JOSH MCDERMOTT: We're going to get started again. So where we stopped, I had just played you some of the results of this text, your synthesis algorithm. We all agreed that they sounded pretty realistic. And so the whole point of this was that this gives plausibility to the notion that you could be representing these textures with these sorts of statistics that you can compute from a model of what we think encapsulates the signal processing in the early auditory system.

And, again, I'll just underscore that the sort of cool thing about doing the synthesis is that there's an infinite number of ways in which it can fail, right. And by listening to it and convincing yourself that those things actually sound pretty realistic, you actually get a pretty powerful sense that the representation is sort of capturing most of what you hear when you actually listen to that natural sound, right?

And for instance, we could design a classification algorithm that could discriminate between all these different things, right. But the point is that they could-- the representation could still not capture all kinds of different things that you would hear. And by synthesizing, because of the fact that you can potentially fail in any of the possible ways, right. And then listen and observe whether the failure occurs. You get a pretty powerful method.

All right. But one thing that you might be concerned about. And this is sort of something that was annoying me, right, is that what we've done here is we've imposed a whole bunch of statistical constraints, right. So we're measuring like this really large set of statistics from the model, right. And then generating things that have the same values of those statistics.

So there's this question of whether any set of statistics will do. And so we wonder what would happen if we measured statistics from a model that deviates from what we know about the biology of the ear. So, in particular, you remember that in this model that we set out, there were a bunch of different stages, right. So we've got this initial stage of bandpass filtering. There's the process of extracting the envelope and then applying amplitude compression. And there's this modulation filtering.

And in each of these cases, there are particular characteristics of the signal processing of that's explicitly intended to mimic what we see in biology. And so in particular as we noted the kinds of filter banks that you see in biological systems are better approximated by something that's logarithmically spaced than something that's linearly spaced. So we remember-- remember that picture I showed at the start, where we saw that the filters up here were a lot broader than the filters down here, all right.

OK, and so we can ask, well, what happens if we swap in a filter bank that's linearly spaced. It's sort of more closely analogous to like an FFT, for instance. Similarly we can ask, well, what happens if we kind of get rid of this nonlinear function here that's applied to the amplitude envelope. And we make the amplitude respond to linear instead. And so we did this. So you can change the auditory model and play the exact same game. So you can measure statistics from that model and synthesize something from those statistics. And then ask whether they sound any different.

And so we did an experiment. So we would play people on the original sound. And from that original sound, we have two synthetic versions. One that's generated from the statistics of the model that replicates biology as best we know how. On the other of it that is altered in some way. And we would ask people which of the two synthetic version sounds more realistic?

And so there's four conditions in this experiment because we could alter these models in three different ways. So we could get rid of amplitude compression-- that's the first bar. We could make the cochlea linearly spaced. Or we could make the modulation filters linearly spaced. Or we could do all three, and that's the last condition. And so it's being plotted on this axis-- whoops I gave it away-- is the proportion of trials on which people said that the synthesis from the biologically plausible model was more realistic.

And so if it doesn't matter what statistics you use, you should be right here at this 50% mark in each of these cases. And as you can see in every case, people actually report them, on average, the synthesis from the biologically plausible model is more realistic. And I'll give you a couple of examples. So here's crowd noise synthesized from the biologically plausible auditory model.

[CROWD NOISE]

And here's the result of doing the exact same thing but from the altered model. And this is from this condition here where everything is different. And you, a little here, it just kind of sounds weird.

[CROWD NOISE]

It's kind of garbled in some way. So here's a helicopter synthesized from the biologically plausible model.

[HELICOPTER NOISE]

And here's from the other one.

[HELICOPTER NOISE]

Sort of-- doesn't sound like the modulations are quite as precise. And so the notion here is that-- we're initializing this procedure with noise. And so the output is a different sound in every case that are sharing only the statistical properties. And so the statistics that we measure and used to do the synthesis, they define a class of sounds that include the original that, in fact, defines a set as well as a whole bunch of others.

And when you run the synthesis, you're generating one of these other examples. And so the notion is that if the statistics are measuring what the brain is measuring, well, then, these examples ought to sound like another example of the original sound. You ought to be generating sort of an equivalence class. And the idea is that when you are synthesizing from statistics of this non-biological model where it's a different set, right?

So, again, it's defined by the original. But it contains different things. And they don't sound like the original because they're presumably not defined with the measurements that the brain is making. I just mentioned to you the fact that the procedure will generate a different signal in each of these cases. Here you can see the result of synthesizing from the statistics of a particular recording of waves. These are three different examples. And if you sort of inspect these, you can kind of see that they're all different, right? They sort of have peaks and amplitude in different places and stuff.

