EPISODE 591

[INTRODUCTION]

[0:00:00.3] JM: A sample of the human voice is a rich piece of unstructured data. Voice recordings can be turned into visualizations called spectrograms. Machine learning models can be trained to identify features of these spectrograms. Using this kind of analytic strategy, breakthroughs in voice analysis are happening at an amazing pace. Much like image recognition and product recommendations through collaborative filtering, breakthroughs in voice technology are really astounding, applying fairly basic machine learning tools.

Rita Singh researches voice at Carnegie Mellon University. Rita's work studies the high volume of latent data that is available in the human voice. As she explains just a small fragment of a human voice can be used to identify who a speaker is. Your voice is as distinctive as your fingerprint. Rita is really excited about her work and in this episode you will see why. Some of the stuff is just mind-blowing.

Your voice can reveal explanations about your medical conditions. Features of the human voice can be strongly correlated with psychiatric symptom severity and potentially things like heart disease, cancer, other illnesses. The human voice can even suggest a person's physique, your height, your weight, your facial features, and if you think about, this kind of makes sense because there are so many muscles that go into the production of the human voice and bones contribute to the bones contribute to the musculature of those muscles that are contributing to the human voice.

It makes sense that there is a very distinctive model of your physique that leads to your voice and it would be unsurprising if things like the level of dampness in your throat, which could be correlated with a cancer or some other illness. Maybe that could affect your voice in a specific way and could give a strong indication of you having a certain illness just by getting a recording of your voice and being able to derive the amount of dampness in your throat from that vocal recording.

In this episode, Rita explains the machine learning techniques that she's using to uncover the hidden richness of the human voice. It was a real pleasure talking to Rita and it's just always great to hear somebody who is as enthusiastic about her work as she is, and you will find out in this episode why she is so enthusiastic, because some of the science is really incredible.

[SPONSOR MESSAGE]

[0:02:00.3] JM: Azure Container Service simplifies the deployment, management and operations of Kubernetes. Eliminate the complicated planning and deployment of fully orchestrated containerized applications with Kubernetes. You can quickly provision clusters to be up and running in no time while simplifying your monitoring and cluster management through auto upgrades and a built-in operations console. Avoid being locked into any one vendor or resource. You can continue to work with the tools that you already know, such as Helm and move applications to any Kubernetes deployment.

Integrate with your choice of container registry, including Azure container registry. Also, quickly and efficiently scale to maximize your resource utilization without having to take your applications offline. Isolate your application from infrastructure failures and transparently scale the underlying infrastructure to meet growing demands, all while increasing the security, reliability and availability of critical business workloads with Azure.

To learn more about Azure Container Service and other Azure services as well as receive a free e-book by Brendan Burns, go to aka.ms/sedaily. Brendan Burns is the creator of Kubernetes and his e-book is about some of the distributed systems design lessons that he has learned building Kubernetes. That e-book is available at aka.ms/sedaily.

[INTERVIEW]

[0:04:23.0] JM: Rita Singh is a research scientist who studies voice processing. Rita, welcome to Software Engineering Daily.

[0:04:28.8] RS: Thank you for having me.

[0:04:30.8] JM: Your work centers around how to process the human voice and how to derive signal from it. There are important problems around this, like identifying fake bomb threats. If somebody calls in a fake bomb threat, you would like to be able to identify that the person is lying on the phone. You could also use voice processing to deduce if somebody is depressed from their voice, and you also study voice impersonation, which is the idea of taking a set of voice signals and then using that to generate an impersonation of somebody else. Why do you focus on this specific field of computer science; the human science?

[0:05:15.4] RS: Okay. Let me give you a little bit of history here. So I've been working on computer speech recognition and audio processing for about 20 years, and in those 20 years, I've been working, I've been looking at search algorithms and all what have you to do with automatic speech processing systems, and I, in that period, in that process studied the human voice very closely. All of us do who are in this field, right? But I never really thought of looking at what the voice signal from the perspective of deriving information about the speaker, or I call it profiling speakers from the human voice. I did not think about that until I was challenged by the US Coast Guard with a real crime that involved a hoax caller who keep calling them over a period of five or six months. These hoax calls are really problematic for the Coast Guard because they have to fire up search and rescue missions, which are very expensive and dangerous at times.

The hoax calls that they get are usually mayday calls, which are very, very short in duration. They are made over the VHF 16 radio channel, and so they are really noisy as well. For the most part, from these short recordings, it's nearly impossible to get a clue as to who the person is and the time. They can kind of locate the person to a very large geographical area, about – We're talking hundreds of square miles and they go out searching for these people, and if they don't find them, it's a loss in many ways.

So anyway, they get a lot of these hoax calls. When this guy started bothering them, this was in 2014, they sent the recordings to a few centers around to find out more about the speaker. The challenge was, "Could you tell us more about this person? This is the only clue we have of the person." The recordings totaled about 10 seconds over a few calls made over five or six months.

So when these calls landed up at my desk, that is when I started looking at the voice signal from the perspective of finding out more about the speaker, and I discovered that in these 20 years that I had been studying the human voice, I could actually derive a lot of information from the signal, and this is because we, as a community, have been – And I'm speaking of the scientific community, have been curious about the human voice since the dawn of time, I think. Humans have been curios about the human voice since the dawn of time.

So there've been lots and lots of scientific studies about the human voice. There are more than 30 fields that have studied the human voice, and we're talking architectural acoustics, drama, medical fields, so on and so forth, psychoacoustics, even developmental psychology, things like that. There are more than 400 scientific journals that have published and continue to publish results on the human voice.

