[SQUEAKING]

[RUSTLING]

[CLICKING]

JOSH MCDERMOTT: Today, we're going to talk about texture perception, which we started talking about last time. So remember, this distinction between things, like objects and stuff. So objects are things. They're kind of individuated. We often think of the goal of vision as being object recognition.

But the world also has a lot of stuff. When you have lots of objects all together, they create what we call a texture, and it's got a certain appearance and you can discriminate it. And we use it for all kinds of different things. You can use texture to tell about material, about shape, to segment scenes.

And we're pretty good at detecting boundaries between textures. And what we initially talked about last time was the fact that when you look at these texture boundaries, in some cases, they're really easy to see, and in other cases, they're are very subtle and nonobvious and don't really pop out at you. So we saw lots of examples of this.

So each of these things has got kind of a region in the middle that is a different texture than the surround. In some cases, it's obvious, in some cases, it's not. Each of these has a texture boundary. Some of them are obvious, some of them are not.

And we spent some time talking about whether we can account for the salience of texture boundaries using pretty standard models of early vision, like the stuff that we kind of learned about earlier in the class. So this is an example of a texture image. So there's a region in the middle. It's a different texture than the rest.

You convolve it with these two Gabor-like filters, stand-ins for filters in V1, this is the result of the convolution. Remember, the convolution means that the filter kind of gets placed down at every possible location in the image. And then we can view the convolution as an image. You can think of this as the response of a big population of neurons that all have the same type of receptive field, just positioned position at different places in the visual field.

And so the filter responses are different in the center region than in the surround. And we talked about how you might be able to make that more explicit by performing some simple operations on the output of the filter. So this is just the raw filter output, the convolution of the filter with the image. And this is a horizontal slice there.

If you square the filter response, you can make everything positive. And then if you average over local regions, you get this quantity that's a lot like the energy measure that we talked about as being a good model of complex cells. And in this case, the energy measure has a different value kind of in the surround than in the center. So it kind of makes explicit the idea that there's a different kind of stuff in the center region of the image than in the surround.

So these energy measures can predict some texture boundaries. And this is that energy measure displayed as a convolution-like operation. And it captures this idea that texture is really defined by the average properties of a local image region. So what makes a texture a texture is what it's like on average. So in this case, there's some kind of orientation energy in one part of the image, and a different type in another.

So then we asked whether we could potentially account for the perception of texture boundaries in some of these more complicated textures with these similarly simple models, and saw a bunch of examples that kind indicate that they do a pretty good job. So specifically, we can look at a whole bunch of different images of this nature, and the extent to which you really see a salient region in the middle is kind of mirrored by the extent to which these simple mechanisms show differences in their responses.

So here's a case where the center region is really obvious and you see a really big difference. Here's a case where the little elements in the texture have been scaled. And at this particular relative scale, the difference doesn't really pop out at you, and it's also not really evident in this energy-based mechanism.

So then we kind of pivoted to this question of whether or not these kind of very simple models that are based on the ingredients that we think of as being present in V1, whether they can actually account for the way that textures look. And the critical idea that we discussed is that if we've correctly identified the brain's representation of texture, then if we measure that representation, for some example natural texture-- some image that you give me-- and then we generate synthetic textures that have the same values of that representation, well, then the synthetic textures, they should look like the example. They should look like the same type of texture. So that's the principle that we're working with here.

And so the key notion here is that a given type of texture can be present in a lot of different examples. So there's lots of examples of any given texture. And we looked at the carpet, and you could look at different regions of the carpet. They're all the same kind of texture even though they're slightly different images. So textures kind of consist of an equivalence class of all these different images that share some kind of property.

And in this particular case, there's an equivalence class of all the things that are generated by the same physical process that kind of created carpet, for instance, or leaves. But if we're really interested in how textures look, then what really matters to us is whether textures are perceptually equivalent, so whether they'll look the same to a human observer. And so the game that we're going to play is to try to synthesize images that will have the same values of some candidate representation that embodies a theory of what is represented in your head when you look at a texture.

And then we're going to look at the textures and actually see whether they look the same. So the idea is that we're going to take noise images, and then try to alter them in various ways to actually cause them to reside in this part of the space that's got the same representation as some example. So this was a case where maybe we got part of the way there, but not all the way there.

And so the initial example that we talked about-- and this is where we ended last time-- was this initial attempt at implementing this methodology using a candidate representation of texture that consists of marginal histograms of filter responses, so the distribution of the filter responses. And so this is the representation that this is based on.

So we're taking an image. We convolve it with a big set of filters that are tuned in orientation and spatial frequency, and you get a bunch of convolutions. These are often referred to as subbands. And so the model of texture perception will be to represent textures with the distribution of the activity in these different subbands, which we take as simulations of the responses of oriented, spatial frequency tuned neurons.

So we talked about the idea that this is often called a pyramid representation of an image. So we've got different spatial scales of images, and when the spatial scale is coarse, the image can be downsampled. And so that's what these things are up here.

These are the low spatial frequencies of the image. And they can be represented with a smaller pixel array. So that's why it's called a pyramid representation. So you have these big subband representations here, and then they get smaller as you move up the pyramid. So it's kind of a classic image processing-based representation.

And so remember that the representation we're trying to evaluate here is one where you take each of these subband representations and you form a histogram. So the histogram is just the distribution of the values in this array. So it's like throwing all the values into a bag and then seeing what the distribution is.

So you're throwing out all of the spatial organization of the image. And so this is really just capturing the average properties of that subband, in this case averaged across space. So it captures the variance as well as other as well as all other marginal statistics. So that's the proposal.

So what we want to do is to take some candid image, generate the subband in representation, and then generate these histograms. And then we want to take a noise image, and then mess with the image so that it has the same histograms in these subbands in the filter responses. And so in this particular case, there was a pretty simple trick to do this called histogram equalization.

So all that you do is you generate the cumulative histogram for an image or a subband. And then there's just a simple kind of mapping function where you take a given pixel in the noise image, and then you change its value depending on where it falls in the cumulative histogram. So in this particular case, this intensity value gives you this probability value. So you map that over here, and that tells you that this thing is going to be 1. And so in this case, we're just doing this on the pixel intensities, and so the noise image ends up getting turned into this thing that's kind of mostly white with a little bit of black because that's the pixel intensity histogram that you have in this original test image.

