analysis of the inference of sources in the world? Well, in order for you to select a source with your attention-- the sources kind of have to be segregated. So it gives us some constraints on where that segregation happens. But I would say we still really know very little about the neural implementation of the inference of individual sources from mixtures.

And part of the challenge in that is that it's not so obvious actually how to go looking for neural correlates of auditory scene analysis. So looking for neural correlates of attention is pretty straightforward because attention is directed to a particular thing. And you can ask whether the representation of that particular thing is different compared to the thing that you're not paying attention to.

But the question of how is it that when the two people are talking, you actually represent in your brain the fact that there are these two things? That's actually something that we don't know very much about still. And my guess is there probably is some of fundamental representational difference when you're listening to one thing versus two in part.

So we saw those examples last time of when you perceive streaming between these two sets of tones, you become unable to discriminate the temporal relations between those tones. So that's some evidence that there's something pretty different about the representation of one thing versus two things. But what that is at this point is still essentially unknown. So that's like a major open problem, I would say, in neuroscience. Any other questions about neural correlates of auditory scene analysis?

The next topic we're going to discuss-- and this will really be the last part of the section of the course on hearing-- is speech perception. And you may wonder why are we talking about speech in a perception class. You often think of language as being a high-level cognitive phenomenon. And the answer is that speech is typically received by the brain as a sound signal, and perceptual processes are needed to transform the sound into a form that semantic and syntactic processes that we associate with language can handle. And this requires solving some challenging perceptual problems.

Speech is also like one of the most important things that we do with our auditory system. It's like an important thing to understand in the context of the auditory system.

So one of the key concepts to take away from today's lecture is the source-filter model. And that refers to how we think of the way in which speech is generated. So the idea is that speech is produced by a source that generates a sound that gets passed through a filter. The sound source is the larynx.

This is a video of somebody's larynx. So what happened was there's a little camera that got stuck down somebody's throat. And then they had to say something like a vowel. And then they slowed down the movie a lot.

So these are the vocal folds or vocal cords. So there are these bits of tissue that kind of open and close. They open and close when air is passed through them. So air gets blown through the vocal folds, and they open and close at a particular rate. The rate at which they open and close depends on the tension in the muscles of the larynx. So that's something that you can modulate.

So I can talk like this, or I can talk like this. So that's just changes in the larynx. So that produces a sound.

And then there are resonators that are part of-- that are essentially the vocal tract. And those are things that you can change. And when you talk, you are constantly changing those resonators. The sound passes through the pharynx or the throat, the mouth, the lips, and the nose, and that filters the sound.

So here's a sideways view of the vocal production apparatus. We've got the vocal folds down here that generate sound. And then you can see the sound kind of passes through these cavities before emerging and traveling to whoever you're trying to talk to.

So this is the spectrum of the sound source from the larynx when the speech is voiced-- so when the larynx is vibrating. And typically, the opening and closing of the larynx will be periodic in time, characterized by a fundamental frequency that produces harmonics. So you see how there are these frequencies that are regularly spaced? That's a harmonic signal.

So that's what kind comes out of the larynx. That then gets passed through your vocal tract. And the vocal tract can be configured in different shapes, and that causes it to filter the sound differently. So these are three shapes corresponding to three different vowels. This is a schematic of the transfer function that is associated with those three shapes. So the transfer function essentially defines the frequency response of the filter.

So this is like the gain that would be applied to the frequency. And this is frequency. And you can see that these transfer functions, they have peaks associated with them. Those peaks are known as formants. And critically, the peaks are in different places. And so that's what makes different vowels sound different.

So the signal that comes out of your mouth you can think of as having the spectrum of the source multiplied by the transfer function of the filter. And that gives you this signal or this one or this one. So source and filter-- that's the source-filter model of speech production.

So speech is typically characterized by having a fundamental frequency. So everybody's vocal folds can produce some range of fundamental frequencies.

In men, that would typically be from around 80 to 240 hertz. In women, it's higher-- from 140 to 500 hertz. In children, it's higher still-- 170 to 600. This is determined by the length and the thickness of the vocal cords. And it's also different obviously across individuals.

