EPISODE 98

[INTRODUCTION]

[0:00:10.8] SC: Hello and welcome to another episode of TWiML Talk, the podcast where I interview interesting people doing interesting things in machine learning and artificial intelligence. I'm your host, Sam Charrington.

Last week, I spent some time at CES, the Consumer Electronics Show in Las Vegas exploring the vast sea of drones, cameras, paper-thin TVs, robots, laundry-folding closets and other smart devices. You name it, it was there. Of course, I was also able to sit down with some really interesting folks working on some pretty cool AI-enabled products.

Head on over to our YouTube channel to check out some behind-the-scenes footage from my interviews and other quick takes from the show. Be on the lookout for our AI and Consumer Electronics Series right here on the podcast coming soon.

The show you're about to hear is part of a series of shows recorded at the RE•WORK Deep Learning Summit in Montreal back in October. This was a great event. In fact, their next event the Deep Learning Summit San Francisco is right around the corner on January 25th and 26th, and will feature more leading researchers and technologists like the ones you'll hear on the show this week, including Ian Goodfellow of Google Brain and Daphne Koller of Calico Labs and more. Definitely check out the event and use the code TWIMLAI for 20% off of registration.

In today's show, I sit down with Erich Humphrey, Research Scientist in the Music Understanding Group at Spotify. Eric was at the Deep Learning Summit to give a talk on advances in deep architectures and methods for separating vocals and recorded music.

We discuss this talk including how Spotify's large music catalog enables such an experiment to even take place, the methods they use to train algorithms to isolate and remove vocals from music and how neural network architectures like Unit and Pix2Pix come into play.

We also hit on the idea of creative AI in general, Spotify's attempts at understanding music content at scale, optical music recognition and much more.

Now, on to the show.

[INTERVIEW]

[0:02:33.8] SC: All right Eric. Why don't you tell us a little bit about your background and how you got interested in machine learning and AI?

[0:02:39.5] EH: Sure. I guess, it all started in high school. I always played music and my dad was an engineer. The compromise over that when I was going to college was electronic engineering. I really wanted to make guitar pedals, I want to learn about how amplifiers worked, and then that got into more signal processing and algorithms, which led to a lot of parameter tuning by hand.

If you want a delay pedal, we have all these different knobs and whatnot. I started doing some more stuff around beat tracking and tempo tracking for running the music. It was like, I spent so much time designing these algorithms by hand that's like a lot of the tradition for a lot of digital signal processing things. That was right around the time that machine learning really started to pick up and it's like, collecting data for these things and then training an algorithm, it just simplified the process so much.

I was like, "Well, ths is a no brainer. I can have the system that I want to do all these cool things," but leveraging this elegant data-driven solutions. That freed me up to go running and play music and doing all these other things and just have these systems work.

[0:03:42.2] SC: This is all why you're in school?

[0:03:43.8] EH: Yeah. I did a masters down at University of Miami, in the really great music tech program down there and got to do a little bit more with running music and worked with the music therapy department. Then parlay that into a PhD at New York University, where I got to work with Juan Pablo Bello and Yann LeCun a little bit, which really pivoted it even further. I took a pretty formational machine learning class with Yann my first semester at NYU. It put me off on that path.

[0:04:14.3] SC: That's a way to get started.

[0:04:15.5] EH: Yeah. Yeah. It was pretty serendipitous. I didn't fully appreciate what I was getting myself into at the time. It was right around 2009, 2010. For me, it was always much more about computer visions, doing a lot of things with images and audio. Had some stuff around speech recognition, but music was always lagging behind.

For me, I always wanted to say tweak similar to computer vision, or ASR recognize chords and music. My guitarist, "Can you show me how to play this song automatically? Can you show me where the beats, the bars, the chords are? Can I have a playlist of just courses?" These kinds of things.

When you really start to think about the opportunities around music, leveraging things like deep learning and machine learning, you can – imagination can run wild, the things you can do with that.

[0:05:06.7] SC: Awesome. Tell us a little bit about what you're up to nowadays at Spotify.

[0:05:10.8] EH: Sure.

[0:05:11.3] SC: Did you go to Spotify right after grad school?

[0:05:13.1] EH: No. I spent some time at a small music and tech startup that was trying to do things around optical music recognition, so in the same way that you could do document scanning. One of the core pieces of technology that we're working on was if you took a picture of a sheet music could you turn it into MIDI, so that say kids could learn how to play any piece of music at their disposal? Then be able to on the longer arc, gives them feedback how they're doing and taking it from there.

[0:05:40.9] SC: That seems like that should be fairly straightforward.