So now we're going to do the same thing with the histograms of the subbands. So this original texture has subband histograms that have these very long tails. The noise image has something that's a lot more Gaussian. And so we do histogram equalization, and the noise subband gets kind of altered in this way, so that it's got a histogram that looks a lot more like the original one.

So we do this for all of the different subbands. And then this subband representation can be inverted. You can take it and then essentially add these things back up in order to get an original image. So this is the result showing you that histograms are all matched.

So we do this, and we've now taken this noise image, and we've altered it so that it now matches the representation for some natural texture. And remember, the whole point of doing this is to actually test whether that representation that we're matching is a good model for the texture representation in our brain. So if the texture representation that we're doing the matching with captures what's in our brain's representation of texture, then the synthetic examples ought to look like the original texture.

Now, there's one other little detail here that I'll tell you about, which is that we're going to actually see this done on color images. So color images are slightly more complicated than the black and white image that I just showed you because there are three color channels. Remember, because there are three cones, three color channels are sufficient to recreate the perception of all colors.

And so there's a little trick that's used to actually do this on color images, which is that you do this in the principal component space of the RGB values. So you essentially work with color axes that are orthogonal or uncorrelated. And that means that you can mess with them independently and then add them back up. So that's a small detail.

But the big picture here is to inspect these pairs of images. One of these is the original, so that's the one on the left. And the right image is a synthetic image that was generated from noise just by matching the subband histograms. And so you can see that in some cases, this produces actually pretty good-looking textures right.

So that one looks pretty good. That one maybe looks pretty good. In some cases, they don't look perfect. You can definitely see some things that are kind of funny about them. But there are a lot that look pretty good. Here's some other examples that look pretty reasonable. So maybe you can get something that looks like a leaf or some kind of carpet or something.

But there's a lot of other cases where the synthesis pretty clearly fails. So for instance, what is this? What material is this?

AUDIENCE: Marble.

JOSH MCDERMOTT: Marble, exactly. Everyone recognizes this. It's marble. It looks like marble. This does not look like marble. So it just has a very different appearance. What is this?

AUDIENCE: Hay.

JOSH MCDERMOTT: Yep. That is not hay. I don't know what these things are. Something more exotic, but it also doesn't look the same.

So what does this tell us? Well, this tells us that when you are looking at these images, your brain is generating a representation that differs from the one that we used to do this matching process. So you're representing something above and beyond these subband histograms.

So what's wrong? Like what's missing here? So one thing that you might notice-- so when you look at the marble, for instance, the marble kind of has these long-- these elongated things in it. Some function of the composition of marble causes that to happen. I don't know what, but that's just part of what makes marble look like marble.

And then when you look at the synthetic image, that's kind of missing. And you get the same issue with the hay. The hay has got all these strands of hay, and those are missing in the image.

And so one of the things that characterizes these kind of long, elongated structures is that there are these statistical dependencies between orientation energy at different locations. And this is a little bit like what we talked about when we were talking about the statistical basis for grouping by good continuation, where you can measure the dependencies between orientations at different positions and images. And so edges tend to-- natural images have a lot of elongated edges. So you have an orientation here, you tend to have an orientation here. And so some of these textures have those properties, and that's not really being captured

And so one way that you might imagine being able to capture that is by measuring the correlation between, say, a filter response here, and a filter response here. And this representation doesn't do that. So what does this representation do? Well, it measures-- the filter response is across this whole image. So you get some numbers. And then it just throws them in a bag. And you get the distribution of the bag.

So whether or not like this is active at the same time as this is active is lost in that process. So it's a marginal distribution. You don't really have any information about the relationship between responses at, say, different points in space or different orientations.

So big picture here. This was a very influential initial attempt at this. It was really like the first instance of the use of texture synthesis. It's very simple and elegant, but it's not great as a theory of texture perception.

So as I said, the intuitive explanation of the failure is that there are these higher-order statistical properties that are present in these natural textures that we might plausibly be sensitive to that are missing from this model. So what if we include them? So what is shown here is a picture that you've seen before. This is a canonical model of early vision.

So you start with an image up there, and then you have several stages of simple operations. So there's some center surround filtering kind of like what we think happens in the retina and the LGN. There's some spatial frequency tuned and orientation tuned filtering here, like what we think happens in V1. And then there's what's called envelope extraction, which is sort of like an energy measure. So that's like a complex cell kind of stage. So this is the energy kind of measure.

So this is our standard model of early vision. And the little, red circles that are stuck on top of it are different statistics that are computed from different stages. So the word M stands for a Marginal statistic, so the average or the variance, things like that. And then the things that have a C in them are different types of Correlations.

So AC is an Autocorrelation, so that's a correlation between different points in space. And then these other Cs are correlations between different orientations or different spatial scales. So it's measuring the joint dependence of the responses along these different dimensions. But essentially, it's just a big set of statistics. And all of them are kind of capturing the average properties over space of the image viewed through a certain type of representation.

So what we'd like to do now is the same kind of thing. So we'd like to test this idea using synthesis. So what's shown here is a model whereby you can measure these statistics. So you take your model of early vision, and then we can measure these statistics. They're just mathematical functions of this.

And so what we want to do is now generate a new image starting from some random image that's going to match these statistics. And so because of the fact that we're using correlations, we can't use the sort of simple trick of histogram equalization. So instead, people use gradient-based optimization, which nowadays is like super standard, and like pretty easy to do.

So you could even think of this as a neural network. This is the output of the neural network. And what you're going to do is solve an optimization problem where you want to minimize the difference between the statistics of a synthetic texture-- which is not shown here, so that would be something that's initialized as noise-- and those of a natural texture.

So the natural texture, you get this set of numbers. And you can then write down a loss function, which tells you about the difference between the numbers from this image and the numbers that you would get from a noise image passed through that same model. And then you can compute the gradient of that loss function with respect to the noise image. And that's going to tell you-- the gradient is going to tell you how to change the noise image so as to minimize that loss function. That is, to make the statistics of the noise image become similar to those of the natural image.

So it's a different method of doing the matching, but the scientific purpose is really the same. You're trying to synthesize an image that has the same values of this candidate representation, which, in this case is a bunch of statistics measured on the image

You can think of these that-- when we talk about correlations and stuff, it may sound kind of abstract, but you can think about building this into a pretty neurally plausible model. So this is showing several-- a few different stages by which you could compute these correlations. So you've got these original images. You convolve it with a set of V1-like filters.

