HYNEK HERMANSKY: So we have this wanted information and unwanted information. Not-- I call-- unwanted information noise and wanted information signal. Not all noises are created equal. There are some noises which are partially understood, and I'm claim this is what we should strip off very quickly, like linear distortions of speaker dependencies. Those are two things I will be talking about. You can easily do it in a future instruction.

There are noises which are expected. But efforts may not be well understood. These should go into machine land, but this is what I'm going-- this is something which we-- whatever you don't know, you better have the machine to learn. It's always better to use the dumb machine with a lot of training data than putting in something which you don't know for sure. But what you know for sure, it should go here.

And then there is an interesting set of the whole noises, which you don't know that they are-- they even exist. These are the ones which I'm especially interested, because they cause us the biggest problems. Basically, the word you don't know that it exists, noises which somebody introduces-- they've never talked about, and so on, and so on. So I think this is an interesting problem. Hopefully, I will get to it towards the end of the talk-- at least a little bit.

So some noise is with known effects. One is like this. When you have a-- you have a speech-- oh.

VOICE RECORDING: You are yo-yo.

HYNEK HERMANSKY: And you have another speech which looks very different.

VOICE RECORDING: You are yo-yo!

HYNEK HERMANSKY: But it says the same thing, right? I mean this is a child. This is a adult. And you can tell this was me. This was my daughter when she was 4-- not 30.

The problem is that different human beings have different vocal tracts. Especially when it comes to children, vocal tract is much, much shorter. And I was showing you the effects of what happens is that you get a very different set of formants like these dark lines, which a number of people believe we should look at if we want to understand what's being said.

We have four formants here, but we have only two formants here. They are in approximately similar positions, but where you had a fourth formant, you have only second formant here.

So what we want-- we want some techniques which would work more like a human perception, not look at the spectral envelopes. But mainly you look at the whole clusters. So this is a technique which has been developed a long time ago, but I still mention it because it's an interesting way of going about the things.

So it uses several things. One is that it suppresses the signals at low frequencies. You basically use this equal loudness curve. So you emphasize the parts of the signal each child heard well.

Second one that uses this is critical bands. Because you say, first step which you want to do is to integrate overcritical band. This is the simplest way of processing within the band-- is you integrate what's happening inside.

So what you do, you take your full ear spectrum. This is the spectrum which has equal frequency resolution at all times and a lot of details with this case, because it's a fundamental frequency. And here you integrate over these different frequency bands. They are narrower at low frequencies and getting broader, and broader, and broader, very much how we learned from the experiment with the simultaneous masking. So this is a textbook knowledge.

So you get a different spectrum, which is unequally sampled. So, of course, you go back into the equal sampling, but you know that there is fewer samples at the high frequencies, because you are integrating more spectral energy at high frequencies than in low frequencies. And you multiply these outputs through these equal loudness curves.

So from the spectrum you get something which has a resolution, which is more like auditory-like. Then you put it through the equal loudness curves, because you know the loudness depends on a cubic root of intensity. So you get a modified spectrum.

And then you find some approximation to this spectrum-- auditory spectrum-- saying, that I don't think that all these details still have to be important. I would like to have some control of how much spectral detail I want to keep in. So the whole thing looks like-- let's start with the spectrum.

You go through-- a number of steps. And you end up with the spectrum, which is, of course, related to the origin of the spectrum, but it's much simpler. So we eliminated information about fundamental frequency. We merged number of foramens and so on, and so on. So we follow our philosophy. Leave out the stuff which you think may not be important.

You don't know how much stuff you should leave out. So if you don't know something, and you are engineer, you run an experiment. You know, research is what I'm doing when I don't know what I'm doing-- supposingly. I don't know where now-- from Brown or somebody was saying that.

So we didn't know what's-- how much smoothing we should do if we want to do our-- speaker independent representation. So we ran an experiment for a number of smoothing, a number of complex poles telling you how much smoothing you get through auto-regressive model. And there was a very distinct peak in the situations where we had a training coming for-- templates coming from one speaker and the test coming from another speaker.

Then we used this kind of representation in a speech recognition-- I mean in a-- to derive the features from the speech. Suddenly, these two pictures start looking much more similar, because what this technique is doing is basically interpreting the spectrum in a way the hearing might be doing. It has much lower resolution that normally people would use. It has only two peaks, right? But they say that it was good enough for speech recognition.

What was more interesting about-- a little bit of interesting science is they also-- found that difference between production of the adults and the children might be just in the length of the pharynx. This is a back part of the vocal tract; that the children may be producing speech in such a way that they are putting already the-- [AUDIO OUT] constriction into right position against the palate.

And because they know-- or they-- well, whatever-- mother nature taught them that pharynx will grow in the lifetime. But the front part of the vocal tract is going to be similar. So it is the front cavity which is speaker independent, and it is the back cavity, the rest of the vocal tract,

which may be introducing speaker dependencies.

