JOSH MCDERMOTT: I'm going to talk about hearing. I gather this is the first time in this class that, really, people have talked about audition. And usually, when I give talks and lectures, I find it's often helpful to just start out by doing some listening.

So we're going to start out this morning by just listening to some examples of typical auditory input that you might encounter during the course of your day. So all you've got to do is just close your eyes and listen. Here we go.

[AUDIO PLAYBACK]

[INAUDIBLE]

[END PLAYBACK]

OK. So you guys could probably tell what all of those things were, in case not, here are some labels.

So the first one is just a scene I recorded on my iPhone in a cafe. The second one was something from a sports bar. Then there was a radio excerpt, and then some Barry White.

And the point is that in all of those cases, I mean, they're all pretty different, but in all those cases, your brain was inferring an awful lot about what was happening in the world from the sound signal that was entering your ears. Right?

And so what makes that kind of amazing and remarkable is the sensory input that it was getting, to first order, just looks like this. Right? So there was sound energy that traveled through the air. It was making your ear drums wiggle back and forth in some particular pattern that would be indicated by that waveform there, that plots the pressure at your eardrum as a function of time.

And so from that particular funny pattern of wiggles, you were able to infer all those things. So in the case of the cafe scene, you could tell that there were a few people talking. You could probably hear a man and a woman, you could tell there was music in the background. You could hear there was dishes clattering. You might have been able to tell somebody had an Irish accent. You could probably tell, in the sports bar, that people were watching football and talking about it, and there were a bunch of people there.

You instantly recognized the radio excerpt as drivetime radio, that sort of, like, standard sound. And in the case of the music excerpt, you could probably hear six or seven, or maybe eight different instruments that were each, in their own way, contributing to the groove. So that's kind of audition at work.

And so the task of the brain is to take the sound signal that arrives at your ears that is then transduced into electrical signals by your ears, and then to interpret it and to figure out what's out there in the world.

So what you're interested in, really, is not the sound itself. You're interested in whether it was a dog, or a train, or rain, or people singing, or whatever it was. And so the interesting and hard problem that comes along with this is that most of the properties that we are interested in as listeners are not explicit in the waveform, in the sense that if I hand you the sound waveform itself, and you either look at it, or you run sort of standard machine classifiers on it, it would be very difficult to discern the kinds of things that you, with your brain, can very easily just report. And so that's really that's the question that I'm interested in and that our lab studies, namely, how is it that we derive information about the world from sound?

And so there's lots of different aspects to the problem of audition. I'm going to give you a taste of a few of them over the course of today's lecture. A big one that lots of people have heard about is often known as the cocktail party problem, which refers to the fact that real world settings often involve concurrent sound. So if you're in a room full of busy people, you might be trying to have a conversation with one of them, but there'll be lots of other people talking, music in the background, and so on, and so forth.

And so from that kind of complicated mixed signal that enters your ears, your brain has to estimate the content of one particular source of interest, of the person that you're trying to converse with.

So really, what you'd like to be hearing might be this.

AUDIO: She argues with her sister.

JOSH MCDERMOTT: But what might enter your ear could be this.

AUDIO: [INTERPOSING VOICES]

JOSH MCDERMOTT: Or maybe even this.

[INTERPOSING VOICES]

JOSH MCDERMOTT: Or maybe even this.

AUDIO: [INTERPOSING VOICES]

JOSH MCDERMOTT: So what I've done here plotted next to the icons are spectrograms. That's a way of taking a sound signal and turning that into an image. So it's a plot of the frequency content over time.

And you can see with the single utterance up here at the top, there's all kinds of structure, right? And we think that your brain uses that structure to understand what was being said. And you can see that, as more and more people are added to the party, that that structure becomes progressively more and more obscured, until, by the time you get to the bottom, it's kind of amazing that you can kind of pull anything out at all. And yet, as you hopefully heard, your brain has this remarkable ability to attend to and understand the speech signal of interest.

And this is an ability that still, to this day, is really unmatched by machines. So present day speech recognition algorithms are getting better by the minute, but this particular problem is still quite a significant challenge. And you've probably encountered this when you try to talk to your iPhone when you're in a car or wherever else.

Another kind of interesting complexity in hearing is that sound interacts with the environment on its way to your ears. So, you know, you typically think of yourself as listening to, say, a person talking or to some sound source, but in reality, what's happening is something like this picture here.

So there's a speaker in the upper right corner from which sound is emanating, but the sound takes a whole lot of different paths on its way to your ears. There's the direct path, which is shown in green, but then there are all these other paths where it can reflect off of the walls in the room.

So the blue lines here indicate paths where there's a single reflection. And the red lines indicate paths where there are two reflections, and so you can see there's a lot of them. And so the consequence of this is that your brain gets all these delayed copies of the source signal. And what that amounts to is really massive distortion of the signal. And this is known as reverberation.

So this is dry speech. Of course, you're hearing this in this auditorium that itself has lots of reverberation, so you're not actually going to hear it dry, but you'll still be able to hear a difference.

AUDIO: They ate the lemon pie. Father forgot the bread.

JOSH MCDERMOTT: And this is that signal with lots of reverberation added, as though you were listening to it in a cathedral or something. Of course, you're, again, hearing it in this room, as well.

AUDIO: They ate the lemon pie. Father forgot the bread.

JOSH MCDERMOTT: And you can still hear a difference. And if the reverb in this auditorium is swamping that, you can just look at the waveforms. And you can see that the waveforms of those two signals look pretty dramatically different, as do the spectrograms. All right?

So the point is that the consequence of all of those delayed reflections massively distorts the signal, all right? Physically, there are two really different things in those two cases. But again, your ability to recognize what's being said is remarkably invariant to the presence of that reverberation. And again, this is an instance where humans really are outperforming machines to a considerable extent.

This graph is a little bit dated. This is from, I think, three years ago, but it's a plot of five different speech recognition algorithms. And the percent of errors they're making when given a speech signal is a function of the amount of reverberation. And so zero means that there's no reverberation. That's the dry case, right? And so speech recognition works pretty well without any reverb.

But when you add a little bit of reverberation, and this is measured in terms of the reverberation time, it's the amount of time that it takes the reverb to fall off by a certain specified amount, and 300 and 500 milliseconds are actually very, very modest. So in this auditorium, my guess would be that the reverberation time is maybe even a couple seconds. So this is, like, what you get in a small classroom, maybe. Maybe even less than that.