And then you could measure products between these filters, and then average those over space, and that will give you a correlation. So correlation is like the average between is the average of a product between two variables, maybe with the mean subtracted out. And so we're going to, again, try to minimize the difference between the statistics of a synthetic texture and those of a natural texture.

So again, it's an optimization problem. And nowadays, we have lots of ways to solve this. Back when this was originally introduced, this was a little bit harder to do. And so what I'm going to show you here are a bunch of examples of synthetic textures that are generated in this way.

So here is an original texture. This is synthesized just using these the marginal histograms of Heeger and Bergen, so it doesn't look that great. And this is synthesized with this newer version from Portilla and Simoncelli that incorporates correlations and other statistics, and it looks like a little bit of a better match.

And here's a bunch of examples. So the top row here shows you three original images. And the bottom row shows you synthetic textures that result from matching the statistics that are measured from the original image. And they look really pretty good.

So you get this thing out that's got all these plus signs in it. You get a pretty respectable-looking brick wall texture. And if you look very closely, you can see that it's not completely faithful. But at a glance, they look really, really similar.

Here's some others. And so in particular, you can see that the synthetic textures that you get out here, it has a lot more of these elongated contours that were missing from the original model. So that's one of the things that those correlations buy you.

Again, it's not perfect. So the beans don't quite look like beans. But again, at a glance, it might look pretty good. And so these models, like one of the strengths of this approach is that you can do this on any sort of image and you'll get something out.

And so you can ask, well, what would the texture representation be for a crowd of people? And it's like this, and that seems like it's kind of missing something. There's definitely cases where maybe it doesn't work quite as well. So you can see some of the remaining inadequacies. But it nonetheless kind of accounts for a lot of what we see when we look at these images.

Here are some other examples where the center thing is like an actual photo of a texture, and then the surround is synthesized from the statistics of the original constrained to agree at the boundary. So you can get this endless New York apartment building on the right, and lots and lots and lots of bell peppers on the left. Any questions about this? Yeah.

AUDIENCE: Is this publicly available software?

JOSH MCDERMOTT: Yes. Although it's probably only available in MATLAB. Yeah. So this particular method was-- the paper came out, I think, in 1999. So yeah, there's a MATLAB toolbox that you can download and try this kind of stuff out. And I mean, there's more recent variants of this that do similar things. But yeah, you can try it out.

And I should emphasize that all of this-- these kinds of things have gotten much easier to implement nowadays. So one of the things that deep learning has resulted in is the widespread availability of really pretty good optimization tools. So things like computing gradients, all that nowadays is kind of automated. It used to be you would actually have to do math if you wanted to do this. Now, you just let the computer do it for you. So yeah, you could probably code this up yourself without an enormous amount of effort.

So the key ideas that I want you to take away from this is the idea that textures are believed to be represented with statistics that capture their average properties over a region of an image. These statistics could plausibly be computed by neurons that are pooling the responses of neurons that have smaller receptive fields, like, again, pooling over some region of space. And we can test theories of texture using synthesis. So that's kind of an important idea.

So we started out the lecture kind of just asking, well, can we account for the perception of texture boundaries? And that turns out to not be that difficult to do, like pretty simple things actually work pretty well for that. But if you actually want to account for the appearance of things, the way that it looks, then having a more complicated representation seems to pay dividends. What questions you got about that? Yep?

AUDIENCE: What exactly do algorithms like texture synthesis actually recreate what you would call texture? Like say, for all the artifacts, which is just the image itself. But if you take a picture of a person and you run a texture synthesis, how much of that is actually texture, and how much is it just like a worse version of the original picture?

JOSH MCDERMOTT: Yeah. So that's a really good question, is-- I think what you're essentially asking is like when we generate the synthetic image, is it possible that we're essentially like recreating the original image? Yeah. So because essentially, like what's happening here is that you can think of this at an abstract level, is you have this image, you make a whole bunch of measurements of that. And now you're generating something that gives you the same values of that measurement. And you might imagine that if there was enough different measurements, that would essentially uniquely specify the original image.

And I think it's true, that if there were enough measurements, that would happen. This is not really close to that regime. I mean, one way to see that is like these models-- I think this one probably has maybe 700 parameters, 700 measurements, so there's 700 statistics. The pixel arrays here, they're probably 256 by 256. What's 256 times 256? I don't know. It's a lot more than 700. But you can also just see from inspection that you're not getting something that's the same as the original.

The other thing that you could do to test this would be to do this procedure with different samples of noise. And different samples of noise will give you different images. So you can essentially generate an infinite number of these samples more or less. Yeah. That's a good question. Any other questions?

So this is where we're going to improvise some audio demos using my microphone. So so far, we've been talking about visual textures, but textures also occur in sound. In sound, textures result from large numbers of acoustic events, and they include lots of things that you hear all the time, like rain and wind or birds in a forest--

[AUDIO PLAYBACK]

[MULTIPLE BIRDS CHIRPING]

[END PLAYBACK]

JOSH MCDERMOTT: --or running water--

[AUDIO PLAYBACK]

[WATER GURGLING]

[END PLAYBACK]

JOSH MCDERMOTT: --or insects.

[AUDIO PLAYBACK]

[MULTIPLE INSECTS CHITTERING]

[END PLAYBACK]

JOSH MCDERMOTT: Crowd noise or applause--

[AUDIO PLAYBACK]

[APPLAUSE]

[END PLAYBACK]

JOSH MCDERMOTT: --or fire. So these things are all over the place in the real world, just as they are for vision. And like what we were talking about when we discussed the sense of hearing in the first part of the class often really had to do with individual events. So these are sound waveforms of like two individual events. I can't remember what this is.

[AUDIO PLAYBACK]

[SQUEAKING]

[END PLAYBACK]

JOSH MCDERMOTT: So that's a baby toy. So these sounds, like they have a beginning and they have an end, and they have a temporal trajectory, and that's what makes them what they are in some sense. By comparison, if we just consider the sound of rain--

[AUDIO PLAYBACK]

[RAIN FALLING]

[END PLAYBACK]

JOSH MCDERMOTT: I mean, at some point the rain started, and hopefully, at some point, it will end. But the beginning and the end, they're not really what make it sound like rain. The qualities that make it sound like rain there, they're just there.