So there are so much information out there that correlates the human voice to one thing or the other, to the environment around us, to the speaker's parameters, many, many things about our physical, physiological, behavioral and medical demographic, lots of different categories of speaker parameters.

So with all that information out there, all you have to do is connect the dots, and I did. When the case came to my desk, I did what I could very quickly and manually for the most part, and I could give them about 14, 15, 16 different pieces of information, which later turned out to be quite accurate when the person was caught. So that was when in that process I realized that, A, there is a lot value in now looking at the speech signal from this perspective of deriving more information about the speaker. B, suddenly I realized that the speech signal is a treasure trove of information about you. It's like your fingerprint or DNA in a sense. It has so much information, and we don't have – I mean, we are in the age of AI. We have powerful techniques to process these signals, but we really haven't developed profiling as a science. So we don't have the algorithms and the knowhow and the pipelines that we need to get this kind of information from the speech signal. There was a science.

In my limited worldview, I would say there was a science waiting to be developed and I'm curious, I was very interested. So I dropped computer speech recognition right there and called it a solved problem, and it is now a solved problem with Google and Amazon coming up with

such good devices for speech recognition, and I moved into this field. Ever since then, I have been trying to develop this field of what I call human profiling, or profiling human from their voice as a science.

[0:10:28.7] JM: I love how enthusiastic and excited you are about this field. You can really tell from – I can tell from your voice how excited you are about this –

[0:10:38.0] RS: There you go! If you look at the spectrogram of my voice as I'm talking now, you sense enthusiasm in me, right? I am enthusiastic. I agree.

[0:10:47.3] JM: Yeah.

[0:10:47.9] RS: However, if you see my spectrogram, you're not going to be able to point to one thing and say, "Hey, that is the signature of enthusiasm." But you know, you and I both agree that the signature is there, right?

[0:11:01.9] JM: Yeah.

[0:11:02.5] RS: Where is it? How do you discover it?

[0:11:03.9] JM: Well, we know that there are some neural network architecture that we could make to identify that enthusiasm.

[0:11:13.1] RS: Not neural networks again. They're not the [inaudible 0:11:16.0]. They're not the answer to everything. They're very powerful mechanisms, but they're not the answer to everything, and certainly not the answer to profiling humans from their voice, way too many challenges.

[0:11:28.2] JM: Let's jump into it. Let's say we have a spectrogram, or we have a collection of spectrograms, we want to identify enthusiasm without using neural nets. What are we doing?

[0:11:39.3] RS: Okay. Let's try to solve that problem. Okay. A, let's you and I both agree that a spectrogram is one of the many, many, many different transformations of the speech signal that you can –

[0:11:50.8] JM: Maybe you could just define, for people who aren't familiar, what a spectrogram is.

[0:11:53.6] RS: Oh! A spectrogram is basically a breakdown of the frequency content of an audio signal against time. So you can think of it as an image. So think of it as a plane with the X-axis being the time and the Y-axis, the frequency, and you will see many patterns on it, and these patterns are – If you think of it as an image on your screen, the color of any pixel at any point is the energy in that frequency at that time. I don't know if that explains it well enough.

[0:12:27.1] JM: No. That's a good explanation. Now, let's take that dataset and solve for enthusiasm.

[0:12:31.6] RS: Okay, for enthusiasm. Okay. So where were we? Okay. So you and I both agree that the spectrogram is only one of the many, many, many different possible representations of the speech signal, right? Now, if these representations, like the spectrogram is a 2D represent – Or a 3D representation, however you want to see it. It's human viewable and human interpretable, and you and I are trying to look for signatures of enthusiasm and this human viewable and human interpretable representation, right? We may not find it. However, you and I both agree that the signature is there, right?

So let me begin with the hypothesis that there are two ways in which I can go about this, okay? I'll talk about the second way after I finish with the first one. Let me begin. The first one is I begin with the hypothesis that there exists a signature of enthusiasm. It is there in the voice and I need to discover it. How do I discover it?

Okay. Let's say I know that the signature is there. I assume that it's in some high dimensional space for very many reasons, high dimensional mathematical space. So I design some mechanisms. Let's talk about neural networks. Maybe I design a neural network mechanism wherein I can transform the speech signal into some high-dimensional latent mathematical

space where I expect to find the signature of enthusiasm. Now, how will I find the signature of enthusiasm in that space just visioning that there is such a space and I can transform it is not enough. I actually have to design the entire architecture to bring out that information, right?

So there are very many other assumptions, then I would have to make and impose in this latent space. I would have to say, A, I mean, I can't arbitrarily have very dimensions. I want the signatures to be in that space, but confined to some lower dimensional manifold, a subspace of that space. Okay, I can impose that condition on that space. Then I can impose other factors, other conditions that allow me to constrains the solution to the problem within that space in a manner that I can then discover it, right?

I can say, "Okay. I want to find these representations, but I want them to have such and such distribution, because I feel that – I mean, it's not entirely intuition here. I know a little bit about speech that they should have such a distribution. They should be confined within this space with this distribution. At the same time, I want these features to reflect the quality of enthusiasm as best as possible.

Now, I have a mechanism and I have a way of designing an entire neural network architecture to impose these conditions and discover this kind of – Or engineer this kind of latent features in that space. So that part. So once I have those representations, I can then use standard machine learning algorithms to map those representations to speech that exhibits enthusiasm versus speech that does not exhibit enthusiasm. It's a standard classifier training paradigm that we then get into after discovery of these features.