It's quite possible if you will-- if you ask people how they've been treated-- like actors-- on how they are being trained to generate the different voices, they are being trained to modify back part of the-- vocal tract. Normally, we don't know how to do. But there is some circumstantial evidence for this might be at least partially true.

So what is nice is that when we synthesize the speech and we made sure that front cavity is always in the same place, even when the foramens were in different positions, we were getting very similar results.

So we have this theory. The message is encoded in the shape of the front cavity. Through speaker-dependent vocal tracts, you generate the speech spectrum with all the formants. But then there comes the speech perception fine point, which extracts what is called perceptual second formant. Don't worry about that. Basic-- [AUDIO OUT] on the-- at most, two peaks from the spectrum. And this is being used for decoding of the signal, speaker independently.

However, I told you one thing, which is don't use the textbook data and be the exact-- [AUDIO OUT]. And so I was challenged by my friend, late Professor Fred Jalinek. He is claimed to say, airplanes don't flap wings, so why should we be putting the knowledge of the hearing in-- actually, he said something quite different. This is what *The New York Times* quoted after he passed away, because that was one of-- supposedly one of his famous quotes. No, he said something else.

Well, if airplanes do not flap wings, but they have a wings nevertheless. They use some knowledge from the nature in order to get the job done. The flapping of the wings is not important. Having the wings is important if you want to create the machine which is heavier than the air and flies. So we should try to include everything what we know about human perception, and production, and so on.

However, we need to estimate the parameters from the data, because don't trust the textbooks and that sort of thing. You have to derive in such a way that is relevant to your task.

What I wanted to say-- you can use the data to derive the similar knowledge. And I want to show it to you. What you can do is to use a technique again, known from the '30s, called Linear Discriminant Analysis. This is the statistician's friend.

For this you need a within-class covariance and between class covariance matrix. You need the labeled data. And you need to make some assumptions, which turns out are not very critical. But I mean, they are approximately satisfied when you are working with the spectra.

So what we did was we would take this spectrogram, and we would generate the spectral vectors from that. So we would always cut part of the spectrum, or short term the spectrum, and we assign it to the label from which part of speech it came from. So this one would have a label, "yo," right?

And so you get the big box full of vectors. All of them are labeled. So you can do LDA. And you can look what they are-- discriminants are telling you. From LDA, you get the discriminant matrix and each of the row of the-- column of the-- column or row-- whatever-- creates the basis on which you should project the whole spectrum, right?

These are the four obvious ones here. You also have the amount of variability present in this discriminant matrix which you started with. What you observe, which is very interesting, is that these bases tend to project the spectrum at the beginning, with more detail, than the spectrum at the end. So essentially, they tend-- they appear-- in the first-- group they appear to be emulating properties of human hearing, with some of the well known properties of human hearing, namely, non-equal spectral resolution being verified in any ways-- many, many ways. Among them was one-- I was showing you this masking experiment of Harvey Fletcher.

There is a number of reasons to believe that this is a good thing. This is what you see. So essentially, if you look at the zero crossings of these bases-- this is the first base-- they are getting broader and broader. So you are integrating more and more spectrum, right?

This is all right-- so that I leave it. Oh, this is from another experiment with a-- very large database, very much and very similar thing. Eigenvalues quickly decay.

And what is interesting-- you can actually formally ask, "What is your resolution?" by doing what is called perturbation analysis. So you take some-- say, some signal here. This is a Gaussian. And you project this on this LDA basis. Then you perturb it. You move it.

And you ask, how much effect this movement of this, say, emulated spectral element of the speech causes the output-- seen as the output of this projection of this many-- on these many bases? And what do you see is, as I was suggesting, spectral sensitivity to the movements of the formant is much higher at the beginning of the spectrum and much less at the end of the

spectrum. You can actually compare it to what we had-- initially in the PLP analysis when we integrated the spectrum based on the knowledge coming from the textbook. And it's very much the same if there were just a plain cosine basis computing mel cepstrum sensitivity is the same at all frequencies.

But these bases-- are from the LDA, they're very much doing the thing which physical bin analysis would be doing. And so this is a-- you can look at it. It was a PhD thesis from Oregon Graduate Institute by Naren Malayath, who is now a big-- you better be friends with him. He's at Qualcomm. I think he's a head of Image Processing department. [COUGHS] We better-- better be good friends with him. [INAUDIBLE].

[LAUGHTER]

OK. Another problem with linear distortions-- linear distortions void was a problem. Now it's not problem anymore, but in the old days, this was a problem. A problem shows up in a rather dramatic way, in following way. Here we have one sound.

VOICE RECORDING: Beat.

HYNEK HERMANSKY: Beat. So "buh ee- tuh." Here is the very distinct E. Every phonetician would agree. This is E-- high foramen, cluster of high of foramens, and so on.

Some vicious person, namely one of my graduate students, took this spectral envelope, designed a filter, which is exactly in reverse of that, and put this speech through this inverse filter so it looked like this. This was-- there is a spectrum where there were nine formants. It's entirely flat. And if you listened to it, you've probably already guessed what you will hear.

