HYNEK HERMANSKY: I'm basically an engineer. And I'm working on speech recognition. And so you may wonder, so what is there to work on? Because you have a cell phone in your pocket, and you speak to it, and Siri answers you and everything.

And the whole thing is working very basic principles. You start with a signal. It goes to signal processing. There's some pattern classification. Of course, deep neural nets as usual. And so this recognizes the message.

This recognizes what you are saying, so the question is, what is it that is fighter of the boat? Why not keep going, and try just to improve the error rates, and improve them basically step by step? Because we have a good thing going. We have something which is already out there. And it's working.

But you may know the answer, so this is, imagine that you are, sort of, skiing or going on a sled. And suddenly, you come somewhere. And you have to start pushing. You don't want to do that, but you do it for the reason, because there may be some-- another kind of slope going out there. And that's the way I feel about the speech recognition. So basically, sometimes we need to push a little bit up and maybe go slightly out of our comfort zone in order to get further.

Speech is not what we are-- it's not the thing which we are using for communicating with Siri. Speech is this. Basically, people speak the way I do. They hesitate. There's a lot of fillers, interruptions. And I don't finish the sentences. I speak with a strong accent, and so on. I become excited, and so on, and so on.

And we would like to put a machine there instead of the other person. Basically, this is what a speech recognition ultimately is, right? I mean, and actually, if you see what government is supporting, what the big companies are working on, this is what we are worried about. We are worried about the real speech produced by real people in the real communications by speech. And you know, I didn't mention all the disturbing things like noises, and so on, and so on, but

we will get into that.

So I believe that we don't only need signal processing, and information theory, and machine learning, but we also need the other disciplines. And this is where you guys are coming in. So that's what I believe in. We should be working together, engineering and life sciences working together. At least we should try. We should at least try to be-- we engineers should be try to inspired by life sciences.

And as far as inspiration is concerned, I have a story to start with. There was a guy who won the lottery by using numbers 1, 2, 3, 6, 7, 49. And they said, well, this is of course unusual sequence of numbers, so they say, how did you ever get to that?

He says, I'm the first child. My mother was my mother's second marriage and my father's third marriage. And I was born on the 6th of July. And of course, 6 by 7 is 49.

And that's sometimes I feel, I'm getting this sort of inspiration from you people. I may not get it right. I may not get it right, but as long as it works, I'm happy. You know, I'm not being paid for being smart and being knowledgeable about biology. I'm being, really, paid for making something which works.

Anyways, so this is just the warm up. I thought that you will still be drinking a coffee, so I decided to start with a joke. But anyway, but it's an inspiring joke. I mean, it's about inspiration. And I would maybe point out to some of the inspiration points, which I, of course, didn't get it right, but still, it was working.

Why do we have audition? Josh already told us-- because we want to survive in this world. I mean, this is a little ferret or whatever, and there is-- he's getting something now. And there is a object. And ferret is worrying, is this something I should be friendly with or I should-- it should be something which I run away.

So what is the message in this signal? Is it a danger or is a opportunity? Well, the same way, how do we survive in this world as human beings? So there is my wife who has some message in her head. And so she wants to tell me, eat vegetables, they are good for you, so she's using speech.

And speech is actually amazing sort of mechanism for sharing the experiences and for-- actually, without speech, we wouldn't be where we are, I can guarantee you, because that

allows us to tell the other people things what they should do without going through much trouble like a ferret with the bird. That we may not have to be eaten, maybe we just die a little early if we don't get this right, if we don't get this message. So she says this thing, and hopefully, I get the message.

So this is what speech is about, but I wanted to say, the speech is an important thing, because it allows us to communicate abstract ideas like good for you. And that's sort of not only vegetable, vegetable is saying, but a lot of abstract ideas can be conveyed by speech. And that's why I think it's kind of exciting.

Why do we work on machine recognition of speech? Well, first one is just like Edmund Hillary said, because it's there. They are asking, why did you climb Mount Everest? He said, well, because it's there.

I mean, it's a challenge, right? Spoken language is one of the most amazing things, I already told you before, of human race so there would be hell if we can't build a machine which understands it. And we don't have an easy time so far yet.

In addition, when you are addressing speech, you are really addressing the generic problems which we have in processing of other cognitive signals. And we touched it to some extent during this panel, because, you know, problems which we have in speech, we have the similar problems in perceiving images and perceiving smells. All these cognitive signals, basically, machines are not very good at it. Let's face it. Machines can add 10 billion numbers very quickly, but they cannot tell my grandmother from the monkey, right? I mean, so this is actually important thing.

There are also practical applications, obviously-- access to information. Voice interaction with machines extracting information from speech, given how much speech is out there now with-- I don't know how much we are adding every second through the YouTube and that sort of things, but there's a lot of speech out there. It would be good if information can actually extract information from that.

And I tell always the students, there is a job security. It's not going to be solved during your lifetime, certainly not during mine. I mean, sort of, if you get into it-- in addition, I mean, I know that this is maybe on YouTube, but also, if you don't like it, you can get fantastic jobs. There is a half of the IBM group ended up on the Wall Street making insane amount of money. So I mean, you know, what skills we should get in recognizing speech, working on the speech, can

be also applied in other areas. Obviously it can be applied in vision, and so on, and so on.