[0:15:58.1] JM: So just to make sure I understand this so far. Let' say we need – First of all, we do need a labeled dataset for this mechanism that you –

[0:16:06.1] RS: We do need a label dataset for this mechanism to work, yes.

[0:16:10.0] JM: So maybe we get a thousand podcast guest recordings. So the guests that have come on my show, let's say we have a thousand guests and we label them as either enthusiastic about their work or not enthusiastic about their work. So we've got labeled

enthusiasm and non-enthusiasm recordings of these guests, and then we process these recordings into a spectrogram and then we use a neural net –

[0:16:37.5] RS: Not spectrogram. I could start with something else. I could start with a 3D video representation of speech. I could start with one of very many standard representations of speech, right? Okay. So that was just a correction. But spectrogram is just as good. Right.

[0:16:54.4] JM: Right. So any high-dimensional set of data, high-dimensional meaning like large, unstructured set of data where there's not – You can look at that dataset from a lot of different angles. I think that's one way of defining that term unstructured data.

[0:17:10.8] RS: Yeah. So that you and I are on the same page, we are talking about feature discovery, feature engineering and feature discovery. That's a vital] part of this. I didn't talk about the second mechanism, but we'll get to it.

[0:17:23.8] JM: Yes, absolutely. So still talking about – So you used the feature discovery to find the features that are correlated –

[0:17:31.1] RS: To custom design the features that –

[0:17:34.4] JM: Custom design the features, and these are not features like we would be able to articulate as humans. They are just mathematical features that happen to be correlated with the enthusiastic recordings, and perhaps anti-correlated –

[0:17:49.1] RS: Or are correlated to the enthusiastic recording in the process.

[0:17:52.9] JM: And you can use some fairly well-defined methods for doing – Like I think random forest would be a good approach for that feature discover, right?

[0:18:00.9] RS: Not for the feature discovery, but yes, it would be a good approach for applying to the mapping, the modeling stage. Once you've discovered the features, you have to come up with predictors and classifiers for the parameter you're looking for. So they would be – All of

these standard machine learning, classifiers, predictors, regressors, what have you. All of these are very well-suited for that stage.

[0:18:25.5] JM: Right. Okay.

[0:18:26.7] RS: A lot of the technology that we need, it's already there. A lot of the knowhow is spread over that many fields that have studied the human voice, and all I'm sitting here and doing is I'm not claiming I'm the first person to have thought of this or anything. I'm just sitting here connecting the dots. It's such a fantastic voyage of discovery. I discover more and more and more about the human voice, and it's amazing.

[SPONSOR MESSAGE]

[0:19:04.1] JM: Every team has its own software, and every team has specific questions about that internal software. Stack Overflow for Teams is a private secure home for your team's questions and answers. No more digging through stale wikis and lost emails. Give your team back the time it needs to build better products. Your engineering team already knows and loves Stack Overflow. They don't need another tool that they won't use. Get everything that 50 million people already love about Stack Overflow in a private security environment with Stack Overflow for Teams. Try it today with your first 14 days free. Go to s.tk/daily. Stack Overflow for Teams gives your team the answers they need to be productive with the same interface that Stack Overflow users are familiar with. Go to s.tk/daily to try it today with your first 14 days free. Thank you Stack Overflow for Teams.

[INTERVIEW CONTINUED]

[0:20:16.9] JM: Okay. Let's go a little bit deeper on this example, and then we'll move on, because we're spending a lot of time on this first example, but I think it's a great example. So let's say we've discovered the features and now we want to figure out how to use these defined features to process this dataset and identify which of these people are enthusiastic. What's the next phase?

[0:20:39.9] RS: Standard machine learning. The next phase is just standard machine learning; enthusiastic, not enthusiastic. You can throw a random forest classifier. You can throw a [inaudible 0:20:49.0] machine, a binary. If you have a binary classification problem, these two will work. If you have a multi-classification problem, if you are grading a psychiatric symptom or something like that, you use different classifiers and regressors, whatever suitable for the parameter that you're trying to predict.

[0:21:07.7] JM: What are the standard tools you would use for that feature discovery part?

[0:21:11.5] RS: So we're engineering the tools. Most of these are based on neural networks, but they're not already there. So you have – Yes, we use some elements of generative adversarial networks, some elements of LSDMs, RNNs and it's an engineering process. We take pieces of different kind of neural networks, different paradigms within those as required, engineer the systems and do the best we can to discover these features. These feature are what I call micro-features. These are often these signatures that are embedded in about 1/20th or 1/40th of a second duration in speech. Many of these actually are – You and I can actually relate to them from standard literature. I don't know if you know about voicing onset time. Have you read about that?

[0:22:05.5] JM: I haven't heard that term.

[0:22:06.1] RS: Okay. Do you want me to explain that? It's very interesting.

[0:22:08.8] JM: Sure.

[0:22:09.7] RS: Okay. So let's say you or I say a word like cat. Now, as I say cat, I have uttered three different sounds, "k" and "a" and "t", and in order to utter these three different sounds and in order for you to hear them, you heard them in a rapid succession and you heard the entire word, but you also heard the sounds, right? How did I produce these sounds? I produce the sounds by moving the articulators in my vocal tract, and when the articulators move, the dimensions of the vocal chambers change and the resonance has changed. So the quality of the sound changes and you hear different sounds. That's how I produce these sounds.