And so in this respect, I mean, textures are what we call stationary. So the essential properties don't change over time. And so you can think of sound textures as the time analog of image textures. In the sense that we think of image textures as being defined by statistical properties over space, sound textures we think of are defined by statistical properties over time. So just to be concrete about what we want to understand here, here's another texture.

[AUDIO PLAYBACK]

[CHATTER]

[END PLAYBACK]

JOSH MCDERMOTT: And here's another one.

[AUDIO PLAYBACK]

[CHATTER]

[END PLAYBACK]

JOSH MCDERMOTT: All right. And it's probably trivially obvious that those are the same kind of thing. Those are just two different excerpts of the same recording of a room full of people talking. But the sound waveforms that correspond to those excerpts are totally different. So there's some property that those sounds share that you immediately extract and that causes you to recognize that those are the same thing. So we'd like to know what that is and what would be different about, say--

[AUDIO PLAYBACK]

[BUZZING]

[END PLAYBACK]

JOSH MCDERMOTT: --which is a very different kind of texture. So what is it that you store and-- what is it that you extract and store about these waveforms to recognize that they're the same kind of thing?

So the key theoretical proposal, and this is echoing what we just talked about with images, is that because they're stationary, textures can be captured by statistics, that, in this case, are time averages of acoustic measurements. And so the proposal is that when you recognize the sound of fire or the sound of rain, you're recognizing these statistics.

So what kinds of statistics might we be measuring? Well, it seems pretty plausible that whatever statistics the auditory system measures are derived from auditory system representations. And this is one pretty standard model of the front end of the auditory system that you probably recall from a month and a half ago or so.

So we've got a sound signal. That's the input to the system. There's a bank of bandpass filters that mimics the frequency selectivity of the cochlea. We then often think that there's a process of extracting the amplitude envelope. There's compression that happens in the cochlea. And then there's often thought to be a second set of filters called modulation filters that operates on the envelope of the cochlear filters. And the output of that are what you might call modulation bands.

So one idea is like, well, maybe we can take this pretty standard model of the front end of the auditory system and measure some statistics from those representations. And so here, again, the M stands for Marginal statistics, and particular moments, so things like the mean and the variance. The C stands for Correlations.

So if these kinds of statistics are going to be useful for recognizing sound textures, minimally, they kind of need to give different values for different sounds. And so we can just take a look at a few of these and get a sense of what they might capture. So let's take a look at what you might get from looking at the marginal statistics in this representation, so the envelopes of cochlear frequency channels.

So this is what we're dealing with here. So this is what we often call a cochleagram. So the y-axis here, you can think of as the cochlear channel, or the place along the cochlea. So remember, cochlea is tuned for frequency. What frequencies are transduced near the base of the cochlea?

AUDIENCE: High.

JOSH MCDERMOTT: High. Yes. Good. What frequencies are transduced near the apex? Low. OK. All right. So we've got cochlea laid out there. Time is on the x-axis. The gray level represents the firing rate or the amplitude in a frequency channel.

If we just take a horizontal slice through the cochleagram, we got the blue curve here, which is like a subband. So that's the response of the filter. And then the red curve here is the envelope of the filter.

And so the marginal statistics are going to capture properties of the marginal distribution of the envelope. The marginal distribution, remember, that's what you get if you take all of these red values, you throw them in a bag, and then you plot their distribution. So this is plotting the frequency of occurrence of different amplitudes here of sound energy in this particular frequency band.

And so what does this tell us? Well, it says that on average, the amplitude is in this particular vicinity, and sometimes it's lower, and sometimes it's higher. So that may not seem super exciting, but if you look at these distributions for different sounds, you see that they look different.

So here, we have the histogram of an envelope for one particular frequency channel at 220 hertz for three different sounds. So the red curve is for noise. The blue curve is for a stream. And the green curve is for geese like honking or-- what do geese do? They quack? Honk?

AUDIENCE: Honk.

JOSH MCDERMOTT: Yeah, they honk. So that's geese honking And so these particular examples, they were chosen so that they would have about the same average value. That's the little vertical line segments at the top. But you can see that the distributions have different shapes.

So in particular, the natural sounds like the stream and the geese, the distributions are wider, so they have higher variance. And they also-- they vary in how skewed they are. So the green one is very positively skewed.

And so if you look at the cochleagrams of these sounds, you can see where these distributions are coming from. So if you look at the distribution for pink noise-- sorry, not the distribution, the cochleagram. The cochleagram is pretty gray. So that means it doesn't deviate a whole lot from the average values.

Whereas the stream, you can see more light and dark gray, fluctuates more in amplitude. And with the geese, there's a fair bit of white and black. So it's varying quite a lot in amplitude. And so in this particular case, this is an indication of the fact that a lot of natural signals are sparser than noise.

And the intuition here is that natural sounds, they contain events, like raindrops or geese calls, and these events are infrequent, or relatively infrequent, but when they occur, they produce large amplitudes. So you get these pretty big swings in the amplitude of the signal. And critically-- so that shows up here in these relatively simple statistical properties of these distributions like the variance and the skew, which are examples of moments.

So the point is that we could measure some fairly simple things right, just variance, for instance, or skew, and differentiate between these different sounds. Let's take a look at what information you might get from correlations. So we talked a little bit about how correlations in images might kind of capture like elongated edges. What about sound?

So this is a cochleagram for fire. So fire has lots of crackles and pops. This is what it sounds like.

[AUDIO PLAYBACK]

[FIRE CRACKLING]

[END PLAYBACK]

JOSH MCDERMOTT: All right. Very crackly. So those crackles and pops are broadband events. So they're like impulsive. So they contain energy at all frequencies.

And so they show up in the cochleagram as these vertical streaks. And those vertical streaks induce dependencies between different frequency channels. So they mean that the amplitude at one frequency is going to be correlated with the amplitude at another frequency.

And so this is a matrix of all those correlation coefficients. So the axes here are the cochlear channels. And so the diagonal here has got to be 1, but the off-diagonals can be kind of whatever. But you can see that for fire, especially kind of either at the very low frequencies or the relatively high frequencies, there's a lot of red. So things are pretty strongly correlated.

