[SQUEAKING]

[RUSTLING]

[CLICKING]

NANCY KANWISHER: We are turning from our various other topics to talk about hearing today. And let's start by thinking about all the cool stuff that you can do just by listening. So just by listening, you can identify the scene that you're in and what's going on in it, like for example, this.

[AUDIO PLAYBACK]

[END PLAYBACK]

OK, so you know what kind of room you're in and roughly what's going on, just from that little bit of sound. You can localize events and people and objects. So close your eyes, everyone. Keep them closed.

And if you just listen to me talking, it's really very vivid, isn't it, exactly how obvious it is where I am. And I will refrain from the temptation of coming up and speaking in somebody's ears because it's just too creepy. OK, you can open your eyes.

It's very vivid. Just from listening, you know where the sound source is. You can recognize sound sources, so for example, sounds like this--

[GLASS BREAKING]

You know what happened there. It's a whole vivid event just unfolded there in a whole second and a half, or a random series of sounds like this.

[AUDIO PLAYBACK]

- It's supposed to either rain or snow.

[RANDOM SOUNDS]

- Hannah is good at compromising.

[RANDOM SOUNDS]

[LAUGHTER]

[END PLAYBACK]

NANCY KANWISHER: Anyway, every one of those sounds you immediately recognize. You know exactly what it is. And that's environmental sounds, things that happen outdoors, speech, what is being said, voices, who is saying it. If you don't know the person, if they're male or female, young or old, much like faces-- if you know them, you'll recognize them pretty fast.

You can selectively attend to one sound among others. Like if you had a little, hidden earphone that I didn't see, and you wanted to listen to your favorite podcast, you could listen to that occasionally when I was getting boring. And then you could turn back and listen to me. And you could just selectively choose which of those different audio inputs to listen to.

And we'll talk more in a moment about this classic problem in hearing, which is known as the "cocktail party effect." I guess it was named in the '50s when cocktail parties were big. And it consists in the fact that when there are multiple sound sources, such as many people talking in a room, you can tune in one channel and then tune in another channel. And you can just selectively attend to one of many sound sources, even though those sound sources are massively overlapping on top of each other in the input. And it's a big computational challenge, as we'll talk about shortly, to do that.

You can enjoy music. And you can determine what things are made of. So close your eyes and I'm going to drop things on the table. Don't look.

I'm going to do various things and you're going to identify them. So let's see-- don't open your eyes. See if you can tell what's being dropped on the table, or at least what it's made of. Close your eyes. That's cheating.

Wood, exactly. Very good. OK, what is this? Keep your eyes closed. What is this made of?

- Plastic.

NANCY KANWISHER: Yeah, good. Keep your eyes closed. What is this made of?

STUDENT: [INAUDIBLE]

NANCY KANWISHER: Yeah. OK, keep your eyes closed. What's this made of?

STUDENT: [INAUDIBLE].

NANCY KANWISHER: Awesome. OK, you can open your eyes. Perfect. You guys are awesome. I just dropped these objects that I found from my kitchen this morning and you guys could tell what they're made of. That's amazing.

OK, all of this that you guys just did happens from the simplest possible signal. We'll talk about what that signal is exactly in a moment, but it's just sound compression coming through the air. And it tells you all this rich stuff about your environment.

So the question is, how do we do that? And the first question is, how do we start to think about how hearing works, how you're able to do all of that? And you guys know we start with computational theory-- considering what the inputs are, what the outputs are, the physics of sound, what would be involved if we tried to code up a machine to take those audio input and deliver the output that you guys all just delivered with no trouble whatsoever.

What cues are in the stimulus? What are the key computational challenges? And what makes those aspects of hearing challenging?

And then after we do all that stuff at the level of computational theory, we can, of course, study hearing in other ways, like studying it behaviorally. What can people do and not do? What's hard? What's less hard? And we can measure neural responses.

So we'll talk about all of that. But let's start with a little more on what sound is. So sound is just a single univariate signal coming into the ears. We'll say more about that in a second, but it's really, really simple.

And from that, you get all this rich experience. And so the question is, what goes on in that magic box in the middle to enable you to extract this kind of information from this really simple signal? So let's start with what is sound.

Sound is just a set of longitudinal compressions and decompressions of the air coming from the source into your ear. So these waves travel from the source to the ear in little waves of compression where the air is just compressed, and rarefaction where the air is spread out. And just to give you a sense of how physical sound is, there's a silly video here.

It's a speaker in a sink with a bunch of paint. And you can just see that the movement of the speaker-- normally, it makes those compressions and rarefactions of air, but if you stick paint on it. It's going to shove the paint up in the air, too, just to show you how physical it is.

There's something called Schlieren photography, which is totally cool, and which is a way to visualize those compressions of the air to show you what's--

[VIDEO PLAYBACK]

[INTERPOSING VOICES]

- --use it to study aerodynamic flow. And sound-- well, that's just another change in air density, a traveling compression wave. So Schlieren visualization, along with a high-speed camera, can be used to see it as well. Here's a book landing on a table, the end of a towel being snapped, a firecracker, an AK-47, and of course, a clap.

[END PLAYBACK]

NANCY KANWISHER: OK, so just compressions of air traveling from the source to your ears-- that's all it is. So natural sounds happen at lots of different frequencies. And one of the ways we describe sounds is by looking at those frequencies.

So there's an awesome website that is here on your slides. You can play with it offline. But meanwhile, we're going to play with it a little bit right now because it is so cool.

So what we're going to do is we're going to look at spectrograms of different sounds. Let's start with a person whistling.

[WHISTLING]

OK, so frequency is on this axis, higher frequencies up here, lower frequencies there. And it's going by in time.

[WHISTLING]

So whistling is unusual in that it's pretty much a single frequency at a time. Many natural sounds are not like that. So you see not single, but a small, narrow band of frequencies at a time. OK, that's enough.

[WHISTLING]

Stop. All right. OK.

[TROMBONE PLAYING]

OK, so you see how with the trombone, there were many different bands of frequencies. In contrast-- this is me talking, by the way. We'll talk about that in a second. But with the whistling, you saw just a single band at a time. With the trombone, it has all of these harmonics, these parallel lines of multiples of frequencies.

Those are called "pitched sounds." Sounds that have a pitch where you could sing back the tune have those bands of frequencies like that. And so that's what you see with the trombone.

You see a little bit of this with natural speech here. You can see sets of bands, but mostly, you see vertical stripes. That's because I'm talking fast and mostly what's coming out is consonants. If I slowed down and stretched out the vowels, you would see more of those harmonics. Fun and games.

OK, so that's what sound looks like. So everybody has to have a sense of this is showing you the energy at each frequency over time in response to natural speech. We'll play with this a little bit more later in the lecture.

So we did all that. We'll do some of that other stuff later. So now that we have some sense of what sound is and what that input is, how are we going to think about how to extract information from it?

What we want to do is think about how is it? Why is it challenging to get to that from this? There are several reasons that's challenging.

First is invariance problems, much like we've discussed in the domain of vision and other domains already in this class. And so the way to think about that here is that a given sound source sounds really different in different situations. So if we have different people saying the same word, that will look very different on those spectrograms. The stimulus is actually different, even though we want to just know what word is being said.