Speech has been produced to be perceived. Here is Roman Jakobson, the great Harvard, MIT guy, passed away unfortunately. He would be now a hundred and something. He says, we speak in order to be heard, in order to be understood. Speech has been produced to be perceived.

And over the millennia of the human evolution, it evolved this way so that it reflects properties of human hearing. And so I'm very much also with Josh. If you build up a machine which recognizes speech, you may be verifying some of the theories of speech perception. And I'll point out that along the way.

How do I know that the speech evolved to fit the hearing and not the other way around? I got some big people arguing over that, because they say, you don't know, I mean, basically, but I know. I think. Well, I think that I know, right?

Every single organ which is used for speech production is also used for something much more useful, like, sort of typically, eating and breathing. So this is the organs of speech production-- lungs, the lips, teeth, nose, and velum, and so on, and so on. Everything is being used for some life-sustaining functions, including speaking.

So I know that it's not the same in hearing. Hearing has evolved to hear, for hearing. Maybe there are some organs of balance, and that sort of thing, but mostly, you do hear. In the speech, everything, what is being used, has been used for-- it's used for something else also, so clearly, we just learned how to speak because we had the appropriate hardware there, and we learned how to use it.

So in order to get the message, you use some cognitive aspects, which I won't be talking much about. So you have to use the common language. You have to have some context of the conversation. You have to have some common set of priors, some common experience, and so on, and so on, but mainly what I will be talking about, you need the reliable signal which carries the message, because the message is in the signal. It's also in your head, but the signal supports what is happening in your head.

So how much information is in speech signal? This is, I have stolen I believe from George Miller, I think. So if you look at Shannon's theory, I mean, there will be about 80 kilobytes per second. And indeed, you can generate a reasonable signal without being very smart about it

just by coding it to 11 bits at 8 kilohertz per second, 80 kilobits per second. This verifies it. So this is how much information might be in the signal.

How much is in the speech is actually very-- it's, sort of, not very clear, but at least we can estimate it to some extent. If you say, I would like to transcribe the signal in terms of the speech sounds, phonemes, so there is maybe about 40 to 49 phonemes, or something, 41 phonemes. So if you look at the entropy of that, it comes to about 80 bits per second. So there is three orders of magnitude difference.

If you push a little bit further-- indeed, I mean, if you speak with about 150,000 words, that means about 80 bits, 30 words per minute, again, it comes to less than 100 bits. So there's, as I said, there's a number of ways how you can argue about this amount of information. If you start thinking about dependencies between phonemes, it can go as low as 10, 20 bits per second. No question that there is much more information in the signal than it is in useful message which we would like to get out. And we'll get into that.

Because what is in the message, there is not only information about the message, but there is a lot of other information. There's information about health of the speaker, about which language the speaker is using, what are-- what emotions, there is who is speaking, speaker-dependent information, what is the mood, and so on, and so on. And there is a lot of noise coming from around, reverberations.

We talk about it quite a lot in the morning, all kinds of other noises. So what I call noise in general, I call everything what we don't want besides the signal, which, in speech recognition, is the message. So when I talk about the noise, it can be information about who is speaking, about their emotions, about the fact that maybe my voice is going, and so on, and so on. To my mind, purpose of perception is get the information which carries-- get the signal which carries the desired information and suppress the noise, eliminate the noise. So the purpose of perception, being a little bit vulgar about it, is how to get rid of most of the information very quickly, because otherwise, your brain would go bananas.

So you basically want to focus on what you want to hear, and you want to ignore, if possible, everything else. And it's not, of course, easy, but we discuss that again in the morning, about some techniques how to go about it. And I will mention a few more techniques which we are working on. But this a key thing, is, purpose of perception is to get what you need and not to get what you don't need, because otherwise, your brain would be too busy.

Speech happens in many-- it's a very simple example. Speech happens in many, many environments. And there is a lot of stuff happening around it, so the very simple example, which I actually used when I was giving a talk to some grandmothers in the Czech Republic is that, what you can already use is the fact that things happen at different levels. And they happen at different frequencies, so perception is selective.

Every perceptual mode is selective and attends only to part of the world. You know, we don't hear the radio-- we don't see the radio waves. And we don't hear the ultrasound, and so does all the lower elements, and so on, and so on. So there are different frequencies, different sound intensities are in the first approximation.

This is what you may use. If something is too weak, I don't care. If something has too high frequencies, I don't care, and so on, and so on.

There are also different spectral and temporal dynamics to speech, which we are learning about that quite a lot. It happens at different locations of the space. Again, this is the reason why we have a spatial directivity. That's why we have two ears. That's why we have a specifically-shaped ears, and so on, and so on.

There are also other cognitive aspects, I mean, sort of, like, the selective attention. Again, we talk about it, that people appear to be able to modify the properties of your cognitive processing depending on what you want to listen to. And my friend Nima Mesgarani with Eddie Chang, who was supposed to be here instead of me, just had a major paper in *Nature* about that, and so on, and so on. There's a number of ways how we can modify the selectivity. We talk about this sharpening the cochlear filters, I mean, depending on the signal from the brain.