And this is not true of all sounds. So in particular, a lot of water sounds are actually pretty not correlated. So here is a stream. I don't have the audio demo for that, but I played you that earlier. So this is a stream, and you can see that it's not quite as streaky. And then if you look at the correlation matrix for a stream, you can see that things are more green, which corresponds to something pretty close to zero.

So the point is that this correlation that you could measure between different frequency channels has got very different values depending on what kind of sound that you listen to. So things that are more crackly are going to have more correlation. Finally, just one more example here. Let's look at what kind of information you might get from measuring a correlation between the outputs of these modulation filters. So this is kind of one stage deeper in the auditory system.

So these correlations are going to be measured between a particular modulation filter applied to different frequency channels. So this guy, and this guy. And so you get these correlation matrices, but now you have a whole bunch of them. Each one kind of corresponds to a particular modulation filter.

And the main thing to take away from this is that here, these statistics are plotted for waves-- that's on the left-- and fire on the right. And the main thing to take away from this is that they look different. In particular, fire has got a lot more of these correlations at fast modulation rates, like 50 hertz and 100 hertz, whereas waves basically doesn't. It's only got them at the slow modulations. So this is just a statistical acoustic fact about these types of natural sounds.

So the big picture here is that this little tour of these statistics that we can measure from our model of the auditory system capture variation across sound. They got different values for different types of sounds. So now what we're going to ask here is whether they can account for the perception of real-world textures.

And so our key methodological proposal, again echoing what we kind of just talked about for visual texture, is that synthesis is a powerful way to test a perceptual theory. And so the logic here is that if your brain represents sounds with a set of measurements, then signals that have the same values of those measurements ought to sound the same to us. And so in particular, sounds that we synthesize to have the same measurements as some particular real-world recording ought to sound like that real-world recording if the measurements that we use to do the synthesis are like the ones that the brain is using to represent sound.

So I'm going to walk you through a really simple example, just to make sure everybody's on board with the logic here. So let's suppose that we have the world's simplest theory of texture perception, which is that our representation of texture consists of something that approximates the power spectrum. So the power spectrum is a super standard thing to measure with sound. On some of your earlier problem sets, you actually plotted the power spectrum.

In this model, the power spectrum is approximated by the mean of each cochlear envelope. So remember, we've got our bank of cochlear filters. The envelope is the instantaneous amplitude. And so if you just average that over time, it's going to tell you about how much power is in that particular frequency channel.

So let's suppose we have 30 cochlear filters. We're going to get 30 numbers, one for each frequency channel. So the procedure here, in order to test this theory, is first, we're going to measure this quantity, this proposed representation-- so in this case, the average value of each envelope in a real-world texture like a recording of water. We're then going to synthesize a random signal that's got the same values for those envelope means.

And so in this particular case, this is actually pretty easy to do. So we're going to start with noise. We will pass the noise through a bank of bandpass filters that mimics the cochlea, so we get noise subbands. We're then going to just rescale each one of those so that it's got the correct average power. And then we can add those back up, and we can get a new signal

And there's a few little details of the procedure for doing this to do it correctly that I'm leaving out. You actually need to do this iteratively, but it's still very, very simple. So now what we want to do is listen to the sounds, and just ask ourselves whether they sound like the real thing. So I'm going to play you some synthetic sounds that are generated from noise just by matching this proposed representation to that of each of the originals. So here's what they sound like.

[AUDIO PLAYBACK]

[MID-FREQUENCY STATIC]

[END PLAYBACK]

[AUDIO PLAYBACK]

[LOWER-FREQUENCY STATIC]

[END PLAYBACK]

[AUDIO PLAYBACK]

[LOW-FREQUENCY STATIC]

[END PLAYBACK]

[AUDIO PLAYBACK]

[VERY LOW-FREQUENCY STATIC]

[END PLAYBACK]

[AUDIO PLAYBACK]

[MID-FREQUENCY STATIC]

[END PLAYBACK]

JOSH MCDERMOTT: Now each of those sounded different, and you may have even actually been able to convince yourself that, yeah, that maybe sounds a little bit like applause. But they don't sound like the real thing.

[AUDIO PLAYBACK]

[APPLAUSE]

[END PLAYBACK]

JOSH MCDERMOTT: Or--

[AUDIO PLAYBACK]

[WATER GURGLING]

[END PLAYBACK]

JOSH MCDERMOTT: So we have ruled out the power spectrum theory of texture with this exercise. And now the question is-- so since this is not realistic, everything pretty much sounds like noise. And the question is, well, we can do better with some additional simple statistics? So more of these red circles.

So just as kind of was the case when we were talking about image textures, so the fact that we're using this big set of statistics means that the synthesis process has to work a little bit differently. And it's, again, going to be gradient-based optimization. So there's this initial step where you take your original sound, you pass it through your auditory model and you measure its statistics.

And then there's a synthesis procedure where you take a noise signal, you initially measure its statistics, and you get an error signal here. So this is the loss function you're trying to minimize, which is the difference between the statistics of the noise and the statistics of rain or fire or whatever it is. And then you're going to try to impose that with the gradient on the original thing

And again, back when we originally did this stuff, this was a little bit more complicated than it is now, and it's pretty straightforward to do nowadays. And there is MATLAB code for this too if you're interested. Lots of MATLAB code from back in the day.

So the result of this procedure is that you get a signal that shares the statistics of some real-world sound. And again, what we're interested in is what they sound like. And again, the reason that we care about what they sound like is that if the statistics account for texture perception, then the synthetic signals that we synthesize in this way, they ought to sound like new examples of the real thing.

And so what I'm going to play you here are a whole bunch of examples of synthetic sounds that are generated from noise just by causing the noise to match the statistics of each of these examples. And in many cases, they actually sound pretty good.

[AUDIO PLAYBACK]

[RAIN FALLING]

[END PLAYBACK]

JOSH MCDERMOTT: So that's pretty good rain.

[AUDIO PLAYBACK]

[WATER GURGLING]

[END PLAYBACK]

JOSH MCDERMOTT: The stream.

[AUDIO PLAYBACK]

[BUBBLES POPPING]

[END PLAYBACK]

[AUDIO PLAYBACK]

[FIRE CRACKLING]

[END PLAYBACK]

JOSH MCDERMOTT: Fire.

[AUDIO PLAYBACK]

[APPLAUSE]

[END PLAYBACK]

JOSH MCDERMOTT: Applause.