VOICE RECORDING: Beat.

HYNEK HERMANSKY: You'll hear the first speech, right? But you--

VOICE RECORDING: Beat.

HYNEK It's OK when this-- oops! That's what you would-- sorry.

HERMANSKY:

VOICE RECORDING: Beat. Beat. Beat.

HYNEK HERMANSKY: But whoever doesn't hear E-- don't spoil my talk. I mean, I think everybody has to hear E, even though any phonetician would get very upset. Because they say this is not E. Because, of course, what is happening, human perception is taking relative-- percept relative to these neighboring sounds, right? And since we filtered everything with the same filter, I mean, relative percepts is still the same. So this is something which we needed to put into our machine.

And we did-- actually, this is a very straight-- signal processing wise, the things are very straightforward, because what you have, the signal is actually signal of the speech convolved with the spectrum of the impulse response of the environment. So in logarithmic domain this is the signal processing stuff. Basically, you have a logarithmic spectrum of the signal plus logarithmic spectrum of the environment, which is fixed.

So-- what we're finding here is that if you remove somehow this environment, or if you make it invariant to this environment thing, then you maybe win-- you may be winning. The problem here is that each frequency you have a different amount of additive constant, because this is a spectrum, right? If it was just a constant at all frequencies, you just subtract it.

But in this case, you can use the trick. You remember what Josh told us this morning. Hearing is doing spectral analysis. And what I was trying to tell you, that each-- at each frequency, each critical band is trajectory of the spectral energy is independent of-- in the first approximation, of the other-- others. You can do independent processing at each frequency band-- and maybe don't screw up that many things.

So this was a step which we took. We said, OK. We will treat each sample trajectory differently, right? But we will filter out stuff which is not changing.

So related to different frequency channel, do the independent processing channel. And processing was that we would take first logarithm, and then we would do-- then we put each trajectory through a bandpass filter, which would-- main thing is which would be suppressing DC, and slowly changing components. Mainly it was suppressing anything which was slower than one hertz. And also, it turned out it was useful to separate things which are higher than

about 15 hertz.

So what you get out-- this was the origin spectrogram. This was the modified spectrogram. It seems it got a little bit-- this trajectory got a little bit smoother. Transitions got smoothed out because there was a bandpass filter. There was a high pass elements to that. Very much what we thought-- well, this is interesting. Maybe this is what a human hearing might be doing.

To tell you the truth, we didn't know. It was-- for the people who are from MIT and who are working in Image, it was inspired by some work on a perception of lightness, what David Marr called lightness. And here was the type-- thing which I told you about-- 6 by-- 6 by 749. David Marr was talking about processing in space. We applied it in processing in time. But it was still good enough. I mean, so that we definitely got rid of the problem.

So here it is. The spectrograms, which look very different, look-- suddenly start looking very similar. I was good seeing what's here. Remember, I'm an engineer. I was working for a telephone company at the time. It was still working better in some problems which we had before. And we had a severely mismatched environment, getting the training data from the labs and testing it in the US west, in Colorado. So it didn't work at all; recognized after this processing everything was cool and dandy.

OK. So now we have a RASTA LDA. And we can do the same trick. How about that?

So we take-- you take the spectral temporal vectors, and you label each of these vector by the label which is of the phoneme, which is in the center of this trajectory. And just for the sake of-- just to have some fun, we took a rather long vector. It was about one second. And we said, well, what kind of projections would these temporal trajectories would go on if we wanted to get rid of a speaker-dependent-- I mean, of an environment-dependent information?

Well, these were the impulse responses. These were the frequency responses. Because in this case, you get a FIR filters. These discriminants are FIR filters, which are to be applied to temporal trajectory-- so spectral energies. Because it's just basically projection of the spectrum on the basis.

And basis is one second long. This is impulse response. It cannot be all that long, because eventually, they become zero, right, if you should do nothing to it. You do do nothing. But you can see active part is about a couple of hundred millisecond, maybe a little bit more. And these are the bandpass filters. So essentially-- passing frequency between the 1 hertz and 10,

15 hertz-- very similar at all frequencies.

There was another thing which we were very interested in. We should really do different things at different frequencies. Answer is pretty much no. And so that was very exciting. What-- what I-- well, anyway, so let me tell you-- yet another experiment which hasn't happened and is going to be presented next week.

Well, we wanted to move in the 21st century, so we did convolutive neural network. And our convolutive network is maybe not what you are used to when you have a 2D convolutions there. But we just said, we will have a 1D filter as a first processing step in this deep neural network.

So we postulated the filter, the input, to the neural network. But in this case, we trained the whole thing together. So it wasn't just LDA and that sort of thing. So we forced all filters at all frequencies be the same, because we expected that's what we want to get. And we were asking how these forces look like, which come from the convolutive neural network.

