Episode Book

EPISODE #6174

What Your Voice Reveals: Accent, Distance, and Movement in Audio AI

PUBLISHED	21 October 2026 at 14:00 UTC
DURATION	23:09
LISTEN	myweirdprompts.com/episode/voice-analysis-accent-movement-27964/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.11
BOOK GENERATED	22 October 2026 at 07:46 UTC

SYNOPSIS

Can a single microphone tell where you're standing, where you're from, or which way you're walking? This episode sorts the real capabilities of audio AI from the claims that don't hold up. From language ID models that nail 107 languages but stumble on regional accents, to bioacoustic models that learned birds and turned out to work on whales, to industrial systems that detect failing machinery with a nearest-neighbour score. Along the way: why text-only models scored 40% on listening benchmarks, what sperm whale codas still can't tell us, and why the cleanest accuracy numbers come from the least realistic recordings. A grounded tour through what audio machine learning genuinely does, what it does with caveats, and what it mostly just sells.

TRANSCRIPT



Corn

Picture the voice memo Daniel is asking about. Someone dictates a shopping list into their phone while walking from the kitchen to the stairwell, and by the last item they're two rooms away with a slight Cork accent and a cough. The transcript says "milk, bread, lemons." Everything else is still sitting in the waveform.



Herman

Which is Daniel's question, more or less. Last time we went through whether models can hear emotion and tone, and that's settled ground, so we're leaving it there. This is the rest of the audio. He mentioned accent detection, and he mentioned detecting a speaker's movement, and wanted both stretched as far as they honestly go.



Corn

Movement is the one I'd like to look at, because I've never heard anyone say "speaker motion detection" as a job title.



Herman

Nobody has, it isn't a named task. The closest real version of that is a few separate things. Estimating how far a speaker is from a single microphone. Locating a sound source and tracking where it goes, which needs a microphone array or binaural input. And identifying someone's gait from the sound of their footsteps. Each of those exists, none of them is quite what you'd picture, and they sit at very different points on what I think is the useful way to sort all of this.



Corn

Which is?



Herman

A sliding scale. At one end, things that work and are deployed, where nobody argues about the accuracy. In the middle, things that work with caveats. And at the far end, things that are claimed, sold, and in some cases just wrong. I'd like to go deep on a few in each rather than recite forty.



Corn

Start at the solid end, then. I'd guess sound event detection.



Herman

Yes, and the plainest case is language identification. Take the standard ECAPA model trained on a corpus called VoxLingua one hundred and seven, which covers a hundred and seven languages, and it gets about six point seven percent error on its dev set. That's from a few seconds of speech. But it's a model that is good at languages, and notice what it's bad at... accents within a language. It degrades on non-native and regional speech. The same property that makes it work is what makes it blunt for Daniel's example.



Corn

So it's good at telling Welsh from Dutch and useless at telling Cork from Kerry.



Herman

Roughly, though I'd say useless is too strong. Hold that for the middle tier. The other end of the solid tier is the one I find most fun, which is bioacoustics. BirdNET, from Cornell, is the one most people have met, you hold up a phone and it names the bird. Google DeepMind's Perch two point zero is trained on something like fourteen and a half thousand species. And the part that surprised me is that embeddings trained on birds transfer to whales and underwater sound with a small probe on top, in some cases beating Google's own whale model on orca subpopulations. That's developer-reported, so take it as promising rather than settled.



Corn

A model that learned birds being good at whales makes sense. Both of them are just things shouting in a noisy room.



Herman

That's not far off. And the whale side is where it stops being a solved problem. Project CETI proposed a kind of phonetic alphabet for sperm whale codas, built from a bit under nine thousand codas from around sixty whales. What it does not give anyone is meaning. You can sort the clicks. Nobody can tell you what a click says.



Corn

So the machine hears the whale perfectly and understands it not at all. I know people like that.



Herman

Then there's the industrial one, which hardly anyone talks about. Listening to machines for faults. The yearly DCASE challenge had a task where systems had to flag abnormal machine sounds on machine types they had never been shown, from single-channel clips of ten to twelve seconds. Thirty-five teams, a hundred and nineteen submissions. And the winning recipe was not a giant model. It was a pretrained audio backbone and a nearest-neighbour score. Take the embedding, ask how far it sits from normal.



Corn

A nearest neighbour. In twenty twenty-six. I want that on a plaque.



Herman

The reason it works is that normal is a tight cluster and broken is anywhere else. You never have to teach the model what a failing bearing sounds like, which is good, because by definition you have never recorded one on this machine. Same logic, I'd guess, is behind most of the general sound event models, the ones trained on AudioSet-style labels. A dog bark, a siren, glass breaking, a baby crying. Small encoders, like BEATs, do that class of task cheaply and on a phone.



Corn

Glass breaking is the one that shows up in security cameras, isn't it? I have always assumed it was sold as magic and was actually a threshold.



Herman

Some of it was, years ago. The learned ones are better, and they're the real deployed end of the scale, the kind of thing running in smart speakers and home alarms. I can't give you a clean accuracy for those, it depends enormously on the room.



Corn

Fine, but is that the point? Small and cheap beating big and clever?



Herman

It's the point, yes. These are small, cheap models and each one does one job well. The audio language models, the ones you talk to, are the flexible generalists, and on the hard skills they're much shakier. On a benchmark called MMAU-Pro, with about fifty skills, the best I saw was Gemini two point five Flash at roughly fifty-nine percent, and some categories were close to random. Corn, what do you make of that? You'd think the chatty model would have eaten the specialists by now.