[AUDIO PLAYBACK]

[WIND BLOWING]

[END PLAYBACK]

JOSH MCDERMOTT: Wind. Here's insects.

[AUDIO PLAYBACK]

[MULTIPLE INSECTS CHITTERING]

[END PLAYBACK]

JOSH MCDERMOTT: Birds.

[AUDIO PLAYBACK]

[MULTIPLE BIRDS CHIRPING]

[END PLAYBACK]

JOSH MCDERMOTT: And crowd noise.

[AUDIO PLAYBACK]

[AMBIENT CROWD NOISE]

[END PLAYBACK]

JOSH MCDERMOTT: OK. So with marginal moments and these pairwise correlations, the synthesis is often pretty compelling. It also works for a lot of things that you might consider to be unnatural sounds. I'll play you a few examples. Here's rustling paper.

[AUDIO PLAYBACK]

[PAPER RUSTLING]

[END PLAYBACK]

JOSH MCDERMOTT: That's holding a piece of paper and going like that. And the jackhammer.

[AUDIO PLAYBACK]

[JACKHAMMER HAMMERING]

[END PLAYBACK]

OK. So the success of the synthesis suggests that these statistics could underlie the representation and the recognition of textures. So there's also this kind of interesting possibility, which is that if we take this idea seriously, that texture is represented with statistics that kind of capture the time-averaged properties, then if really, your brain is just representing those time-averaged statistics, then I think the conclusion that follows from that is that different exemplars of the same texture should be difficult to tell apart.

So these are two different excerpts of fire that were synthesized from two different samples of noise. If you look closely, you can see that the patterns of the vertical streaks, so the clicks and the pops, are different in the two excerpts, but they also kind of look sort of the same. They've got the same texture.

And so this is a graph that is plotting the standard deviation of one particular statistic-- you would see a similar graph regardless of which statistic you measured. But this is one example statistic measured in different excerpts of different lengths. And so what this shows is that if you take very short excerpts of like 40 or 80 milliseconds, and you measure the statistic in different excerpts, the standard deviation is high.

So the value of the statistic that you will measure from different excerpts will be different. Because you're getting these small samples, and sometimes maybe you have a couple of clicks, and other times maybe you don't, for instance. But what this shows is that as the excerpts become longer and get out to a second or two, the standard deviation becomes quite small.

So what that means is that the statistics that you measure from these different excerpts of a texture, they converge to the same values. So again, the idea is that if you do an experiment where you ask people to try to tell these things apart, well, the prediction here is that if they're just representing the statistics, they should be able to tell them apart when the excerpts are short, but not when they're long.

So first, we're going to do kind of a vanilla experiment, just measuring the ability to discriminate different textures. So these are signals that have different long-term statistics. So here's the task. So people are hear three excerpts of sound-- examples are shown up top.

Two of the excerpts-- in this case, it's the second and the third-- are different excerpts of the same texture. So it could be rain. The third excerpt, which in this case is first, but it can either go first or last, is an excerpt of a different texture. So it could be fire. And the task that the person has to perform is to say which sound was produced by a different source. In this case, the answer would be first, because it's a different texture.

And this graph is going to show you how accurately people can perform this task. So the y-axis plots proportion correct as a function of the duration of the stimulus excerpts. And what you see is not super surprising, is that as the excerpts get longer, people get better. It's not surprising in the sense that as the stimulus excerpts get longer, you're giving people more information with which to do the task in which to tell that these two things have something in common, and that first one is different.

So this is kind of standardly what you would expect to get on most psychophysical tasks, is that when the stimuli are longer, people are going to be better. And in this particular case, it's certainly consistent with the idea that people are using statistics because, as the excerpt gets longer, the statistics of these two will kind of converge to the same values and be reliably different from the statistics of the first one.

So now, the really interesting experiment is this one. So in this particular case, people hear three excerpts of the same texture. Two of them, in this case, it is the second and third, are physically identical. So it's the exact same excerpt of rain. The first one is a different excerpt of rain, for instance.

And the task is just to say which sound was different from the other two. So this is what's called exemplar discrimination. And so the idea here is that if, when you listen to these texture sounds, you're representing them with statistics, well, then, as the excerpts get longer, the statistics of these three sounds should all converge to very similar values. And it should be difficult to tell which one is different from the other two. Whereas when the excerpts are short, as we just saw, the statistics will tend to be different because you have small samples, and so you ought to be able to tell them apart.

So there's this kind of wacky prediction that for this particular task, if the hypothesis is correct that you're representing these sounds just with statistics, then performance ought to get worse as the sounds increase in duration. And that's, in fact, what happens. So you can see that people are very good at this task at 90% correct when the excerpts of the textures are very short. But then as they get longer, up to a few seconds, people get worse and worse.

So let's just think about what this means. So you might imagine that you can't really tell one example of rain from another example, just because they all sound the same. Maybe they don't actually have discriminable detail that you could use to tell them apart. But we can rule that out, because when you just give people these very, very short excerpts of rain, or fire, or whatever it is, people can very accurately discriminate them. They can tell whether they're the same or different.

So that says that these sounds do contain discriminable detail. It's just somehow, when the duration gets long, you're not able to access it. You might also imagine that-- so this is a discrimination task with 22.5 second-long stimuli. You might think it's really hard to do tasks with long stimuli. But on the other hand, when it's a texture discrimination task, and one of the stimuli has different statistics from the other, people are really good at that. So it's not like there's some issue with doing tasks with things that are long.

Moreover, the fact that performance on this texture discrimination task gets better with duration indicates that the details that are present in these sounds, they do get accumulated. So the longer the sound is, the better you are at the task. So all that details that's kind of streaming into your ear, you're able to use that for performance on a texture discrimination task.

So it seems like they contribute to your estimate of the statistics. And when the task is based on differences in statistics, you get better and better. But it seems that those details are not otherwise retained. So that when you have different excerpts of the same texture, you have a hard time telling them apart.

So all of this is consistent with the idea that when you hear texture-like things, your brain turns them into a statistical representation that kind of averages information over time, and you lose access to the detail, even though that detail is used to compute the statistics. What questions you got about this? Yeah.

AUDIENCE: Well, it's about discrimination, is it strictly decreasing with the excerpt duration, or is that supposed to be a turning point where it initially gets better and then it gets worse?