And conversely, if we have the same person saying two different words, that will look really different. And even if we want to know just who's speaking, we have to deal with the invariance of generalizing across those very different ways, very different sounds that they produce when they say different things.

So those are kind of flips of each other. To recognize voices, we want invariance of the voice with respect to the words. To recognize words, we want invariance for the words independent of the voice. And those are all tied up together. So we need to appreciate the sameness of those stimuli across those changes.

Here's another reason that hearing is challenging-- I mentioned this briefly-- in normal situations-- it's pretty quiet in the room. There's some background noise, but not a whole lot of other noise, so it's mostly just me making the noise in here. But in many situations, there are multiple sound sources. For example, listen to this.

[AUDIO PLAYBACK]

[INTERPOSING VOICES]

- All right, Debbie Whittaker, Sterling James, wrapping things up.

[END PLAYBACK]

NANCY KANWISHER: OK, little segment of radio, there's music, and a person speaking both at once. And you had no problem hearing what the person was saying and knowing something about the gender and age of that person. You recognize the voice, the content of the speech, even though the music is right on top of it.

So the music might be like this and the speech like that. And what you hear is this, with those things right on top of each other. So you need to go backwards to hear these things, even though that's all you get. Everybody see how that's a big challenge? If you had to write the code to take this and recover that, best of luck to you. Yeah, question?

STUDENT: How does intensity or volume come into this picture again?

NANCY KANWISHER: It's not really well depicted on these diagrams. This is just showing you the entire source. So the intensity I showed before essentially takes this and does a Fourier analysis of it so that it gives you the energy at each of those frequencies. So you could just do a Fourier analysis on this and you get a spectrogram.

So the listener's usually interested in individual sources even though they're superimposed on other sources. And that's a real problem. So this is the input. They get added together, and the brain has to pull them apart.

So this is a classic, ill-posed problem. That means just given this, we have no way to go backwards to that if that's all we have, because there's multiple possible solutions. It's like saying, "x plus y equals 9, now solve for x and y."

And whenever we're in that situation of an ill-posed problem with multiple possible solutions, only one of which is right in any situation in the world, the usual answer is that we need to bring in some other assumptions or world knowledge or something, to constrain that problem and narrow that large, usually infinite, space of possible answers down to the one correct one.

So this is a classic problem that people have talked about in audition for many decades. Josh McDermott in this department does a lot of work on it. And you can solve it in part by knowledge of natural sounds, which I won't talk about in detail here.

One more challenge for solving problems in audition comes from the fact that real world sounds, including the sound of my voice right now, have reverb. So "reverb" means-- this is an aerial view. That's a person, kind of hard to see in an aerial view.

And that's a sound source. And some of the sound comes straight from the sound source to the person's ears. But a lot of the sound goes and ricochets off the walls, god knows how many times, before it hits the ears.

And all of those different paths of sound are all kind of superimposed at the ears. And they arrive at different times, making a hell of a mess of the input sound. So instead of that nice, clean, straightforward input, you have the input plus a slightly delayed input, a more delayed input, another delayed input, all superimposed on top of each other. That's reverb.

Is that clear, what the problem is? So now we have this really messed-up signal that we're trying to go backwards and understand what the input is. So I'll give you an example.

This is a recording of what's known as "dry speech." That means speech with no reverb. Sorry, question?

STUDENT: I'm just having a little trouble understanding why reverb poses a problem. The stimulus isn't changing, it's just delayed over time.

NANCY KANWISHER: Yeah, OK. Let's do a vision example. This is a little crazy, but let me just try this.

Suppose we had a photograph of my face and you have to recognize it. OK, fine. Various visual algorithms can do that. But now suppose we took that photograph and we moved it over 10%, and we superimposed it and added them together, and then we moved it over again and added them together, and moved it over again and add them together. Pretty soon you have a blurry mess.

And those things are all on top of each other, just as two people talking at once are on top of each other. And so you have a real problem going backwards. Does that make sense? OK. OK, so here's dry speech with no reverb.

[AUDIO PLAYBACK]

- They ate the lemon pie. Father forgot the bread.

[END PLAYBACK]

NANCY KANWISHER: OK, here's the same speech but with lots of reverb.

- They ate the lemon pie. Father forgot the bread.

[END PLAYBACK]

NANCY KANWISHER: OK, now you can still hear it because your auditory system knows how to solve this problem. But look what happens to the spectrogram. Here-- this is time this way, frequency this way, and the dark bits are where the energy is, where the power is.

In the dry speech, you see all these nice, vertical things, and here you see a blurry mess. Nonetheless, you can hear it fine. And further, what else could you tell from the reverb?

STUDENT: The size of the room.

NANCY KANWISHER: Yeah, it's in a cathedral or something, right? So it's not just that it causes a problem. Reverb also tells us something about the location we're in, if we know how to extract it, which you guys' visual-- auditory systems do. You can see I'm a vision scientist.

So how to study this? There's a very beautiful paper that Josh McDermott published a few years ago. And I'm going to try to give you the gist of the paper without all the technical details, because I think it's just brilliant.

So they wanted to characterize what exactly is reverb. And reverb is going to vary for different sounds. You heard the reverb in that cathedral-like space. That's very different from the reverb in this room, which also happens. It's harder to hear because it's less obvious.

But you can tell a lot about the space you're in because the reverb properties are different. The distance to the walls are different. The reflective properties are different. And so there's information there.

So you can characterize the nature of the reverb in any one location by making an instantaneous, brief click sound in that environment and recording what happens after that. And then you can collect all the reverberant reflections of that sound off the walls. So what they did is they went around to lots of natural locations and they played a click like this.

[CLICK]

That's it, just a click. And then they recorded. So this is the initial click, but this is what you record in a single location, all this stuff. And those are all the reverberant reflections of that sound off the walls-- make sense? For one location.

So then they did that in a whole bunch of locations. And the idea is that here is a description of the basic problems, just the same thing I said before, but slightly more detailed. So a sound source would be something like this. This looks like a person speaking, with those nice, harmonic, parallel bands, like you saw when I was speaking. Maybe it's a trombone.

So that's time. That's the source. That's what you want to know.

This is now the impulse response function for the location where that sound is being played, determined by doing that click and recording. I showed you just you do a Fourier analysis of that black curve in the previous slide and you get something like this. And that shows you all the echoes that happen in that sound in that location. And there are different time delays, and different intensities, and frequency dependence.

What comes to your ear is basically this times that. So you're given this and you have to go backwards and solve for that. Everybody see the problem?

So what McDermott and Traer showed is that-- just to state the problem a little more clearly-- you're interested in the source and/or the environment. You might want to know what kind of room am I, if somebody is dragging you around blindfolded. You might want to know if you're outside, or inside, or in a cathedral, or a closet, or what.

And now this should seem very analogous to the problem of color vision. Remember the problem of color vision? We want to know the color of this object right here.

So this little, purple patch here, we want to know that, but all we have is the light coming to our eyes from that patch. And the light coming to our eyes from the patch is a function not just of the property of the object, but whatever light happens to be coming onto it and then reflecting to our eyes. And so in color vision, we have one set of tricks to try to solve that problem and recover the actual properties of the object, even though it's totally confounded in the input with the properties of the incident light.