But on the other hand, they all kind of look the same in a sense that they have the same textural properties, right? And that's what's supposed to happen. And so the fact that you have all of these different signals that have the same statistical properties raises this interesting possibility, which is that if the brain is just representing time average statistics, we would predict that different exemplars of a texture ought to be difficult to discriminate.

And so this is the thing that I'll show you about next is an experiment that attempts to test whether this is the case to try to test whether, really, you are, in fact, representing these textures with statistics that summarize their properties by averaging over time. And in doing so, we're going to take advantage of a really simple statistical phenomenon, which is that statistics that are measured from small samples are more variable than statistics measured from large samples.

And that's what is exemplified by the graph that's here on the bottom. So what this graph is plotting is the results of an exercise where we took multiple excerpts of a given texture of a particular duration. So you're 40 milliseconds, 80, 160, 320. So we get a whole bunch of different excerpts of that length. And then we measure a particular statistic from that excerpt. So in this case it's a particular cross correlation coefficient for the envelopes of a pair of sub-bands.

So we're going to measure that statistic in those different excerpts. And then we're just going to try to see how variable that is across excerpts. And that's summarized with a standard deviation of the statistic. And that's what's plotted here on the y-axis. And so the point is that when the excerpts are short, the statistics are variable. So you measure it in one excerpt and then another and then another. And you don't get the same thing, all right. And so the standard deviation is high.

And as the excerpt duration increases, the statistics become more consistent. They converge to the true values of the station underlying stationary process. And so the standard deviation kind of shrinks. All right, and so we're going to take advantage of this in the experiments that we'll do.

All right, so first to make sure that or to give plausibility to the notion that people might be able to base judgments on long-term statistics, we ask people to discriminate different textures. So these are things that have different long-term statistics. And so in the experiment, people would hear three sounds, one of which would be from a particular texture like rain. And then two others of which would be different examples of a different texture like a stream.

So you'd hear a rain-- stream one-- stream too. And the task was to say which sound was produced by a different source. And so in this case, the answer would be first-- all right. And so we gave people this task. And we manipulated the duration of the excerpts. And so the notion here is that while given this graph, what happens is that the statistics are very variable for short excerpts. And then they become more consistent as the excerpt duration gets longer.

And so if you're basing your judgments on the statistics computed across the excerpt, well, then you ought to get better at seeing whether the statistics are the same or different as the excerpt duration gets longer. All right, and so what we're going to plot here is the proportion correct that this task is a function of the excerpt duration. And, indeed, we see that people get better as the duration gets longer. So they're not very good when you give them a really short clip. But they get better and better and as the duration increases.

Now, of course, this is not really a particularly exciting result. When you increase the duration, you give people more information. And pretty much on any story, people ought to be getting better, right? But it's at least consistent with the notion that you might be basing your judgments on statistics.

Now the really critical experiment is the next one. And so in this experiment, we gave people different excerpts of the same texture. And we asked of them to discriminate them. So again on each trial, you hear three sounds. But they're all excerpts from the same texture. But two of them are identical. So in this case, the last two are physically identical excerpts of, for instance, rain. And the first one is a different excerpt of rain. And so you just have to say, which one is different from the other two?

All right, now but maybe the null hypothesis here is what you might expect if you gave this to a computer algorithm that was just limited by sensor noise. And so the notion is that as the excerpt duration gets longer, you're giving people more information with which to tell that this one is different from this one. So maybe if you listen to just the beginning, it would be hard. But as you got more information, it would get easier and easier.

If in contrast you think that what people represent when they hear these sounds are statistics that summarize the properties over time. Well, I've just shown you how the statistics converge to fixed values as the duration increases. And so if what people are representing are those statistics, you might paradoxically think that as the duration increases, they would get worse at this task-- all right. And that's, in fact, what we find happened.

So people are good at this task when the excerpts are very short on the order of, like, 100 milliseconds. So they can very easily tell you whether you're-- which of the two excerpts is different. And then as the duration gets longer and longer, they get progressively worse and worse.

And so we think this is consistent with the idea that when you are hearing a texture, once the texture is a couple of seconds long, you're predominantly representing the statistical properties averaging the properties over time. And you lose access to the details that differentiate different examples of rain so that the exact positions of the rain drops or the clicks of the fire, what have you.