The sound signal is typically harmonic. And then there are resonators. These are cavities that amplify certain frequencies and dampen others. In general, bigger cavities have lower resonant frequencies and produce lower-frequency sounds. Smaller cavities have higher resonant frequencies and produce higher sounds, So the filter resonance is those peaks in the transfer function are known in speech known as formants, and they will be abbreviated as F1, F2, F3, where F1 would be the lowest peak. F2 would be the next one up. F3 would be the next one up, and so on and so forth. And these result from the size and the shape of the resonating cavities.

And so sound is modulated by manipulating the articulators. So these are the movable parts of the vocal apparatus-- tongue, lips, palate. And these movements change the resonant properties, and they also change the airflow.

So phonemes are kind of an important idea in speech. We often think of them as the goal of speech perception. So phonemes are typically defined as the smallest unit of sound that can make a difference in the meaning of speech-- so the thing that differentiates pot and dot. It's that initial sound-- the "puh" and the "duh."

So phonemes often have a rough correspondence with letters when you write words down. But they're not letters. There's not a one-to-one correspondence. They're a part of sound.

So we can think of the problem of speech perception as the problem of extracting a string of phonemes from the speech signal. And the issue is that this is hard to do. And that's because the sound waveform that results from a given phoneme can be highly variable across different conditions.

So English is typically thought to have around 40 phonemes. Different languages have different numbers of phonemes. They vary from as low as 11 to as high as 140. The total inventory pooled across all the languages of the world is thought to be in the thousands, but there are some phonemes that are very common across most languages.

So one type of phoneme are vowels. Vowels are phonemes that are produced by an unrestricted vocal tract, and different vowels are distinguished by how the sound produced by the vocal cords is filtered.

And typically, the resonances can be altered in two main ways. One is by which part of the tongue gets moved, either the front or the back, and the other is the position of the tongue in the mouth-- can be high or low. So let's practice. Everybody say bet, bet, bet, bet, and pay attention to what's happening to your tongue-- bet, bet, bet. Now say buh, buh, buh, bet, but, bet, but, bet, but.

So you feel your tongue kind of moving around? Yeah? Yeah. So you're normally not aware of this stuff. But that's what's happening. Let's do this one-- beat, beat, beat, bat, bat, bat, beet, bat, bat, beet, bat, bat, beet, bat, bat. Yeah.

So here are three examples. So we talked about how the key feature of these transfer functions are these peaks, the formants-- so F1, F2, F3. Here you can see F4. And the first two formants do a pretty good job, actually, of distinguishing different vowels.

So this is a graph that is plotting the frequencies of the first formant and the second formant for a whole bunch of different vowels-- heed, hid, head, had, hawed, hood, who'd. So when I just did that, my vocal tract was rapidly reconfiguring itself, causing these resonant frequencies to move around in this particular way. So heed happens to have a very low first formant and a very high second formant. Who'd has a pretty low first formant but a relatively low second formant, and so on and so forth.

So it's pretty common to plot vowels in this space of the first formant versus the second formant. And to first order, the first formant is kind of mostly determined by the position of the tongue on the high-low axis, and the second formant is predominantly determined by the position of the tongue on the front-back axis.

So let's do boot versus beat-- boot, beet, boot, beet. So you feel like when you're saying "ooh," the tongue is low in the mouth, and "EE" gets higher-- ooh ee-- or sorry, we messed this up. Ooh, ooh-- the tongue is more towards the back. Ee is more towards the front-- ooh, ee, ooh, ee.

Feel your tongue moving from front to back? Yeah. So now if we do a comparison between boot and bot, so we should be able to detect the tongue moving from low to high-- ooh, aah, ooh, aah, ooh, aah. Yeah, that works. You can have lots of fun with this.

These are spectrograms of these vowels. These look differently from spectrograms that you have seen before. And this is because of something that we haven't really talked about, which is that when you are generating a spectrogram, what happens is you take the sound signal, and you take a little window of that signal at a given point in time. And you take the Fourier transform of that, and then you plot that.

And the Fried parameter, when you're generating the spectrogram, is the size of the window. And because of the uncertainty principle between the time and frequency domains, when the time window is very short, the frequency resolution will be poor. When the time window gets longer, the frequency resolution will get better. But your time resolution will get worse because you're kind of averaging the signal over a longer period of time. And so that's a choice that gets made.