So speech happens like this, start with a message. You have a linguistic code, maybe 50 bits per second. There are some motor controls. Speech production comes to a speech signal, which has three orders of magnitude larger information content. Through speech perception and cognitive processes, we get, somehow, back to the linguistic code and extract the message, so this is important-- from the small, low bit-rate, to high bit-rate, to the low bit-rate.

In production, actually, I don't want to pretend it happens in such a linear way. There are also feedbacks, so there is a feedback from you listen to yourself when you are speaking. You can control how you speak. And you can also actually change the code, because you realize, oh, I should have said it somehow differently.

In speech perception, again, we just talked about it, you can, if the message is not getting through, you may be able to tune the system in some ways. You may be changing the things, you know? And you may also use the very mechanical techniques, as I told you, close the window, or walk away. There is also feedback through the dialogue, so from-- between message and message, depending what I'm hearing, I may be asking a different kind of question, so which also modifies the message of the sender.

How do we produce speech? So we speak in order to be heard, in order to be understood. So very quickly, I want to go back to something which people already forgot a big way, which is Homer Dudley. He was a great researcher at Bell Laboratories before the Second World War. He retired I think sometime early in '50s. He passed away in the '60s.

He was saying message is in the movements of the vocal tract which modulates the carrier, so message in the speech is not in fundamental frequency, it's not the way you are exciting your vocal tract. Message is how you shape the organs of speech production. Proof for that is that you can whisper and you can still understand, so you don't-- how you excite the vocal tract is secondary.

How do you generate this audible carrier is secondary. You know, you can use the artificial larynx, so there is this idea, there's a message. A message is being-- goes through modulator into carrier, comes out as speech.

So this modulation actually has been used a long time ago, and excuse me for being maybe a little bit simplistic, but it's actually, in some ways, interesting. So this was the speech production mechanism which was developed in some time in the 18th century by the guy Johannes Wolfgang Ritter von Kempelen. And he actually had it right.

The only problem is nobody trusted him, because he also invented Mechanical Turk, which was playing the chess. And so he was caught as a cheater, so when he was showing his synthesizer, nobody believed him. But anyways, he was definitely a smart guy.

So he used already the principle which is now used. This is a linear model of speech production developed actually before the Second World War, really, again, Bell Laboratories should get the credit. I believe this is stolen from Dudley's paper. So there is a source, and you can change it. It periodic signals out random noise, if you are producing voice signal or unvoice signal. And then there is a resonance control which goes into amplifier, and it produces the speech.

So this is the key here, this a key to the point that Dudley developed for this called a voder. And he trained the lady who spent a year or something to play it. It was played like a organ.

And she was changing the resonance properties of this system here. And she was creating excitation by pushing on a pitch pedal and switching on the big-- on the wrist bar. And if it works well, we may even be able to make the sound. This is a test.

[AUDIO PLAYBACK]

- Will you please make the voder say, for our Eastern listeners, good evening, Radio--

HYNEK HERMANSKY: This is a real--

- --audience.

HYNEK HERMANSKY: --speech.

- Good evening, radio audience.

HYNEK HERMANSKY: This is--

- And now, for our Western listeners, say, good afternoon, radio audience.

- Good afternoon, radio audience.

[END PLAYBACK]

HYNEK HERMANSKY: Good enough, right? I mean, sort of-- so already, 1940s, This was the demonstrated at a trade fair. And the lady was trained so well that, in the '50s, when Dudley was retiring, they brought her in. She was already retired a long time ago. And she still could play it.

How the speech works-- I mean, maybe-- oh, I wanted to jump this, but anyways, let's go very quickly through that. So this is a speech signal. This is a acoustic signal. It changes in-- this is a sinusoid, high pressure, low pressure, high pressure, low pressure.

If you put somewhere in the in the path, some barrier, what happens is you generate a

standing wave. A standing wave stands in a space. And there are high pressures, low pressures, high pressures, low pressures.

And the frequency depends on the frequency-- I mean, the size of this standing wave depends on the frequency of the signal. So if I put it into something like a vocal tract, which we have here-- so this is a glottis. This is where it gets exciting. This is a very simple model of vocal tract. And here I have a lips. So it takes certain time to provide this through the tube.

And the tube will have a maximum velocity at certain point for-- so that it will be resonating in a quarter wavelength of the signal, 3/4 of the wavelength of the signals, in 5/4 of the wavelength of the signal, and so on, and so on. So we can compute which frequencies this tube will be resonating. This is a very simplistic way of producing speech, but you can generate reasonable speech sounds with that.

So if we start putting a constriction there somewhere, which emulates the way, very simple the way how we can speak by moving the tongue against the palate or making of constrictions in the speech-- so when the tube is open like this, it resonates at 500, 1,500, 2,500 if the tube is 17 centimeters long, which is a typical length for the vocal tract. So if I put a constriction here, everything moves down because there is such a thing like perturbation theory, which says that, if you are putting a constriction through the point of the maximum velocity, which is, of course, at the opening, all the modes will go down.

And as you go on, basically, the whole thing keeps changing. The point is that, almost in every position of the, say, this tongue, all the resonance frequencies are changing, so the whole spectrum is being affected. And that may become useful to explain something later. But we go like this. At the end, you end up, again, in the same frequencies.