But you can see that it causes major problems for speech recognition. And it's because the information in the speech signal gets blurred out over time, and again, it's just massive distortion. So your brain is doing something pretty complicated in order for you to be so robust to the presence of the reverberation.

So I run a research group where we study these kinds of problems. It's called the Lab for Computational Addition in the Department of Brain and Cognitive Sciences at MIT. We operate at the intersection of psychology, and neuroscience, and engineering, where what we aspire to do is to understand how it is that people hear so well in computational terms that would allow us to instantiate them in algorithms that we might replicate in machines. And so the research that we try to do involves hopefully symbiotic relationships between experiments in humans, auditory neuroscience, and machine algorithms.

And the general approach that we take is to start with what the brain has to work with. And by that, I mean we try to work with representations like the ones that are in the early auditory system.

And so here's the plan for this morning. And this is subject to change, depending on what kind of feedback I get from you guys. But my general plan was to start out with an overview of the auditory system, because I gather there's sort of a diversity of backgrounds here, and nobody's talked about audition so far. So I was going to go through a little overview.

And then there's been a special request to talk about some texture perception. I gather that there were some earlier lectures on visual texture, and that might be a useful thing to talk about. It's also a nice way to understand auditory models a little bit better.

I was then going to talk a little bit about the perception of individual sound sources and sort of the flip side to sound texture, and then conclude with a section on auditory scene analysis, so what your brain is able to do when it gets a complicated sound signal like you would get normally in the world, that has contributions from multiple causes and you have to infer those.

OK. And so we'll take a break about halfway through, as I guess that's kind of standard. And I'm happy for people to interrupt and ask questions. OK.

So the general outline for hearing, right, is that sound is created when objects in the world vibrate. Usually, this is because something hits something else, or in the case of a biological organism, there is some energy imparted to the vocal cords. And the object vibrates. That vibration gets transmitted to the air molecules around it, and you get a sound wave that travels through the air.

And that sound wave then gets measured by the ears. And so the ear is a pretty complicated device that is designed to measure sound. It's typically divided up into three pieces.

So there's the outer ear, consisting of the pinna and the eardrum. In functional terms, people usually think about this as a directional microphone. There's the middle ear. They're these three little bones in between the eardrum and the cochlea that are typically ascribed the functions of impedance matching and overload protection. I'm not going to talk about that today. And then there's the inner ear that the cochlea, which in very coarse engineering terms, we think of as doing some kind of frequency analysis.

And so again, at kind of a high level, so you've got your ears here. This is the inner ear on each side. And then those send feedforward input to the midbrain, and there's a few different way stations here. The cochlear nucleus is superior olivery complex, the inferior colliculus, and then the mediagenic nucleus or the thalamus. And the thalamus then projects to the auditory cortex.

And there's a couple things at a high level that are worth noting here. One is that the pathways here are actually pretty complicated, especially relative to the visual system that you guys have been hearing lots about it. Right? So there's a bunch of different stops on the way to the cortex. Another interesting thing is that input from the two ears gets mixed at a pretty early stage. OK.

All right, so let's step back and talk about the cochlea for a moment. And I realize that some of you guys will know about this stuff, so we'll go through it kind of quick.

Now one of the signature features of cochlear transduction is that its frequency tunes. So this is an unwrapped version of the cochlea. So if we step back to here, all right-- so we've got the outer ear, the ear canal, the eardrum, those three little bones that I told you about that connect to the cochlea. And the cochlea is this thing that looks like a snail.

And then if we unroll that snail and look at it like this, you can see that the cochlea consists of these tubes separated by a membrane. The membrane is called the basilar membrane. That's worth knowing about.

So sound enters here at the base and sets up a traveling wave along this membrane. So this is really a mechanical thing that happens. So there's actually, like, a physical vibration that occurs in this membrane. And it's a wave that travels along the cochlea.

And one of the signature discoveries about the cochlea is that that traveling wave peaks in different places, depending on the frequency content of the sound. And that's schematized in these drawings here on the right.

So if the ear were to receive a high frequency sound, that traveling wave would peak near the base. If it were to receive a medium frequency sound, the wave would peak somewhere in the middle. And a low frequency sound would peak near the apex.

And the frequency tuning, it's partly mechanical in origin, so that membrane, it varies in thickness and stiffness along its length. And there's also a contribution that's non-linear and active, that we'll talk briefly about it in a little bit.

So this is a close up. This is a cross-section. So imagine you took this diagram and you kind of cut it in the middle. This is a cross-section of the cochlea.

So this here is the basilar membrane, and this is the organ of Corti that sits on top of the basilar membrane. And so if you look closely, you can see that there's this thing in here. This is the inner hair cell. And that's the guy that does the transduction that takes the mechanical energy that's coming from the fact that this thing is vibrating up and down, and turns that into an electrical signal that gets sent to your brain.

And so the way that that works is that there is this other membrane here called the tectorial membrane. And the hair cells got these cilia that stick out of it. And as it moves up and down, there's a shearing that's created between the two membranes. The hair cell body deforms, and that deformation causes a change in its membrane potential. And that causes neurotransmitter to be released. All right, so that's the mechanism by which the brain takes that mechanical signal and turns it into an electrical signal that gets sent to your brain.

The other thing to note here, and we'll return to this, is that there are these other three cells here that are labeled as outer hair cells. And so those kind of do what the inner hair cell does in reverse. So they get an electrical signal from your brain, and that causes the hair cell bodies to deform, and that actually alters the motion of the basilar memory. So it's like a feedback system that we believe serves to amplify sounds and to sharpen their tuning. So there's feedback all the way to the all the way to the cochlea.

OK. So this is just another view. So here's the inner hair cell here. As this thing vibrates up and down, there's a shearing between these membranes. The inner hair cell membrane potential changes, and that causes neurotransmitter release.

OK, and so here's the really important point. So we just talked about how there's this traveling wave that gets set up that peaks in different places, depending on the frequency content of the sound. And so because to first order only, part of the basilar membrane moves for a given frequency of sound, each hair cell and the auditory nerve fiber that it synopsis with, signals only particular frequencies of sound. And so this is sort of the classic textbook figure that you would see on this, where what's being plotted on the y-axis is the minimum sound intensity needed to elicit a neural response. And the x-axis is the frequency of a tone with which you would be stimulating the ear.