This is extremely analogous. We're trying to solve for what is the sound source. And we have to deal with this problem that is completely confounded with the reverberation of the room it's in. Does everybody see that analogy? They're both classic ill-posed problems in perception.

So here's another way of putting it-- we're given that and we want to solve for at least one of these, ideally both of those. And you can't do that with just this. So we need to make assumptions about the room.

And what Traer and McDermott showed is that first, they measured those impulse response functions in natural environments to characterize reverb in different environments. And they found that there's some systematic properties of reverb having to do with the decay function as a function of frequency. And those systematic properties are preserved across different environments.

And then they showed that your auditory system knows about the way reverb works, in the sense that if you make up a different, non-physical reverb property and you play it to people, it sounds weird, number one. And two, they can't recover the sound source. And what that means is, built into your auditory system is knowledge of the physics of sound, and in particular about the particulars of the decay function of reverb, such that you can use that knowledge of how reverb works in general to undo this problem, and constrain this problem, and solve for the sound source.

I didn't give you all the details. But I want you to get the gist. Do you get the kind of idea? OK.

But as I said, that's only true for reverb that has the reverb properties of real-world sound. If you make up fake reverb, it doesn't work. And people can't solve this problem.

That tells you they're using their knowledge. Doesn't tell us whether that knowledge is built in innately, or whether they learned it, or what. All right, good. So in other words, we solve the ill-posed problem of recovering the sound source despite reverb by building a knowledge of the physics of the world into our auditory system and using it to constrain the problem.

So we just said, why is this computationally challenging? Invariance problems, appreciating the sameness of a voice across different words, appreciating the sameness of a word across different voices. Separating multiple sound sources that come in simultaneously and are just massively superimposed on the input-- the cocktail party problem, also ill-posed-- and the reverb problem. So everybody see how these are three really big challenges for audition? Yeah.

STUDENT: So was brain imaging as well a part of the [INAUDIBLE]?

NANCY KANWISHER: Nope. One could do that and ask questions about where that's solved in the brain. But the beauty of that study is that in a way, who cares where it's solved? I mean, it's kind of interesting, but it's such a beautiful story already just from actually, a big part of their study was measuring reverb.

Nobody had done it before. They sent people out with speakers, and recording devices, and little random timers on their iPhones. And at random times-- how did this go-- oh, yeah, they had people had to mark the location they were in using their iPhone GPS and then that's right-- they didn't send people out with recording devices. It's too hard.

And so then they sampled what kind of places do people hang out in. And then they went back with their impulse sound source and the recording device, and they measured that impulse response function in lots and lots of different natural sounds in order to characterize what is the nature of reverb in the world. Nobody had done that before.

So that's why I tell you this, is that to me, it's just one of the most beautiful examples of computational theory-- no measurement in the brain. A big part of the study was just characterizing the physics of sound, and then some psychophysics to say actually, do people use that knowledge of how reverb works in the world? And yes, they do.

So I've been talking about hearing in general, but let's talk about one of the most interesting examples of hearing, the one you're doing right now-- speech perception. So what do speech sounds look like? You saw a few of them briefly before.

Here are a few spectra. So just to remind you, each one of these things has time going along the x-axis, frequency here. And the color shows you the intensity of energy at that frequency band.

So this is a person saying, "hot," and "hat," and "hit," and "head." That's the same person saying these four things, a person with a high-pitched voice. And here's a person with a slightly lower-pitched voice saying the same things.

So what do we notice here? Well, first of all, we see that vowels have regularly spaced harmonics. That's the red stripes. This is a vowel sound right there.

See those perfectly regularly spaced harmonics? That makes a pitchy sound, so voices are pitchy. You may not think that there's a pitch to my voice right now because I'm talking, not singing, but there is a pitch. And you use that, actually, in the intonation of speech, as you guys read about in the assigned reading for yesterday.

So each of these things with the stacked harmonics is a vowel sound. It's got a pitch and it lasts over a chunk of time. And the consonants are these kind of muckier things that happen before and after. And consonants don't have pitch. They don't have harmonics. They have kind of muck.

So there are certain band-- people who study speech spend a lot of time staring at these things and characterizing them. And they like to talk about bands of frequency, of power. And so this band down here that's present in all of these speech sounds here is called a "formant." It's just a chunk of the frequency spectrum that you hear with speech. So that's a formant.

And some of those frequency bands or formants are particularly diagnostic for different vowels. So if you look in this range here, only in that mid-range here, only for "hat" and a little bit for "F" sound do you get an energy in that frequency band, not for "hot" or "hit." And that's true both for the high-pitched voice and the low-pitched voice. This frequency band here is really diagnostic to which of those vowels you're hearing.

So we're going to play with that spectrogram again a little bit more, although I now have learned avoidance. So this is me speaking again, as you saw before. So I'm going to say an A, an E, an I-- look how different that one is, O, and U. And there's lots of other vowels. Do you see how that energy moves around for the different vowels?

Now as I said before, if I do a long vowel like this, it makes a big, long bunch of harmonics. But a lot of the time, they're just these vertical lines. The vertical lines are consonants, t, p, k, r. If I don't say a vowel, you just see a vertical line. It's not quite a vertical line. They are different from each other in ways you can tell.

So the consonants are those bands of energy that go vertically. And the vowels are the big, long harmonic structures that stretch between them.

Now, I'm not sure you'll be able to do this. I'm going to need a volunteer in a second, and I'm going to pick on [? Iadun, ?] because he's most accessible right there. So come on up here. You know it won't be horrible or embarrassing. So you can stand here for a second.

I'm going to say "ba's" and "pa's" and I'll tell you in a moment what to say. I'm not sure this is going to work. I tried it before. We're going to look at two different formants when I say "ba."

Actually, I'm going to do it rising-- ba, pa, ba, pa. So there's two different formants, here and here, with both of those. I'm going to do it again.

And there's just a tiny, little difference between a ba and a pa. And it has to do with the interval between the consonant, which is the first vertical thing, and the vowel, which is the horizontal stuff. So let's see if we can see it again. Here we go.

Ba, pa-- do you see how the pa starts earlier there and the ba is slightly delayed? I'll show you diagrams that show you more clearly. OK, great. Don't go away.

So we're going to do the cocktail party thing with the recording devices here. What is this? This is just some boring administrative thing you can just read. I actually brought it to crumple and make a crumpling sound, but we'll do that afterwards.

Right now, you will read from that and I will recite something boring. And we'll just do it simultaneously. So just focus on what you're doing.

And everybody watch. You can see my voice here. And let's see what happens when we're both talking at once. OK, here we go.

Four score and seven years ago-- oh, geez, I forget how it goes after that, so I'll just have to make up some other random garbage.

[INTERPOSING VOICES]

STUDENT: --outstanding--

NANCY KANWISHER: OK.

STUDENT: --review the student's course--

[INTERPOSING VOICES]

NANCY KANWISHER: That's great. That's great. Don't go away. I don't know if you could tell that it got muckier when we were both talking. Maybe it's mucky enough with me talking fast to begin with.