These are called nomograms. And they will be heavily worked on at the Speech Group at MIT and at Stockholm. So you can see how the formants are moving. And you can see that, for every position of the [INAUDIBLE], here we have a distance of a constriction from the lips. For every position, we are having all the formants moving, so information about what I'm doing with my vocal organs is actually at all frequencies, all audible frequencies in different ways, but it's there everywhere.

It's not a single frequency which would carry information about something. All the audible frequencies carry information about speech. That's important. You can also look at it and you can say, you know, what is the-- where the front cavity resonates, the back cavity resonates.

Again, this front cavity resonance may become interesting a little bit later if we get to that. But this is a very simplistic model of the speech production, but pretty much contains all the basic elements of the speech.

Point here is that, depending on the length of the vocal tract, even when you keep the constriction at the same position-- this is how long is this front part before the construction is-- so all the resonances are moving, but a shorter vocal tract, like the children's vocal tract, or even in a number of females, they typically have a shorter vocal tract than the males, there's a different number of resonances. So if somebody is telling you the information is in the formants of speech, question it, because it's actually impossible to generate the same speech being two different people, especially having two different lengths of the vocal tract. And we get into it when we talk about the speaker dependencies.

Second part is-- of the equation is hearing. So we speak in order to be heard, in order to be understood. And again, thanks to Josh, he spent more than sufficient time explaining you enough what I wanted to say. I will just add something-- some very, very small things.

So just to summarize, Josh was telling you the theory works basically like a bank of bandpass filters with a changing frequency and output depending on sound level intensity. There are many caveats to that, but I mean, in a first approximation, I 100% agree this is enough for us to follow for all the rest of the talk.

Second thing which Josh mentioned very briefly, but I want to stress it, because it is important, firing rates-- because you know the cochlea communicates with the rest of the system through the firings, through the impulses. Firing rates on the auditory nerve are of the order of 1 kilohertz every one millisecond. But as you go up and up in the system, already here on the colliculus is maybe order of magnitude less. And the order in the level of auditory cortex is 2 orders of magnitude less.

So of course, I mean, you know, this is how the brain works. I mean, so here we have from periphery up to cortex, but also, I think it was mentioned very briefly, if you look at it, number of neurons increase more than actually decrease in the firing rates, because if we have-- again, those are just orders of magnitude-- 100,000 neurons maybe on the level of auditory nerve, or colliculus nucleus, and you have 100 million neurons maybe on the level of the brain. And this can become handy later, when, if I get all the way to the end of the talk, I will recall this piece of information.

Another thing which was mentioned a number of times is that there are not the only connections from ear, from the periphery to the brain, but there is, by some estimates, many, many more-- I mean, again, I mean the estimates vary, but this is something which I have heard somewhere-- maybe there is maybe almost 10 times more connections going from brain to the ear than from the ear to the brain. And typically, the nature hardly ever builds anything without a reason, so there must be some reason for that. And perhaps we will get into that.

Josh didn't talk much about the level of the-- on the cortex. So what's happening on the lower levels, on the periphery? They are just these simple increases of auditory-- of firing rate. There is a certain enhancement of the changes. So at the beginning of the tone-- this is a tone-- the beginning of the tone, there is more firing on auditory nerve. At the end of the tone, there is some deflection.

But when you look at a higher level of the cortex, all these wonderful curves, which are sort of increasing with intensity like it would if you had a simple bandpass filter, start looking quite differently. So we measure majority-- what I heard, the majority of the cortical neurons are selective to certain levels. Basically, the firing increases to a certain level, and then it decreases again. And they are, of course, selective at different levels.

Also, I mean, you don't see, just these simple things like here, that your firing starts as a tone starts. But they are neurons like that, but there are also neurons which just are interested at the beginning of the signal. There are neurons which are interested in beginning and ends. There are neurons which are interested only at the ends of the signals, and so on, and so on.

Receptive fields, again, has been mentioned already before. Just as we have a receptive field in the visual cortex, we have also receptive fields in auditory cortex. Here we don't have the-- here we have a frequency and a time, unlike x and y, receptive fields which are typical, sort of, first thing you are hearing about when you talk about visual perception.

They come in all kinds of colors. They tend to be quite long, meaning they can be sensitive for about quarter of the second-- not all of them, but certainly, there are many, many different cortical receptive fields. So some people are suggesting, given the richness of the neurons in auditory cortex, it's a very legal thing to suggest that maybe the sounds are processing in following way, not only that you do the frequency analysis in the cochlea, but then, on the higher levels, you are creating many pictures of the outside world.

And then, of course, only the question is here, if answer, this is Murray Sachs' paper from their labs, from Johns Hopkins in 1988. They just simply said pattern recognition, but I believe there is a mechanism which picks up the best streams and leaves out not so useful things, but the concept was here around for a long time. So this was physiology 101.

Psychophysics is saying that you play the signals to listeners, and you ask them what they hear. But we want to know what is the response of the organism to the incoming stimulus, so simply, you play the stimulus and you ask what is the response. First thing which you can ask, do you hear something or not? And you already will discover some interesting stuff.

Hearing is not equally-sensitive everywhere. It's selective. And it's more sensitive in the area somewhere between 1 and 4 kilohertz. It's much less sensitive at the lower frequencies.