There are many categories or sounds that we produce as we speak continuously. We produce vowels and consonants and different kinds of consonants, different kinds of vowels, right? So when I said cat, the first "k" is this top consonant. What would that from articulatory phonetics, you would say that that particular sound, "k", does not require your vocal folds to vibrate. There is no excitation signal in the vocal tract to product that sound. The way you produce that sound is by creating an obstruction in your vocal tract, building up the air pressure behind that obstruction and suddenly releasing it, right?

Now, the very next sound is "a", and when you say "a", if you keep your fingers on your Adam's apple, you'll feel the vibrations, right? In order to utter the sound "a", your vocal folds have to vibrate at full potential. Now, when I say cat, I'm taking my vocal folds from a state of complete rest from "k" to "a" in a very short time. It's like accelerating a car from 0 to 60 miles per hour, and that time is different for different cars. The more higher-end your car is, the shorter the time and so on and so forth, right?

So our vocal tract, our speech production apparatus has a certain inertia, and it takes some micro-duration of time to go from a state of – For the vocal folds to go from a state of complete rest to a state of full motion, right? That time, that short time is called a voicing onset time and it's different for different people. It's not only different for different people because their vocal tract inertia is different, because the vocal tract is a very complicated physical – I don't have the right word for it, apparatus for want of a better word. So it's quite complicated in its structure, musculature and everything.

[0:24:55.1] JM: Your point here is that this is an extremely high signal component of human speech.

[0:24:58.0] RS: Yes. So it is something that is so individual to you that it's not only different for different people. It's different for different combinations of sounds that you produce. If you were able to actually measure this micro-feature accurately, and there have been groups around the world, especially my colleague, [inaudible 0:25:17.0] University in Israel has done a lot of work in accurately measuring voicing onset time.

If you were able to do that, you could come up with a set of voicing onset times for you, for Jeffrey, that would be like – I don't know want to go into hyperbole here, but something equivalent to a bar code for you.

[0:25:39.0] JM: How big of a sample do you need to develop that fingerprint?

[0:25:41.7] RS: I could measure it manually on the spectrogram. So I just need one sample from you.

[0:25:46.9] JM: Like how long of me speaking? Just a minute?

[0:25:49.7] RS: Yeah. No, one word, cat. You say cat. I can measure your voicing onset time.

[0:25:53.3] JM: Oh my God!

[0:25:55.4] RS: If you say other words that have different combinations of stops and voiced sounds, I could measure the voicing onset time for different combinations of sounds that you produce, and I can put all of these together. Because you have no control over them, it doesn't matter how hard you try to disguise your voice, I might be able to actually identify you from just this set of features. This, by the way, is not even one of the discovered features that I've been talking about. This is a standard feature that has been studied in the past in other fields. There's been lots papers written about it. Very good algorithms device to measure it, and so on and so forth, and there are many, many, many searchable features about the human voice that could be put together and used to profile you.

[0:26:45.9] JM: The different applications of this, you wrote this paper about deducing the severity of psychiatric symptoms from the human voice. So I thought this was a perfect example of how this can be applied. So this is the idea that you can grade the severity of psychiatric symptoms from somebody's voice on the phone. So if you already know somebody is schizophrenic, for example, or has some degree of schizophrenia . Schizophrenia is like, I think, is graded on some kind of severity scale. If you just take a sample of their voice, you can grade the severity of their symptoms, or if you assume that they are schizophrenic, then you would – If

you were trying to develop some system for somebody calling in to 911 and grading their severity of schizophrenia, or bipolar or –

[0:27:38.5] RS: I could start by just trying to – I mean, I cannot already do this. So take this with a pinch of salt. But I want to be able to do this. Hypothetically, anything that affects your voice, and all of these symptoms or these conditions that you're talking about do affect a person's voice, and there's literature out there saying that could be identified or differentiated from voice alone.

So if you call in, and I don't know that you have schizophrenia, or that you're depressed, or whatever. I, a few years down the line, hope to have an automated system that might be able to first understand that you have such and such a condition or such and such a set of conditions, and then go ahead and try to grade the severity of that condition. It's important in the current crime cases that I look at often have voices that sound intoxicated, and I would very much like to be able to tell whether the person is intoxicated with alcohol, or cocaine, or heroin, or whatever. I would like to be able to differentiate between the intoxicant from the voice, and it'd doable. It'd doable, because at least in law enforcement cases, it gives the detectives some clue as to what places to look for the perpetrators – Yeah.

[0:29:10.9] JM: And the degree or erratic behavior to expect. I mean, if you're talking about somebody that's on coke versus somebody that's just drunk a little bit, there are very different extremes of the behavior that you can expect.

[0:29:22.4] RS: Very different extremes of the behavior, yes. So it's all now, today, as I sit here and talk to you, Jeffrey. It's all doable. It's a function of data.

[0:29:34.2] JM: One way of looking at this from your paper is the fact that human speech takes into account so many different physical elements. You've got teeth, lips, tongue, uvula, which is the thing that dangles down the back of your throat, your pharynx and all of these – These are physical components, and just like your heartbeat, your other elements of your physique are influenced in distinctive ways by alcohol, or cocaine.