Let's try a few other things. Let's have me say words and you say words. And let's see how different they look. OK, so I'm going to say "mousetrap."

STUDENT: Mousetrap.

NANCY KANWISHER: You can see some similarity there, can't you? Let's do it again. Mousetrap.

STUDENT: Mousetrap.

NANCY KANWISHER: OK, that's good. It's funny, I see more low-frequency band here. I'm sure your voice is lower than mine.

Pitch, interestingly, isn't just about how low the energy goes. It's an interesting, complicated property of the lowest common denominator of that whole frequency stack. So I'm not going to do pitch. It's complicated.

What else do we want to do? Let's try some ba's and pa's. But let's stick them on the fronts of words. Maybe that'll work better-- pat, bat.

STUDENT: Pat, bat.

NANCY KANWISHER: Oh, I could see the commonality there. Could you guys see that? Let's do it again. Pat, bat.

STUDENT: Pat, bat.

NANCY KANWISHER: Well, yours look more similar. All right. Anyway, thank you. That's good. That's all I need, just to show you how hard this is, and how there's variability across speakers saying the same thing, and very, very subtle differences between sounds that sound totally different to us.

So back to lecture. So you saw the harmonics, those red stripes, during the vowels. You noticed that I showed the consonants and the ba's and pa's.

So here's a diagram. I'm sorry, this is very abstracted away from those spectrograms, which are messy, as you can see. The idea is that a consonant vowel sound, a single syllable like ba or pa-- this is time this way-- has this big, long formant which is a band of energy that's the vowel, the ah sound. And it's these transitions that happen just before that that make the difference for different consonants.

And in particular, the difference between a ba and a pa-- this is a ba, that's a pa-- the difference we were looking for that didn't show up that clearly, but you can try it at home, maybe you can get it clearer than I just got it now-- has to do with that transition onto the first formant. So with a ba, the transitions happen in parallel. And with a pa, this transition happens before that lower formant. So that tiny, little-- it's a 65 millisecond delay in the case of pa that you don't have in the case of ba, is how you tell that difference. It's very, very subtle.

So there's lots of different kinds of phonemes. We've been talking about vowels and consonants. Each vowel or consonant sound is called a "phoneme" if a distinction in that sound makes the difference between two different words in your language.

And that means that what counts as a phoneme in one language may not be a phoneme in another language, because it won't make a distinction between different words. Many of the phonemes are shared across languages, but not all. We've talked about R and L that aren't distinguished in Japan, and two different D sounds that sound the same to me that are distinguished in Hindi, and lots of others.

And so those are just variations across natural languages on which of those phonemes, which of those sounds, are used to discriminate different words, and hence count as phonemes in that language. So there's some particularly awesome phonemes that use a particular kind of consonant known as a click consonant. And these are common in some Southern African languages.

And a year ago, I was traveling in Mozambique, which was just hit by a devastating flood. It's really awful. But anyway, I was there visiting a game park seeing all kinds of animals.

And I met this guy, Test. And he's amazing. I mean, his knowledge of the natural history was mind blowing, but he also speaks, I think, six different languages fluently, one of which is Xhosa, or as he would say, [SPEAKING Xhosa] or something like that. You'll hear him say it in a moment.

And so he was illustrating click languages. And I'll play this for you in a second. And he says there's a sentence in Xhosa which is a little bit crazy, but has all the different clicks.

And it means, basically, "the skunk was rolling and accidentally got cut by the throat." Doesn't mean a whole lot, but listen to Test saying the sentence, first in English and then in Xhosa.

[AUDIO PLAYBACK]

- The phrase in English, it says skunk was rolling and accidentally got cut by the throat. In Xhosa, or in east Xhosa, [SPEAKING Xhosa].

[END PLAYBACK]

NANCY KANWISHER: Isn't that awesome? I think we just have to crank it up a little bit and hear him again.

[AUDIO PLAYBACK]

- The phrase in English, it says the skunk was rolling and accidentally get cut by the throat. In Xhosa or in east Xhosa, [SPEAKING Xhosa].

[END PLAYBACK]

NANCY KANWISHER: OK, for the most part, we don't have click consonants in English that count as phonemes in the sense of distinguishing different words. But we do have click consonants that we use in other domains. Anybody know what we use click consonants for? There's at least two.

Know any click consonants?

STUDENT: [INAUDIBLE]

NANCY KANWISHER: Yeah, what?

STUDENT: [INAUDIBLE] That's--

NANCY KANWISHER: Like what?

STUDENT: [INAUDIBLE]

NANCY KANWISHER: Yes, but that's a regular consonant. It's actually not a click. It's just a regular consonant. Well, one is when you go, tsk, tsk, tsk, the scolding sound. It's not a phoneme. It's not a word, but it has a very particular meaning.

Another one is how you get a horse to giddy up. (CLICKS) So those are the click consonants we have in English. They're not phonemes, but we have them, and he's got a whole lot more. That was just for fun.

So why is speech perception challenging? Well, one is the essence of it is that a given speech sound is highly variable. One way it's variable is that when you speak at different rates, all the frequencies go up and down and haywire, making them very different across different talking rates.

Another is the context. So a given phoneme, like a ba, or a pa sound, or a vowel, sounds totally different depending on what phonemes come before and after it. They're not little punctate, one at a time things. They all overlap and affect each other in a big mess.

And the third is one we've already mentioned, which is the big differences across speakers in the language. So you have to recognize a ba sound even though it sounds quite different when spoken by different speakers. So all of these things make it very computationally challenging to understand speech.

Here's an illustration of that talker variability. So what's shown here is not a whole spectrogram, but just the intensity of the first formant and the second formant, those bands of energy that I showed you in the spectrogram. And so each dot here is a different person pronouncing a vowel.

And each color-- this is one vowel here in green, in that green ellipse, with lots of different people saying that vowel. Here's another vowel up here in red, with lots of different people saying that vowel. And what you see is they're really overlapping.

So that means you can't just go from the energy at those two formants, a point in that space, and know what the vowel is. What if you were right there? Well, then it could be any of four different vowels. So that's the problem of talker variability illustrated with vowels. Does that make sense?

I think I just said all of this, blah, blah, blah-- another classic ill-posed problem in perception. You're given a point in this space. How do you tell which vowel it is?

So one way we solve that is that we learn each other's voices. And we know how a given person pronounces a given set of vowels or words. And we use that to constrain what they're saying.

Have you ever noticed, especially if you meet somebody new-- well, actually, you just experience this with Test. When he first speaks, his English is beautiful, but he's from Zimbabwe and he has kind of Zimbabwe, British-type accent. And at first it's hard to understand what he's saying.

Did you all experience that briefly? I mean, that's why I put the text on the slide, so you would get used to his English and understand it. If I hadn't, you probably wouldn't have understood that sentence he spoke first.

That's because we don't know his voice yet. But did you notice, after even just a few words, you start to like tune right in and you can understand him? So learning about an individual's voice helps you pull apart the properties of the voice, and unconfound them from the sound so you can understand what that person is saying.