Why should people be unable to discriminate two examples of rain? Well, you might think, well, these textures are just homogeneous, right? There's just not enough stuff there to differentiate them. And we know that that's not true, because if you just chop out a little section at random, people can very easily tell you whether it's the same or different, right. So at a local time scale, the details is very easily discriminable.

You might also imagine that what's happening is that over time, maybe there's some kind of masking in time. Or that the representation kind of gets blurred together in some strange way. On the other hand, when you give people sounds that have different statistics, you find that they're just great, right. So they get better and better as the stimulus increases.

And, in fact, the fact that they continue to get better, that seems to indicate that the detail that is streaming into your ears is being accrued into some representation that you have access to. And so we think what we think is happening is that those details come in. They're incorporated into your statistical estimates. But the fact that you can't tell apart these different excerpts means that there's these details are not otherwise retained. All right, so they're accrued into statistics, but then you lose access to the details on their own.

The point is that the result as it stands, I think, provides evidence for a representation of time average statistics. So that is that when the statistics are different, you can tell things are distinct. When they're the same, you can't, and relates to this phenomenon of the variability of statistics as a function of sample size. So a couple control experience that are probably not exactly addressing the question you just raised but maybe are related.

So one obvious possibility is that the reason that people are good at the exemplar discrimination here when the excerpts are short and bad here might be the fact that maybe your memory is decaying with time. All right, so the way that we did this experiment originally was that there was a fixed inner stimulus interval, so that was the same. It was couple of hundred milliseconds in every case.

And so to tell that this is different from this, the bits that you would have to compare are separated by a shorter time interval than they are in this case, right, where they're separated by a longer time interval. And if you might just imagine that memory decays with time, you might think that would make people worse. So we did a control experiment where we equated the inner onset interval. All right, so that the elapsed time between the stuff that you would have to compare in order to tell whether something was different was the same in the two cases.

And that basically makes no difference, right. You're still a lot better when the things are short than when they're long. And we went to pretty great lengths to try to help people be able to do this with these long excerpts. So you might also wonder, well, given that you can do this with the short excerpts. With the short excerpts are really just analogous to the very beginning of these longer excerpts. All right, so why can't you just listen to the beginning?

And so we tried to help people do just that. So in this condition here, we put a little gap between the very beginning excerpt and the rest of the thing, right? And we just told people, all right, there's going to be this little thing at the start-- just listen for that. And people can't do that. So we also did it at the end-- so if at the gap at the end.

So again, you get this little thing of the same length as you have in the short condition. And this is performance, in this case, some people are good when it's short and a lot worse when it's longer. And the presence of a gap doesn't really seem to make a difference, right. So you have great trouble accessing these things.

Another thing that's sort of relevant and related was these experiments that resulted from our thinking about the fact that textures are normally not generated from our synthesis algorithm, but rather from the super position of lots of different sources. And so we wondered what would happen to this phenomenon if we varied the number of sources in a textures. So we actually generated textures by superimposing different numbers of sources.

So in one case we did this with speakers. So we wanted to get rid of linguistic effects. And so we used German speech and people that didn't speak German. So it was like a German cocktail party that we're going to generate. So we have one person like this.

[FEMALE VOICE 1] [SPEAKING GERMAN]

And then 29--

[GROUP VOICE] [SPEAKING GERMAN]

All right, room full of people speaking German, all right? And so we do the exact same experiment where we give people different exemplars of these textures. And we ask them to discriminate between them. And so what's plotted here is the proportion correct is a function of duration. Here, we've reduced it to just two durations-- short and long. And there's four different curves corresponding to different numbers of speakers in that signal, right.

So the cyan here is what happens with a single speaker. And so with a single speaker, you actually get better at doing this as the duration increases. All right, and so that's, again, consistent with the null hypothesis that when there's more information, you're actually going to be better able to say whether something is the same or different. But as you increase the number of people at the cocktail party-- the density of the signal in some sense-- you can see that performance for the short excerpts doesn't really change.

So you retain the ability to say whether these things are the same or different. But there's this huge interaction. And for the long excerpts, you get kind of worse and worse. So impairment at long durations is really specific to textures-- doesn't seem to be present for single sources.

To make sure that phenomenon is not really specific to speech, we did the exact same thing with synthetic drum hits. So we just varied the density of a bunch of random drum hits. Like, here's five hits per second.

[DRUM SOUNDS]

Here's 50.

[DRUM SOUNDS]

All right, and you see the exact same phenomenon. So for the very sparsest case, you get better as you go from the short excerpts to the long. But then as the density increases, you see this big interaction. And you get selectively worse here for the long duration case.