[0:05:45.7] EH: You think that one of the really, really interesting things about music and all of its forms, and I'll mention this tomorrow is that it's so fundamentally intelligent that when you have even just a piece of music for a single voice, monophonic instrument, you don't have polyphony or these other things, you actually have instead of OCR for being generally one-dimensional, it's linear and you'll span a vertical axis.

Music, you actually have a two-dimensional grid that moves linearly with these really complicated links backward and forward. If you have DSL signals or quotas or repetitions or multiple endings, it turns into death by million papercuts, because one of the things you'll find is triplets will be notated in the first measure. Then non-after, because it's cheaper that way to not print, needs additional trees on all these things.

[0:06:36.8] SC: There are any musician will know that they don't –

[0:06:38.5] EH: Any type of musician would know that that was there.

[0:06:40.5] SC: Got it. Got it.

[0:06:42.1] EH: Or you'll run into these other really interesting common cases. There aren't even edge cases, where for children's music, they won't notate rests, because they're trying to simplify the musical surface.

It's in four, but you'll have these two notes here and then those two notes in the upper staff. The machine is like, I guess they're the same. Having all these – Then it doesn't have a ton of data for supervised training. It creates this really, really interesting challenging but fundamentally intelligent problem.

[0:07:13.1] SC: Interesting.

[0:07:15.6] EH: It's one of those things that to crack from a very human interest level. There's a ton of music that just hasn't been brought into the 21st century yet. Everything that was notated from Gregorian Chants, all the way up to now where we've started to shift away from notated music to more recorded music, or digital audio workstations/Ableton-styled project files. Where there isn't really an artifact of the music, except the recording. There's all these older stuff that could be brought into the future for the creative purposes, musicology and more anthropological considerations.

[0:07:52.9] SC: Okay. Interesting. I did an interview with Doug Eck at Google Brain Project Magenta. He had an interesting presentation at another conference a few months ago that talked about – even the step beyond what you're just describing, like once you have the music or you can describe the music to the computer accurately, how do you then get it to play

expressively? They've been doing some interesting things there. I don't know if you're familiar with it, but it's interesting stuff.

[0:08:20.2] EH: Yeah. I mean, expressivity and music creation and composition it's so interesting to really dig into, because I think it cuts to the core of humanity. There's been so much amazing work around game playing at AI recently and Alpha Go, Atari, but you have these well-defined objective functions; make the score high, win the game.

You have these extrinsic motivators that fit pretty well into a reinforcement learning formulation. Music, the most interesting things are internal, they're intrinsic. It's the novelty and the surprise of, "Oh, I didn't see that chord coming." When you're sitting in your room when Hendrix was really just digging in and just sinking his teeth and deep to a solo, it's for him, it's his heart.

You could think about having these feedback loops for listens on Spotify, or revenue generated, but these aren't really what cuts the heart of creativity and expression. What are you getting after? I think that's going to be one of the really interesting challenges as we start to move into that next stage of really autonomous or things that are self-directed in the AI space. Why is it doing it? What motivates it? Does it have this notion of self?

When you think about elements of – I think music, humor, sarcasm, these kinds of things start to encroach on that in a way that a lot of the recent history of machine learning hasn't gotten to yet.

[0:09:47.1] SC: Interesting that you put sarcasm in that bucket as a New Yorker living in the Midwest. I find that sarcasm is under-appreciated in a lot of places and I would love to dig deep into AI and sarcasm.

[0:09:59.9] EH: I think it's a East Coast-North East thing. I joke off and nothing to back this up that it's probably related to just local climates and it's a way to deal with a great day, or great weather we're having, right? It's just gray snow and the sludge and all the street corners are backed up and whatnot. You don't have that same thing in the West Coast, where it's beautiful every day. Sarcasm doesn't land in the same way with native Californians.

[0:10:29.6] SC: So true. At Spotify, you worked – what aspect of this problem are you working on there?

[0:10:34.0] EH: Sure. I think at a higher level, to the extent that I can delve into, I work on a team of researchers where we are building algorithms that can understand music content at scale. Some of the obvious applications would be to fit into can we better understand users, can we help provide better recommendations? A lot of recommended systems right now have gotten very far by looking at how users, consumers, listeners interact with content; whether it's purchased at the same time, or in the same catalog, or they've been grouped into the same basket or playlist or things like that. You can get really far without having to look inside the box.