So that's part of how we solve this ill-posed problem. And so evidence that we do that is that if you have people listen to voices they don't know or voices that are changing from word to word, it's much harder to understand speech. So you imagine you took the sentence I'm saying right now, and you spliced in a different person saying each word.

Actually, I should make that demo. One of you guys send me an email-- make that demo of a different person speaking each word in a sentence. It'd be really hard to understand.

Because you wouldn't have been able to fix this was a property of the voice, now we can kind of separate that from everything else. Because the damn voice will be changing on each word. It'll be a mess. So that's one problem.

So it turns out that the opposite is true, as well. And that is, your ability to recognize somebody's voice is a function of what you know about that language. So you can recognize voices better in a language you know than a language you don't because you're doing the opposite. You're using knowledge of the language and its speech properties that you already know to constrain the problem of figuring out who is this person's voice.

So does everybody get this? These two things are affecting each other-- the speaker and what's being said. And because they're so confounded, massively confounded in the stimulus, to solve that, the more you know about the speaker, the better you can understand what's being said. And the more you know about the language and its properties, the more you can recognize the voice. Each one is a source of information about one of those two confounded variables.

And so people have shown that psychophysically. And I think I have time to do this. Here's a kind of cool corollary of this, and that is, it's commonly thought that dyslexia is most fundamentally a problem of auditory speech perception, not a visual problem.

There may also be a bit of a visual problem, but it's thought that at core, it's a problem of auditory speech perception. So if that's true, then you might think that this ability to use knowledge of the language and its sounds to constrain voice recognition would be reduced in people with dyslexia, because they are less good at processing speech sounds. And it turns out that's true.

So here's a beautiful study from Gabrielli Lab a few years ago. So first look at the bars in blue. So this is accuracy at voice recognition, which person is speaking.

And this is native English speakers who don't speak Chinese. They are much more accurate recognizing who's speaking when they're speaking English than when they're speaking Chinese. So that's kind of cool. That shows you the way in which you use knowledge of the language to constrain recognition of the voice.

But now look what happens in the dyslexics-- no effect, exactly as they predicted. Given that the dyslexics have a problem with speech perception, they're apparently not able to use that knowledge of the phonemes of the language to constrain the problem of voice recognition. They're just as bad at voice recognition-- I'm sorry, they're no better at voice recognition in their native language than in a foreign language.

They can't use that knowledge to constrain voice recognition. Does that make sense? Yeah, I love that study.

So we haven't done any brain stuff so far. We were just thinking about the problem of hearing and speech perception, and what we know from behavior. And we've learned a lot already, but we'll learn more by looking at the brain, and the meat, and all of that.

So let's start with the ear. Again, remember, compressions of air come into the ear. They travel through the ear canal. They hit the tympanic membrane.

They go through a whole series of transducers, these three little ear bones here that connect to this snail-shaped thing, which is called the "cochlea." Cochlea is really important. You should remember that word. It's the place where you transduce incoming sound into neural impulses, way in there.

And the cochlea is really cool. It's this, as I said, a snail-shaped thing. And there are nerve endings all the way along this thing. And because of the physics of the cochlea, there are different resonant frequencies at different parts of this snail.

So basically, here are some low-frequency sound waves. This is the cochlea stretched out with the base and the apex. This is the base. That's the apex.

And what you see is the low frequencies have transduced some energy at the base of the cochlea, and also at the apex. But midway range frequencies and high frequencies do nothing at the apex. This business, there's only physical fluctuations happening up here for low frequency sounds.

So there's little nerve endings here that detect those fluctuations up there and send those signals up into the brain through the auditory nerve. And so in the middle, here or something, you have sensitivity to mid-range frequencies, not high or low. And at the base, it's sensitive more to high frequencies than mid or low. So everybody get that?

So basically, the cochlea is doing a Fourier transform on the acoustic signal. It's taking these compressions of air, and it's just saying, let's separate those out into different frequencies, just with this physical device. It's like a physical Fourier transform that's saying, let's just physically separate the energy at each frequency range along the length of the cochlea. Does that make sense?

And then once you get different parts of the cochlea that are sensitive to different frequencies oscillating to different degrees, then you stick some nerve cells there to pick up those oscillations, go up the auditory nerve, and travel into the brain. Everybody have a gist of how this works? So that's cool.

But now, let's go up to the brain. So now, this is a view like this. And so here are the cochleae-- I guess that's the plural-- on each side-- ears, ear canal, cochleae.

And the first thing to know, which is important, is that the path between the cochlea and the first step up in the cortex is much more complicated in hearing than it is in vision. Look at all these nuclei deep down in the basement of the brain. In contrast, in vision, how many synapses do you have to make between the retina and primary visual cortex? Sorry. One synapse. Right?

STUDENT: Well, I was thinking--

NANCY KANWISHER: Yeah, two, that's right, so retinal ganglion cells send their axons straight into the LGN in the thalamus, make a synapse. And then those LGN neurons go straight up to primary visual cortex, just one stop on the way. Look at all the stops on the way here.

So audition is a really different beast from hearing in many ways. Next time, we'll talk about how audition-- not these parts of it, but after you get up to the cortex-- audition, we in my lab and a few other labs are really starting to suspect, is profoundly different in humans from any non-human animal. And I think that's for very interesting reasons, but this part is pretty similar in animals, just getting information up to the cortex. And audition is already very different from vision just in the number of relays going up to the brain.

So those structures down there do all kinds of awesome things. And last year, I talked at great length about how we detect the locations of sounds. It's absolutely beautiful work, and elegant, and fun, but I decided that was a little too much behavior. We should get on to the brain.

But I recommend 9.35 if you want to learn more about audition-- awesome course. Did you take it? Really awesome course. Yeah, exactly. And so you'll learn more about all that stuff.

So instead, we will just skip all that and go straight up to cortex. So the first place that auditory information hits the cortex coming up from the cochleae is primary auditory cortex, just like the first place visual information hits the cortex coming up from the eyes is primary visual cortex. So you can see in here that in a cross-sectional view like that, this is primary auditory cortex.

It's in that sulcus right there. That's kind of a drag, because when we get occasional opportunities to test patients who have grids of electrodes on the surface of their brain, the grids don't usually go in there and we can't see primary auditory cortex. Although there are new methods where they stick depth electrodes, which is surprisingly, apparently, better on the patients.

And right now your TA, Dana [? Bobinger, ?] is over at Children's Hospital recording from a 19-year-old who has bad epilepsy and who has depth electrodes in his brain. And he's listening to all kinds of sounds. And she's recorded his neural activity with depth electrodes. And so we are hopeful, one, that we can find some information that will be relevant to the neurosurgeons-- I don't know about that-- but two, that we'll get some information from those deep structures that you can't usually see when you have just grids sitting on the surface.

So back to functional MRI-- so this is primary auditory cortex. It's quite stylized. Let me remind you where you are.

This is an inflated view of the right hemisphere-- back of the head, front of the head, temporal lobe, all funny looking because it's been mathematically unfolded so you can see stuff in the sulcus where I just showed you. Primary auditory cortex is in the sulcus. But we've inflated it so you can see it.