So we have a little pure tone generator with a knob that allows you to change the frequency, and another knob that allows you to change the level. And you sit there recording from an auditory nerve fiber, varying the frequency, and then turning the level up and down until you get spikes out of the nerve fiber.

And so for every nerve fiber, there will be some frequency called the characteristic frequency, at which you can elicit spikes when you present the sound at a fairly low level. And then as you change the frequency, either higher or lower, the level that is needed to elicit a response grows. And so you can think of this as like a tuning curve for that auditory nerve fiber. All right?

And different nerve fibers have different characteristic frequencies. Here is just a picture that shows a handful of them. And so together, collectively, they kind of tile the space. And of course, given what I just told you, you can probably guess that each of these nerve fibers would synapse to a different location along the cochlea.

The ones that have high characteristic frequencies would be near the base. The ones that have low characteristic frequencies would be near the apex.

OK. So in computational terms, the common way to think about this is to approximate auditory nerve fibers with bandpass filters. And so this would be the way that you would do this in a model. Each of these curves is a bandpass filter, so what you see on the y-axis is the response of the filter. The x-axis is frequency. So each filter has some particular frequency at which they give a peak response, and then the signal is attenuated on either side of that peak frequency.

And so one way to think about what the cochlea is doing to the signal is that it's taking the signal that enters the ears, this thing here-- so this is just a sound signal, so the amplitude just varies over time in some particular way. You take that signal, you pass it through this bank of bandpass filters, and then the output of each of these filters is a filtered version of the original signal. And so in engineering terms, we call that a subband.

So this would be the result of taking that sound signal and filtering it with a filter that's tuned to relatively low frequencies, 350 to 520 Hertz, in this particular case. And so you can see that the output of that filter is a signal that varies relatively slowly. So it wiggles up and down, but you can see that the wiggles are in some sort of confined space of frequencies.

If we go a little bit further up, we get the output of something that's tuned to slightly higher frequencies. And you can see that the output of that filter is wiggling at a faster rate. And then if we go up further still, we get a different thing, that is again wiggling even faster.

And so collectively, we can take this original broadband signal, and then represent it with a whole bunch of these subbands. Typically, you might use 30, or 40, or 50.

So one thing to note here, you might have noticed that there's something funny about this picture and that the filters, here, which are indicated by these colored curves, are not uniform. Right? So the ones down here are very narrow, and the ones down here are very broad. And that's not an accident. That's roughly what you find when you actually look in the ear.

And why things are that way is something that you could potentially debate. But it's very clear, empirically, that that's roughly what you find. I'm not going to say too much more about this now, but remember that because it will become important a little bit later on.

So we can take these filters, and turn that into an initial stage of an auditory model, the stuff that we think is happening in the early auditory system where we've got our sound signal that gets passed through this bag of bandpass filters. And you're now representing that signal as a bunch of different subbands, just two of which are shown here for clarity.

And the frequency selectivity that you find in the ear has a whole host of perceptual consequences. I won't go through all of them exhaustively. It's one of the main deterrents of what masks what. So for instance, when you're trying to compress a sound signal by turning it into an mp3, you have to really pay attention to the nature of these filters. And, you know, you don't need to represent parts of the filters that would be-- sorry, parts of the signal would be masked, and these filters tell you a lot about that.

One respect in which frequency selectivity is evident is by the ability to hear out individual frequency components of sounds that have lots of frequencies in them. So this is kind of a cool demonstration. And to kind of help us see what's going on, we're going to look at a spectrogram of what's coming in. So hopefully this will work.

So what this little thing is doing is there's a microphone in the laptop, and it takes that microphone signal and turns it into a spectrogram. It's using a logarithmic frequency scale here, so it goes from about 100 Hertz up to 6400. And so if I don't say anything, you'll be able to hear the room noise, or see the room noise.

All right, so that's the baseline. And also, the other thing to note is that the microphone doesn't have very good bass response. And so the very low frequencies won't show up. But everything else will. OK.

[AUDIO PLAYBACK]

- Canceled harmonics. A complex tone is presented, followed by several cancellations and restorations of a particular harmonic. This is done for harmonics 1 through 10.

[LOUD TONE]

[TONE]

[TONE]

END PLAYBACK]

OK. And just to be clear, the point of this, right, is that what's happening here-- just stop that for a second-- is you're hearing what's called a complex tone. That just means a tone that has more than one frequency. All right? That's what constitutes complexity for a psychoacoustician. So it's a complex tone.

And each of the stripes, here, is one of the frequencies. So this is a harmonic complex. Notice that the fundamental frequency is 200 Hertz, and so all the other frequencies are integer multiples of that. So there's 400, 600, 800, 1,000 1200, and so on, and so forth. OK?

Then what's happening is that in each little cycle of this demonstration, one of the harmonics is getting pulsed on and off. All right? And the consequence of it being pulsed on and off is that you're able to actually hear it as, like, a distinct thing. And the fact that that happens, that's not, itself, happening in the ear. That's something complicated and interesting that your brain is doing with that signal it's getting from the ear.

But the fact you're able to do that is only possible by virtue of the fact that the signal that your brain gets from the ear divides the signal up in a way that kind of preserves the individual frequencies. All right? And so this is just a demonstration that you're actually able to, under appropriate circumstances, to hear out particular frequency components of this complicated thing, even if you just heard it by itself, it would just sound like one thing.

So another kind of interesting and cool phenomenon that is related to frequency selectivity is the perception of beating. So how many people here know what beating is? Yeah, OK.

So beating is a physical phenomenon that happens whenever you have multiple different frequencies that are present at the same time. So in this case, those are the red and blue curves up at the top. So those are sinusoids of two different frequencies.

And the consequence of them being two different frequencies is that over time, they shift in and out of phase. And so there's this particular point here where the peaks of the waveforms are aligned, and then there's this point over here where the peak of one aligns with the trough of the other. It's just because they're two different frequencies and they slide in and out of phase.

And so when you play those two frequencies at the same time, you get the black waveform, so some linearly. That's what sounds do when they're both present at once.

And so the point at which the peaks align, there is constructive interference. And the point at which the peak and the trough align, there is destructive interference. And so over time, the amplitude waxes and wanes.

