**CRAIG:** Hi, I'm Craig Smith, and this is Eye on AI. This week, I talk to Tom Denton, a software engineer at Google working on new algorithms that can separate out individual bird songs from the cacophony of the forest and identify the species. I’m a bird lover and so the research fascinates me.

Before we begin, MLOps, or machine-learning operations, encompasses tools and platforms to build machine-learning models. It is a fast-growing space, speeding development and allowing engineers to focus on creative work rather than building the hammers to drive the nails. I was recently given a demonstration of the open-source MLOps platform ClearML, which sponsors this podcast. I’ve looked at such platforms before and was impressed at how intuitive and comprehensive ClearML is. You can try it for free. Simply register at clear.ml. Tell your developer friends about it, too.

**Craig:** Meanwhile, I hope you find the conversation with Tom as fun as I did.

**CRAIG:** Why don't we start by having you introduce yourself. What's your educational background, employment background, how you got involved in machine learning and then we'll start talking about the project.

**TOM:** Yeah, totally. So, my name's Tom Denton. I'm a software engineer at Google. I've been at Google for about seven years. Before that I was a mathematician in a subfield with too many syllables and I got interested in machine learning specifically on audio, maybe about, I don't know, four or five years ago and then joined up with the bioacoustics group at Google.

**TOM:** And since then, have transitioned into working on audio machine learning full-time and have been doing that for a few years now.

**CRAIG:** Yeah. Is working with audio signals very different from other kinds of machine learning, computer vision, for example.

**TOM:** Yes and no. So, there's a number of techniques that are picked up from the vision domain.

**TOM:** So often we'll convert things into a spectrogram, which is a picture that plots time against frequency and then you can just use your usual image things on it for classification. But there are a lot of specific techniques as well. So, one of the things we're going to talk about is audio separation and that uses some properties of audio that you don't really have in the visual domain.

**CRAIG:** Yeah. And we're talking specifically about birdsong, and I'm interested because I have three different bird song apps on my phone, and I sit in my backyard and none of them really work. So, there's one that kind of works. It'll come up with three options, but I can't see the birds.

**CRAIG:** So that does really help me. So yeah. Can you talk about signals separation and how that applies to this domain? And is this research and you're using birdsong as a way to explore isolating certain elements in the audio signal or is this a product effort where you're building an application that'll eventually be public.

**TOM:** Yeah. So maybe I'll start with the second question. I've been helping on the research side specifically for bioacoustics and birdsong. There's another team, the sound separation team within Google that's made a general-purpose sound separator. And then I realized, hey, this could be really interesting and useful on this birdsong domain.

**TOM:** The bioacoustics group, our mission is to support scientists, ecologists, conservationists help them use machine learning to help the world. So that's really my interest, build better models for them, open source them, and then let them pick up and run with it.

**CRAIG:** Yeah. And the general sound separator that Google has, why was that developed?

**TOM:** So, one of the things that the sound separation team works on is denoising for Google Meet. You've got someone speaking and then maybe got a motorcycle going by, and then you want something that's able to automatically pull the voice out and get rid of everything else.

**TOM:** So, their bread and butter is working with speech, and normally with speech, you can take people into a recording studio and get isolated recordings of them. And then when you're training the model, you'll take that speech and then mix in some other noise and then ask the model to pull them back apart again.

**TOM:** And that works really well. But we don't have that for birdsong. So, birds, if you try to pull them into a recording studio, they're either going to make really different noises or they won't make sound at all. I haven't really tried it. So, you have to start with recordings of them when they're in the wild.

**TOM:** And when they're out in the wild, you've got all of your usual sort of background noise, but you've also got other birds singing as well. And they're often singing over one another. And so, there's a real question of how do you start with this sort of noisy, mixed-up recording, where lots of things are vocalizing and there's background noise and pull it into parts.

**TOM:** And Scott Wisdom and John Hershey, my collaborators on the sound separation team, developed a new algorithm that's able to do that when you don't have the clean audio to start with.

**CRAIG:** And the apps that are out there, the public facing software, how are they doing it and why don't they work very well?

**TOM:** First I'll say that they're getting a lot better over the last few years. They just take the raw audio stream, whatever's coming into your microphone and then run that through a classifier and what we've done in the new paper is we first separate the audio with one model, and then we take the output from those separated tracks and then feed each separated track into the classifier.

**TOM:** So, it's an extra step in front.

**CRAIG:** The classifier, is that out there? Are there databases that have been annotated. And so, you don't have to do that step. Your focus is really on getting a clean signal.