And so often in the spectrograms that we have seen previously in the class, we will be choosing the settings of the spectrogram such that the frequency resolution is pretty good so you can see the individual harmonics. However, it's pretty common with speech to actually set things up so that the frequency resolution is not so good. And so when you do that, you get these stripes here that really correspond to the formants. And so the reason that choice is made is because in the analysis of speech, you often don't care so much about the harmonics, but you really care about the formants because that differentiates phonemes. So that's why these look the way that they do.

Here's a spectrogram that's made a little bit more like what you're used to. So this is just a recording that I made of two different vowels.

[AUDIO PLAYBACK]

- La, looh.

[END PLAYBACK]

JOSH MCDERMOTT: And so you can see that there is a fundamental frequency here that's around 250 hertz. You've got harmonics of that fundamental frequency. But the relative amplitude of those harmonics is pretty different. So the "ooh" sound kind of has all the energy down here. The "ahh" sound kind of has peaks there and there.

But this is supposed to make the point that you can maintain the fundamental frequency and keep the pitch the same and still have the vowel sounds be totally different because the formants will be moving around. So in this case, the filter is changing, and the source is not.

So that's kind of the story with vowels. They're made with an unrestricted vocal tract. They're characterized by the formants. That's the classical textbook story.

Consonants are the other main type of phonemes, and consonants are generated when the vocal tract is restricted in some way. And the different types of consonants that you can make are distinguished by differences in what's called the place of articulation. So that's where the restriction is occurring.

And so it could be dental. That refers to the restriction happening at the teeth. We'll see some examples in a second. Velar-- that means the roof of the mouth.

They can also be distinguished by the manner of articulation. So the restriction can be complete. So in other words, the vocal tract-- the airflow will be completely cut off and then released. So that's a stop.

You can get nasal articulation where the airflow is forced up through the nose. Fricatives are where the airflow kind of gets blown through a narrow gap. So that's like where you force the air through.

And then they also differ in the voicing. So consonants can be voiced or unvoiced. And that refers to whether the vocal cords are vibrating at the start of the consonant. We'll see some examples of that.

So here are examples of stops that differ in the place of articulation as well as whether they are voiced. So let's look at the first two. So "buh" and "puh"-- everybody go buh, puh, buh, puh. So you can tell when you were doing that, your lips were opening and closing. That's a labial stop-- buh puh, buh puh.

So what causes "buh" and "puh" to sound different? Well, the difference is that when you say buh, buh, like at the moment when the stop is released, the vocal cords vibrate. Whereas with puh, puh, there's this little moment where the vocal cords are not vibrating, and so you just get turbulent airflow being forced out. And then the vocal cords are vibrating, and you get the periodic voiced signal coming through-- buh, ph.

So let's do the second one-- deh, teh, deh, teh, deh, teh. So that's a dental stop. So your tongue is coming up against your teeth and then releasing. And again, though the difference between deh and teh is whether or not they are voiced-- so whether the vocal cords vibrate right at the start of the release or not.

And so when you are learning how to talk, you're picking up on these distinctions that exist in your language and figuring out how to do this stuff. And then it all happens just kind of unconsciously. So it's like incredibly precise motor control of this stuff-- coordination between what your lips and mouth are doing and what the vocal folds are doing.

Finally, let's do the third one-- guh, kuh, guh, kuh, guh, kuh. So there the stop is happening at the back of the roof of the mouth. And again, we get the distinction between voiced and unvoiced consonants. So this is making the same point that I was just saying earlier.

So these kinds of spectrograms are conventional for speech analysis where they use small time windows that gives you coarse frequency resolution. So the horizontal stripes here are formants. And then the thing that you can see here that wasn't so evident on the ones we saw earlier is that the source, which are these vocal folds that are opening and closing, is evident as thin vertical lines. Each of those vertical lines, it corresponds to one pitch pulse-- so one opening and closing of the vocal cords.

So it's kind of cool that you can look at it with these small time windows. And the periodicity of the source shows up in the time domain. Or you can look at it with longer time windows, and it would actually show up as harmonics in the frequency domain. So that's the thing to take away from this is that those small kind of thin vertical stripes are the voicing. That's the vocal cords oscillating periodically.