OK, so, again, it's worth pointing out that the high performance with the short excerpts indicate that all the stimuli have discriminable variation. So it's not the case that these things are just like totally homogeneous. And that's why you can't do it. It seems to be a specific problem with retaining temporal detail when the signals are both long and texture-like.

OK, so what does this mean? Well, go ahead. Here's the specular framework. And this sort of gets back to these questions about working memory, and so forth. And so this the way that I make sense of this stuff. And each one of these things is pure speculation or almost pure speculation. But I actually think you need all of them to really totally make sense of the results. It's at least interesting to think about.

So I think it's plausible that sounds are encoded both as sequences of features and with statistics that average information over time. And I think that the features with which we encode things are engineered to be sparse for typical natural sound sources. But they end up being dense for textures. So the signal comes in-- you're trying to model that with a whole bunch of different features that are in some dictionary you have in your head.

And for a signal like speech, your dictionary features include things that might be related to phonemes and so forth. And so for like a single person talking, you end up with this representation that's relatively sparse. It's got sort of a small number of feature activations. But when you get a texture, in order to actually model that signal, you need lots and lots and lots of feature coefficients, all right, in order to actually model the signal.

And my hypothesis would be that memory capacity places limits on the number of features that can be retained. All right, so it's not really related to the duration of signal that you can encode, per se. It's on the number of coefficients that you can retain that you need to encode that signal. And the additional thing I would hypothesize is that sound is continuously, and this is critical obligatorily encoded.

All right, so this stuff comes into your ears. You're continuously projecting it onto this dictionary of features that you have-- all right. And you've got some memory buffer within which you can hang onto some number of those features. But then once the memory buffer gets exceeded, it gets overwritten. And so you just lose all the stuff that came before. So when your memory capacity for these future sequences is reached, the memory is overwritten by the incoming sound. And the only thing you're left with are these statistics.

So I'll give you one last experiment in the texture domain, and then we'll move on. So this is an experiment where we presented people with an original recording, and then the synthetic version that we generated from the the synthesis algorithm. And we just ask them to rate the realism of the synthetic example.

And so this is just a summary of the results of that experiment where we did this for 170 different sounds. And this is a histogram of the average realism rating for each of those 170 sounds. And there's just two points to take away from this, right. The first that there's a big peak up here. So they rate it as the realism on a scale of 1 to 7. And so the big peak looks centered at 6 means that the synthesis is working pretty well most of the time. And that's sort of encouraging.

But there's this other interesting thing, which is that there's this long tail down here, right. And what this means is that people are telling us that this synthetic signal that is statistically matched to this original recording doesn't sound anything like it. And that's really interesting because it's statistically matched to the original. So it's matched in all these different dimensions, right. And, yet, there's still things that are perceptually missing. And that tells us that there are things that are important to the brain that are not in our model.

This is a list of the 15 or so sounds that got the lowest realism ratings. And just to make things easy on you, I'll put labels next to them. Because by and large, they tend to fall into sort of three different categories-- sounds that have some sort of pitch in them. Sounds that have some kind of rhythmic structure. And sounds that have reverberation.

And I'll play you these examples, because they're really kind of spectacular failures. Here, I'll play the original version and then the synthetic.

[RAILROAD CROSSING SOUNDS]

And here's the synthetic. I'm just warning you-- it's bad.

[SYNTHETIC RAILROAD CROSSING SOUNDS]

Here's the tapping rhythm-- really simple but--

[TAPPING RHYTHM SOUNDS]

And the synthetic version.

[SYNTHETIC TAPPING RHYTHM SOUNDS]

All right. This is what happens if you-- well, this is not going to work very well because we're in an auditorium. But I'll try it anyways. This is a recording of somebody running up a stairwell that's pretty reverberant.

[STAIR STEP SOUNDS]

And here's this synthetic version. And it's almost as though like the echoes don't get put in the right place, or something.

[SYNTHETIC STAIR STEP SOUNDS]

And now it sound even worse if this was not an auditorium. Here's what happens with music.

[SALSA MUSIC PLAYING]

And the synthetic version.

[SALSA MUSIC PLAYING]

And this is what happens with speech.

[MALE VOICE 1] A boy fell from the window. The wife helped her husband. Big dogs can be dangerous. Her-- [INAUDIBLE].

All right, OK, so in some sense, this is the most informative thing that comes out of this whole effort, because, again, it makes it really clear what you don't understand-- right. And in all these cases, it was really not obvious, a priori, that things would be this bad. I actually thought it was sort of plausible that we might be able to capture pitch with some of these statistics. Same with reverb and certainly some of these simple rhythms. I kind of thought that some of the modulation filters responses and their correlations would give this to you.