And so physically, that's what's known as beating. And the interesting thing is that the audibility of the beating is very tightly constrained by the cochlea. So here's one frequency.

[TONE]

Here's the other.

[TONE]

And then you play them at the same time.

[TONE]

Can you hear that fluttering kind of sensation? All right. So that's amplitude modulation.

OK. And so I've just told you how we can think of the cochlea as this set of filters, and so it's an interesting empirical fact that you only hear beating if the two frequencies that are beating fall roughly within the same cochlear bandwidth. OK? And so when they're pretty close together, like, one semi-tone, the beating is very audible.

[TONE]

But as you move them further apart-- so three semi-tones is getting close to, roughly, what a typical cochlear filter bandwidth would be, and the beating is a lot less audible.

[TONE]

And then by the time you get to eight semi-tones, you just don't hear anything. It just sounds like two tones.

[TONE]

Very clear. So contrast that with--

[TONE]

All right. So the important thing to emphasize is that in all three of these cases, physically, there's beating happening. All right? So if you actually were to look at what the eardrum was doing, you would see the amplitude modulation here, but you don't hear that. So this is just another consequence of the way that you're cochlea is filtering sound.

OK. All right, so we've got our auditory model, here. What happens next?

So there's a couple of important caveats about this. And I mentioned this, in part, because some of these things are-- we don't really know exactly what the true significance is of some of these things, especially in computational terms. So I've just told you about how we typically will model with the cochlea is doing as a set of linear bandpass filters. So you get a signal, you apply a linear filter, you get a subband.

But in actuality, if you actually look at what the ear is doing, it's pretty clear that linear filtering provides only an approximate description of cochlear tuning. And in particular, this is evident when you change sound levels.

So what this is a plot of is tuning curves that you would measure from an auditory nerve fiber. So we've got spikes per second on the y-axis, we've got the frequency of a pure tone stimulus on the x-axis, and each curve plots the response at a different stimulus intensity. All right?

So down here at the bottom, we've got 35 dB SPL, so that's, like, a very, very low level, like, maybe if you rub your hands together or something, that would be close to 35. And 45-- and you can see here that the tuning is pretty narrow here at these low levels. So 45 dB SPL. So you get a pretty big response, here, at looks like, you know, 1700 Hertz. And then by the time you go down half an octave or something, there's almost no response.

But as the stimulus level increases, you can see that the tuning broadens really very considerably. And so up here at 75 or 85 dB, you're getting a response from anywhere from 500 Hertz out to, you know, over 2,000 Hertz. That's, like, a two octave range, right? So the bandwidth is growing pretty dramatically.

And this is very typical. Here is a bunch of examples of different nerve fibers that are doing the same thing. So at high levels, the tuning is really broad. At low levels, it's kind of narrow.

And so mechanistically, in terms of the biology, we have a pretty good understanding of why this is happening. So what's going on is that the outer hair cells are providing amplification of the frequencies kind of near the characteristic frequency of the nerve fiber at low levels, but not at high levels. And so at high levels, what you're seeing is just very broad kind of mechanical tuning.

But what really sort of remains unclear is what the consequences of this are for hearing, and really, how to think about it in computational terms. So it's clear that the linear filtering view is not exactly right. And one of the interesting things is that it's not like you get a lot worse at hearing when you go up to 75 or 85 dB. In fact, if anything, most psychophysical phenomena, actually, are better in some sense.

This phenomena, as I said, is related to this distinction between inner hair cells and outer hair cells. So the inner hair cells are the ones that, actually, are responsible for the transduction of sound energy, and the outer hair cells we think of as part of a feedback system that amplifies the motion of the membrane and sharpens at the tuning. And that amplification is selective for frequencies at the characteristic frequency, and really occurs at very low levels.

OK. So this is related to what Hynek was mentioning and the question that was asked earlier. So there's this other important response property of the cochlea, which is that for frequencies that are sufficiently low, auditory nerve spikes are phase locked to the stimulus.

And so what you're seeing here is a single trace of a recording from a nerve fiber that's up top, and then at the bottom is the stimulus that would be supplied to the air, which is just a pure tone of some particular frequency. And so you can note, like, two interesting things about this response.

The first is that the spikes are intermittent. You don't get a spike at every cycle of the frequency. But when the spikes occur-- sorry, they occur at a particular phase relative to the stimulus. Right? So they're kind of just a little bit behind the peak of the waveform in this particular case, right, and in every single case.

All right, so this is known as phase locking. And it's a pretty robust phenomena for frequencies under a few kilohertz. And this is in non-human animals. There's no measurements of the auditory nerve in humans because to do so is highly invasive, and just nobody's ever done it, and probably won't ever.

So this is another example of the same thing, where again, this is sort of the input waveform, and this is a bunch of different recordings of the same nerve fiber that are time aligned. So you can see the spikes always occur at a particular region of phase space. So they're not uniformly distributed.

And the figure here on the right shows that this phenomenon deteriorates at very high frequency. So this is the plot of the strength of the phase locking as a function of frequency. And so up to a kilohertz, it's quite strong, and then it starts to kind of drop off, and above about 4k there's not really a whole lot.

OK. And so as I said, one of the other salient things is that the fibers don't fire with every cycle of the stimulus. And one interesting fact about the ear is that there are a whole bunch of auditory nerve fibers for every inner hair cell. And people think that one of the reasons for that is because of this phenomena, here. And this is probably due to things like refractory periods and stuff, right?

But if you have a whole bunch of nerve fibers synapsing under the same hair cell, the idea is that, well, collectively, you'll get spikes at every single cycle.

So here's just some interesting numbers. So the number of inner hair cells per ear in a human is estimated to be about 3500. There's about four times as many outer hair cells, or roughly 12,000, but that's the key number here. So coming out of every ear are roughly 30,000 auditory nerve fibers. So there's about 10 times as many auditory nerve fibers as there are inner hair cells.

And it's interesting to compare those numbers to what you see in the eye. So these are estimates from a few years ago, from the eye of roughly 5 million cones per eye, lots of rods, obviously. But then you go to the optic nerve, and the number of optic nerve fibers is actually substantially less than the number of cones. So 1 and 1/2 million. So there's this big compression that happens when you go into the auditory nerve-- sorry, in the optic nerve, whereas there's an expansion that happens in the auditory nerve.