And so this is primary auditory cortex, this whole thing here. And it shows you a property we've talked about before. It's got a map, but the map in primary auditory cortex is not a map of space like it is in the retina for visual information. It's a map of frequency.

And that makes sense because the input transducer is a cochlea, which already physically creates a map of frequency. And so that gets traveled through all those intermediate stages down in the basement, and it comes up to the brain, and makes a map of frequency space. So what this means, actually-- so here's sensitivity to different frequencies.

And so the classic structure of primary auditory cortex in humans is high, low, high-- high frequencies, low frequencies, high frequencies, in that V-shaped pattern. So this is the right hemisphere. This is the left hemisphere that's been mirror flipped so you can compare them directly.

And you can see this highly stereotyped pattern of high, low, high. That's a tonotopic map. Everybody clear on what a tonotopic map is? And we've just discretized it into two chunks, but it's actually a gradient of high to low to high, which you can kind of see by those intermediate colors in there. Yeah.

STUDENT: [INAUDIBLE] why does the [INAUDIBLE]?

NANCY KANWISHER: Yeah, everything in the brain rearranges everything in the input in multiple ways. So we didn't talk about this, but in visual cortex, you have-- I don't know what the latest count is, at least 10, probably more than that, separate retinotopic maps in different patches of cortex-- map, map, map, map, loads of them. And so there's all kinds of transformations.

And so much less is known about the functional responses and functional organization of auditory cortex than visual cortex, especially in humans where we really don't know a lot, in fact. So there's no real answer to that, other than it's not that shocking, in a way, because you see that in vision and in other domains anyway, with multiple maps that differentially represent different parts of space.

And so yeah, I didn't say this, but many of those dozen or so maps in visual cortex have differential representation of different parts of space. Some focus on the upper visual field, some on the lower visual field. And the whole question of is that really one thing or is it two-- this is all now getting into the kind of cutting-edge, ambiguous state that we don't know. All right, everybody clear on tonotopy, primary auditory cortex? OK, good.

All right, the standard view from recording neurons in primary auditory cortex in animals-- monkeys, ferrets are big in auditory neuroscience, other animals-- is that the receptive fields of individual neurons in primary auditory cortex are linear filters in the following sense-- so here's a spectrogram of a sound. This is just a description of the stimulus. As usual, time, frequency.

So it looks like it could be a speech sound with some vowels there. Or it might be something else. Who knows. So that's a sound.

So now, imagine an electrode sitting next to a single neuron in primary auditory cortex in, say, a ferret listening to that sound, and characterizing what does that neuron respond to. Well, the typical finding is that neurons in primary auditory cortex are what's known as spectral temporal receptive fields, or STRFs to their friends. So what does that mean?

Here's an example of the receptive field that is the response dependence of a given auditory cell, again, with time on this axis and frequency on that axis. So what kind of sound does that cell like? Can you see just by looking at this? What kind of sound?

STUDENT: Increasing frequency.

NANCY KANWISHER: Increasing frequency, yeah, something like that right. Here's one that also likes increasing frequency, but slower, shallower increasing frequency. Here's one that likes decreasing frequency.

Now, you may be wondering what the stripes are. We didn't talk about this in visual cortex, but this is a common property, that it likes this particular set of frequencies here, but is inhibited by adjacent frequencies. So you also see something like that with orientation tuning in primary visual cortex.

And so here, these ones are changing faster, both increasing and decreasing. So the idea is primary auditory cortex in animals, and presumably in humans, is full of a bunch of cells that are basically spectrotemporal filters like this. They are picking out changes in frequency over time that happen to different degrees, and at different rates, and in different frequency ranges. Does that make sense, more or less? Yes, [INAUDIBLE]

STUDENT: I have a question.

NANCY KANWISHER: Yeah.

STUDENT: [INAUDIBLE] how would you tell that was [INAUDIBLE]?

NANCY KANWISHER: Yeah, how do they figure that out? I usually spend all this time talking about the design of the experiment. I just skipped straight to the answer here.

Well, I don't know exactly what you do, but you probably-- I mean, this has been a whole thing that went on for decades for people to get at this. So I'm guessing that somehow, they got into that general space, and then they generated stimuli that make all these different sounds. And they just run through them, and they find, for a given cell, you play all these different sounds. You go-- [MAKES SOUNDS], et cetera.

I'll spare you more imitations. You play all these different sounds to the animal and you record the response of that neuron. And you would find, for example, that it responds much more when you play that sound than any of the others. Does that make sense?

STUDENT: No, it makes sense.

NANCY KANWISHER: But how do they ever hit on that?

STUDENT: No, what I was asking is that are they using separate [INAUDIBLE]?

NANCY KANWISHER: Oh, the red and the blue? How exactly they got-- rather than just the simple thing with just that-- how exactly they arrived on that, I'm not totally sure. I mean, there are mathematical reasons why it makes sense to have that whole thing rather than just a single stripe, that I think are beyond the scope of this lecture for the moment. But anyway, it wasn't just a totally arbitrary thing to try. Those are particularly useful kind of receptive fields for representing the input.

So everybody sort of clear, approximately, what this idea is? So it's very low-level basic, just are the frequencies going up or down, and which range, and how fast? That's what primary auditory cortex does organized in this map, this tonotopic map.

So think of primary auditory cortex as just this bank, this big set of linear filters for particular frequency changes over time. So that's all based on data from animals, from recording individual neurons. But we want to know about humans, not just because that's what this course is about, but we want to know about humans. I mean, ferrets are nice, but really!

So is that true for humans. Well, Josh McDermott and Sam Norman-Haignere just published a paper a few months ago in which they addressed this question in a really interesting way. So here's the logic-- this is a little bit technical. I'm trying to give you the gist. I hope it works. Give it a try.

So they generated synthetically, computationally, what they call "model-matched stimuli." So the idea is this-- the idea is if you present a natural sound-- like a dog barking, or a person speaking, or a toilet flushing, just some sound that you would hear in life-- and then what they do is they make a synthetic signal that matches that sound with respect to those STRFs I just showed you. That is, if you fed the original sound and you fed this synthetic sound into the STRFs, you'd get the same thing in the STRFs.

So this is a way of saying, we're assuming that those STRFs are a good description of what goes on in A1, so let's test that by taking a big, fancy, real-world sound that has meaning and people know what it is, and let's make a control sound that matches the in-STRF properties. And let's see if we get the same response in the brain in that region. If that model is a good description of what that region does, then you should get a very similar response when you give the synthetic sound and the original sound that you recorded in the world.

So they tested this on a STRF-like model, like this thing I just described before. And so just to show you what these sounds are like-- so here's an original sound just recorded in the world of somebody typing.

[AUDIO PLAYBACK]

[TYPING]

[END PLAYBACK]

OK, OK, OK, that's enough. I know it's riveting, but so then they run that through their STRF model. They get a STRF description and they generate a matched stimulus from their STRF description. And it sounds like this.

[AUDIO PLAYBACK]

[TYPING]

Pretty good. It's kind of hard to tell them apart. Sorry, enough.

[END PLAYBACK]

All right. And you can see their spectrograms are really similar. So for a textury thing like typing, it really captures the essence of what's being heard. We're just telling you what these control stimuli sound like.