**TOM:** Yeah. So, there's a whole big conversation about how the classifiers work as well.

**TOM:** One of the things that happened is that we've got a lot more data to work with over the last few years. There's a really lovely public repository called Xeno Canto where just anybody can upload their bird songs. They record that in the wild.

**TOM:** Xeno Canto, which is XENO hyphen CANTO, which is Latin for other voices. But the sort of difficulty with a lot of the training data that we have for birds is that it'll be like, here's a two-and-a-half-minute audio file. And it has been labeled as a blue jay. So somewhere in there, there's a blue jay, but there's often a lot of other birds vocalizing as well.

**TOM:** And you don't know exactly where. So, what I found working closely with the classifiers is that really, they try to solve two problems at once. The first problem is, what species of bird is vocalizing, but the second problem is, has anybody bothered to annotate this bird? So, the training examples we have, it's often somebody pointing the microphone in the direction of the bird that they're recording.

**TOM:** And so, you'll have this clear vocalization out front, and then a lot of other things going in the background. So, if the classifier thinks that something looks more like a background vocalization, it'll heavily discount it and won't give you the activation for it.

**TOM:** So then with the separation algorithm what we're able to do is pull it apart and now we've got the blue jay, or what have you, sitting in its own track and then the classifier says, ‘oh, I've got this nice, clear vocalization, must be a blue jay. I'm going to tell you what it is.

**CRAIG:** Okay.

**CRAIG:** Birds have different calls and songs. Does the classifier go that deep or is it simply, ‘this is a blue jay,’ or ‘this is a blue jay’s mating call,’ or ‘this is a blue jay’s defensive vocalization,’ or something?

**TOM:** Yeah, that's a really fantastic question. So, the training data we have is mostly of the type, ‘this is a blue jay.’ Some of the data we've got is song versus call, but especially when you've got researchers that are interested in one specific species, they'll often be really interested in like, ‘is this a nesting call or is this something just flying overhead?’

**TOM:** And that's something that the classifiers we have aren't necessarily great at right now, but I think is a really interesting feature area for work.

**CRAIG:** But you guys are not working on the classification. You're using an existing annotated database here.

**CRAIG:** Is that right?

**TOM:** I've been building a classifier as well.

**TOM:** So that classifier in the paper is one that I built. Exploring, better techniques for the classification algorithm is what brought me to the separation domain in the first place.

**CRAIG:** And on the separation algorithm, what kind of an algorithm is that? And what is the technique behind separating one vocalization out from the crowd.

**TOM:** We use what we call a mixture of mixtures. We take two sort of noisy pieces of audio, mix them together, maybe changing the levels or this sort of thing. And then the job of the neural network is to produce some masks that separate it out into different channels.

**TOM:** Usually we say four or eight channels, and then after it's produced those four channels of output, during training, we say, okay, which of these four channels contributes best to those two original mixtures that we put in, and then we look at how much distortion there is from the mixed together model outputs versus the original inputs. And it's a little bit crazy that this works as well as it does, but the way to think of it is, if you've got two independent sounds that the network is seeing, if it puts those in the same track, they could have come from different source recordings.

**TOM:** And so, if it has happened to put them in the same track, then it's going to get penalized heavily. So, then it really wants to pull it apart into different tracks so that they can be in the two different source recordings.

**CRAIG:** What's the architecture of the neural network. Is this a convolutional neural network or something?

**TOM:** So, the architecture is a little bit different from what you typically see with an image classifier. This is a 1D convolutional network. 1D convolutional because it's using audio signals instead of these two-dimensional things. We use what we call a learned filter bank, which sort of breaks up the audio signal into different frequencies, and then it's a unit architecture to sample level and then pulls things up into deeper representations and then comes back down again to produce these masks closer to the sample level again.

**TOM:** That architecture starts from this kind of audio domain, pulls out a representation and then pulls it back closer to the audio domain to create those sort of output masks.

**TOM:** And then the masks are applied.

**TOM:** It's like an autoencoder, but the difference is that we're really trying to produce these masks for the audio at the end. And then we're able to apply this directly to the filter bank audio, audio that's broken up into the different frequencies.

**CRAIG:** Then that mask is applied to the original audio, and it filters out everything except what you want to hear.

**TOM:** Yeah, exactly. And it produces four of these different masks and then the different masks are creating the four different channels.