And it's not until you actually test this with synthesis that you realize how bad this is, right? And so this really kind of tells you that there's something very important that your brain is measuring that we just don't yet understand and hasn't been built into our model. So it really sort of identifies the things you need to work on.

OK, so just take home messages from this portion of the lecture. So I've argued that sound synthesis is a powerful tool that can help us test and explore theories of addition and that the variables that produce compelling synthesis are things that could plausibly underlie perception. And, conversely, that synthesis failures are things that point the way to new variables that might be important for the perceptual system.

I've also argued that textures are a nice point of entry for a real world hearing. I think what's appealing about them is that you can actually work with actual real world-like signals and all of the complexity that at least exists in that domain. And, yet, work with them and generate things that you feel like you can understand. And I've argued that many natural sounds may be recognized with relatively simple statistics of early auditory representation.

So the very simplest kinds of statistical representations that you might construct that capture things like the spectrum. Well, that on its own is not really that informative. But if you just go a little bit more complex and into the domain of marginal moments and correlations, you get representations that are pretty powerful.

And finally, I gave you some evidence that for textures of moderate length, statistics may be all that we retain. So there are a lot of interesting open questions in this domain. So one of the big ones, I think, is the locus of the time-averaging. So I told you about how we've got some evidence in the lab that the time scale of the integration process for computing statistics is on the order of several seconds. And that's a really long time scale relative to typical time scales in the auditory system.

And so where exactly that happens in the brain, I think, is very much an open question and kind of an interesting one. And so we'd like to sort of figure out how to get some leverage on that. There's also a lot of interesting questions about the relationship to scene analysis. So usually you're not hearing a texture in isolation. It's sort of the background to things that, maybe, you're actually more interested in-- somebody talking or what not.

And so the relationship between these statistical representations and the extraction of individual source signals is something that's really open, and, I think, kind of interesting. And then these other questions of what kinds of statistics would you need to account for some of these really profound failures of synthesis.

OK, so actually one-- I think this might be interesting to people. So I'll just talk briefly about this. And then we're going to have to figure out what to do for the last 20 minutes. But one of the reasons, I think, I was requested to talk about this is because of the fact that there's been all this work on texture in the domain of vision. And so it's sort of an interesting case where we can kind of think about similarities and differences between sensory systems.

And so back when we were doing this work-- as I said, this was joint work with Eero Simoncelli. I was a post-doc in his lab at NYU. And we thought it would be interesting to try to turn the kind of standard model of visual texture, which was done by Javier Portia and Eero a long time ago, into sort of the same kind of diagram that I've been showing you. And so we actually did this in our paper.

And so this is the one that you've been seeing all talk, right. So you've got a sound-wave form-- a stage of filtering. This non-linearity to extract the envelope and compress it. And then another stage of filtering. And then there are statistical measurements that kind of the last two stages of representation. And this is an analogous diagram that you can make for this sort of standard visual texture model.

So we start out with images like beans. There's centers surround filtering of the sort that you would find in the retina or LGN that filters things into particular spatial frequency bands. And so that's what you get here. So these are sub-bands again. Then there's oriented filtering of the sort that you might get via simple cells and V1. So then you get the sub-bands divided up even finer into both spatial frequency and orientation.

And then there's something that's analogous to the extraction of the envelope that would give you something like a complex cell. All right, and so this is sort of local amplitude in each of these different sub-bands-- right. So you can see, here, the contrast is very high. And so you get a high response in this particular point in the sub-band. So, again, this is in the dimensions of space. So that's a difference, right, so it's an image. So you got x and y-coordinates instead of time.

But, again, there are statistical measurements, and you can actually relate a lot of them to some of the same functional form. So there's marginal moments just like we were computing from sound. In the visual texture model, there's an auto correlation. So that's measuring spatial correlations which we don't actually have in the auditory model. But then these correlations across different frequency channels. So this is across different spatial frequencies to things tuned to the same orientation. And this is across orientations and in the energy domain.

So a couple of interesting points to take from this if you just sort of look back and forth between these two pictures. The first is that the statistics that we ended up using in the domain of sound are kind of late in the game. All right, so they're sort of after this non-linear stage that extracts amplitude. Whereas in the visual texture model, the nonlinearity happens here. And there's all these statistics that are being measured at these earlier stages before you're extracting local amplitude.