Well, again, I wouldn't be showing it if it wasn't really somehow supportive of what I want to say. They don't look all that different for what we were guessing from LDA. They definitely are enhancing the important modulation frequencies around four hertz, right? They are passing a number of them. I'm showing three here, which are somehow arbitrary.

They are passing-- and most of them look like that. And we will use the 16 of them. Passing between in 1 and 10 hertz in modulation spectral domain, so changes which are 1 to 10 times a second. It's coming out in a paper, so you can look it up if you want.

Last thing which I still wanted to do-- I said, well, maybe it has something to do with the hearing after all. We were are deriving everything from speech. There was no knowledge about hearing, except that we said we think that we should be looking at long segments of the signal, and we expect that this filtering will be very much the same at all frequencies. Actually, not even-- earlier it come out automatically.

There wasn't much of the knowledge from human hearing-- [AUDIO OUT] on. In the first one, when I was showing you critical band spectral resolution, we started this full ear spectrum. We didn't tell anything about human hearing.

And what comes out is a property of human hearing. I mean, tell me if there is yet another strong evidence that speech is processed in such a way that fits human hearing, because the

only thing which was used here was the speech, labor, into certain-- into classes which we are using for recognizing it-- speech sounds.

So what we did was-- that was with Nema, with Garani, and my students. We took a number of these cortical receptive fields-- which we talk about it before a little bit-- about 2,000, 3,000, whatever-- we spread-- we basically spread to the floor at the University of Maryland and computed principal components from these fields in both spectral and temporal domain. But here I'm showing the temporal domain.

How they look-- they are very much like the rest of filter. This is what is happening. It's a bandpass. Peak is somewhere around four hertz. Essentially, I'm showing you here what I understood might be a transfer function of the auditory cortex derived with all the usual disclaimers, like, this is a linear approximation to the receptive fields. And there might have been problems with collecting it, and so on, and so on. But this is what we are getting as a possible function of auditory cortex.

I'm doing fine with the time, right?

So you can do experiment in this case. You can actually generate the speech which has certain rates of change eliminated. By doing all this, computing the cepstrum, do the filtering of each trajectory, and reconstruct the speech. And ask people what do they hear? How do they recognize speech?

You can also ask, "Do you recognize it?" For this you don't have to regenerate the speech. But you just use therapy C cepstrum.

This is the full experiment with the-- this is for-- this is called a residual-excited LPC vocoder. But it's modified in such a way that you artificially slow down or modify the temporal trajectories, which are being-- if there is no filter, you cannot make a replica of the origin, the signal here.

So just the bottom line of the experiment here is what-- if you start removing components which are somewhere between 1 and 16 hertz, you are getting hurt significantly. Most you are getting hurt in performance when they are removing component between 2 and 4 hertz. This is a-- that's how you are getting biggest hit.

Here we are showing how much is-- how much these bands contribute to recognition and

performance by humans. This is a white bars; and to speech recognize it. Those are the black bars.

So you can see that in speech recognition, machine recognition, you can safely remove stuff between 0 and 1 hertz. It's not going to hurt you. It's only helps you in this task.

Speech perception, there is a little bit of hit, but certainly not as much hit as you are getting when you are moving to the part where you hear the-- [AUDIO OUT]. And certainly the component that is higher than 16 or 20 hertz are not important. Then, already, Homer Dudley, he knew, in 1930, when he was designing his vocoder.

But it was a nice experiment. It came out in just-- so you can look it up and-- if you want to have a go.

Just to summarize what I told you so far, Homer Dudley was telling us information from the-- information about the message in the slow modulation, slow movements of the vocal tract, which modulates the carrier; information about the message in slow modulations of the signal-- slow changes of speech signal in individual frequency bands.

Slow modulations imply long impulse responses, right? So 5 hertz, I sense something around 200 millisecond. My magic number of what physically needs to allow, which we have observed in this summation of sub-threshold-- the signals and temporal masking. And so I have to hear a number of things which I listed.

Frequency discrimination improved. If you are longer than 200 millisecond, below 200 milliseconds of signal, you don't get such a good frequency discrimination.

Loudness increases up to 200 millisecond, then it stays constant. It depends on amplitude.

Effect of forward masking-- I was showing you-- asked about 200 millisecond independent of the amplitude of the masker.

And sub-threshold integration is also showing you. So I'm suggesting there seem to be some temporal buffer in human hearing on some level. I suspect it's cortical level, which is processing. Whatever happens within this buffer, it's a fair thing to treat as a one element.

So you can do filtering on it. You can integrate it; any-- basically, all kinds of thing. If things are happening outside this buffer, these parts should be treated-- [AUDIO OUT] in parts.

So how does it help us? You remember the story about the phonemes. You remember that phonemes don't look like this, but they look like this. Length of the coarticulation pattern is about 200 millisecond, perhaps more.

So what is a good thing about it is that if you look at sufficiently long segment of this signal, you will get whole coarticulation pattern in. And then you have a chance that your classifier is getting all the information about the speech sound for finding the sound.