**CRAIG:** And then do you blend the channels together again for the classification or is a classifier looking at each of the channels separately.

**TOM:** So, for the classifier, it looks at each channel separately, and then it turns out to also be helpful to give it a look at the original audio as well.

**TOM:** So, it's really just the more that the classifier is able to see, it does a little bit better. So, we give it the original audio, too, so it ends up having five tracks.

**TOM:** And then it produces these sorts of scores for every species, for every track. And then we just take the maximum over all of the different tracks.

**CRAIG:** Yeah. This signal processing is a very old domain. How much of this is already out there? How much is original in what you're doing?

**TOM:** The unsupervised sound separation, that's really new. So that's Scott Wisdom and John Hershey's work from about a year and a half ago. It's out there in the sense that the models are open sourced.

**TOM:** There's a version of it that was trained for general YouTube data that you can download and then we've also opened-sourced the birdsong version of it as well. So that part is new and unique. And then a few of the tweaks that I've made for the classifier also, I think, are pretty new and new to the literature.

**CRAIG:** So, you got a paper out on this. What's the next step.

**TOM:** We have a Kaggle competition that's just launched. We've been doing this for a few years now. This is the third year. And the piece that this is really focused on for this year is what a number of us think is the next big thing, which is understanding how to identify specific species.

**TOM:** And that might sound like a subset of identifying all of the species, right? But this sort of problem that we have is that when you get into rare and endangered species, especially, you often have much less training data. So, if you're trying to build a big general-purpose classifier, that you'll just want to see the score go up at the end of the day, that often just means doing better at the classes that you're already good at. So, we're working with the University of Hawaii for this particular competition and the Cornell Lab of Ornithology. And it's specifically looking at rare and endangered Hawaiian species. Hawaii just had a number of different birds go extinct last year.

**TOM:** To understand habitat and understand what's going on in a forest, audio is super, super rich. If you stand in a forest and look around, you'll see some trees and leaves, but you won't necessarily see that many birds and they might be way off in the distance and hard to tell what they are.

**TOM:** But they're constantly vocalizing and trying to talk to one another. And this gives you a lot of information about the animals, but also gives you lots of information about what the animals eat that might be present and this sort of thing. But experts who can identify birds are pretty rare.

**TOM:** And especially when you get into rainforests, you have fewer experts and you also have a greater species diversity, which is harder to keep track of. So, you put on your hat and say, ‘oh yeah, we should use machine learning for this.’ And then it turns out, problems are always harder than you think they are when you start.

**TOM:** So, it's been a number of years of work to try to get things to where they are today.

**CRAIG:** The Hawaii Kaggle competition is specifically working with a smaller data set, how to augment data sets. What's the task?

**TOM:** The task is there's 21 species that are the targets, and for those species, we've got a really variable amount of training data.

**TOM:** For hardest one, we only have one training example ranging up to 50. Then we've got lots of soundscape data that was collected and annotated by the students and faculty at the university of Hawaii that we're using for the test data. That data is actually hidden from the participants, so you put together your model with the scant training data we have, and then we evaluate it in secret to see how it does against the real data.

**CRAIG:** Your image of standing in the forest and listening to the sound, in my backyard there's this cacophony of sound. I know that some of them are tree frogs. Some of them are crickets. Some of them are something else.

**CRAIG:** Would this lead to a model that could count the number of organisms making sound and say, there are 30 tree frogs or this particular species of tree frog. There are a hundred crickets of this particular species. That sort of thing.

**TOM:** That's a great question.

**TOM:** In the ecology literature, we'd call that the abundance question. So how many of them are there and that's even harder to get training data for than the raw species question so that is a little bit of the holy grail.

**CRAIG:** Although if you can separate signals, that would be a start, with different filters, you could identify this particular species and that particular species in an otherwise very noisy environment. Beyond the basic research, what's going to be the output.

**CRAIG:** Is there going to be an app at some point that this is embodied in or some other service where you can upload a bird song and it would identify it for you.

**TOM:** So, for me, I build my classifiers and we're open source to the separation algorithm. And from my perspective, there's already great people out there building this intermediate software.

**TOM:** So, the Cornell Lab of Ornithology, they've got the BirdNet app and Merlin as well and it's not using our separation algorithm yet. But it may down the line. And likewise, one of the things I really love about the Cornell Lab is they produce this software called Raven that gets used by ecologists and conservationists in the field all over the world.