And just for fun, these are rough estimates of what you find in the cortex. So in primary auditory cortex, this is a very crude estimate I got from someone a couple of days ago, of 60 million neurons per hemisphere. And in v1, the estimate that I was able to find was 140 million. So these are sort of roughly the same order of magnitude, although it's obviously smaller in the auditory system.

But there is something very different happening here, in terms of the way the information is getting piped from the periphery onto the brain. And one reason for this might just be the fact that phenomena here, where the spiking in an individual auditory nerve fiber is going to be intermittent because the signals that it has to convey are very, very fast, and so you kind of have to multiplex in this way.

All right, so the big picture here is that if you look at the output of the cochlea, there are, in some sense, two cues to the frequencies that are contained in a sound. So there's what is often referred to as the place of excitation in the cochlea. So these are the nerve fibers that are firing the most, according to a rate code. And there's also the timing of the spikes that are fired, in the sense that, for frequencies below about 4k, you get phase locking. And so the inner spiked intervals will be stereotyped, depending on the frequencies that come in.

And so it's one of these sort of-- I still find this kind of remarkable when I sort of step back and think about the state of things, that this is a very basic question about neural coding at the very front end of the system. And the importance of these things really remains unresolved. So people have been debating this for a really long time, and we still don't really have very clear answers to the extent to which the spike timing really is critical for inferring frequency.

So I'm going to play you a demo that provides-- it's an example of some of the circumstantial evidence for the importance of phase locking. And broadly speaking, the evidence for the importance of phase locking in humans comes from the fact that the perception of frequency seems to change once you kind of get above about 4k.

So for instance, if you give people a musical interval, [SINGING] da da, and then you ask them to replicate that, in, say, different octaves, people can do that pretty well until you get to about 4k. And above 4k, they just break down. They become very, very highly variable. And that's evident in this demonstration.

So what you're going to hear in this demo, and this is a demonstration I got from Peter Cariani. Thanks to him.

It's a melody that is probably familiar to all of you that's being played with pure tones. And it will be played repeatedly, transposed from very high frequencies down in, I don't know, third octaves or something, to lower and lower frequencies.

And what you will experience is that when you hear the very high frequencies, well, A, they'll be kind of annoying, so just bear with me, but the melody will also be unrecognizable. And so it'll only be when you get below a certain point that you'll say, aha, I know what that is. OK?

And again, we can look at this in our spectrogram, which is still going. And you can actually see what's going on.

[MELODIC TONES]

OK. So by the end, hopefully everybody recognized what that was. So let's just talk briefly about how these subbands that we were talking about earlier relate to what we see in the auditory nerve, and again, this relates to one of the earlier questions.

So the subband is this blue signal here. This is the output of a linear filter. And one of the ways to characterize a subband like this that's band limited is by the instantaneous amplitude and the instantaneous phase. And these things, loosely, can be mapped onto a spike rate and spike timing in the auditory nerve.

So again, this is an example of the phase locking that you see, where the spikes get fired at some particular point in the waveform. And so if you observe this, well, you know something about exactly what's happening in the waveform, namely, you know that there's energy there in the stimulus because you're getting spikes, but you also actually know the phase of the waveform because the spikes are happening in particular places.

And so this issue of phase is sort of a tricky one, because it's often not something that we really know how to deal with. And it's also just empirically the case that a lot of the information in sound is carried by the way that frequencies are modulated over time, as measured by the instantaneous amplitude in a subband.

And so the instantaneous amplitude is measured by a quantity called the envelope. So that's the red curve, here, that shows how the amplitude waxes and wanes.

And the envelope is easy to extract from auditory nerve response, just by computing the firing rate over local time windows. And in signal processing terms, we typically extract it with something called the Hilbert transform, by taking the magnitude of the analytic signal. So it's a pretty easy thing to pull out in MATLAB, for instance.

So just to relate this to stuff that may be more familiar, the spectrograms that people have probably seen in the past-- again, these are pictures that take a sound waveform and plot the frequency content over time. One way to get a spectrogram is to have a bank of bandpass filters, to get a bunch of subbands, and then to extract the envelope of each subband, and just plot the envelope in grayscale, horizontally. All right?

So a stripe through this picture is the envelope in grayscale. So it's black in the places where the energy is high, and white in the places where it's low. And so this is a spectrogram of this.

[DRUMMING]

All right, so that's what you just heard. It's just a drum break. And you can probably see that there are sort of events in the spectrogram that correspond to things like the drumbeats.

Now, one of the other striking things about this picture is it looks like a mess, right? I mean, you listen to that drum break and it sounded kind of crisp and clean, and when you actually look at the instantaneous amplitude in each of the subbands, it just sort of looks messy and noisy.

But one interesting fact is that this picture, for most signals, captures all of the information that matters perceptually in the following sense, that if you have some sound signal and you let me generate this picture, and all you let me keep is that picture. We throw out the original sound waveform. From that picture, I can generate a sound signal that will sound just like the original. All right? And in fact, I've done this. And here is a reconstruction from that picture.

[DRUMMING]

Here's the original.

[DRUMMING]

Sounded exactly the same. OK? And so-- and I'll tell you in a second, how we do this, but the fact that this picture looks messy and noisy is, I think, mostly just due to the fact that your brain is not used to getting the sensory input in this format. Right? You're used to hearing as sound. Right? And so your visual system is actually not optimally designed to interpret a spectrogram.

I want to just briefly explain how you take this picture and generate a sound signal because this is sort of a useful thing to understand, and it will become relevant a little bit later.

So the general game that we play here is you hand me this picture, right, and I want to synthesize a signal. And so usually what we do, is we start out with a noise signal, and we transform that noise signal until it's in kind of the right representation. So we split it up into its subbands.

And then we'll replace the envelopes of the noise subbands by the envelopes from this picture. OK? And so the way that we do that is we measure the envelope of each noise subband, we divide it out, and then we multiply by the envelope of the thing that we want. And that gives us new subbands. And we then add those up, and we get a new sound signal.

And for various reasons, this is a process that needs to be iterated. So you take that new sound signal, you generate its subbands, you replace their envelopes by the ones you want, and you add them back up to get a sound signal. And if you do this about 10 times, what you end up with is something that has the envelopes that you want. And so then you can listen to it. And that's the thing that I played you. OK?