JOSH MCDERMOTT: Yeah, I mean, if you made this short enough, you're going to get worse, because at some point they'll-- like if they get short enough, they'll probably all just sound like clicks, is my guess, and they'll get harder to discriminate. Yeah. But the experiment didn't go down that short. Yeah.

Yeah. I mean 40 milliseconds is still-- that's like a blink of an eye. So that's pretty brief. But it's possible that if you went down to 10, at some level, if you just end up with one sample, they're all going to be clicks.

So you might also be worried about that maybe what's going on here is related to memory. So this is what the stimuli actually look like on a long-duration trial versus a short-duration trial right. So if the inner stimulus interval is the same, you might worry about the fact that, well, when the stimuli are really long, the bits that you have to compare are kind of separated like further in time, and so maybe it's hard to remember, and that kind of messes you up.

So you can do a control experiment where you just kind of space the short ones out in time, and that has very little effect. So you're still much, much better when they're short than when they're long. So the take-home message here is that the results suggest that people are using a representation of time-averaged statistics. And again, based on this phenomenon that the statistics will converge as the sample size increases. In this case, the sample size is determined by duration.

So textures in the real world are normally generated from a superposition of sources. So you have rain, it consists of a lot of raindrops, or insects, it's a lot of insects, things like that And so, for instance, like here's one person talking.

[AUDIO PLAYBACK]

- [GERMAN]

[END PLAYBACK]

JOSH MCDERMOTT: And here's 29 people talking. This is a German cocktail party.

[AUDIO PLAYBACK]

[MULTIPLE CONVERSATIONS IN GERMAN]

[END PLAYBACK]

JOSH MCDERMOTT: And that's a texture. So we can generate, in principle, kind of a continuum of sounds, where on one end you have a single source, and on the other hand, you have a lot of sources. And we can ask, well what happens to this phenomenon of being dependent on statistical representations as a function of the extent to which the sounds are really texture-like?

And so here's what happens. So this is a graph that-- it's the same task. This is exemplar discrimination, but with four types of signals. So here, you have something that's very texture-like-- it's 115 speakers at this party-- all the way up to one speaker. So the graph plots proportion correct versus excerpt duration. So here, there's just two durations, one that's short and one there's one that's long.

And so the black curve is just kind of a replication of the thing that we saw previously for textures, which is that you have a hard time discriminating excerpts when they're long, consistent with the idea that you become reliant on statistics. But what's pretty interesting is that when the density of this mixture is varied, the behavior really changes. And so in the extreme case where you have one speaker, you actually get better at the task as the duration increases.

So with a single speaker, you're not completely reducing the representation to these time average statistics, which at some level, is not very surprising, because speech has got a lot of temporal structure, and allows you to understand what someone says, is the temporal sequence of the phonemes and syllables that they utter. But you can also see that there's this gradual transition. So as the thing becomes more and more texture-like, you seem to become more and more dependent on this statistical representation. And in all of the cases at the short durations, you're about equally good at discriminating the sounds.

So we can ask whether this is specific to speech. So here's the same idea, but with drum hits. So this is a very sparse sequence of drum hits.

[AUDIO PLAYBACK]

[SPARSE DRUM HITS]

[END PLAYBACK]

JOSH MCDERMOTT: And here's, it's more dense.

[RAPID DRUM HITS]

[END PLAYBACK]

JOSH MCDERMOTT: And you get the same phenomenon. So the very sparse signal, five hits per second, shows an improvement in this discrimination task with duration, whereas the dense ones show this decrease in discrimination.

So in all these, cases, again the same points apply. So the high performance with these short excerpts indicates that all of the stimuli contain discriminable variation. It's not the case that you can't discriminate these things, just because there isn't discriminable detail that's contained in the sound. So they all contain discriminable variation, but the temporal detail is not retained when the signals are texture-like.

So that seems to be a fundamental thing about the representation of textures. And we think there's probably similar principles that would apply for visual textures, but only across space. So as the size of a texture would increase, you'd probably see similar decreases in the ability to discriminate examples. Questions about that?

So one other question you might wonder about is, will any set of statistics do? So we just kind of made up this set of statistics, but it was based on our best bet for a model of the early stages of the auditory system. So there are these three stages of this model. There are these filters that are based on the cochlea. There's this compressive nonlinearity that is motivated by the compression that we think happens in the ear. And then we've got these modulation filters, the second stage of filters.

And we could ask, well, what happens if we actually change this in a way to make it less consistent with what we know about biology? So this is called a log-spaced filter bank. You all know that the filter bank in the ear is not exactly log spaced, but so it's kind of approximately log spaced. But remember, the filters in the ear, they get broader at high frequencies than at low frequencies.

So you can ask, well, what happens if we use an incorrect auditory model where the filters are linearly spaced, like they're all the same bandwidth? So this violates what we know about the ear. Similarly, we could take that compressive nonlinearity that happens in the ear and get rid of that and end up with different auditory models, measure their statistics, and then ask if it makes a difference in how the textures sound.

And so this was measured with an experiment where people heard the original texture, and then two synthetic versions. One of the synthetic versions came from the biologically correct model, or at least our best bet at what's biologically correct. And the other one came from an alternative model.

And so there were four alternative models, one where we got rid of the compression, one where we made the cochlea linearly spaced, one when we made the modulation filters linearly spaced, and the one where we did all three. And so the prediction here is that it doesn't really matter what model you use and what statistics you're measuring. If you just need to measure a big set of statistics, then people shouldn't have, on average, a preference for the textures from one model over another. And so they should be right at 50%, that dashed line.

But instead what happens is that in every case, people systematically prefer the biologically plausible model. And here's just a few examples. So this is crowd noise generated from the biologically plausible model.

[AUDIO PLAYBACK]

[CROWD NOISE]

[END PLAYBACK]

JOSH MCDERMOTT: And from the non-biological one.

[AUDIO PLAYBACK]

[DISTORTED CROWD NOISE]

[END PLAYBACK]

JOSH MCDERMOTT: Probably here it sounds kind of weird and a little bit garbled or something. And here's one more example, a helicopter.

[AUDIO PLAYBACK]

[HELICOPTER WHIRRING]

JOSH MCDERMOTT: Can you hear that?

AUDIENCE: Yeah.

[END PLAYBACK]

JOSH MCDERMOTT: And then here's from the nonbiological model.