And that's an important difference, I think, between sounds and images and that a lot of the action and sound is in the kind of the local amplitude domain. Whereas there's a lot of important structure and image-- images that has to do with sort of local phase that you can't just get from kind of local amplitude measurements.

But at sort of a coarse scale, the big picture is that we think of visual texture as being represented with statistical measurements that average across space. And we've been arguing that sound texture consists of statistical computations that average across time.

That said, as I was alluding to earlier, I think it's totally plausible that we should really think about visual texture as something that's potentially dynamic if you're looking at a sheet blowing in the wind or much of people moving in a crowd. And so there might well be statistics in the time domain as well that people just haven't really thought about.

OK, so auditory scene analysis is, loosely speaking, the process of inferring events in the world from sound, right. So in almost any kind of normal situation, there is this sound signal that comes into your ears. And that's the result of multiple causal factors in the world. And those can be different things in the world that are making sound. As we discussed, the sound signal also interacts with the environment on the way to your ear. And so both of those things contribute.

The classic instantiation of this is the cocktail party problem where the notion is that there would be multiple sources in the world that the signals from the two sources sum together into a mixture that enters your ear. And as a listener, you're usually interested in individual sources, maybe, one of those in particular like what somebody that you care about is saying. And so your brain has to take that mixed signal-- and from that to infer the content of one or more of the sources.

And so this is the classic example of an ill-posed problem. And by that I mean that it's ill-posed because many sets of possible sounds add up to equal the observed mixture. So all you have access to is this red guy here, right? And you'd like to infer that the blue signals, which are the true sources that occurred in the world. And the problem is that there are these green signals here, which also add up to the red signal.

In fact, there's lots and lots and lots of these, right? So your brain has to take the red signal and somehow infer the blue ones. And so this is analogous to me telling you, x plus y equals 17-- please solve for x. And so, obviously, if you got this on a math test, you would complain because there is not a unique solution, right. That you could have 1 in 16, and 2 in 15, and 3 in 14, and so on and so forth, right?

But that's exactly the problem that your brain is solving all the time every day when you get a mixture of sounds. And the only way that you can solve problems of these sorts is by making assumptions about the sound sources. And the only way that you would be able to make assumptions about sound sources is if real-world sound sources have some degree of regularity. And in fact, they do. And one easy way to see this is by generating sounds that are fully random.

And so the way that you would do this is you would have a random number generator-- you would draw numbers from that. And each of those numbers would form a particular sample and a sound signal. And then you could play that and listen to it, right. And so if you did that procedure, this is what you would get.

[SPRAY SOUNDS]

All right, so those are fully random sound signals. And so we could generate lots and lots of those. And the point is that with that procedure, you would have to sit there generating these random sounds for a very, very long time before you got something that sounded like a real world sounds, right? Real world sounds are like this.

[ENGINE SOUND]

Or this--

[DOOR BELL SOUND]

Or this--

[BIRD SOUND]

Or this--

[SCRUBBING SOUND]

All right, so the point is that the set of sounds that occur in the world are a very, very, very small portion of the set of all physically realizable sound-wave forms. And so the notion is that that's what enables you to hear it. It's the fact that you've instantiated the fact that the structure of real world tones is not random. And such that when you get a mixture of sounds, you can actually make some good guesses as to what the sources are.

All right, so we rely on these regularities in order to hear. So one intuitive view of inferring a target source from a mixture like this is that you have to do at least a couple things. One is to determine the grouping of the observed elements and the sound signal. And so what I've done here is for each of these-- this is that cocktail party problem demo that we saw that we heard at the start. So we've got one speaker-- two, three, and then seven.

And in the spectrograms, I've coded the pixels either red or green, where the pixels are coded red if they come from something other than the target source, right. So this stuff up here is coming from this additional speaker. And then the green bits are the pixels in the target signal that are masked by the other signal. Or the other signal actually has higher intensity. And so one notion is that, well, you have to be able to tell that the red things actually don't go with the gray things.

But then you also need to take these parts that are green, where the other source is actually swamping the thing you're interested in, and then estimate the content of the target source. That's at least a very sort of naive intuitive view of what has to happen. And in both of these cases, the only way that you can do this is by taking advantage of statistical regularities in sounds.

So one example of irregularity that we think might be used to group sound is harmonic frequencies. So voices and instruments and certain other sounds produce frequencies that are harmonics, i.e., multiples of a fundamental. So here's a schematic power spectrum of somebody of what might come out of your vocal chords. So there's the fundamental frequency here. And then all the different harmonics. And they exhibit this very regular structure.

Here, similarly, this is A440 on the oboe.