So it's this iterative procedure, where you typically start with noise, you project the signal that you want onto the noise, collapse, and then iterate. OK. And we'll see some more examples of this in action.

So I just told you how the instantaneous amplitude in each of the filter outputs, which we characterize with the envelope, is an important thing for sound representation. And so in the auditory model that I'm building here, we've got a second stage of processing now, where we've taken the subbands and we extract the instantaneous amplitude.

And there's one other thing here that I'll just mention, which is that another important feature of the cochlea is what's known as amplitude compression. So the response of the cochlea as a function of sound level is not linear, rather it's compressive. And this is due to the fact that there is selective amplification when sounds are quiet, and not when they're very high in intensity. And I'm not going to say anything more about this for now, but it will become important later.

And this is something that has a lot of practical consequences. So when people lose their hearing, one of the common things that happens is that the outer hair cells stop working correctly. And the outer hair cells are one of the things that generates that compressive response. So they're the nonlinear component of processing in the cochlea. And so when people lose their hearing, the tuning both broadens, and you get a linear response to sound amplitude because you lose that selective amplification, and that's something that hearing aids try to replicate, and that's hard to do.

OK. So what happens next? So everybody here-- does everybody here know what a spike triggered average is? Yeah. Lots of people talked about this, probably. OK.

So one of the standard ways that we investigate sensory systems, when we have some reason to think that things might be reasonably linear, is by measuring something called a spike triggered average. And so the way this experiment might work is you would play stimulus like this.

[NOISE SIGNAL]

So that's a type of a noise signal. Right? So you'd play your animal that signal, you'd be recording spikes from a neuron that you might be interested in, and every time there's a spike, you would look back at what happens in the signal and then you would take all the little histories that preceded that spike. You'd average them together, and you get the average stimulus that preceded a spike.

And so in this particular case, we're going to actually do this in the domain of the spectrogram. And that's because you might hypothesize that, really, what the neurons would care about would be the instantaneous amplitude in the sound signal, and not necessarily the phase.

So if you do that, for instance, in the inferior colliculus, the stage of the midbrain, you see things that are pretty stereotyped. And we're going to refer to what we get out of this procedure as a spectrotemporal receptive field. And that would often be referred to as a STRF for short. So if you hear people talk about STRFs, this is what they mean.

OK. So these are derived from methods that would be like the spike triggered average. People don't usually actually do a spike triggered average for various reasons, but what you get out is similar.

And so what you see here is the average spectrogram that would be preceding a spike. And for this particular neuron, you can see there's a bit of red and then a bit of blue. And so at this particular frequency, which, in this case, is something like 10 kilohertz or something like that, the optimal stimulus that would generate a spike is something that gives you an increase in energy, and then a decrease.

And so what this corresponds to is amplitude modulation at a particular rate. And so you can see there's a characteristic timing here. So the red thing has a certain duration, and then there's the blue thing. And so there is this very rapid increase and decrease in energy at that particular frequency. So that's known as amplitude modulation. And so this is one way of looking at this in the domain of the spectrogram.

Another way of looking at this would be to generate a tuning function as a function of the modulation rate. So you could actually change how fast the amplitude is being modulated, and you would see in a neuron like this, that the response would exhibit tuning. And so each one of the graphs here, or each one of the plots, the dashed curves in the lower right plot-- so each dashed line is a tuning curve for one neuron. So it's a plot of the response of that neuron as a function of the temporal modulation frequencies. So that's how fast the amplitude is changing with time.

And if you look at this particular case here, you can see that there's a peak response when the modulation frequency is maybe 70 Hertz, and then the response decreases if you go in either direction. This guy here has got a slightly lower preferred frequency, and down here, lower still.

And so again, what you can see is something that looks strikingly similar to the kinds of filter banks that we just saw when we were looking at the cochlea. But there's a key difference here, right, that here we're seeing tuning to modulation frequency, not to audio frequency. So this is the rate at which the amplitude changes. That's the amplitude of the sound, not the audio frequency.

So there's a carrier frequency here, which is 10k, but the frequency we're talking about here is the rate at which this changes. And so in this case here, you can see the period here is maybe 5 milliseconds, and so this would correspond to a modulation of, I guess, 200 Hertz. Is that right? I think that's right. Yeah. OK.

And so as early as the midbrain, you see stuff like this. So in the inferior colliculus, there's lots of it. And so this suggests a second, or in this case, a third stage of processing, which are known as modulation filters. This is a very old idea in auditory science that now has a fair bit of empirical support.

And so the idea is that in our model, we've got our sound signal. It gets passed through this bank of bandpass filters, you get these subbands, you extract the instantaneous amplitude, known as the envelope, and then you take that envelope and you pass it through another filter bank. This time, these are filters that are tuned in modulation frequency.

And the output of those filters-- again, it's exactly conceptually analogous to the output of this first set of band pass filters. So in this case, we have a filter that's tuned to low modulation rates, and so you can see what it outputs is something that's fluctuating very slowly. So it's just taken the very slow fluctuations out of that envelope. Here, you have a filter that's tuned to higher rates, and so it's wiggling around at faster rates. And you have a different set of these filters for each cochlear channel, each thing coming out of the cochlea.

So this picture here that we've gradually built up gives us a model of the signal processing that we think occurs between the cochlea and the midbrain or the thalamus. So you have the bandpass filtering that happens in the cochlea, you get subbands, you extract the instantaneous amplitude in the envelopes, and then you filter that again with these modulation filters.

This is sort of a rough understanding of the front end of the auditory system. And the question is, given these representations, how do we do the interesting things that we do with sound? So how do we recognize things and their properties, and do scene analysis, and so on, and so forth. And this is still something that is still very much in its infancy in terms of our understanding.

One of the areas that I've spent a bit of time on to try to get a handle on this is sound texture. And I started working on this because I sort of thought it would be a nice way in to understanding some of these issues, and because I gather that Eero was here talking about visual textures. I was asked to feature this, and hopefully this will be useful.

So what are sound textures? Textures are sounds that result from large numbers of acoustic events, and they include things that you hear all the time, like rain--

[RAIN]

--or birds--

[BIRDS CHIRPING]

--or running water--

[RUNNING WATER]

--insects--

[INSECTS]

--applause--

[APPLAUSE]

--fire, so forth.