This is a threshold. On the threshold level-- here's another interesting thing. If you just apply the signals in different ears, as long as the signals happen within a certain period, about a couple of hundred millisecond, and if couple of hundred millisecond you hear from your ear would be more often, the thresholds are half.

Basically, neither of these signals would be heard if you applied only a single one, but when, if you apply both of them, basically you hear them. If you play the signals of different frequencies, if these signals are close enough, close so that, as Josh mentioned about the beats, they happen within one critical band, again, neither blue or green signal would be heard on its own. But if you play them together, you hear them.

But if they are further in frequency, you don't hear them. Same thing if these guys are further in time, you wouldn't hear them. So this subthreshold perception actually is kind of interesting. And we will use it.

Which we didn't talk much about is that there are obvious ways how you can modify the threshold of hearing. Here we have a target. And since it is higher than threshold of hearing, you hear it. But if you play another sound called masker, you will not hear it, because your threshold basically is modified.

It's called the mask threshold. And this part is suddenly not-- this target is not heard. The target can be something useful, but in mp3, it can be pretty annoying, because it's typically noise. You try to figure out how you can mask the noise by the useful signal. You're computing these masked thresholds on the fly.

The initial experiment with this, what is called simultaneous masking, was following, and, again, was Bell Labs, Fletcher, and his people. They would figure out what is the threshold of certain frequency without the noise. But then they would put noise around it, and the threshold had to go up, because there was a noise, so there was masking. Then they made a broader noise, and threshold was going up, as you would expect. There was more noise, so you had to make the signal stronger.

And you made it to a certain point, when you start making the band of noise too wide, suddenly it's not happening anymore. There is no more masking anymore. That's how they came with this concept of critical band.

Critical band is what happens inside the critical band matters, basically, influences the decoding of the signal within a critical band. But if it happens outside the critical band, it doesn't. So essentially, if the signals are far away in frequency, they don't interact with each other. And again, this is a useful thing for speech recognition people. They didn't much realize that this is the main outcome of the masking.

Critical bands, I mean, again, I mean, discussions are here, but this is a Bark scale which has been developed in Germany by Zwicker and his colleagues. It's pretty much regarded to be from about 600, 700 hertz up. And it's approximately constant up to 600, 700 hertz. If you talk to Cambridge people, Brian Moore, and that sort of logarithmic it's pretty much regarded to be pretty much everywhere.

But not really, but the critical bands, remember, critical bands from the subthreshold things? Again, the critical band is masking. It's starting it with things happen within the critical band. They integrate. They happen outside the-- each of them outside the critical band, they don't interact.

Another masking is temporal masking. So you have a signal-- and of course, if you put a mask on it, it's simultaneous masking. You have to make it much-- the signal much stronger in order for you to hear it. But it also influences things in time.

This is what is called forward masking. And this is the one which is probably more interesting and more useful. It's also backward masking, when a masker happens after the signal, but it probably has a different origin, more like cognitive rather than earlier.

So there is still a masker. You have to make the signal stronger up to a certain point. When the distance between masker and the signal is more than 200 milliseconds, there is like there's no masker anymore. Basically, there is no temporal masking anymore, but it is within this interval of 200 milliseconds.

If you make mask stronger, masking is stronger initially, but it also decays faster. And again, it decays after about 200 milliseconds. So whatever happens outside this critical interval, about a couple of hundred millisecond, it doesn't integrate. But if it happens inside this critical interval, that seems to be influencing-- these signals seem to be influencing each other. And again, I mean, you know, I talk about the subthreshold perception-- if there were two tones which happen within 200 millisecond, neither of them would be heard in isolation, but they are heard if you play them together.

Another part which is kind of interesting is that loudness depends, of course, on the intensity of the sound, but it doesn't depend linearly on that. It depends with about cubic root, so in order to make a signal twice as loud, you have to make it about 10 times more in intensity for stimuli which are longer than 200 milliseconds.

Equal loudness curves, this is a threshold curve, but these equal loudness curve are telling you what the intensity of the sound-- sorry-- would need to be in order to hear it equally loud. So it's saying that, if you have a 40 dB signal at 1 kilohertz, and you want to make it equally loud at 100 hertz, you have to make it 60 dB, and so on.

These curves become flatter and flatter, most pronounced at the threshold at lower levels, but they are there. And they are actually kind of interesting and important. Hearing is rather non-linear. Properties depend on the intensity. Speech of course is happening somewhere around here where the hearing is more sensitive. That was the point here.

Modulations, again, we didn't talk much about that, but modulations are very important. Since 1923, it's known that hearing is the most sensitive to certain rate of modulations around 4, 5 hertz. These are experiments from Bell Labs repeated number of times. So this is this for a, a modulations.

This experiment, what you do is, that you modulate a signal, and change the depth, and change the frequency. And you are asking, do you hear the modulation or don't you hear the modulation? Very interesting-- interesting thing is, if you look at-- again, I mean, I refer to what Josh was telling you in the morning. If you just take one trajectory of the spectrum, you treat it

as a time domain signal, remove the mean and compute its Fourier components-- frequency components, they peak somewhere around 4 hertz, just where the hearing is the most sensitive.