[0:30:03.9] RS: Yes, they are, and just your vocal folds are moved by six different muscles. Those little vocal folds are controlled by six different muscles. Anything that affects your brain, your neurotransmitters, your physiology, all of these have some influence and some way on some part of your vocal tract, and we are talking about so many different parameters within the vocal tract. We're talking about starting from skeletal proportions; tissue elasticity, muscular thickness, moisture levels, the shape or your lip, nose, neck, the length of your vocal folds, the density of the cartilages inside your – What are called laryngeal cartilages, even your lung volume, your trachea diameter, your lunch capacity. You name it, there are so many, many different factors influence your voice that if you put all of these together, it's impossible for any of these things not to have an influence on your voice.

[0:31:05.7] JM: Right. Let's take this a step further. I did a show recently with somebody from the TensorFlow team. You probably saw this result. There was this result that came out of, I think, this Google project where they're looking at a lot of pictures of eyeballs basically to diagnose diabetic retinopathy. Have you seen this class of research?

[0:31:28.6] RS: No, I'm afraid not. No, I haven't come across that one.

[0:31:31.0] JM: Basically, they have this big set of data of just close up scans of human eyeballs and they're using this to diagnose diabetic retinopathy. Yeah, and you can actually get to the degree of an ophthalmologist, or I think ophthalmologist, yeah, by just a well-trained neural net. But the other thing that they discussed, the other thing they were able to predict was actually – I think it was heart conditions. They were able to predict heart conditions and things like – I think they were able to predict age and gender based off of just looking at a human's eyeball, which is kind of –

[0:32:11.7] RS: Fantastic. Yes, it's totally believable. Yes.

[0:32:16.3] JM: But the idea here, if you extrapolate this, is basically aspects of humans. Humans have so many different moving parts within them and so many different distinctive lower level components that you could really derive a lot of information from these high signal imaging pieces of human data, whether you're talking about an eyeball or a human's voice. So what I want to ask you –

[0:32:42.1] RS: I think with the technology we have now and with Google results and with what I have been looking at, I firmly believe that we're in a stage where we're only just scratching the surface of the amount of information that be gleaned from all kinds of signals that humans generate.

[0:33:02.9] JM: You could imagine, I download some app on to my phone and it processes everything that I say. I mean, let's ignore the privacy implications for a moment, but it hears everything I say, it might be able to predict if I'm going to have a stroke, or if I'm going to have a heart attack in the near future, or if I'm going to have cancer. These things are actually fathomable. Would you agree?

[0:33:25.6] RS: Yes. I agree. Conceivable, fathomable, doable, it's inevitable. These things are going to come up in the next few years.

[0:33:36.2] JM: Incredible. What are you working on right now? As we volley towards this future, what are the problems that you're trying to solve that will get us closer to those kinds of solutions?

[0:33:48.7] RS: My entire group here at Carnegie Mellon is working towards one goal, and that's my dream. That goal is to be able to recreate the human persona in 3D from voice alone. We have been able to generate the human face. Back in December, we got our first results on that, and the results are not perfect, of course. But we are getting there. Yes, the human form is one of the things.

I am working towards, as I said, developing the signs of micro- [inaudible 0:34:24.6] and profiling. I would like to be able to get as much information about the human body and the human mind and the human persona as possible to get from to get from the human voice. I want to be able to design the right algorithms in order for us to do that.

If I'm successful down the line, I don't know how many years it will take. I think we're looking at a future, and not just me. Of course, all my very capable colleagues as well around the world. We can accelerate this and we won't be looking at the world where machines could understand

humans better than humans can conceivably, because machines might be able to glean more information from your voice than you can. Our hearing is not so perfect. Our hearing has evolved to a point where it just serves us to survive in this world. So we are looking at a future where machines might be able to understand us better than we can ourselves, maybe able to help doctors diagnose diseases much before actual severe physical symptoms show up, which would be very useful for most of the world that is still struggling for medical facilities.

[0:35:49.2] JM: Right. I mean, this is the same kind of – Like the environment where this diabetic retinopathy technology was deployed was, I think, in a place where most people who might have diabetic retinopathy can't afford to go an ophthalmologist.

[0:36:02.9] RS: Yeah, and there are so many other use that one can think of. Who knows? People could think of many different users for the information that you can get from voice.

[0:36:14.5] JM: So to be clear, your goal, your moonshot goal of you and your team is to be able to take a sample of a human voice and recreate the physical visage of that person.

[0:36:24.6] RS: Among other things, yes.

[0:36:26.3] JM: What would be some other things?

[0:36:27.9] RS: The other things is getting accurate information about their physical health status, mental health status, their environment, social background. Your environment also influences your voice, and it's a very, very longshot, but at some point I hope we'll get to a stage where we might be able to recreate some part of the physical environment around you from your voice.

If you're in a room, maybe gauge the dimensions of the room. What the ceiling is made of? What the walls are made of? What's around you that's reflecting sound and things like that? Very longshot. I mean, I know, don't scoff at this, but this is a future. I'm working to it. Yeah. More information about the person and persona from the voice.

[0:37:15.0] JM: And the steps to getting there, what are the projects that you're focused on specifically right now to get to that future?

[0:37:23.0] RS: Okay. My biggest source of funding is from the law enforcement government. So in the context of law enforcement, I have to deal with particular nuance of the human voice and human personality, which is the fact that we, as a raise, are pretty devious of where we commit crimes with our voice. As the soul tool for that crime, we often try to disguise our voices. Even though in most of the cases of hoax calls especially that I've studied, even though these hoax callers are not very technically savvy and have no idea how potent their voice is in giving away their identity, they instinctively try to not sound like themselves.