[FIRE BURNING]

OK. So these are sounds that they, typically, are generated from large numbers of acoustic events. They're very common in the world. You hear these things all the time. But they've been largely unstudied. So there's been a long and storied history of research on visual texture, and this really had not been thought about very much until a few years ago.

Now, a lot of the things that people typically think about in hearing are the sounds that are produced by individual events. And if we have time, we'll talk about this more later, but stuff like this.

AUDIO: January.

JOSH MCDERMOTT: Or this.

[SQUEAK]

And these are the waveforms associated with those events. And the point is that those sounds, they have a beginning, and an end, and a temporal evolution. And that temporal evolution is sort of part of what makes the sound what it is, right?

And textures are a bit different. So here's just the sound of rain.

[RAIN]

Now of course, at some point, the rain started and hopefully at some point it will end, but the start and the end are not what makes it sound like rain, right? The qualities that make it sound like rain are just there. So the texture is stationary. So the essential properties don't change over time.

And so I got interested in textures because it seemed like stationarity would make them a good starting point for understanding auditory representation because, in some sense, it sort of simplifies the kinds of things you have to worry about. You don't have to worry about time in quite the same way.

And so the question that we were interested in is how people represent and recognize texture. So just to make that concrete, listen to this.

[CHATTER]

And then this.

[CHATTER]

So it's immediately apparent to you that those are the same kind of thing, right? In fact, they're two different excerpts of the same recording. But the waveform itself is totally different in the two cases. So there's something that your brain is extracting from those two excerpts that tells you that they're the same kind of thing, and that, for instance, they're different from this.

[BEES BUZZING]

So the question is, what is it that you extract and store about those waveforms that tells you that certain somethings are the same and other things are different, and that allows you to recognize what things are?

And so the key theoretical proposal that we made in this work is that because they're stationary, textures can be captured by statistics that are time averages of acoustic measurements. So the proposal is that when you recognize the sound of fire, or rain, or what have you, you're recognizing these statistics. So what kinds of statistics might we be measuring if you think this proposal has some plausibility?

And so part of the reason for walking you through this auditory model is that whatever statistics the auditory system measures are presumably derived from representations like this, right, that constitute the input that your brain is getting from the auditory periphery. And so we initially asked how far one might get with representations consisting of fairly generic statistics of these standard auditory representations, things like marginal moments and correlations.

So the statistics that we initially considered were not, in any way, specifically tailored to natural sounds, and really, ultimately, what we'd like to do would be to actually learn statistics from data that, actually, we think are good representations. That's something we're working on. But these statistics are simple and they involve operations that you could instantiate in neurons, so it seemed like maybe a reasonable place to start to at least get a feel for what the landscape was like.

And so what I want to do now is just to give you some intuitions as to what sorts of things might be captured by statistics of these representations. And so at a minimum, to be useful for recognition, well, statistics need to give you different values for different sounds. And so let's see what happens. So let's first have a look at what kinds of things might be captured by marginal moments of amplitude envelopes from bandpass filters. OK?

So remember, the envelope, here, you can think of as a stripe through a spectrogram, right, so it's the instantaneous amplitude in a given frequency channel. So the blue thing is the subband, the red thing here is the envelope. And the marginal moments will describe the way the envelope is distributed. So imagine you took that red curve and you collapsed that over time to get a histogram that tells you the frequency of occurrence of different amplitudes in that particular frequency channel. And that's what it looks like for this particular example.

And you might think that, well, this can't possibly be all that informative, right, because it's obviously got some central tendency and some spread, but when you do this business of collapsing across time, you're throwing out all kinds of information. But one of the interesting things is that when you look at these kinds of distributions for different types of sounds, you see that they vary a fair bit, and in particular, that they're systematically different for a lot of natural sounds than they are for noise signals.

So this is that same thing, so it's this histogram here, just rotated 90 degrees. So we've got the probability of occurrence on the y-axis, and the magnitude in the envelope of one particular frequency channel on the x-axis, for three different recordings. So the red plots what you get for a recording of noise, the blue is a stream, and the green is a bunch of geese. Geese don't quack. Whatever the geese do.

AUDIENCE: Honk.

JOSH MCDERMOTT: Yeah, honking. Yeah. All right. And this is a filter that's centered at 2,200 Hertz. In particular, these examples were chosen because the average value of the envelope is very similar in the three cases. But you can see that the distributions have very different shapes.

So the noise one, here, has got pretty low variance. The stream has got larger variance, and the geese larger still, and it's also positively skewed. So to got this kind of long tail, here. And if you look at the spectrograms associated with these sounds, you can see where this comes from.

So spectogram of noise is pretty gray, so the noise signal kind of hangs out around its average value most of the time, whereas the stream has got more gray and more black, and the geese actually has some white and some dark black. So here, white would correspond to down here, black would correspond to up here.

And the intuition here is that natural signals are sparse. In particular, they're sparser than noise. So we think of natural signals as often being made up of events like raindrops, or geese calls, and these events are infrequent, but when they occur they produce large amplitudes in the signal. And when they don't occur, the amplitude is lower. In contrast, the noise signal doesn't really have those.

But the important point about this from the standpoint of wanting to characterize a signal with statistics is that this phenomenon of sparsity is reflected in some pretty simple things that you could compute from the signal, like the variance and the skew. So you can see that the variance varies across these signals, as does a skew.

All right. Let's take a quick look at what you might get by measuring correlations between these different channels. So these things also vary across sound. And you can see them in the cochleogram here, in this particular example, as reflected in these vertical streaks. So this is the cochleogram of fire, and fire's got lots of crackles and pops.

[FIRE BURNING]

And those crackles and pops show up as these vertical streaks in the cochleogram. So a crackle and pop is like a click-like event. Clicks have lots of different frequencies in them, and so you see these vertical streaks in the spectrogram, and that introduces statistical dependencies between different frequency channels.

And so that can be measured by just computing correlations between the envelopes of these different channels, and that's reflected in this matrix here. So every cell of this matrix is the correlation between a pair of channels. So we're going from low frequencies, here, to high, and low to high. The diagonal has got to be one, but the off-diagonal stuff can be whatever.

You can see for example of fire, there's a lot of yellow and a lot of red, which means that the amplitudes of different channels tends to be correlated. But this is not the case for everything. So if you look at a water sound, like a stream--

[STREAM]

--there are not very many things that are click-like, and most of the correlations here are pretty close to zero.