Now this slide is supposed to show you the difference between voiced consonants and voiceless or unvoiced consonants. And that shows up at the very start of the consonant. So you can see that in all of the voiced cases, you see those pitch pulses kind of right from the get-go. Whereas in the unvoiced cases, you don't see the pitch pulses.

Well, the time axis is not labeled here, but it would be like 40 milliseconds or something. It's a pretty short amount of time. So that's like voiced versus unvoiced.

Another there are a couple other types of consonants that we'll talk about. Fricatives are cases where there's a partial restriction, a narrowing. And typically, there'll be some turbulent energy that gets forced through this thin aperture. So let's do "zuh" versus "suh"-- zuh, suh, zuh, suh.

So there's a thin gap between your teeth. But they differ in voicing. So zuh is voiced. Your vocal cords vibrate from the get-go. Suh, suh is unvoiced. There's this little moment at the start where the vocal cords aren't doing their thing. So that's an example of fricatives, where air is forced through a narrow gap between two articulators.

And then there are nasal consonants where air is forced up through the nose, and that kind of just makes it sound a certain way. So let's do "muh," "nuh"-- muh, nuh, muh, nuh, muh, nuh. And so the difference between those is again where the stoppage in the mouth is occurring. And then the air gets forced up through the nose. And that changes the resonances. It just creates different sounds.

So those are examples of phonemes and how they are produced. So you can see a lot of structure if you look at spectrograms of speech. So this is an old-school spectrogram of somebody saying I can see you.

So this is the fricative of the S. You can see this broadband high-frequency noise. Here are the vowels-- eeh, euu. These are formants. You can see that they change as you move from the eeh, euu.

The other thing that's kind of important to note about spectrograms of speech is that the things that we perceive to be words are typically not well-separated in the actual signal. So "I can see you" has four words in it. The gap between the C and the U is just not there in the signal. There's this continuous thing. And so that's kind of an important kind of hard problem that also has to get solved and figuring out how this continuous signal actually consists of these discrete words. So I just said all this stuff.

So whispering is another interesting thing that happens with speech. (WHISPERING) So can talk like this. Everyone can probably understand me. But it's very different.

So whispering-- what happens with whispering is that the vocal cords don't vibrate. So how does that work? How would you be able to understand me talk when my vocal cords don't vibrate? Anybody have any idea? Yeah.

STUDENT: You can hear the noise in the air. It's not going to be vocal.

JOSH MCDERMOTT: Yeah, you can hear the noise, but how would you be able to detect the structural features of these phonemes? Yeah.

STUDENT: By seeing you speak, like seeing--

JOSH MCDERMOTT: (WHISPERING) But I can go like this, and you can still understand me just fine.

STUDENT: The noise is still shaped by the resonances of your mouth area. Yeah.

JOSH MCDERMOTT: So that's probably like a big part of it is that the filters are doing more or less the same thing. And so you can probably detect the formants. They're just excited by a noise signal rather than by a periodic harmonic. But William is correct that it definitely helps to be able to see what people are doing with their mouths as well. And we'll see some examples of that.

So naively, you might imagine that you could build detectors for the phonemes, and speech reception would be solved. And the challenge that's classically associated with speech perception is that phonemes are produced differently depending on lots of factors. And so this results in a constancy problem, a lot like the others that we often encounter in perception, like the problem of recognizing objects across viewpoints.

And so one factor that's kind of important in this is what's called coarticulation. And this refers to the fact that phonemes are produced differently depending on what comes before and after them. And the crazy thing is that even though the underlying acoustic signature that corresponds to the phoneme kind of differs, they sound like the same thing to us. So somehow, they're recognized as the same despite producing different patterns of energy.

And so this is an example where We have three syllables that start with the same consonant-- so "baah," "boo," "bee" or "dah," "doo," "dee," or "gah," "goo," "gee." And these are spectrograms of each of those syllables. And the point is that-- so the consonant kind of lives at the start of the syllable. And if you look at what is actually going on in that initial little bit of syllable, it's quite different in the three cases because it varies depending on what vowel is going to come after the consonant. And that's true in all of these cases.