They're making a hoax call. They want the emergency, the rescue services to respond, so they have to sound like a real person, but they still try to disguise their voice in a way that they don't sound like themselves. This comes instinctively to us.

Now the big question is, I want to derive all these information, but I have no idea what the extent of – I mean, how much humans can actually vary their voice? There has been no systematic study about this. What is the extent to which you can vary your voice? How much can you disguise your voice? If you do disguise your voice, what is it that is within your voluntary control and what is it that is not in your voluntary control? Can I use what is not in your voluntary control to profile you then? So I have to discover all of that, right?

There are so many, many fundamental questions that remain to be answered in this context. What happens to your voice when you scream, when you whisper? What changes? What aspects change? I mean, what can I home in to accurately profile you? What must I discard?

[0:39:25.8] JM: Right. If you had a dataset of one person saying 20 different things while they're whispering and then they say the same 20 different things in their normal voice, you might be able to derive which features remain constant between those two sets of data and use that signal.

[0:39:44.6] RS: Under those conditions. Whispering is not the only way this person can change their voice. They can disguise their voice by changing the style. What is style? It's such a fuzzy

concept. What is style? Can you measure it? If you can't measure it, I have to find a way to measure it. If I want to automate anything, I have to measure those things, right?

[0:40:04.7] JM: Right. This is another thing you talked about in one of your papers, your voice impersonation paper, where you talked about style – Like the style transfer stuff that came out of – I mean, it came out of the – Did it come out of the Deep Mind?

[0:40:18.5] RS: Not the Deep Mind. So this was all that paper. I know which you're referring to. I have been looking at style transfer in order to discover the elements of style automatically in a manner that they could measure.

[0:40:30.7] JM: Right. If you took a set of Van Gogh paintings and applied style transfer to derive what is the style of Van Gogh? You can articulate what is the style of Van Gosh, but a machine can present several mathematical properties of Van Gogh paintings and then apply those mathematical properties to other images in order to get this style transfer of images where if you apply the Van Gogh style transfer transform –

[0:40:58.3] RS: You've articulated this very, very well. That's exactly what I'm trying to do.

[0:41:03.5] JM: Yeah. If you take this – Just to drive it home a little bit further. So if you take this Van Gogh transform and apply it to a random picture of a cloud or a cat, suddenly that cloud or cat will look like a Van Gogh painting. You'll say, "Oh! It's Van Gogh." Again, I don't know how to describe it, but that is a Van Gogh picture of a cloud or a cat.

[0:41:24.6] RS: That is Van Gogh. I don't know how to describe it. Look at all the – I mean, I'm moving the topic a little bit sideways. If you take up all the papers that have studied the correlation of human voice to one thing or the other and one thing or the other relating to human parameters, you will find that people are often talking about "voice quality". Voice quality is related to this. Voice quality is related to that. But what is voice quality?

If you look deeper into it, voice quality is made up of 23 or 24 different sub-qualities and they have very subjective names, like breathiness, raspiness, hardness, nasality. How do you measure these? If you show me a spectrogram, and if I say Jeffrey's voice is breathier than

mine, if I compare the two spectrograms, I don't think I can point out to one thing and say, "Hey, that is breathiness, and I can measure that."

So how do I measure breathiness? I have to figure out a way to do that. I will figure out a way to do that, right? So we do that using proxy features, but I'm not going to go into that. But the idea with proxy features, it's just skimming the surface, is to – If you can't measure something, find something else that is highly correlated with that and is measureable and measure that other thing.

[0:42:52.3] JM: Then those proxy features abstract into something that averages to the higher level feature you're looking for.

[0:43:00.6] RS: No. Let me get back to deviate it from your Van Gogh, and you were – I wrote a poetry about it sometime back. But very good example that you picked up. So now, if I were to try to discover style or try to quantify style, I could do that in two different ways. One is I might want to go bottom up where I study many different instances of disguised voices and figure out what the specific patterns are that I could call style and how they could be measured. This has been the traditional way of describing style, [inaudible 0:43:39.3], this, that, whatever things that are measurable, entities that are measurable.

But now we have the AI and revolution and neural network revolution going on and everybody is throwing data at neural networks. So you can actually afford to go top down. You can take a neural network. You can device a neural network to transform one kind of style or one kind of voice to another kind of voice, my voice to your voice, right? Once I'm able to do that, then I take the neural network and I dissect it. I try to find out from its internal representations what might be the element of style or elements of style that it might have captured in the process of doing that, learning to do that transformation, right? That's another way of discovering style.

Now, we all understand that style that is discovered in this manner might not be – It might not be something that you and I can readily name or relate to, but we could still identify it and we can still measure it and we can, in fact, measure it more than we could measure the bottom up features that we were trying to come up over the past decades in relation to style. So we

suddenly have all these tools and mechanisms at our disposal that allow us to do top down discover. That's very valuable, I think.

[SPONSOR MESSAGE]

[0:45:15.9] JM: Users have come to expect real-time. They crave alerts that their payment is received. They crave little cars zooming around on the map. They crave locking their doors at home when they're not at home. There's no need to reinvent the wheel when it comes to making your app real-time. PubNub makes it simple, enabling you to build immersive and interactive experiences on the web, on mobile phones, embedded in the hardware and any other device connected to the internet.

With powerful APIs and a robust global infrastructure, you can stream geo-location data, you can send chat messages, you can turn on sprinklers, or you can rock your baby's crib when they start crying. PubNub literally powers IoT cribs.