So again, this is a pretty simple thing that you can measure, but you get different values for different sounds. Similarly, if we look at the power coming out of these modulation filters, you also see big differences across sounds.

So that's plotted here for three different sound recordings, insects, waves, and a stream. And so remember that these modulation filters, we think of them as being applied to each cochlear channel. So the modulation power that you would get from all your modulation filters is-- you can see that in a 2D plot. So this is the frequency of the cochlear channel, and this is the rate of the modulation channel. So these are slow modulations and these are fast modulations.

And so for the insects, you can see that there are these, like, little blobs up here. And that actually corresponds to the rates at which the insects rub their wings together and make the sound that they make, which kind of gives it this shimmery quality.

[INSECTS]

In contrast, the waves have got most of the power here at very slow modulations.

[WAVES]

And the stream is pretty broadband, so there's modulations at a whole bunch of different rates.

[STREAM]

All right. So just by measuring the power coming out of these channels, we're potentially learning something about what's in the signal. And I'm going to skip this last example.

All right, so the point is just that when you look at these statistics, they vary pretty substantially across sound. And so the question we were interested in is whether they could plausibly account for the perception of real world textures.

The key methodological proposal of this work was that synthesis is a potentially very powerful way to test a perceptual theory. Now, maybe the kind of standard thing that you might think that you might try to do with this type of representation is measure these statistics, and then see whether, for instance, you could discriminate between different signals or maybe [AUDIO OUT].

And for various reasons, I actually think that synthesis is potentially a lot more powerful. And the notion is this, that if your brain is representing sounds with some set of measurements, then signals that have the same values of those measurements ought to sound the same to you.

And so in particular, if we've got some real world recording, and we synthesize a signal to cause it to have the same measurements, the statistics in this case, as that real world recording, well, then the synthetic signal ought to sound like the real world recording if the measurements that we use are like the ones that the brain is using to represent sound. And so we can potentially use synthesis, then, to test a candidate representation, in this case, these statistics that we're measuring, are a reasonable representation for the brain to be using.

So here's just a simple example to kind of walk you through the logic. So let's suppose that you had a relatively simple theory of sound texture perception, which is that texture perception might be rooted in the power spectrum. It's not so implausible. Lots of people think the power spectrum is sort of a useful way to characterize it [AUDIO OUT]. And you might think that it would have something to do with the way textures would sound.

And so in the context of our auditory model, the power spectrum is captured by the average value of each envelope. Remember, the envelope's telling you the instantaneous amplitude, and so if you just average that, you're going to find out how much power is in each frequency channel. So that's how you would do it in this framework.

So the way this would work, if you get some sound signal, say, this--

[BUBBLES]

--you would pass it through the model, you'd measure the average value of the envelope, so you get a set of 30 numbers. Say, if you have 30 bandpass filters there, at the output of each one of those, you get the average of all of the envelopes, so you get 30 numbers. And then you take those 30 numbers and you want to synthesize the signal, subject to the constraint of it having the same value for those 30 numbers.

And so in this case, it's pretty simple to do. So we take a noise signal, we want to start out as random as possible, so we take a noise signal, we generate it's subbands, and then we just scale the subbands up or down so that they have the right amount of power. And we add them back up, and we get a new sound signal. And then we listen to it and we see whether it sounds like the same thing. OK?

Here's what they sound like. And as you will hear, they just sound like noise, basically, right? So this is supposed to sound like rain--

[RAIN]

--or a stream--

[STREAM]

--or bubbles--

[BUBBLES]

--or fire--

[FIRE BURNING]

--or applause.

[APPLAUSE]

So you might notice that they sound different, right? And you might have even been able to convince yourself, well, that sounds a little bit like applause, right? So there's something there. But the point is that they don't sound anything like--

[APPLAUSE]

--or--

[FIRE BURNING]

--so on, and so forth. OK?

All right, so the point is that everything just sounds like noise, and so what this tells us is that our brains are not simply registering the spectrum when we recognize these textures. Question is whether additional simple statistics will do any better.

And so we're going to play the same game, right, except we have a souped up representation that's got all these other things in it. And so the consequence of this is that the process of synthesizing something is less straightforward.

So I was in Eero Simoncelli's at the lab at the time, and spent a while trying to get this to work, and eventually we got to work. But conceptually, the process is the same.

So you have some original sound recording, rain, or what have you, you pass it through your auditory model, and then you measure some set of statistics. And then you start out with a noise signal, you pass it through the same model, and then measure its statistics, and those will in general be different from the target values. And so you get some error signal here, and you use that error signal to perform gradient descent on the representation of the noise in the auditory model.

And so you cause the envelopes of the noise signal to change in ways that cause their statistics to move towards the target value. And there's a procedure here by which this is iterated, and I'm not going to get into the details. If you want to play around with it, there's a toolbox that's now available on the lab website if you want to do that, and it's described in more detail in that paper.

All right. So the result, the whole point of this, is that we get a signal that shares the statistics of some real world sound. How do they sound? And remember, we're interested in this because if these statistics, if this candidate representation accounts for our perception of texture, well, then the synthetic signals ought to sound like new examples of the real thing. And the cool thing, and rewarding part of this whole thing, is that in many cases, they do.

So I'm just going to play the synthetic versions. All of these were generated from noise, just by causing the noise to match the statistics of, in this case, rain--

[RAIN]

--stream--

[STREAM]

--bubbles--

[BUBBLES]

--fire--

[FIRE BURNING]

--applause--

[APPLAUSE]

--wind--

[WIND BLOWING]

--insects--

[INSECTS]

--birds, oops--

[BIRDS CHIRPING]

--and crowd noise.

[CHATTER]

All right. It also works for a lot of unnatural sounds. Here's rustling paper--

[RUSTLING PAPER]

--and a jackhammer.

[JACKHAMMER]

All right. And so the cool thing about this is you can put it whatever you want in there, right? You can measure the statistics from anything, and you can generate something that is statistically matched. And when you do this with a lot of textures, you tend to get something that captures some of the qualitative properties.

And so the success of this, and this is in this case, the reason why this is scientifically interesting in addition to fun, is that it lends plausibility to the notion that these statistics could underlie the representation and recognition of textures.