And then you may have a chance to get a good estimate of the speech sounds. But you need to use these long temporal segments. And here I can say it even to YouTube. I think we should claim the full victory here, because most of the speech recognition systems do it nowadays. They use the long segments of the signal as a first step of the processing.

So I can happily retire telling my grandchildren, well, we knew it. We were the only ones. But I mean, we were certainly using it for a long time and probably for a long time, in such a way that we even designed several techniques that you do there.

So this is a classifying speech recognition from the temporal patterns directly. So we would take these long segments of the speech through some processing, put neural nets on every-- every temporal structure, trying to estimate the sound at each frequency-- each carrier frequency. And then we would fuse all these decisions from different frequency bands. And then we would use the final vector of posterior probabilities.

Unlike what people do very often-- most often-- that they just take the short term spectra, and then they maybe now take the longer segment of this block of these short term spectra. We say, short term spectrum is good for nothing. We just cut it into pieces. And we classify each temporal trajectory individually in the first step. Tell now that it was used-- it may be useful later when I will be telling you about dealing with some kind of noises.

But you understand what we did here, right? Instead of using the spectral temporal blocks, we would be using temporal trajectories at each critical event, very much along the lines of what we think that hearing is doing with the speech signal.

First thing is, hearing is doing. It takes the signal, sub divides it into individual frequency bands, and then it treats each temporal trajectory coming from each of these cochlear filters to extract the information. And then it tries to figure out what to do with this information later, right?

Well, we have another technique called MRASTA, just for people who are interested in cochlear-- I mean, in-- cortical modeling. You take this data. We project a number of projections with variable resolution. So we get a huge vector of the data coming from different parts of the spectrum. And then we feed it into speech recognizers.

The first test looked like this. I mean, they have a different temporal resolution, spectral resolution. We are pretty much integrating or differentiating over three critical bands following some of the filters, coming from these three-- I mean, old PLP low order model and three bark three bark critical event integration.

So these ones look a bit like what people would call Gabor filters. But they are just put together, basically, from these two places in time and in frequency-- different temporal resolution enhancing different components of moderating the spectrum. Again, you may be claiming that this is something which resembles the Thorston-- [AUDIO OUT] Josh was mentioning in the morning. It's cochlear filter banks-- [CLEARS THROAT] auditory, of course. I mixed up cochlear and cortical-- cortical filter bank, modulation filter banks.

So there are some novel aspects in this type of processing I want to impress. It was novel in 1998. That is, as I said, this is fortunately becoming less novel 15 years later.

Use is rather long temporal context of the signal as a input. It uses already hierarchical neural nets. So deep neural network processing, which wasn't around in 1998. The only thing was that there was independent processing of frequency of neural net estimator at frequencies.

The only thing which we didn't do at the time, and I don't know how important it is. I don't think it doesn't hurt-- it hurts anybody. They were training these parts of the system, this deep neural net individually. And it's just concatenated output. So we never did training all together as we do now in convoluted nets and that sort of thing. Because simply, we didn't even dream about doing that, because we didn't have the hardware.

That was one thing which I tried to point out during this panel. A lot of progress in neural nets research and success of neural nets comes from the fact that we have very, very powerful hardware, which we didn't have. So we didn't dream about many things doing, even when they might have made sense. So, OK.

Where are we? Oh, I see-- one more thing.

Coarticulation. This is a problem which is known since people started looking at the spectrograms. There's some consonants like a "kuh", or "huh." They are very dependent on what's following. So "kuh" in front of "ee, kee, koo, kah, koo-kah," we [AUDIO OUT] has a burst here. In front of "ooh" has a burst here. And in front of "ah" there's a burst here.

So the phonemes are very different depending on the environment. When you start using these long temporal segments, you know all the tricks, or some of the tricks, I showed you about, what comes out are the posteriogram in which the "kuh" almost looks the same as a "kuh." It basically recognizes this-- recognizes that since it looks at the whole coarticulation pattern or group of the phonemes, in order to recognize this sound, it does the right thing. So I suspect that success of these long temporal context which people are using now with speech recognition, comes from the fact that this partially compensates for the problems with-- by problems for the coarticulation.

And what I also want is to say-- coarticulation is not really a problem. It just spreads the information for a long period of the time. If you know how to suck it out, it can be useful. But it's a terrible thing if you start just looking at individual frequency events, even with your frequency's slices of the short term spectrum. So it's another deep net-- deep net from, I don't know the name. Sorry. It was already almost legal deep net.

You do the-- you estimate the posteriorgram from the short window in the first step for about 40 mm length window. And then you take the long-- I mean, big window of the posteriors' tree. Another neural net you get much better, which also work better. Again, the mainstream technique nowadays is being used in most of the DARPA systems.

Oh, yes-- one more thing. I want to stress this one. [LAUGHS] So, I'm sorry I didn't want to show it all at the same. But anyways, I don't think that there's anything which is terribly special about short term spectrum of speech. I think what really matters is how you process the temporal trajectories of the spectra energies. This is what the human hearing is doing that seems to be doing a good job on our speech recognizers.