Let's take another sound, a person walking in heels. And you can see all those verticals. Those are the clicks.

Clicks have energy across lots of different frequencies. And that's what a vertical line means-- it means all those different-- remember, this is frequency on this axis. So a vertical line means energy at lots of different frequencies not organized in harmonics, so it's not pitchy. Here we go.

[AUDIO PLAYBACK]

[HEELS CLICKING]

[END PLAYBACK]

OK, here's the STRF version, the control stimulus.

[AUDIO PLAYBACK]

[CLICKING]

[END PLAYBACK]

So it captures some of it, but not all of it. It captures the sound of each click, but not the spacing between. So it's getting the local properties, but not all of the properties. Yeah.

STUDENT: How did you say-- like just the [INAUDIBLE]?

NANCY KANWISHER: How do they make it? I didn't tell you because it's complicated. They basically start with pink noise, or white noise, or some kind of noise. They run it through their STRF thing.

They run the original sound through the STRF thing. They compare them. And they say, how are we going to adjust the noise to make it more like that? And they just iterate a lot, and they end up with these stimuli.

And you can see just looking at it, they ended up with something that's pretty similar in terms of the spectrogram. Let's listen to a person speaking. Here's the original sound.

[AUDIO PLAYBACK]

- Is that art offers a time warp to the past, as well as insight.

[END PLAYBACK]

NANCY KANWISHER: OK, now I'm going to turn it off. Here's the synthetic version.

[AUDIO PLAYBACK]

[INAUDIBLE]

[END PLAYBACK]

OK, now we've lost something. So does everybody see how with keyboard typing, it really sounds the same, the synthetic version? With walking in heels, kind of, sort of, at least locally, but not globally, and with speech, we've just totally lost it.

The stuff that you can capture with a STRF model does not capture the full richness of speech. There's something more in a speech stimulus than you can capture with that just simple STRF model. OK, let's listen to a violin.

[AUDIO PLAYBACK]

[MUSIC PLAYING]

[END PLAYBACK]

OK, what does the STRF model do with that?

[AUDIO PLAYBACK]

[MUDDY MUSIC PLAYING]

[END PLAYBACK]

I love that. It sounds like a sea lion colony. Anyway, so what you see is the STRF model totally fails to capture speech and music, but it captures textury sounds like that. And it loses some of the broader temporal scale information.

So that's the stimuli. Then you scan people listening to these sounds. Just pop them in the scanner and play those sounds.

And so then what they do is they just ask. So this is, again, the white outline is primary auditory cortex where you have that frequency map, mapped in a separate experiment, and just plunk down on the brain here. We're zooming in on that part of the top of the temporal lobe.

And so what's shown here is, for each voxel, they're showing the correlation of the response of that voxel to the original sound and the synthetic, STRF-y sound. And what you see is those correlations are really high in primary auditory cortex. In other words, primary auditory cortex responds pretty much the same to the original sound and the synthetic sound. It doesn't detect that difference.

But as soon as you get outside of primary auditory cortex, you get something totally different. And so that was exactly the prediction, is that model that's being tested here is a model of how they thought primary auditory cortex worked-- a bank of linear filters. They test that model by generating a new set of stimuli that are matched for those linear filters, and they get pretty much the same response in primary auditory cortex.

So check-- that's a good model of primary auditory cortex. But also, the blue shows you much lower correlation out here. It is not a good model of stuff outside of auditory cortex. Josh.

STUDENT: So isn't this kind of self-fulfilling, in the sense that I build my synthetic stimuli based on these kind of models, and then--

NANCY KANWISHER: It is, except the models were all based on animal work and this is human brains. So this is a way-- but that's exactly right. It's a way of saying all this work from animals precisely characterizing response properties of individual neurons, which you can do in animals and mostly not in humans, do we think that's true of human primary auditory cortex?

And yes, it is. Does everybody get at least the gist of that? I realize I skipped over lots of details because I want you to get the general picture. Yeah.

STUDENT: What are they trying to achieve by doing this type of [INAUDIBLE]? I mean, the hypothesis is that the human and the animal auditory cortex is the same?

NANCY KANWISHER: Primary auditory cortex, yes. Yes. They're basically testing-- you derive that model from the animal work, then you design a test of it, which is making those synthetic stimuli. And I left this out because actually, I don't think they've done that, but presumably, if you test those stimuli with single units in ferrets, you get the same thing. You get very, very similar responses in primary auditory cortex to the original sound and the synthetic version of it based on the STRF model.

STUDENT: It's predicated on the assumption that both of them are structurally the same.

NANCY KANWISHER: Well, it's testing. It's asking that question. It's asking that question.

Because I've occasionally in here lamented about how crappy our methods are in human cognitive neuroscience. I mean, they're fun. We can do something, but we hit a wall pretty fast.

We want to see the actual neural code. We don't have spatial and temporal resolution at the same time. We pretty much only get that in animals.

We can pretty much only do really careful causal tests in animals. We can pretty much only see connectivity in a precise way. And all these things we can do only in animals.

And so we need to know if those animal models are good models for humans. And this is a way to test it. And it passed with flying colors. Make sense?

So primary auditory cortex seems in humans that it's much like it is in ferrets, a bank of linear filters with STRF-y properties. What about everything else? After all, you guys can hear the difference between the original version and the synthetic version of the woman talking and the violin. And if I played you all the other stimuli of real-world sounds, you could hear the differences in many of the other ones as well.

So what are you doing? Well, there's lots of auditory cortex beyond primary auditory cortex that could represent that difference. And what this is suggesting is, whatever's going on out here is doing something really different with those sounds.

It is not fooled. It does not think the synthetic thing is the same thing as the original thing. That's what the low correlation means.

So I'll tell you about just one little patch of cortex out there. And that is-- again, this is just for reference. We've zoomed in again on this is the little code for separate mapping of high, low, high, primary auditory cortex right there. And what the yellow bands are is selective responses to speech. So you compare a whole bunch of speech sounds to a whole bunch of non-speech sounds, and you get a band of activation right below primary auditory cortex. Yes.

STUDENT: I thought the separation was low, high, medium [INAUDIBLE].

NANCY KANWISHER: High, low, high-- I probably said it backwards. That would be like me. But it's-- wait, wait. What the hell is it?

I'm pretty sure it's high, low, high. Let's go back and look. I might have screwed it up on the slide or said it backwards, but I'm pretty sure it's high, low, high.

STUDENT: So the low frequency is the [INAUDIBLE].

NANCY KANWISHER: Yeah, just like that's the code for frequency, right there. But ask me those questions because I'm very capable of getting things backwards, as you've probably already noticed. So there is a band of speech-selective cortex just outside of primary auditory cortex, in that region that we just saw responds differently to the original sound and the model-matched synthetic sound.

So that's pretty cool. What do I mean by "speech-selective cortex?" What I mean is-- this is some of our data. I tried to find you someone else's data and I went down a 45-minute rabbit hole trying to find a nice slide. And I just couldn't find a good picture.

I finally said, screw it, I'll show you my data, even though I'm trying to-- we're not the only ones who've shown this. We just have the best data. Other people had tested it with four, five, six conditions. We tested it with 165 sounds.