So hearing is not very sensitive, obviously, to when the signal is non-modulated, but also there is-- there are almost no components in the signal which would be non-modulated, because when I talk to you, I move the mouth. I mean, I change the things. And I change the things about four times a second, mainly.

When it comes to speech, you can also compute-- music, you can also figure out what are the natural rhythms in the music. I stole this from, I believe, the Munich group, from [INAUDIBLE]. He played 60 pieces of music. And then he asked people to tap to the rhythm of the music.

And this is the histogram of tapping. Most of the people, for most of the music, tapping was about four times a second. This is where the hearing is most sensitive. And this is modulation frequency of this music. So people play music in such a way that we hear it well, that it basically resonates with the natural frequency which we are perceiving.

You can also ask the similar thing. So, in speech, you can play the speech sentences. And you ask people to tap in to the rhythm of the sentences. Of course, what gets out is the syllabic rate. And syllabic rate is about 4 hertz. Where is the information in speech?

Well, we know what the ear is doing. It analyzes signal into individual frequency bands. We know what Homer Dudley was telling us. When messages and modulations of these frequencies-- as a matter of fact, that was the base of his vocoder. What he also did was that he designed-- actually, it wasn't only him.

There was another technique. This one is, kind of, somehow cleaner thing, which is called the spectrograph, which tells you about the spectrum of frequency components of the acoustic signal. So you take the signal. You put it through a bank of bandpass filters. And then here, you basically display, on the z-axis, intensity in each frequency band.

This was, I heard, used for listening for German submarines, because they wanted to-- they knew that acoustic signatures were different for friendly submarines and enemy submarines. People listen to it-- for it, but also people realized it may be useful to look at the signal-- acoustic signal somehow. Waveform, it wasn't making all that much sense, but the spectrogram was.

Danger there was that the people who were working in speech got hold of it. And then they start, sort of, looking at the spectrograms. And they say, haha, we are seeing the information here. We are seeing the information in waves. The spectrum is changing, because not only that this was the way the origin of the spectrogram was developed, that you were displaying changes in energy in individual frequency bands, but you can also look at this.

This when you get to what is called a short-term spectrum of speech. And people said, oh, this short-term spectrum looks different for R than for E, so maybe this is the way to recognize speech. So indeed, I mean, those are two ways of generating the spectrograms. I mean, this was the original one, bank of bandpass filters. And you were displaying the energy as a function of time.

This is what your ear is doing. That's what I'm saying. This is not what your ear is doing, that if you take a short segments of the signal, and you compute the Fourier transform, then you display the Fourier transform one frame at a time, but this is the way most of the speech recognition systems work. And I'm suggesting that maybe we should think about other ways.

So now we have to deal with all these problems. So we have a number of things coming in in the form of the message with all these chunk around it. And machine recognition of speech would like to transcribe the code which carries the message. This is a typical example of the application of speech recognition. I'm not saying this is the only one. There are attempts to recognize just some key words. There are attempts to actually generate the understanding of what people are saying, and so on, but we would be happy, in most cases, just to transcribe the speech.

Speech has been produced to be perceived. We already talked about it. It evolved over millennia to fit the properties of hearing. So this is-- I'm sort of seconding what Josh was saying. Josh was saying, you can learn about the hearing by synthesizing stuff. I'm saying you of learn about hearing by trying to recognize the stuff.

So if you put something in and it works, and it supports some theory of hearing, you may be kind of reasonably confident that it was something which has been useful. Actually there's a paper about that, which, of course, I'm co-author, but I didn't want to show that. I thought I would leave this one, but I didn't do it at last minute.

Anyways, speech recognition-- speech signal has high bit-rate, recognizer comes in,

information, low bit-rate. So what you are doing here, you are trying to reorganize your stuff. You are trying to reduce the entropy. If you are reducing the entropy, you better know what you are doing, because otherwise, you get real garbage. I mean, that's, kind of, like, one of these common sense things, right?

So you want to use some knowledge. You have plenty of knowledge in this recognizer. Where does this knowledge come from? We keep discussing it all the time. It came from textbooks, teachers, intuitions, beliefs, and so on.

And its a good thing about that, that you can hardwire this knowledge so that you don't have to learn it, relearn it next time based on the data. Of course, problem is that this knowledge may be incomplete, irrelevant, can be plain wrong, because, you know, who can say that whatever teachers tell you, or textbooks tell, or your intuitions or beliefs is always true?

Much more often now, what people are using is that they-- that knowledge comes directly from the data. Such knowledge is relevant and unbiased, but the problem is that you need a lot of training data. And it's very hard to get architecture of the recognizer from the data, at least, I don't know quite well how to do it yet.

So these are two things. And again, I mean, let me go back to '50s. First knowledge-based recognizer was based on the spectrograms. There was Richard Galt.

And he was looking at spectrograms and trying to figure out how this short-term spectrum looked like for different speech sounds. Then he thought he would make this finite state machine, which will generate the text. Needless to say, it didn't work too well.

He got beaten by data-driven approach, where people took a high-pass filter speech, low-pass filter speech, displayed energies from these to two channels on, at the time it was oscilloscope. And they tried to figure out what are the patterns. They tried to memorize the patterns, make the templates from the training data. And they tried to match it for the test data was recognized, which was recognizing ten digits.