Here's another case that's pretty interesting. So these are spectrograms of somebody saying "eebah," "oobah," "eedah," "oodah." And the point of this is that the acoustic signature of the "buh" sound when you're seeing eebah looks a lot like the acoustic signature of the duh sound when you're seeing oodah. So the physical signal in this little bit of sound is pretty similar for the buh and the dub, but eebah sounds like a buh, and oodah sounds like a duh. Yeah.

STUDENT: Is this only for consonants? Are vowels also changing based on preceding?

JOSH MCDERMOTT: Most of the examples that I know about are with consonants, yeah. I'm not aware of this happening a whole lot with vowels, although there may be some effects of that too. Yeah. Yeah. So co-articulation-- phonemes are produced differently. So they have distinct acoustic signatures depending on what comes before and after them. And the reason for this, we think it really has to do with mechanical constraints on how quickly you can change your vocal tract shape.

So you got to get from point A to point B if you want to make one sound and then another. And that kind of forces you to go through certain trajectories in the kind of space of the state of your vocal tract. And you can't just make arbitrary sounds with your vocal tract, and that's probably why this happens.

But there's lots of other factors that kind of affect the speech signal. So there's prosody. So you can vary stress to convey emphasis.

We can say permit or permit. So you need a permit to skip class, or I permit you to skip class. Those mean very different things. And it's conveyed by the stress on the different parts of the word.

There's intonation. Dad wants me to mow the lawn. Dad wants me to mow the lawn. Those mean different things. And it results in the speech having different acoustic characteristics.

There's also differences in emotional state that can change the way people talk. Different speakers sound different, so people vary in accents. It's a function of gender, age.

So the point is that the same phoneme can be realized in many different ways. And sometimes the variation in phoneme acoustics produces ambiguity in what is said. So this was an example that everybody probably knows because it kind of went viral like a while ago. So this speech signal that people disagree about. I'll play it for you.

[AUDIO PLAYBACK]

- Laurel, Laurel, Laurel, Laurel.

[END PLAYBACK]

JOSH MCDERMOTT: How many people think that says Yanny? How many people think it says Laurel? Yeah, so it's about half and half. And the presumptive explanation here is that there is some ambiguity in what was said. And different people interpreted it a little bit differently.

So there was lots of experiments that were done on this like when this was first kind of discovered. This was like now five years ago. And there are definite effects of filtering the stimulus.

So for instance, if you low-pass filter the stimulus, people are a lot more likely to say that they hear Laurel. If you high-pass filter the stimulus, people are more likely to say that they hear Yanny. But some people consistently hear one or the other, no matter what. And I think the explanation for why for the individual differences remains a little bit unclear.

So here's just a continuum where the-- so I think this is the high-low ratio. So this one is probably pretty low.

- Laurel.

JOSH MCDERMOTT: So that should mostly sound like Laurel to a lot of you. Is that right? How many people say that's Laurel? Now, let's just do the controlled experiment.

- Yanny.

JOSH MCDERMOTT: How many people say that's Laurel? Nobody-- so everybody thinks this one's Yanny, and a lot of people think that one is Laurel. This is a sequence where the thing is the cutoff of the filter. I don't know exactly how they did this, but they're changing the frequency content. And it goes from low to high. So that means that it should tend to sound more like Laurel initially and then eventually maybe transition.

- Laurel,
Laurel, Laurel,
Laurel, Laurel,
Laurel, Laurel,
Laurel, Laurel,
Laurel, Laurel.

JOSH MCDERMOTT: Never transitioned for me-- how about you? Laurel all the time? Some of you. Now, here's the reverse.

- Laurel,
Laurel, Laurel,
Laurel, Laurel,
Laurel, Laurel,
Laurel, Laurel,
Laurel, Laurel,
Laurel, Laurel.

JOSH MCDERMOTT: So what you probably experience here is that the perception of these two continua is different. And this illustrates something called hysteresis. So hysteresis refers to the fact that there tends to be some persistence of the perceptual interpretation. So in other words, if you started out hearing the thing as Laurel, you'll tend to hear it as Laurel for a while. If you start out hearing the thing as Yanny, you'll tend to hear it as Yanny for a while.