So essentially, this is one message which I want to say. Don't be afraid to treat different parts of the spectrum different. Individually you may get some advantages from them. It started with your stub, but it shows up over and over again.

So away from the short term spectrum, go away, they start using what hearing is doing-- start using a temporal trajectories of the spectra energies coming from your analysis.

To the point that we did this work on real [INAUDIBLE], on the-- going directly, don't do this. And don't get your time frequency patterns from the short term spectra. I think about always how to get directly what you want.

It turns out that there is a nice way of doing-- for estimating directed hilbert envelopes of the signal in the frequency bands called frequency domain linear prediction.

[STATIC]

Mario says-- there's his PhD thesis. And we were working together for a couple of years.

So what you do, instead of using the time trajectory, so use this case, autoregressive modeling-- LPC modeling-- and put the windows on a time to get the frequency-- frequency vectors. You do it on a cosine transform of the signal.

So you move the signal into a frequency domain. And then you put the windows on this cosine transform of the signal. And you derive directly the-- all polar approximations to hilbert envelopes of the signal in the sub bands. You don't ever do the hilbert transform. You just use the usual techniques from autoaggressive modeling. The only difference is-- [AUDIO OUT] on the cosine transform of the signal. And your windowing determines which frequency range you are looking.

So, of course, what you typically do, you can use the longer windows at higher frequencies, shorter window of lower frequencies. You do all these things. But this is a convenient way. [COUGHS] It's convenient, and this is more and more like fun. But maybe somebody might be interested in that.

So essentially, what you do-- oops. Sorry. What you do is that you take the signal, and you eliminate modulation component out of that AM component which the signal is being modulated. So this carries the information about the message. And this is the carrier itself.

And you can do what is called channel vocoder, which we did. And you can listen to the signal. So this is-- in some ways it's interesting-- original signal.

VOICE RECORDING: They are both trend-following methods.

HYNEK Oops. I tried to make it somehow-- [AUDIO OUT].

HERMANKSY:

VOICE RECORDING: They are both trend-following methods.

HYNEK HERMANKSY: Somebody may recognize Jim Glass from MIT in that.

VOICE RECORDING: In an ideological argument, the participants tend to dump the table.

HYNEK HERMANKSY: So this is silly, right? Now you can look at what you get if you just keep the modulations and excite you know, with the white noise. Oops. Sorry. Oops! What am I doing? Oh, here.

VOICE RECORDING: (WHISPERING) They are both trend-following methods.

HYNEK HERMANSKY: Do you recognize Jim Glass? I can.

VOICE RECORDING: (WHISPERING) In an ideological argument the participants tend to dump the table.

HYNEK HERMANSKY: And then you can also listen to what is left after you eliminate the message.

VOICE RECORDING: Mm-hmm. Ha, ha.

[LAUGHTER]

HYNEK HERMANSKY: Maybe it's a male, right?

VOICE RECORDING: Mm-mm [VOCALIZING]

HYNEK HERMANSKY: Oh, this is fun. This is [CHUCKLES] fun. It may have some implication for speech recognition. But certainly, if I have seen one verification of what old Homer Dudley was telling us-- where

the message is-- I mean, this is it. All right?

Anyways, what is good in-- for that, is that once you get an open-- [AUDIO OUT] it's relatively very easy to compensate for your ear distortions. Because main effect of linear distortions is basically shifting the energy a different frequency by-- bends by different amounts. But all this information is in the gain of the model. It's one parameter which you essentially ignore after you do this frequency domain linear prediction. And you get a very similar trajectory for both.

This is a telephone speech and clean speech which differed quite a bit. And I hope that I have-- oh, this is for reverberant speech. There seem to be also some advantage, because reverberation is in the first information, it is a convolution with the impulse response of the room.

So you make the-- if you use a truly long segments-- in this case, we used about 10 seconds of the signal approximating by this open model, and eliminated the DC from that. You know, it seems to be getting some advanced-- [AUDIO OUT].

So known noise with unknown effects. I say train the machine on that.

Here is the one example, right? You have a phoneme error rates, noise estimate. If everything is good, clean, trading clean test, you have about 20% phoneme accuracy. This is a stage of the result-- reasonable result. But once you start adding a noise, things quickly go south.

Typical way of dealing with it is if you train multi-style. So if you know which choices you are going to deal with, you train on them. And things get better, but you pay some price. I mean, certainly, you pay the price on clean, because you recognize your model became must mushier, basically. It's not a very sharp model.

So here we had a wonderful 21%. We paid 10% relative price for getting this better performance on the noises.

What we observe is that you get much better results, most noticeably better results, if you would have different recognizers for each type of noise.

But of course, the problem is that you have different types of noise. So you have this number of recognizers. But now you need to pick up the best stream. And how do you do that?