**TOM:** And as we're able to open source more of these models and make things that are useful for those on the ground conservationists, they'll be able to pick it up, put it in Raven and make it immediately usable by the people that really need it.

**TOM:** For ecology, there's kind of two big tracks. The first is as an ecologist, you want broad understanding of what's going on in the habitat. Tell me the big list of all of the species that are here. And that's what the existing classifier techniques are really good for. And then the other is I really want to track the specific species and I want to get all of the places where there's nesting calls, but maybe I don't care so much about the flight calls, and I don't have very much training data.

**TOM:** Maybe I can find like a few training examples, but not that many. How do I help that researcher? How do I help that conservationist? And it turns out that also has the biggest potential for on the ground impact because there's, the US especially, you've got things like the Endangered Species Act where if you're able to identify, ‘hey, yeah, there's an endangered species here,’ then you get some protection for the habitat that it's living in. There's similar sort of frameworks in other parts of the world. Identifying endangered and rare species is often like one of the best levers that you've got in terms of identifying the critical habitat and protecting it.

**TOM:** So, finding better ways to do that and help conservationists do that on the ground in as easy a way as possible is the next piece I'm thinking about.

**CRAIG:** Yeah. You mentioned nesting calls and flight calls. How many different calls or songs does a species typically have?

**TOM:** It's highly variable.

**TOM:** Some will just make the same sound forever. And then there's others where you'll have lots. I was talking to researcher in Australia, looking at the glossy black cockatoo and was listening to the example of those. And I was like, ‘okay, yeah, there's a long call and a short call here.’

**TOM:** And she said, ‘oh yeah, they have 16 calls.’ And it's oh, okay.

**TOM:** You'll have variation where it's geographic, maybe it's the same species, but they've got different calls on one side of the mountains versus the other. Sometimes it’ll be at the flock level, you'll have variation where individuals are, you might have like a marker within this song that says, it's this species.

**TOM:** But then the rest of it is sort of jazz where they're trying to like impress potential mates by doing the most complicated thing that they can. These are definitely challenges.

**CRAIG:** So, the algorithms that you've built, could a researcher studying the glossy black cockatoo in Australia take this and use it with their own data to classify the 16 different calls or songs that that bird makes?

**TOM:** Yeah, so this is the difficulty, right? The training data we have for classifiers is the form like, oh yeah, this is a glossy black cockatoo, but the labels don't have that level of distinguishing feature. So, the directions I'm thinking about are how do we get a good embedding or representation of that audio, where you could match different things.

**TOM:** So, this can look like clustering. Take all of the things that the classifier says is a glossy, black cockatoo, and then cluster them, maybe I'll be able to see those 16 different types and then pick out the ones that I'm interested in. These sorts of approaches.

**CRAIG:** And that's something that you do now, or it's a research direction?

**TOM:** This is a research direction.

**TOM:** These questions of not just, how do we do that? Because there's a lot of ways we can do that, but also, how do we make that easy and accessible for conservationists and ecologists that are actually out there working? We say, okay, yeah, we ran this clustering algorithm and here's 5,000 hours of audio to listen to you to figure out where the clusters are good or not.

**TOM:** We haven't necessarily helped them in that case, but we're able to do it in a smart way where they're able to get what they want and in easy and efficient way, then I would consider it a win.

**TOM:** I do have some audio examples.

**CRAIG:** I would love that.

**TOM:** So, if you imagine that you've dropped into a forest and open your eyes and you'll see trees and leaves, but if you open your ears, you might hear something.

**TOM:** And so, the separation algorithm is then able to take that sort of cacophony of interesting activity and break it apart.

**TOM:** So, three pieces there, first was insects, which are a little bit high-pitched, and I think there's some cricket in there that's intermittent.

**TOM:** And then a cackling bird was the second one I played. And then the third one is a different bird that’s sort of vocalizing at the same time as the cackling bird. And that is actually, I think, pretty hard to notice the first time you listen to the original audio.

**TOM:** So, here's the original audio from a Costa Rican rain forest.

**TOM:** And then when we break that apart so the first thing, we'll hear is the insect noise.

**TOM:** And then the second thing we'll hear is a cackling bird. That was probably the main thing.

**TOM:** And then finally, we've got a third bird that's high pitch, that maybe you didn't even notice the first time we played the original audio.

**TOM:** And just so you believe me, I'll play the original audio one more time. So, you can try to listen for the high-pitched one.