And so this is a graph that shows the proportion of times that people respond with the thing being Laurel. And this is as a function of the high-to-low ratio. And what this is showing is that-- so there's some people here-- this is the blue line-- who basically always hear Laurel, kind of independent of where the cutoff is set. There's some people, as the yellow line, who pretty much always say they hear Yanny. But then there's a group, a very substantial group of people, that transition from hearing Laurel to Yanny.

And this is a plot that shows how confident people are in the judgment. And the point is that people are very confident. So you're like, usually pretty sure that it's either Laurel or Yanny, and then it just kind of switches at some point.

So there are these occasional kind of ambiguities in the speech signal that kind come out in these particular ways. And now that there's social media and the internet, and it's a lot easier to discover these things because people start talking about them. And then they become known. And we don't so far actually have a way to discover these things automatically. They just kind of happen by accident.

So the point is that speech perception is thought to be a challenging perceptual problem because of all of the variation in how particular phonemes are produced across conditions. So what are some potential solutions to this?

Well, it's often thought that there are probably some phonemes that are pretty invariant-- that is, whose acoustic signature is fairly consistent across conditions. So stops pretty much always produce bursts of energy. Turbulence are typically associated-- sorry, fricatives are typically associated with turbulence, so broad-spectrum energy. Vowels usually have steady-state formants, and the relations between the formants is often pretty consistent. Nasals are also fairly consistent.

There's also other things that have been proposed, though, and one of the most famous ideas that has come out of speech perception is this idea of categorical perception. And this is the idea that we learn and impose categories on physically continuous stimuli. So we'll do a quick demonstration of this before we break.

And the way that this works is we'll take two phonemes, in this case "cah" and "gah." So those differ in voicing-- cah, gah, cah, gah. And you can define that difference in terms of the voice onset time. That is how long after the onset of the consonant does the voicing kick in? And that will be in like tens of milliseconds, typically.

And we can artificially create a continuum by manipulating the voice onset time using a speech synthesizer. So when the voice onset time is small, the thing will be voiced. When it's large, it will be unvoiced.

So we can create a physical continuum of stimuli where the voice onset time is going to vary from 0 to 80 milliseconds. And then we can ask you what you hear. And the phenomenon of categorical perception is that people tend to hear one phoneme or the other, even though you're drawing these things from a continuum.

So these are stimuli that are drawn from the continuum that will be presented in random order. So just listen to these. And so I'm pretty sure this is the labeling of the stimuli, and then this is telling you what people normally hear. But if you don't want to be biased, just kind of close your eyes.

[AUDIO PLAYBACK]

- Gah, kah, kah, gah, gah, gah, kah, kah.

[END PLAYBACK]

JOSH MCDERMOTT: So this is the answer key here. And that was fairly consistent with what I heard. So there's maybe one or two that's uncertain, but people are usually fairly confident. And so if you ask people to categorize the phoneme as being either gah or kah and you measure the categorization judgments as a function of the voice onset timing, you find that there's this very rapid transition between them saying that they hear gah and then switching to saying that they hear kah. So that's one feature of categorical perception.

The second is this idea that discrimination will be best at the category boundary. And so the kind of naive idea around that is that if you've taken this speech signal and mapped it onto a category, then if I give you two examples that are kind of within that same category, if you've thrown out like the physical differences between them and you're just representing them as a category, then it should be difficult to tell the difference between those two examples, even though they're physically distinct. And so the idea is that we can do an experiment where we measure discrimination between pairs of stimuli along this continuum.

The category boundary's here kind of in the middle. And the idea is that if you're within the category, so over here or over here, your discrimination will be poor. But if the pair of points straddles the category boundary, your discrimination is good. And so this is the discrimination function that would be predicted.

So those are the two characteristics of classical categorical perception. So the first is steep categorization functions. So if you ask people to categorize the stimulus, there's this very rapid transition from people hearing it as one thing to hearing it as another. And the second is that if you do a discrimination experiment, discrimination is best at a category boundary.

And when we resume next week, I'll show you some empirical data that's kind of in support of this. And then we will finish talking about speech perception. Have a good weekend.