This is something, again, which I was mentioning also earlier. This is something which we are

struggling with and we don't know how to do that. If you are a human being, maybe you can just look at the output. And you can see, just keep switching after your message starts looking reasonably well. But if you want to do it fully automatically, I don't know why we want to only build a fully automatic recognizers, but that's what we are doing.

So you want to pick up the-- you want a system to pick up the best stream. So how do we do that? First thing is, of course-- one way is to recognize type of noise. This is a typical system nowadays. You recognize type of the noise, and you use the appropriate recognizers. BBN is doing it.

My feeling is that it's somehow cleaner and more elegant to be able to figure out what is the right output, because what-- neither. It's not what is the signal, but what is the signal interacting with the classifier? So for this we have to figure out what the best means.

So here we have two posteriograms. If you look at it, if you know that these are trajectories of the posteriors-- of the speech sounds, you know this one is good. This one is not so good. Because the word is nine-- "ne-ine," "ne," right? Here is a lot of garbage.

So I will know that-- I will do it automatically. So ideally, I would pick up the stream which gives me the lowest error. But I don't know what the lowest error is, because I don't know what the correct answer is. That's the problem, right? So one is to maybe try to see what I-- what my I did? Try to figure out which posteriogram is the cleanest.

Another one is following thinking. When I trained the-- I trained the neural net on something. It's going to work well on the data on which it was trained. So I have some gold standard output.

And then I will try to see how much my output differs if the test data, which are not the same as the data on which the recognizer was trained. So both of these tricks we were using.

So first one uses a technique which is like this. You look at the differences between posteriors or KL divergence areas a certain distance from each other, and you will slice this window. You cumulatively cover as much data as you possibly can.

And what you observe is that if you have a good, clean data, this cumulative divergence keeps increasing. And after you cross the point where there is a coarticulation pattern or coarticulation ceases, suddenly you start getting pretty much the fixed high tail divergence-- cumulative KL divergence.

If we have a noisy data, the noise start dominating this KL divergences and differences RS. Because the signal in the first place carries the information, and the information is in the changes. But noise is creating the information which it doesn't have these segments, or something. So this is one technique which we use.

Another technique which is even more, now, popular, in at least in my lab, training of another unit. We trained this autoencoder on the output of a classifier. And we say-- so we-- on the output of the classifier as it's being used on the training data, and we say autoencoder. Then we learn how in average the output from the classifier used on its training data looks like. And if-- and then we use it on the output from the classifier used to unknown data. And if the autoencoders then knew, then we tried to predict input on its output. This is how it is being trained.

So if the output-- if the prediction is not very good, then we say we are probably dealing with the data for which the classifier is not good. It's how it works. I mean, you know, it's honest.

If you are looking at the output of the neural net which has been applied to a-- towards training data, the test is-- or the test is maybe-- the training data is matched. Pretty much the error is very similar as it goes on the training data. When you apply to a data for which the classifier wasn't trained, your error is, of course, much larger.

Prediction is prediction of the output of the-- so there's a double-- there is a double deep net. One is classifying, and then another one is predicting its output. One-- and the one which predicts output is trained to predict its best output it can possibly have, which is the output on the training data of the previous classifier.

I don't know if you follow me or if it is becoming too complicated. [CHUCKLES] But essentially, we are trying to figure out if anything like it is looking on a training-- and applied on the training data.

So it seems to be working to some extent. Here we have a multi-style results. Here we have a matched result. This is what we would like to achieve. Of course this is oracle. This is what we-- it will be ideal if we knew which stream is best.

This is what we would be getting. But what we are getting is not that terribly bad. I mean, certainly, it is typically better than a multi-style training.

All right. And we have some ways to go to oracle like-- not too far from the matched trace. Sometimes it's even there, because it's making a decision on every utterance. So sometimes it can-- going to do quite well. So we were capable of picking up the good streams and leaving out the best bad streams, even at the output of the classifier.

How does it work on previously unseen noise? Fortunately, for this example, we still seem to be getting some advantage. We are using noise which has never been seen by the classifiers. But still, it was capable of picking up the good classifier, which is actually better than any of these classifiers. Sort of a, so this seems to be good.

Another technique of dealing with unseen noises is actually one which I like maybe even a bit more, which is you do the pre-processing, and in some processing in frequency bands, hoping the main effect of the different noises is in the shape, spectral shape, of different noise. So if you are doing recommission in the sub bands, then the noise start looking-- in each sub band starts looking more like a white noise except of different levels.

So meaning, it's here, that maybe the signal to noise ratio is higher. Here it is more miserable. But if I have the classifier, which is strained on multiple levels of the white noise, in each frequency band perhaps I can get some advantage. So I do what in the cochlear might be doing, which is I divide each-- the signal into a number of frequency bands, and then I have one fusion DNN which will try to put these things together. But each of these nets are going to be trained on multiple levels, but this time of the white noise. And it's going to get back to noises which are not white, or the noise which is not white.

So how it works, you can figure it out to see if it was done in the case of Aurora. So here we have the examples here-- how it works on a-- in matched situations. Here is multi-style training. But here it is what you are getting if you apply this technique. This is what you get now, multi-style training. But in the sub band recommendation, you are getting 1/2 of the error rate.