The metaphor I like to make is algorithms for content work as well as they do for say Amazon, when you can just – you don't have to look inside the box. But when you really want to say more deeply understand a user and what they're after, it's like what color is the show? Is it felt? Is it vinyl? How does someone interact with music? Do they really gravitate toward an artist, or is it lyrical content? Is it about the harmonic content? Because one of the things that's really interesting about music in particular is your ability to enjoy is tightly coupled to how well you can understand it and to what extent you're surprised by it.

For example, there is one artist called Marisabel. For the untrained listener, it's going to sound a little bit like noise art. But if you go to a Marisabel concert, there will be people that are just totally rocking out and they're in it and they get it, because they have a model and an understanding for what's going on. You can use how people interact with content as a proxy for what they understand, what they can relate to.

[0:12:20.3] SC: How explicit is that understanding for them though?

[0:12:22.8] EH: A lot of people can't – Well, I can't speak for Marisabel fans. One of the really interesting things about music is that a lot of ways that we describe it are so personal and occasionally they're cultural or they're niche. You'll find that in certain sub-genres, certain words or certain ways, and you'll see things pop up. Playlist is on flick and it's like, you'll find that some of those language and the semantics, their course intermediaries to describe the thing you actually mean.

I love the part of this song where it really just makes my heart pound. What is it? I have no idea. I have no language to describe these things. Which makes education and visualization and

interaction with the content a really interesting area for some of these content understanding at scale.

Spotify, why do I like this song? Are there other songs out there that you don't know about that would make you feel and have such a strong physiological reaction in a similar way?

[0:13:20.0] SC: Am I going to be able to ask Spotify that, why do I like this song?

[0:13:22.9] EH: Wouldn't that be a great question to have answered?

[0:13:24.2] SC: That would be awesome. Yeah. I mean, especially because as I hear you describe this I am not a music sore by any stretch of the imagination, but it resonates for me that a lot of that is not really having – I guess there is – when we talked about understanding right there is the impact that music has on you and its ability to move you. Then your ability to understand it moving you and how and why and when. Then the next level is your ability to articulate all that.

I feel like for me, if I was able to come at it from the back and be able to articulate and understand these things at a conceptual level, that might help me to connect to other types of music and it would be really cool if I could ask if Spotify could basically teach me this.

[0:14:13.2] EH: I would actually take it to the logical conclusion, because beyond that then you could say, all these music was composed by another person. Generally, when you're composing as a composer you're pulling upon all of your experience and your own surprise and novelty, but that's going to relate in land with an audience in a very particular way.

When you think about an artist composing for their fan base, or in a genre, or a style, or trying to achieve a certain outcome, you might say, I really want to drop, I want to averse this minor so I can step into a major chorus, because people will feel that as this release in tension. The only way that that can be conveyed appropriately is if everyone has that similar expectation.

You're playing off of certain kind of musical behaviors that are in cultured in certain ways, which also makes music at the level of a global culture really interesting, or micro-culture as you start to be able to connect the dots across really new genres that didn't have any bandwidth in more of a mainstream music era.

[0:15:15.9] SC: Are you a musician personally? Do you play?

[0:15:17.6] EH: Yeah.

[0:15:18.6] SC: What do you play?

[0:15:19.1] EH: I play everything I can get my hands on. I grew up playing saxophone for about a decade. Switched over to guitar, then guitar into voice, and I've been learning drums for the last couple years. Anyway that I can express myself with sound is a good time.

[0:15:34.0] SC: When you're expressing yourself with sound, do you think about it in the way that you previously described like, "I'm going to try and hit this chorus." I don't even have – I can't even – my ability with the words is so poor. I can't even repeat what you just said. At least the impression that I have from popular and media TV, whatever is that they just go into a room and then music comes out. Not like, "I'm going to nail them with this crescendo right here."

[0:16:00.7] EH: I think everyone's process is different. I mean, my PhD was in a music program, so I had to take some gradual level theory courses, which actually gave me not a perfect vocabulary for it, but a better understanding of the things that I had developed intuitive feel for.

I don't think about it that way in the moment. When I'm playing music, I'm very much in it. You may have heard of the idea of creative flow, whereas like time flies and whatnot. I think for myself coming back at it as an editor, I can think about like, "This is a really good raw idea." I can massage this in a way that if I piece these things together, here is a really interesting musical pun. More in the more moment, it's a little bit more like I surprise myself. That was neat. I have to record everything and then go back through with a little bit more of a higher planning process.

[0:16:50.7] SC: I'm maybe getting away here an turning this into This Week in Music and –

[0:16:54.8] EH: They're related.

[0:16:55.5] SC: - and the arts.

[0:16:56.1] EH: I contend they're related.

[0:16:58.0] SC: But you're speaking here at the conference. What's your talk on?