**TOM:** So, this is the one that when we were playing around with the algorithm initially really blew my mind that it worked. So, here's the original audio. This is from the High Sierra in California, and it was an audio collected by some of our friends at the California academy of sciences.

**TOM:** So, the first track that we'll hear is the separated American pippen.

**TOM:** And then the second track here is a gray-crowned rosy finch.

**TOM:** And what's really impressive about this is that the two birds are really vocalizing in the same frequency range and they're both very rhythmic. And when you look at the spectrograms for these here, it's, you can tell that there's a second thing going on. You’ve really got to look for it. And secondly, most people I play the original file for they don't even register that there's two birds until they hear the separated audio. It's really doing quite a good job.

**CRAIG:** Play the original audio one more time so people can hear it.

**TOM:** Yeah, sure thing.

**CRAIG:** Yeah, that's amazing. Yeah. You would not realize there's the finch in there.

**CRAIG:** And so, this is open source, but it's not something that a consumer could upload or use.

**TOM:** Depends on your level of tech savvy, right? It's open source.

**TOM:** It's on GitHub so you can download it. You can apply it to your own audio. And I'm hopeful that we'll be able to have it available via something like Cornell's Raven software in the near future. Which will be a little bit easier to work with. I hope.

**CRAIG:** Yeah. That's inspiring. I love it.

**TOM:** A lot of this work with driven by collaboration with the California Academy of Sciences, Jack Dumbacher, Durrell Kapan, and Mary Clapp. And they're looking, especially, at the impact of fire on ecosystems. So, they have thousands of hours of data that they've collected from up in the Sierras.

**TOM:** And we get the sort of before and after picture for what happens with prescribed burns and with the Caldor fire last year. We're going to have some picture of what happens after a big uncontrolled fire as well. So, they've been just absolutely great providing all sorts of evaluation data and giving lots of feedback on how the classifiers and separation is working.

**TOM:** And one of the big things we see from that is, say, after the fire, you've got more dead wood standing. And so, you get more woodpeckers, as just a very direct observation.

**CRAIG:** You're focused on birdsong, but presumably these algorithms could be applied to any species

**TOM:** I focus on birds mainly because it's what I have time for. Other colleagues in bioacoustics at Google work on whales, in particular, and doing things like coral reef health and this sort of thing. And absolutely the same sort of separation technique should be helpful there.

**TOM:** This idea of being able to separate signals where you don't have clean training data, there's a lot of different areas where that could be helpful.

**TOM:** One is for EEG, brain recordings, you're trying to capture neurons firing, and that's an electrical signal. But when you try to record that, there's all sorts of other electrical signals going on, things like muscle spasms or eye twitches and this sort of thing. So having a way to separate that out easily into different isolated signals and then classify, it should be helpful.

**TOM:** And then another one is seismology. Record the shaking of the earth. And that's going to be like a mixture of plate tectonics with trucks driving by or oil fracking, or what have you.

**CRAIG:** Yeah. Yeah. That's interesting. And for the human voice, presumably this could be applied to a crowded room, and you could separate out one voice and identify who it is.

**TOM:** There's specific cases where that's something to look at. So, like in a meeting, when you have multiple people speaking in a meeting and you want to be able to distinguish one from another. This is also called the cocktail party problem. So, you're like in a cocktail party and you want to focus on one person and it's also really important for hearing aids.

**TOM:** One of the first things that goes with hearing is the ability to make out an individual person that you're trying to talk in a crowded space. So having improvements in ways that we can focus on the person that you're trying to listen to you and isolate that out, I think is a very helpful thing.

**CRAIG:** Is this kind of technology, is its footprint light enough that it could be embedded in a hearing aid.

**TOM:** Things are getting lighter all the time and we've got some existing software that'll run on your phone and do this sort of thing.

**TOM:** But things are getting better all the time. We're starting to get these like smaller and smaller tensor processing units sort of things. So, I could definitely imagine a day when things do run in a hearing aid. A bit that people maybe don't think about as much on that front is that we've also got a lot of algorithmic improvement in speed as well.

**TOM:** So, we're able to make the neural networks for doing this sort of thing, lighter and lighter, even at the same time that the hardware is getting better.

**CRAIG:** That’s it for this episode. I want to thank Tom for his time. If you want to read a transcript of our conversation, you can find one on our website: eye-on.ai. And please take a moment to check out our sponsor, ClearML, at clear.ml. If you’re a data scientist or developer, they may be just what you’re looking for.

And remember, the singularity may not be near but AI is about to change your world. So, pay attention.