70 SDKs for web, mobile, IoT, and more means that you can start streaming data in real-time without a ton of compatibility headaches, and no need to build your own SDKs from scratch. Lastly, PubNub includes a ton of other real-time features beyond real-time messaging, like presence for online or offline detection, and Access manager to thwart trolls and hackers.

Go to pubnub.com/sedaily to get started. They offer a generous sandbox tier that's free forever until your app takes off, that is. Pubnub.com/sedaily. That's pubnub.com/sedaily. Thank you, PubNub for being a sponsor of Software Engineering Daily.

[INTERVIEW CONTINUED]

[0:46:59.0] JM: One thing you touched on earlier was generative adversarial networks, GANS, and these are tool that you used in your voice impersonation technology, the voice impersonation paper and explaining GANS in detail is probably not something that's like scope to a podcast or something that you want to discuss over a podcast, but maybe you could explain the problem domain that GANS is useful for. Why have I been hearing so much about generative adversarial networks?

[0:47:29.8] RS: I will try to explain it in one way and I'm sure that's not going to work, and then I'm going to come back to the traditional explanation. I'll try the first way. You are an intelligence agent and you're trying to devise a strategy to beat your enemy at something, or you're playing a game against an adversary and you're trying to device an automated strategy to beat that other person's moves or typical moves of that player in that game.

The strategy that you device would have to take into account the skills of your adversary, and I think you will agree with me that if you start off by assuming that your adversary is as smart as you are, if not smarter, your strategy eventually will turn out to be much better and more effective at defeating your adversary. Without taking the adversary into account, any strategy that you come up with based just on rules may not be as effective.

So if you have any situation where you are trying to device something that you want to engineer to do something very important for you, you have to bring it into a situation where you allow it to "do its best" to confound you in your quest, and then you try to device the strategy to beat it. I don't think I did a good job there.

Let's bring it down to the worse case. The adversarial networks, that's the idea, any strategy that you design. Adversarial networks are generally used to design strategies in many different context, and you may not want to call them strategies or whatever, but algorithms in many different context. They allow you to take both sides of the coin into account and play them against each other so that the end result, which is algorithm or the engineered result or the strategy is much better than it would have otherwise been.

Now, in the voice context, I talked about feature discovery and I talked about imposing certain distributions in the latent space. Let's just take that one little example, okay? I want to discover a feature in the latent space that not only captures all the information you have in your voice, but also imposes a certain distribution on it, okay? Just this little, "I just want that."

Now, if I did not want to impose a distribution on it and all I wanted to do was to make sure my feature in the latent space did capture every bit of information that wasn't your voice. How would I do that? Just that part I could do by simply testing the feature in the latent space by decoding it

or reconstructing your speech from it and making sure that the difference between your original speech and the reconstructed space is minimal. I could come up with a situation where I can define a loss function to minimize the difference, and then when it's properly trained and if it's a neural network set up, I could say that, yes, the feature, if I see the speech that I reconstruct from the latent variable is very, very close to, if not indistinguishable from your original speech, the feature does capture all the information that is there in your speech.

Now, I want to impose a distribution on it. I want it to not only capture all the information. I want it to look like Mickey Mouse or something like that. Okay, how will I do that? So this is where I this whole set up in an adversarial situation. I am going to come up with a module that comes up with another feature, let's say a parallel feature that has distribution, okay? It just doesn't have the information that you have, but it does have that information. Think of a random variable that I'm shaping to some distribution through some mechanism.

Now, what I want to do is I want to make sure that one way of making sure that my latent variable has the same distribution is to make sure that the difference between the distributions of these two features that I have now is minimal. Now, what I can do is I can make my encoder something that – The encoder is the part that takes your voice and brings it into that latent space, right? I can make it such or engineer it such that the difference between the latent feature that I come up with and my other feature that has a distribution I want is minimized.

At the same time, I can have a discriminator or some other modules sitting there and trying to, all the time, distinguish as best as possible between the two feature sets. This latter thing that I talked about, the discriminator, is trying to maximally differentiate between the features I have while my encoder is trying to confuse the discriminator, like bring the two as close as possible. Now, I have an adversarial situation and you can see that if I train this set up well, at the end of the day, I am going to impose the exact distribution I want to impose in the latent space on the feature that I want to engineer and I'm going to be able to do it very well, because I have put the whole set up in an adversarial situation, right?

[0:53:18.7] JM: So feature engineering, this is one example of framing a situation as an adversarial situation and being able to use generative adversarial networks to asymptote towards.

[0:53:31.0] RS: Yeah, feature engineering is one example that I gave. There are many other situations, yeah.

[0:53:36.0] JM: What would be a few other applications of GANS?

[0:53:38.5] RS: People are using them everywhere, in mapping on entity to another. I can't think of anything specific, but it's everywhere. One of the very, very, very important arenas where GANS are being used are in generating adversarial attacks on existing AI systems. So think of a face recognition system, for example, a face identification system that's deployed on the border. Takes a photograph or a face and identifies who the person is. If the person is not found, if it's a new person, puts that person in the database so that the next time the person can be identified, right? You've seen face recognition systems at work.

Now, it turns out there are many papers now on how you can actually fool these face recognition systems in many different ways with or without even if you don't have access to insides of the system. It is a pre-trained system. Not accessible at all to you. There are things you can change about the face where make up in some strategic way or wear glasses with some certain patterns on them. People have algorithmically been able to come up with minimal changes to the appearance of a person's face in a manner that they are able to make even the best face recognition systems come up with the wrong conclusions, and those wrong conclusions can be engineered.