[AUDIO PLAYBACK]

[WARBLY HELICOPTER WHIRRING]

[END PLAYBACK]

JOSH MCDERMOTT: So you can hear the modulations don't quite sound right. And that's probably because the modulation filter bank is kind of not spaced in the way that ours is. And so our sensitivity to modulation is mismatched with that of the model.

And so the idea here is that the procedure is initialized with noise. So we get a different sound every time, sharing only the statistical properties. And the statistics define a class of sounds. This is same idea we talked about with visual textures. There's an equivalence class that's defined by these statistics.

And so the idea is that if the statistics measure what the brain is measuring, then the samples should sound like another example of the original sound. But if the statistics don't measure what the brain is measuring, then we get a different equivalence class. So in this case, we think the statistics of that nonbiological model, they define a different class of sound. And when we run the synthesis procedure-- so both sets of sounds, like the set for what our brain is using, which is symbolized in blue, and the set for this incorrect model in red, both of them contain the original, because the original defines the set of statistics along with the model that defines that set.

And when we run the synthesis, we're pulling an example from that set. And so if we've got the wrong set of statistics, we get the wrong equivalence class, and you get something that just doesn't sound right to us. So the point is that it does actually kind of matter to use a model that we think is correct.

Last thing I want to show you, this is an experiment where people were played an original sound, and then a synthetic sound, like generated from the biologically plausible model. And they were just asked to rate the realism of the synthetic sound like on a scale of 1 to 7. And they did this for 170 sounds.

So this is a histogram that just plots the distribution of the average realism ratings for each one of those 170 sounds, like averaged across a bunch of participants. There's just two things to take away from this. The first is that there's a big bump up here kind of at the top. So that is encouraging, because it means that for a lot of sounds, people are saying that the synthesis is fairly realistic.

But the second thing is that there's this tail down here. And this is quite interesting because these are cases where, despite the fact that we have matched these sounds along this big set of statistics, people are saying that they don't sound anything like the original. So it means that our brain is measuring something that the model is not. And so these failures, again, always give us kind of interesting clues for how we can make our theories better.

And so this is a list of the 15 or so sounds that got the lowest rating. And just to make it easy on you, I'll put some labels next to them, because they tend to have one of three things. So they tend to either be sounds that involve pitch, sounds that have some kind of rhythm, or sounds that have reverberation. And so I'll play you both the original and the synthetic version, just to give you a flavor.

[AUDIO PLAYBACK]

[RAILROAD BELLS]

[END PLAYBACK]

[AUDIO PLAYBACK]

[DISTORTED RAILROAD BELLS]

[END PLAYBACK]

[AUDIO PLAYBACK]

[BANGING]

[END PLAYBACK]

[AUDIO PLAYBACK]

[DISTORTED BANGING]

[AUDIO PLAYBACK]

[TAPPING]

[END PLAYBACK]

[AUDIO PLAYBACK]

[DISTORTED TAPPING]

[END PLAYBACK]

JOSH MCDERMOTT: So there's some taps there, but the rhythm is kind of lost. And I should emphasize that it wasn't really obvious before doing the whole exercise of synthesizing the sounds that these statistics would not capture rhythm. We actually thought there was a good chance that they might. But then you go and do it, and you're forced to face the truth. It really doesn't do a great job with music.

[AUDIO PLAYBACK]

[MAMBO MUSIC]

[END PLAYBACK]

JOSH MCDERMOTT: And here you go.

[DISTORTED MAMBO MUSIC]

[END PLAYBACK]

JOSH MCDERMOTT: So these failures are kind of exciting in the sense that they point the direction to things that we need to understand. And so this is the value of having models that can be applied to any sound. That's what we aspire to in perceptual science, because they force you to face the truth.

[AUDIO PLAYBACK]

[DISTORTED SPEECH]

[END PLAYBACK]

JOSH MCDERMOTT: So take-home messages. We talked about the idea that synthesis can help us test and explore theories of perception, and we saw that applied to texture in both vision and audition. Variables that produce compelling synthesis are things that could plausibly underlie perception. Synthesis failures are really interesting because they point the way to new variables that might be important for the perceptual system.

We saw evidence that many natural sounds may be recognized with relatively simple statistics of early auditory representations. So the very simplest statistics capturing things like the spectrum and are not that informative, but if you go a little bit more complex, you get representations that are pretty powerful. We saw evidence that for sufficiently texture-like sounds, statistics may be all you have, which actually makes it difficult to discriminate different excerpts of different textures.

And I'll just highlight that textures are one domain where I think we have models that do a pretty good job of accounting for perception. And we're not quite this far in really any other aspect of perception at this point. So I'm going to end there, but I'm happy to take questions. What questions do you have? Yeah.

AUDIENCE: This might be kind of like a weird question, but I feel like texture makes a lot of sense in the modalities of like hearing and seeing, but are there other sensory textures? Are there like touch textures?

JOSH MCDERMOTT: Oh yeah. I mean, it's super important for touch. I mean, it's one of the most important thing. I mean, just think about feeling different materials. I mean, a lot of what happens with touch is you're sweeping usually your fingertips like over some surface, and getting some statistical sense of the textural properties. Yeah, so texture is really important for touch.

I mean, it's really important for eating and drinking as well, but that's also touched just inside your mouth, so maybe interacts with flavor and taste in some kind of complicated way. But yeah, vision, hearing, and touch, it's important in all of them. Yeah.

And I think all of these same ideas, they could be applied to touch. It's just a lot harder because it's much harder to generate the stimuli. And this is one of many reasons why there's lots of work on vision and hearing and less on touch. It's just like just manipulating stimuli with touch is challenging-- not impossible, but hard. What other questions you got? Yeah?

AUDIENCE: Can you play the English one, the last one?

[AUDIO PLAYBACK]

- A boy fell from the window. The wife helped her husband. Big dogs can be dangerous. Her shoes were very dirty. The--

[END PLAYBACK]

[AUDIO PLAYBACK]

[DISTORTED SPEECH]

[END PLAYBACK]

JOSH MCDERMOTT: And that's also interesting, because on the one hand, it's probably totally obvious that is supposed to be speech. So speech definitely has distinctive properties in this kind of texture representation. But it's also missing a whole bunch of stuff that is present in normal speech. So speech and music in particular involve more complicated measurements. And I think if we had the right model, we could probably generate speech and music textures in the same way.