And it was working reasonably well, better than 90% of the time for a single speaker, and so on, and so on. But it's interesting that, already in '50s, the data-driven approach got beat by the knowledge-based approach, because knowledge maybe wasn't exactly what you needed to use. You were looking at the shapes of the short-term spectra basically.

Of course, now, we are in 21st century, finally. Number of people say, this is the real way of

recognizing speech. You take the signal as it comes with the microphone. You take the neural net. You put a lot of training data, which contain all sources of unwanted variability, basically, what you-- all possible ways of-- you can disturb the speech and comes out the speech message.

The key thing is, I'm not saying that this is wrong, but I'm saying that, maybe this is not the most efficient way of going about it, because, in this case, you would have to retrain the recognizer every time. It's a little bit like, sort of, you know, if you look at the hearing system, or the simple animal system-- this is a moth here. Here it changes in acoustic pressure to changes in firing rate. It goes to very simple brain, very small one.

You know, this is not the way the human hearing is working. Human hearing is much more complex. And again, Josh already told us a lot about it, so I won't spend much time. The point here is, the human hearing is frequency-selective. It goes through a number of levels.

This is very much along the deep net and that sort of things. But still, there is a lot of structure there in the hearing system. So it makes at least some sense to me, if you want to do what people are doing more and more, and there will be a whole special session next week at Interspeech on how to train the things directly from the data, probably you want to have highly-structured environment.

You want to have a convoluted pre-processing recursive structures, and so on, and long, short-term memory. Yeah, here are actually some, but all these things are being used. And I think this is the direction to go. But I still argue that maybe it's a better-- there's a better way to go about it. A better way to go about it is that you try first to do some pre-processing of the signal and derive some way of describing the signal more efficiently, using the features, and so on, and so on.

Here you put all the knowledge which you possibly may want to-- you already have. This knowledge can be derived from some development data, but you don't want to use directly the speech signal every time you are using-- you don't want to retrain, basically, every time, directly from the speech signal. You want to reserve your training data, the task-specific training data, to deal with the effects of the noise which you don't understand. This is where the machine learning comes. I'm not saying that this is not a part of machine learning, but, I mean, this is-- there are two different things which you are going to do.

I was just looking for some support. This one came from Stu Geman from Brown University and his colleagues. Stu Geman is a machine learning person, definitely, but he says, we feel that meat is in the features rather than in the machine learning, because they go overboard, basically, explaining that, if you just rely on machine learning, sure, you have a neural net which can approximate just about any function, given that you have infinite amount of data an infinitely large neural net. And they say, infinite is a kind of not useful engineering concepts.

So they feel like that, if representations actually are-- I hope they still feel the same. I didn't talk to them now, but it seems like that there is some support in this notion, what I'm saying. But of course, problem with the features is following, whatever you stripped on the features, this is a bottleneck.

Whatever you decide that is not important is lost forever. You will never recover from it, right? Because I'm asking for feature extraction. I'm asking for this emulation of the human perception, which strips out a lot of information, but I still think that we need to do it if we want to design a useful engineering representations.

The other problem, of course, is whatever you leave in, the noise, the information which is not relevant to your task, you will have to deal with it later. You will need to train the old machine on that, so you want to be very, very careful. You are walking a thin line here. What is it that I should leave out? What is it that I should keep in? It's always safer to keep a little bit more in, obviously.

But this is the goal which we have here. And I wanted to say, features can be designed using development data. And when I'm saying use the development data, design your features and use them. Don't use this development data anymore. We have a lot of data for the designing of good features. And I think that, again, is happening in the field-- good.

How the speech recognition was done in 20th century, this is what I know, maybe, the most, so we'll spend some time. And it's still done largely in-- there are some variants of this recognition that's still done. You take the signal. And you derive the features.

In the first place, you derive what is called short-term features, so you take short segments of the signal, about 10 to 20 milliseconds. And you derive some features from that. That was in 20th century. Now we are taking much longer segments, but we'll get into that.

But you derive it with about 100 hertz sampling every 10 millisecond, so you turn the one-

dimensional signal into two-dimensional signal. And here, typically, the first step is the frequency, so those may be-- imagine those are frequency vectors, or something derived from frequency vectors, gets through or stuff like that. Those are just tricks, signal processing tricks which people use-- but one-dimensional to two-dimensional.

Next thing is, you estimate the likelihood of the sounds each 10 millisecond. So here, what I-- imagine that here we have different, say, speech sounds, maybe 41 phonemes, maybe 3,000 context-dependent phonemes, and so on, depends on-- but those are parts of speech which makes some sense. And they come, typically, from phonetics theory. And we know that you can generate different words putting phonemes together in different ways, and so on, and so on.

So suppose for the simplicity that they are-- there's 41 phonemes. And so if there is a red one, red means that, probably, posterior probability of the-- actually, we need them more. We need the likelihoods rather than posteriors, so with your posteriors, we just divided it by priors to get the likelihoods, so meaning that this phoneme has a high likelihood and white ones don't have a likelihood at this time.

So next step is that you do the search on it. This is a painful part. And I won't be spending much time on that. I just want to give you some flavor of this. You try to find the best path through this lattice of the likelihoods. And if you are lucky, the best part, then, is going to present your speech sounds.