If I want an existing face recognition system to come up with your face every time I am in front of it, I could engineer the system, or I could engineer algorithmically engineer patterns that I could put on my face or wear on my glasses or something in order to do that, and that engineering is done using GANS.

[0:55:38.3] JM: Okay, perfect example.

[0:55:39.5] RS: And it's being applied to speech recognition systems. There's work going on at Berkeley about it. They have websites about it. How they were able to fool the Google speech recognizer, [inaudible 0:55:49.5] University, cyber security, people have done that successfully

just a few months ago. They can make Google speech recognizer come up with any hypothesis that they want, any given hypothesis in response to any other speech utterance.

[0:56:08.1] JM: Okay, great example. So wrap up, I want to talk a little bit about the consumer implications of this technology that we're going to see in the near future, and one example that I saw that's not exactly in the same domain that we're talking about but is closely related is this Google Duplex work, where you probably have – Did you see this demo, this Duplex demo?

[0:56:32.8] RS: I did. The speech recognizer that was being demonstrated. Something equivalent to Google Home.

[0:56:39.1] JM: Well, it basically this thing where you can ask Google to call restaurants and –

[0:56:45.1] RS: Oh, yeah. Yes, I saw that.

[0:56:47.5] JM: Yeah. That's not exactly related to what you're working on, but I thought it was a great example of how these incremental machine learning breakthroughs lead eventually to a really mind-blowing consumer application. So do you have any ideas – Over the next couple of years, what are the consumer applications that we're going to see? How is world going to change?

[0:57:12.6] RS: With Google Duplex, there are multiple technologies that we are talking about. One is distant speech recognition. So there is this technology that has to do with the hardware, with the microphone arrays used in the actual device that's capturing your speech, right? The right kind of [B forming]. It has to capture your speech with very high fidelity in order to be able to process it well, right? Then there is a backend technology, the Google Speech Recognizer and Apple and Amazon and all these companies who have very good groups working on speech recognition have come up with better and better and better performing speech recognizers over the past several years based on neural network technology, right? That part is getting better.

The third aspect of it is robust speech recognition in many different ways. It used to be that conversational speech has many nuances that are difficult to model in terms of language, right? The speech recognition systems of the past, and my past, I mean just a few years ago, were

very dependent on good language models to perform well. With neural networks and better engineering and all that, those hurdles have been slowly overcome. So now we have recognizers that are able to recognize conversational style of speech much better, much, much better than they could before. So put all of these together, and also they're able to perform much better in the presence of noise. So put all of these together, you have an app like – You have a demo like Google Duplex that works very well or Alexa, which works in most noisy situations that I have seen.

Most of the applications that I think we are going to see in the next years will have to – Will have an element of distant speech recognition, because that is a part that has alluded the community for a long time and now it's at a point where it's almost just going to be a solved problem very soon. So we may see – You walk into a doctor's office, maybe I don't know how cognizant you're about, how difficult medical transcription used to be, but now there could be an assistant that could just listen to your speech and all the medical transcription part of it can be done by the assistant very accurately. You could have all kinds of forms on the application taken over by machines. So you could do bank transactions through voice and you would not get the amounts wrong. Let's not worry about the password and the security part of it. But you might be able to do accurate transactions, sensitive transactions.

I don't think it's going to be ready for military use because of – Again, keep bringing it back to adversarial systems, but there is other side of research that is very worrisome at this point, where while all of these advances are going to lead to voice systems that do a lot of complicated actions and response to voice, they're also likely to be fooled more and more easily by these adversarial advances and algorithms. So there's always going to be a playoff between the two, which is why I'm saying these systems would, in my opinion, not in the next 5 or 6 or 10 years be ready for military use or use in situations where people's lives are at stake.

Other than that, I mean, if your money at stake and you really care about it, yes, you can't use it even there. But in other cases, I think, with the appropriate caution and the appropriate security measures in place, the systems could find their use in all the complicated tasks that humans perform by hand that they would like to perform by voice now. I can't think of any magical use for this, but the day-to-day use will certainly – We'll see more of it in the future.

[1:01:15.3] JM: All right. Okay. Rita, thanks for coming on the show. This has been a fantastic conversation. I really enjoyed talking to you.

[1:01:20.6] RS: Thank you for having me. I enjoyed it too.

[END OF INTERVIEW]

[1:01:25.3] JM: GoCD is a continuous delivery tool created by ThoughtWorks. It's open source and free to use, and GoCD has all the features you need for continuous delivery. Model your deployment pipelines without installing any plug-ins. Use the value stream map to visualize your end-to-end workflow, and if you use Kubernetes, GoCD is a natural fit to add continuous delivery to your project.

With GoCD running on Kubernetes, you define your build workflow and let GoCD provision and scale your infrastructure on-the-fly. GoCD agents use Kubernetes to scale as needed. Check out gocd.org/sedaily and learn about how you can get started. GoCD was built with the learnings of the ThoughtWorks engineering team who have talked about building the product in previous episodes of Software Engineering Daily, and it's great to see the continued progress on GoCD with the new Kubernetes integrations. You can check it out for yourself at gocd.org/sedaily.

Thank you so much to ThoughtWorks for being a longtime sponsor of Software Engineering Daily. We are proud to have ThoughtWorks and GoCD as sponsors of the show.

[END]