[0:17:00.7] EH: I'm talking about one project that we recently published out of our newly minted music understanding group at Spotify around primarily singer, vocal separation from recorded music. I lovingly refer to it hot tip to a colleague from Miami, but it's unbaking the cake in a way. In recorded music, you have these –

[0:17:23.0] SC: This is basically acapella from any song, any track?

[0:17:25.9] EH: Or instrumental from any song. You can isolate the vocalist, you could remove the vocalist. It's a little bit of the audio processing wizardry that if you look to computer vision, there have been some amazing really interesting things with style transfer, or texture mapping. We took some recent advances with the Pix2Pix and the Unit architecture and have adapted that to music processing.

The thing that really made it work for us is that we have this really large music catalog. One of the big bottlenecks for source separation for a long time has been data. We were brainstorming one day and it was like, deep learning is great when you have data for training.

You have two options when it comes to data; you can curate it, or you can get clever and try to harvest it. A lot of computer vision stuff has gotten really far by using text around images and leveraging these other serendipitous that occur as a bi-product of other kinds of things. Spotify did a similar thing with play listing.

One of the signals we were able to harvest is that instrumental versions actually occur with a non-negligible frequency. If you have say a web scale music collection, you can actually end up with about a month or twos worth of straight audio for training these kinds of algorithms. We are able to do some –

[0:18:43.5] SC: What does that work out to on like a percentage base? What percentage of your catalog has instrumental versions?

[0:18:50.2] EH: A very small portion. Probably couldn't give you a number, either from memory or other reasons. It's small. When you have a large enough catalog, it becomes sufficient. We end up with a couple months of audio and we're able to train these algorithms to both isolate

and remove vocals from the mix. We nudged the state of the art a little bit versus some other models that have been published.

[0:19:13.1] SC: Well, let's jump into that. You mentioned Pix2Pix and another one, Unit?

[0:19:17.5] EH: Yeah. The Unit architecture proceeded the Pix2Pix work.

[0:19:21.7] SC: Tell us about those two.

[0:19:23.0] EH: Sure. The Unit architecture, actually taking a step back, one of the ways that a lot of – you can generally call it signal processing. In machine learning it has worked for a while as the idea that if you could have some compressing autoencoder reducing the dimensionality, you'll preserve the attributes that are most important and you can back it out.

Then by minimizing some loss over the reconstruction, then you could start doing some interesting fiddling with your intermediary representations. What ends up happening especially for audio, a lot of the high-frequency detail in the outputs is just lost.

It works pretty well for general shapes, but sharp edges, these kinds of things fall away in these autoencoder butterfly style architectures. The unit takes this butterfly style architecture, folds it over into however you want.

[0:20:12.4] SC: Butterfly is like the visual, you got this wide input. You're compressing down the dimensionality and you're fanning it back out?

[0:20:18.1] EH: Exactly. Like a bow tie. If you were to take the butterfly and the bowtie and fold the wings on itself, what you can do is you can take the – this is a convolutional architecture. You can take the feature maps from the forward path and can catenate them with a feature maps from the reverse path, or the inverse path.

What that ends up doing is providing a lot more detail and this more fine-grain granularity, so that when you – in source separation what you generally try to do is produce a mask, and then you apply that mask over say an input, time frequency representation, like a spectrogram. Spectrograms are like an equalizer curve drawn out over time.

You can wait the relative contribution of each frequency bend in time, stats as zero to one. We use this unit architecture to produce a mask over the input representation.

[0:21:06.0] SC: Let's back up a second, because I'm losing a – What does it mean to map the – what does it mean to take the feature map and circle it back on itself? I think that's the way you said it.

[0:21:16.6] EH: Generally in a convolutional architecture, you'll have a bank of kernels. You'll take each kernel and you can fold it with an input representation, and you'll get out – generally you got three-dimensional tensor or feature maps. If your input is 2D, you'll have take kernels by X by Y.

On the inverse path, you're also producing this feature map tensors. With the corresponding layer in the inverse path, you can concatenate the input feature maps with the reverse feature maps along that cathe dimension, which becomes then an input you can evolve another kernel matrix, or kernel tensor, I guess.

It's like, you have your forward path features and your inverse path feature being processed at the same time. It allows the composition of parts intuition for a deep architecture. That information is propagating all the way to what is effectively is higher level representation of the network, while preserving some of that fine-grain detail on the way back.

What we find is that the masks that are produced by this architecture, you have a lot more detail in what they're able to pinpoint in terms of the frequencies. It's able to really dialed in the components that are contributing to say singing voice.