So this is the magnitude of response in that yellow region to 165 different sounds, color coded by condition shown down here. And so what you see if you look at it is all the top sounds are light green and dark green. Speech-- notice, importantly, that the response is very similar to English speech and foreign speech which our subjects do not understand.

So that tells us that this is not about language. This is not about the meaning of a sentence, or syntax, or any of that stuff. This is about phonemes, the difference between a ba and a pa, which you can do on a foreign language, even if there's a few phonemes that are different. You get most of them.

Does everybody get the difference between speech and language? Amazingly, the senior author of the paper you read for last night does not understand that difference. He published a beautiful paper. Every time he comes here to speak, he talks about language, language, language, language.

And I say, Eddie, have you ever presented a stimulus that's in a foreign language? He's, like, oh, no, that'd be really interesting. It's like, Eddie, until you do that, you don't know if you're studying language or speech. Oh, yeah, really interesting.

And then he comes back four years later and he doesn't seem to know the difference between language and speech. I'm, like, hello. Anyway, he does beautiful experiments, but it's just-- it's a blind spot, or it's a misuse of a word. I don't know what it is, but it drives me nuts. Can you tell? Anyway, you guys get that difference even if Eddie doesn't.

Let's look at some other things. How about all this light blue stuff? There's a lot of light blue stuff that's almost as high.

Oh, that's music with people singing. That also has speech. The speech is slightly less intelligible because it's singing, and there's background instrumental music, so it's a little bit lower.

Oh, what's next? We've got some light purple stuff and some dark purple stuff. This is non-speech vocalizations. That's stuff like laughing, and crying, and sighing-- pretty similar to speech but not speech. It's the next highest thing, but it's well down from the speech sounds.

And then we have dogs barking, and geese, and stuff like that, that are yet further down. And then we have all kinds of other stuff down there-- sirens, and toilets, and stuff like that. Yeah.

STUDENT: Is instrumental music perceived as speech? I mean, I can't make out the colors.

NANCY KANWISHER: No. The instrumental music is way down in here. Yeah, it's a little hard to see. That stuff up there is non-speech vocalizations. It's not a perfect slide. So that's pretty strong evidence that that band of cortex is pretty selective for speech. Everybody get that? Yeah.

STUDENT: So you're saying it's not like it doesn't process like the other one, so the violin stuff would still be that [INAUDIBLE]

NANCY KANWISHER: Yeah, right. OK, good point. Remember when I first showed you the fusiform face area, I showed you that time where it's faces are like this, staring at dot is like that, looking at objects is like this. So I said, OK, there's a little bit of a response to things that aren't faces. It's just much more to faces.

Now, you guys may not have noticed this because it went by kind of fast, but when I showed you intracranial data from the fusiform face area in that patient who got stimulated there, and saw the illusory faces, the intracranial data showed zero response to things that are not faces. So I think that that's because functional MRI is the best we have in spatial resolution in the human brain, except when we have intracranial data.

But it's still blurry. It's blurry because there's blood flow and all of that. So I would guess the same thing here. In fact, I guess it isn't in the paper you read because he didn't have any non-speech sounds, but I will show you. Dana's recording them right now at Children's Hospital, and we have some other ones that I will show you next time, of intracranial electrodes. And they will be even more selective than that. But this is pretty good already. Yeah, Nava.

STUDENT: What's the human non-vocal?

NANCY KANWISHER: I didn't hear. What?

STUDENT: The human non-vocal?

NANCY KANWISHER: Oh, that's like clapping, and footsteps, and I forget what else, things where you hear it and you know that's a person, but it doesn't sound at all like speaking or speech. So if it was about the meaning, it could have been all about the meaning of people, could be something telling you there's a person there. Deal with it. But no, apparently not.

So we're not the first ones to see this. We've just tested it with more conditions. So our evidence for selectivity is stronger than everyone else's.

Given what I've told you today, can you think of a stronger way to test this? For example, suppose I was worried, maybe the frequency composition of the speech is different than the non-speech. After all, those are just recordings of natural sounds in the world that we went out and made, or mostly got off the web, someone else made.

And maybe they differ in really low-level properties. And so how do we know that that's really speech selectivity, not just selectivity for certain frequencies or frequency changes? Yes.

STUDENT: You could run it with the McDermott generate--

NANCY KANWISHER: Bingo, absolutely. Everybody get that? So then we'd know, because those are beautifully designed to match all those acoustic properties, match the spectrogram for all those lower level properties. And McDermott and Norman-Haigenere have done that. And this region does not respond strongly to the model-matched version, so it's not just the acoustic properties. Yeah.

STUDENT: Can we also do something like [INAUDIBLE] play speech backwards?

NANCY KANWISHER: Yes, people have done that, too. It's a little bit complicated, because speech backward sounds a lot like speech. It's kind of in the intermediate zone.

So it balances many things, but one, it doesn't balance all the acoustic properties. So speech has certain onset properties. I forget how it goes, but if you play it backwards, there's lots of-- [MAKING SOUNDS] You've heard backward speech played, right? And so the STRF model would respond differently to forward and backward speech, whereas the STRF model responds the same to the original and the synthetic speech. Make sense?

So there's a very speech-selective patch of cortex. And it's speech selective, not language selective. And of course, we want to know-- speech is lots of different things.

It's what words you're saying. It's who's saying it. It's your intonation-- are you making a statement, or a question, or what are you emphasizing in the sentence? And it's lots of other things.

And the paper you read asked that question. What's coded here about speech? And so I made a whole bunch of slides to explain what the paper said because I thought people would have trouble with it. And everyone nailed it, so I'm not even going to go through them. Maybe I'll just show one in closing.

So one thing a few of you got wrong-- and I totally get why, it didn't matter-- is that here is this is one patient, and this is the bank of electrodes placed on the surface of the brain. The red bits are the bits where you could account for the neural responses in terms of any of those models-- intonation, speaker identity, sentence, or any of the interactions between those things. And so that just says that's where the action is, is those electrodes there.

And that graph down here is from only three different-- each one is a single electrode, just so you get this. So this critical graph here, that shows electrode E1. That's one of those electrodes in one patient.

An electrode is typically 2 millimeters on a side. It's probably listening to a few tens of thousands of neurons. So it's one or two orders of magnitude better than a voxel with functional MRI, but it's still averaging over lots of neurons, not a single nerve.

STUDENT: The question [INAUDIBLE] averaging over [INAUDIBLE] but it's averaged over [INAUDIBLE].

NANCY KANWISHER: Yeah, that was the response of one electrode listening to male and female. I forget which is which. But other than that, you guys totally nailed it.

And notice how precise, and specific, and fascinatingly separated the responses of those electrodes are, segregated for pitch contour, or speaker identity, or what sentence was being spoken. Those things seem to be segregated spatially in the brain at a fine grain. Whether you'd see it with functional MRI-- you might, might not.

Many of you pointed out we might have not have the resolution. Think about other methods you might use to look for that, even if we didn't have the resolution with a simple binary contrast. And it's 12:26 and I'm going to stop. I will see you guys on Wednesday, and we will talk about music.