So then the next thing is only that you look and transcribe to go from phonemic representation from into lexical representation, basically, because you know there is typically one-to-one relations-- Well, should be careful, one-to-one, but it is a relation, known relation between phonemes and the transcription. So we know what has been said. So this is how the speech recognition is done.

Talking about this part, I mean, here we have to deal with one major problem, which is, like, the speech doesn't come out this way. It doesn't come out as a sequences of individual speech sounds, but, since I'm talking to you, I'm moving the mouse. I'm moving the mouse continuously.

There is a thing that first I can make certain sounds longer, certain sounds shorter. And then I add some noise to it. Finally, because of what is called co-articulation, each target phonemes gets spread in time, so you get a mess. But people say-- sometimes, people like to say,

speech recognition, this is our biggest problem.

I claim to say this is not the problem. It is a feature. And feature is important, because it comes quite handy later. Hopefully, I will convince you about it. But what we get is a mess, so this is not easy to recognize, right? We have co-articulations. We have speaker dependencies, noise from the environment, and so on, and so on.

So the way to deal with it is to recognize that different people may sound different, communication and environment may differ, so the features will be dependent on a number of things, on environmental problems, on who are saying things, and so on. People say same things in different speech. I can speak faster, I can speak slower, still, the message is the same.

So we use what is called the Hidden Markov Model, where you try to find such a sequence of the phonemes which optimizes the conditional probability of the model, given the data. And models you generate, on the fly, as many models as possible, actually, an infinite number of models, but, of course, again, you can't do it infinitely, so you do it in some smart ways. And this is being computed through modified Bayes' rule.

Modified is because, for one, I mean, you would need a prior probability of the signal, and so on. We don't use that. But also, what we are doing, we are somehow arbitrarily scale the things which are called the language model, because this is a prior probability of the particular utterance. This is the likelihoods coming from the data, combining these two things together, and finding the best match, you get the output which best matches the things.

Model parameters are typically derived from the training data. Problem is, how to find the unknown utterance. You don't know what is the form of the model. And you don't know what is the data.

So we are dealing with what is called the Doubly stochastic model, a Hidden Markov model. Speech is a sequence-- it's a sequence of hidden states. You don't see this hidden state. And also, you don't know what comes from any state.

So it somehow-- so you don't know for sure in which state you are on. You don't know for sure what comes out, but you know that-- well, you know, you assume that this is how the speech looks like. So here I have a picture a little bit. I apologize for being trivial about this, but imagine that you have a string of-- group of people.

They are-- some are female, some are male. They are groups of males, groups of females. And each of them says something. He says, hi. And you can measure something. This is a fundamental frequency.

You get some measurement out of that, but you don't see them. But what you know is that they interleave, basically. For a while, there is a group of males, then there is a-- then the speech is to a group of female. And then you stay for a while in a group of female, and so on, and so on.

So basically-- and you know what is the probability of the fundamental frequency for males, so some distribution. So you know what is the path on the fundamental frequency for females. You know what is the probability of the first group being male. Subsequently, you also know what is the probability of the

[AUDIO OUT]

Because, to me, the features are the important as I told you, in which we don't need, but we don't want to take out stuff that you may need. I told you that that one important role of the perception is to eliminate some of this information. Basically that's to so, eliminate irrelevant focus on irrelevant stuff.

So this is where I feel the properties of perception can come in very strongly, because this is what emulates this basic process of the speech, of the extraction of information [INAUDIBLE]. Especially about the Hidden Markov models, that speech consists of the sequences of sounds and they can be previously different speed, and other things. It's important. But here, we can use a lot of our model.

Those features which can be also designed based on the data. And what comes out is probably going to be irrelevant to speech perception, so this is my point for how you can use your engineering to verify our theories of speech perception. We use largely, nowadays, the neural nets to derive the features.

So how we do it is that we sort of-- because we know that best set of features are posteriors of the class we want to recognize our speech sounds, maybe it's going to be useful. If you do a good job, actually you can do the reasonable sound. So if you take a signal, you do something processing-- and I will be talking about signal processing quite a lot.

But then it goes into neural net, nowadays, deep neural net, and you estimate a posterior use of different speech sounds. And then what comes out, whatever, it's always is the high posterior probability of the phoneme, so we have you do it [INAUDIBLE] sequence of the phoneme.

As the classes you can use, directly context independent in this example, small number. You can use context dependent phonemes, which I use quite a lot, because they've tried to optimize this despite that, if the phoneme is produced depends on what happened inside, in the neighborhood, [INAUDIBLE]

These posteriors can be directly used with our research. This is the search through the lattice of the likelihoods in recognition. And again, I mean, it's coming back. This was the late 1990, but this is the way that most of this recognizers work.

This is the major way now how you do this feature cognition. There's another way, which is called bottleneck or tandem-- we were involved in that too-- which was the way to make the neural nets friendly to people who were used to old generative HMM models, because you basically convert it, your outputs from the posteriors, into some features which your generative HMM model would like for you. What you did for you de- correlated them, you coercionized them so that they have a normal distribution and use it as a features. And bottom line is, if you get the good posteriors, you will get the good features. And we know how to use them. And this is pretty much the mainstream now.