The way this works is you can train one of these unit architectures for pinpointing the voice, or you can train a separate one for isolating the voice. If you had different data, you could imagine doing something similar for say drums, or other source specific architectures.

[0:22:49.5] SC: When you're training to pinpoint the voice, are you using like an acapella version as your training data, or are you still using your instrumental and somehow inverting it or something like that?

[0:23:01.5] EH: That's a great question. For the work that we publish and we'll be presenting at the Izmir Conference in Zhuzhou, China in the near future, what we actually did was – there were far more instrumental versions than acapella versions. We estimated what the vocals would be by looking at the positive difference between a full mix, and then the corresponding instrumental.

It's like a proxy for what the voice would be, which is not surprising why we get much better vocal removal results, because we're training on the actual instrumental spectra than the vocal which is estimated. But it's a really interesting point for future work to say what other kinds of content do we have at our disposal that we can fold in to this work? We got some really encouraging results. I've got some demos that I'll be sharing a little bit later. I imagine that will be out on the internet in not too long.

[0:23:51.3] SC: What else? Are there any other things that you talked about during your talk, or that you're planning to talk about during your talk you haven't talked yet?

[0:23:58.1] EH: I guess, the only other thing I would mention is being that we're in a really pivotal time for a lot of machine learning, and artificial intelligence research is we really start to frame this conversation.

It's been an interest of mine for a while now, like what we discussed earlier this idea of creative AI. I do want to take the time tomorrow and just have that shameless plug. Music is a really, really interesting domain. It doesn't necessarily have the same societal impacts that autonomous vehicles could have in terms of saving lives. In a lot of ways, music does have that human power in much more of a emotional and cultural way.

As we start to think about what other ways do we want to tackle artificial intelligence, like can we really study music without inherently studying humanity and the intelligence that we've been in doubt with.

[0:24:51.0] SC: One of the questions that comes up for me in this conversation is a little bit out of left field, but what in your view is the open source as a concept has swept over a bunch of different industries and areas of human activity. Is there an open source – what is the open source for music? The thought that preceded that was it would be interesting if in some period

of time, a musician delivered not this one final thing, but there was like – almost like a project file for your digital workstation that you can pull out the drums if you wanted to, you can pull out – you could isolate all different kinds of things, because it followed some open source ideology or something like that. Is there an analog like that for music, or what's the closest we get?

[0:25:38.6] EH: I think that's a really, really fascinating idea. I have to smile a little bit when you say what's the analog to that, in a lot of ways playing music together is that analog for a while ever since there was music publishing, or sound recording, it's been really interesting what sampling, remixing, reuse means for music. Because everything that you compose, create, share is an amalgamation of all your prior experiences.

[0:26:04.7] SC: A little bit of an inherent open sourceness to music from a perspective.

[0:26:08.1] EH: Yeah, exactly. Larry Lessig wrote a really great book called *Remix.* It touches on a lot of these interesting things, both from a legal, but also philosophical and cultural perspective.

[0:26:17.9] SC: Focused particularly on music, or generally on –

[0:26:21.7] EH: He touches on things. There have been some famous mashup artists, like [inaudible 0:26:25.7] and Girl Talk. I'm testing my memory at this point, but there is a famous composer who decried the rise of sound recordings and that it was going to like change what reuse and remixing really meant. We've seen that get a little bit murky as copyright laws change and whatnot. Yeah, music wants to be open source by nature. I think the analog is analog music. It's the acoustic signal.

[0:26:51.5] SC: Interesting. Now it would be interesting to see how music evolves along with machine learning and AI.

[0:26:56.6] EH: Yeah. I think they have a bright future together.

[0:26:59.4] SC: Absolutely. Well, thanks so much Eric.

[0:27:00.3] EH: Thank you so much, Sam.

[END OF INTERVIEW]

[0:27:03.7] SC: All right everyone, that's our show for today. Thanks so much for listening and for your continued feedback and support. Thanks to you, this podcast finished the year as a top 40 technology podcast on Apple Podcasts. My producer says that one of his goals this year is to crack the top 10, and to do that we will need your help.

Please head on over to the podcast app, rate the show, hopefully we've earned your five stars, leave us a glowing review and share it with your friends, family, co-workers, Starbucks baristas, Uber drivers, everyone. Every review and rating and share goes a long way, so thanks in advance.

For more information on Eric or any or the topics covered in this episode, head on over to twimlai.com/talk/98.

Of course, we'd be delighted to hear from you either via a comment on the show notes page or via Twitter at @twimlai.

Thanks once again for listening, and catch you next time.