Just a simple trick which I think is reasonable, which is you do the sub band recognition. There's a number of power recognized, each of them paying attention to part of the spectrum. And then you-- each of them is being trained to head the white noise, a simple white noise. But you turn, in some ways, arbitrary additive noise, this car noise, into white-like noise in each sub band. And that what's you get.

So in general, dealing with unexpected noise is-- you want to do the adaptation. You want to modify your classifier on the fly. You want to have parts of the classifier or some streams which are doing well. And some of the parts-- so parts of the classifier are still reliable. And you want to pick up these streams which are reliable on unseen situation.

So this is what we call this multi-stream recognition-- adapt to multi-stream adaptation to unknown noise. So you assume that not all the streams are going to give you good result, but you assume that at least some of the streams are going to have good results. And all these streams are being trained on, say, clean speech or something.

So this is this multi-band processing, all right? So this is what we do. We do different frequency ranges. And then we use our performance monitor to pick up the best stream.

So here is the experiment which we did. So you would have the 31 processing streams created from all combinations of fine frequencies. And one stream was looking at full spectrum. And the other things were only looking at the parts of the spectrum.

So more black ones, there is a more spectrum; more white, some of them are looking only at a single frequency band. So we have a decent number of processing channels, and we would hope that if the noise comes here, maybe this one is going to be good, because this one is not going to be-- not worth picking up this noise, or recognize that basically which only uses the bands which are not noisy. It's going to be good.

So this is the whole system. It was published in 230-- [STATIC] like a entire speech. We had this sub band recognition, fuse, and performance monitor, and the selecting of a stream. This is how it works. This is again, showing for car noise. Car noise is very nice, because it mainly, it corrupts the low frequencies.

So all these sub band techniques work quite well. But you can see that it's pretty impressive, and it's-- if you didn't do anything, you get 50% error. With this one you get 38% error. If you knew which bands to pickup you would be getting oracle experiment-- cheating experiment. You would be getting about 35%. So that was, I thought, quite nice.

Just to conclude-- so, auditory system doesn't only look like this, that it starts with a signal as the analysis, and then it reduces the-- reduces the bit rate; but it also increases number of views of the signal. And this is based on the fact that there is a massive increase in number of the cortical neurons on the level of the cortex.

So there is many ways of describing information at high level of perception. Essentially, the signal doesn't go through one pass, but it goes through many, many passes. And then we need to have some means, or we have some means to pick up the good ones and ignore the other ones-- maybe switch them entirely off. This is like vision.

So this is the path of processing, in general, signal. You get the different probability estimates for different streams. And then you need to do some fusion and decide on-- based on the level of the fusion.

How you can create the streams-- but we were showing you differently trained probability estimates on different noises-- different aspects of signal, that is, different parts of the spectrum of the signal. But you can go wild. You can start thinking about different modalities, because there-- as we talked about it also in the panel, you know, very often audiovisual model, if it carries the same information about the same things, [STATIC] infusion of audio visual streams.

You can also imagine the fusion from streams with different levels of priors, different levels of hallucinations. So basically, this is what I see human beings are doing very often. If the signal is very noisy, you are at a cocktail party, you are guessing, because that's the best way to get through if the communication is not very important. It's not about your salary increase, but about the weather-- so basically, guessing what the other people are saying, especially if they speak the way I do, right, with a strong accent or something.

So the priors are very important. And streams with priors are very important. We use this to some extent, I was mentioning, by comparing the streams of different-- with different prior to discover if the signal is biased in the wrong way by priors.

So stream formation-- there is a number of PhD theses right here, right? I think. Fusion-- oh, why? It's select the best probability estimates. I tell you. This is the problem which I was actually asking for, please help me to solve it. Because we still don't know how to do that. I have a-- I suspect that especially in human communications, people are doing it like this, which starts making a sense if they use the certain processing strategy.

So people can tell if the output of our-- this perceptual system makes sense or not. Our machines don't know how to do it yet.

Conclusion. So some problems with the noise are simple. You know, you can deal with it on a

signal processing level by filtering the spectrum, filtering the trajectories, because these effects are very predictable. And if you understand them, you should do it. Because there's no need to train on that. They said there is-- you just do it. And things may be working well.

Unpredictable effects of noise can be-- typically are being handled by now by multi-style training. And these amounts of training are enormous nowadays. You know, if you talk to Google people, they say we are not deeply interested in this-- what you are doing, because we can always collect more data from new environments. But I think it's not-- I shouldn't say dishonest. I'm sorry. Scratch it.

[LAUGHTER]

It's not the best engineering way of dealing with these things, because I think the good engineering way of dealing with those things is to get away with less training and that sort of thing; and maybe follow what I believe that human beings are doing. So we have a lot of parallel experts working with the different aspects of the signal, giving us different pictures. And then we need to pick up the good ways of being.