[OBOE SOUND]

So the fundamental frequency is 440 hertz. That's concert A. But if you look at the power spectrum of that signal, you get all of these integer multiples of that fundamental.

All right, and so the way that this happens in speech is that there are these-- your vocal chords, which open and closed in this periodic manner. They generate a series of sound pulses. And in the frequency domain, that translates to harmonic structure. Not going to go through this in great detail. Hynek's going to tell you about speech.

All right, and so there's some classic evidence that your brain uses harmonicity as a grouping cue, which is that if you take a series of harmonic frequencies and you mistune one of them, your brain typically causes you to hear that as a distinct sound source once the mistuning becomes sufficient. And here's just a classic demo of that.

[MALE VOICE 2] Demonstration 18-- isolation of a frequency component based on mistuning. You are to listen for the third harmonic of a complex tone. First, this component is played alone as a standard. Then over a series of repetitions, it remains at a constant frequency while the rest of the components are gradually lowered as a group in steps of 1%.

[BEEPING SOUNDS]

[MALE VOICE 2] Now after two--

OK, and so what you should have heard-- and you can tell me whether this is the case or not-- is that as this thing is mistuned, at some point, you actually start to hear, kind of, two beeps. All right, there's the main tone and then there's this other little beep, right. And if you did it in the other direction, it would then reverse.

OK, so one other consequence of harmonicity is-- and somebody was asking about this earlier-- is that your brain is able to use the harmonics of the sound in order to infer its pitch. So the pitch that you hear when you hear somebody talking is like a collective function of all the different harmonics. And so one interesting thing that happens when you mistune a harmonic is that for very small mistunings, that initially causes a bias in the perceived pitch.

And so that's what's plotted here. So this is a task where somebody hears this complex tone that has one of the harmonics mistuned by a little bit. And then they hear another complex tone. And they have to adjust the pitch of the other one until it sounds the same. All right, and so what's being plotted on the y-axis in this graph is the average amount of shift in the pitch match as a function of the shift in that particular harmonic. And for very small mistunings of a few percent, you can see that there's this linear increase in the proceeded pitch.

All right, so the mistune that harmonic causes the pitch to change. But then once the mistuning exceeds a certain amount, you can actually see that the effect reverses. And the pitch shift goes away. And so we think what's happening here is that the mechanism in your brain that is computing pitch from the harmonics somehow realizes that one of those harmonics is mistuned and is not part of the same thing. And so it's excluded from the computation of pitch.

So the fact that you segregated those sources then somehow happened prior to or in at the same time as the calculation of the pitch. Here's another classic demonstration of sounds variations related to harmonicity. This is called the Reynolds-McAdams Oboe-- some collaboration between Roger Reynolds and Steve McAdams. There's a complex tone-- and what's going to happen here is that the even harmonics-- two, four, six, eight, et cetera, will become frequency modulated in a way that's coherent.

And so, initially, you'll hear this kind of one thing. And then it will sort of separate into these two voices. And it's called the oboe because the oboe is an instrument that has a lot of power at the odd harmonics. And so you'll hear something that sounds like an oboe along with something that, maybe, is like a voice that has vibrato.

[OBOE AND VIBRATO SOUNDS]

Is that work for everybody? So all these things that are being affected in kind of interesting ways by the reverb in this auditorium, which will-- yeah, but that mostly works.

So we've done a little bit of work trying to test whether the brain uses harmonicity to segregate actual speech. And so very recently, it's become possible to manipulate speech and change its harmonicity. And I'm not going to tell you in detail how this works. But we can resynthesize speech in ways that are either harmonic like this. This sounds normal.

[FEMALE VOICE 2] She smiled and the teeth gleamed in her beautifully modeled olive face.

But we can also resynthesize it so as to make it inharmonic. And if you look at the spectrum here, you can see that the harmonic spacing is no longer regular. All right, so we've just added some jitter to the frequencies of the harmonics. And it makes it sound weird.

[FEMALE VOICE 2] She smiled and the teeth gleamed in her beautifully modeled olive face.

But it's still perfectly intelligible, right. And that's because the vocal tract filtering that I think Hynek is probably going to tell you about this afternoon remains unchanged. And so the notion here is that if you're actually using this harmonic structure to kind of tell you what parts of the sound signal belong together-- well, and if you've got a mixture of two speakers that were in harmonic, you might think that it would be harder to understand what was being said.

So he gave people this task where we played them words, either one word at a time, or two concurrent words. And we just asked them to type in what they heard. And then we just score how much they got correct. And we did this with a bunch of different conditions where we increased the jitters. So there's harmonic--

[MALE VOICE 3] Finally, he asked, do you object to petting?

I don't know why my Rh has this example. But, whatever-- it's taken from a corpus called TIMIT that has a lot of weird sentences.

[MALE VOICE 3] Finally, he asked, do you object to petting? Finally, he asked, do you object to petting? Finally, he asked, do you object to petting?

All right, so it kind of gets stranger and stranger sounding than a bottom's out. These are ratings of how weird it sounds. And these are the results of the recognition experiment. And so what's being plotted is the mean number of correct words as a function of the deviation from harmonicity. So 0 here is perfectly harmonic, and this is increasing jitter. And so the interesting thing is that there's no effect on the recognition of single words, which is below ceiling, because these are single words that are excised from sentences. And so they are actually not that easy to understand.

But when you give people pairs of words, you see that they get worse at recognizing what was said. And then the effect kind of bottoms out. So this is consistent with the notion that your brain is actually relying, in part, on the harmonic structure of the speech in order to pull, say, two concurrent speakers apart. And the other thing to note here, though, is that the effect is actually pretty modest, right. So you're going from, I don't know, this is like, 0.65 words correct on a trial down to 0.5. So it's like a 20% reduction.

And the mistuning thing also works with speech. This is kind of cool. So here we've just taken a single harmonic and mistuned it. And if you listen to that, I think this is this-- you'll basically-- you'll hear the spoken utterance. And then it will sound like there's some whistling sound on top of it. Because that's what the individual harmonic sounds like on its own.

[FEMALE VOICE 3] Academic Act 2 guarantees your diploma.

So you might have been able to hear-- I think this is the--

[WHISTLING SOUND]

That's a little quiet. But if you listen again.

[FEMALE VOICE 3] Academic Act 2 guarantees your diploma.

Yeah, so there's this little other thing kind of hiding there in the background. But it's kind of hard to hear. And that's probably because it's particularly in speech there's all these other factors that are telling you that thing is speech and that belongs together.

And, all right, let me just wrap up here. So there's a bunch of other demos of this character that I could kind of give you about-- I could tell you about. Another thing that actually matters is repetition. So if there's something that repeats in the signal, your brain is very strongly biased to actually segregate that from the background. So this is a demonstration of that in action.

So what I'm going to be presenting you with is a sequence of mixtures of sounds that will vary in how many there are. And then at the end, you're actually going to hear the target sound. So if I just give you one--

[WHOPPING SOUND]

All right, it doesn't sound-- the sound at the end doesn't sound like what you heard in the first thing. But, here, you can probably start to hear something.

[WHOPPING SOUND]

And with here, you'll hear more.

[WHOPPING SOUND]

And with here, it's pretty easy.

[WHOPPING SOUND]

All right, so each time you're getting one of these mixtures-- and if you just get a single mixture, you can't hear anything, right. But just by virtue of the fact that there is this latent repeating structure in there. Your brain is actually able to tell that there's a consistent source and segregates that from the background.

I started off by telling you that the only way that you can actually solve this problem is by incorporating your knowledge of the statistical structure of the world. And, yet, so far the way that the field has really moved has been to basically just use intuitions. And so people would look at spectrograms and they say, oh yeah, there's harmonic structure. There's common onset. And so then you can do an experiment and show that has some effect.

But what we'd really like to understand is how these so-called grouping cues relate to natural sound statistics. We'd like to know whether we're optimal given the nature of real world sounds. We'd like to know whether these things are actually learned from experience with sound-- whether you're born with them. The relative importance of these things relative to knowledge of particular sounds like words.

And so this-- I really regard this stuff as in its infancy. But I think it's really kind of wide open. And so the sort of take-home messages here are that there are grouping cues that the brain uses to take the sound energy that comes into your ears and assign it to different sources that are presumed to be related to statistical regularities of natural sounds. Some of the ones that we know about are, chiefly, harmonicity and common onset and repetition.

I didn't really get to this. But we also know that the brain infers parts of source signals that are masked by other sources, again, using prior assumptions. But we really need a proper theory in this domain, I think, both to be able to predict and explain real world performance. And also, I think, to be able to relate what humans are doing this domain to the machine algorithms that we'd like to able to develop to sort of replicate this sort of competence.

And the engineering-- there was sort of a brief period of time where there were some people in engineering that were kind of trying to relate things to biology. But by and large, the fields have sort of diverged. And I think they really need to come back together. And so this is going to be a good place for bright young people to work.