Corn

I'd have guessed so, yes. But it makes sense if the chatty one is mostly reading. Transcribe it, reason over words, done.



Herman

There's actually evidence for that. Text-only models, which hear nothing, score in the high twenties to low forties on the older audio benchmarks, because the questions leak. The newer one is harder to cheat, text-only drops to sixteen to thirty. So some of what looked like hearing was language prior.



Corn

A deaf model getting forty percent on a listening test is a pretty good review of the test.



Herman

I think the useful reading is not that audio language models are bad. It's that they're a front end. If I wanted to know what's in a ten second clip, I'd run the small specialists first and hand their labels to the chat model to talk about.



Corn

That's a nice division of labour. The specialists listen, the generalist takes the credit.



Herman

A lot of software works that way, to be fair to it.



Corn

Languages, birds, machines. A modest crowd for the top tier.



Herman

So to wrap up the solid tier, the real capabilities are out there, but they come from dedicated models for each job.

L

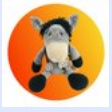
Larry

Friends, lend me your ears, and then lend me your voice, your accent, your dialect, your speaker profile, your vocal health biomarkers, your sound events, and your precise location while moving, because the all new Voiceometer Pro Max Ultra understands absolutely all of it, no dedicated models required, it simply knows. And here is the deal of a lifetime, the price starts at four thousand nine hundred and ninety nine, now it is three thousand, now it is nine hundred, now it is forty, I am basically paying you, and I am throwing in a free hat, a commemorative spoon, a small pony, and a timeshare in a very nice hallway. Wait, the offer just ended. It ended while I was describing the extras. The Voiceometer Pro Max Ultra is gone forever, along with the pony. I am Larry, and I have no idea what happened.



Corn

A timeshare in a very nice hallway.



Herman

Please don't read it back in that voice again, Corn.



Corn

Moving on to the middle. Accent.



Herman

Accent is where the framing matters. The best known open work is CommonAccent, a SpeechBrain recipe fine-tuned on Common Voice with the same ECAPA and XLSR encoders, and the authors report up to ninety-five percent on English accents. That's their own number, on their own data. When someone tested a similar setup on YouTube clips in the wild, it returned percentage mixes, say sixty percent this, thirty percent that. And the language ID model I just praised struggled with English, Welsh and Dutch on noisy audio.



Corn

The percentage mix seems more honest to me than a label. Accent isn't a box.



Herman

I think so too. A person who grew up in Dublin and lived in Jerusalem for ten years has a distribution, not a category. A label makes that sound like a fact about the person. The mix at least admits it's a fact about the audio. And I haven't found a clean benchmark of how the big audio language models do on accent, so I'd be careful claiming they've got it.



Corn

There's the other side of it, though. A tool that says where you're from, applied to a call centre or a hiring line.



Herman

I'd go further. Where I'd be comfortable is accent used by the speaker for the speaker. Language learners getting feedback on their own pronunciation, or a voice assistant picking a better recognition model. The moment the same output is used by a third party to sort the speaker, you've got a tool for discrimination that comes with a decimal point.



Corn

Herman, would you trust a percentage mix if you were the one being sorted?



Herman

No. Which I realise is an answer that argues against my own enthusiasm for the technique.



Corn

Same worry for the rest of the speaker analysis, I'd guess.



Herman

It's all profiling, and a good deal of the speaker analysis in this middle tier is. Gender from voice is basically solved on clean read speech, about ninety-nine and a half percent on the classic TIMIT corpus. Age is off by roughly five years on average. Height gets reported at around four and a half to five centimetres of error, though I'm quoting that from a secondary source on studio read speech, and what the model mostly learns is to guess close to average.



Corn

Five years either way is fine for a bartender. Not for a legal decision.



Herman

And the clean numbers are the trap. Those figures come from read speech, people reading sentences into good microphones. Put the same model on a phone call with a bad codec and a person who is ill, or shouting, or whispering, and the ninety-nine becomes something you can't quote.



Corn

So the cleanest result in the field is on the least realistic recordings.



Herman

Usually, yes. That's almost the rule for the middle tier.



Corn

Humans guess age and height from a voice too, presumably, and worse?



Herman

Right, and human listeners do worse and are more stereotyped about it. But the audio language models are interesting here. There's a recent preprint where the weak spot was pitch, something like thirty-eight percent for open models and fifty-one percent for proprietary ones. Meaning the thing most obviously in the signal is the thing the chat models are worst at.



Corn

Because by the time it reaches them it has been turned into something word-shaped.



Herman

That's my read. And it lines up with the movement part of Daniel's question. Let's take distance first. From a single microphone you can estimate how far away someone is using the ratio of direct sound to reverberation, early reflections, and the shape of the spectrum. A convolutional recurrent network was around eleven centimetres off on clean synthetic data, and about half a metre off on real recordings. Noise is what kills it.



Corn

Half a metre on a recording of a person. That's the difference between the kitchen and the doorway.



Herman

And then you could in principle repeat it over time and say the speaker's getting farther away. But I want to be straight about this: I haven't found a paper that detects walking direction from a single speech microphone using Doppler shift. A walking pace is so slow



Herman (continued)

that the shift is tiny. Tracking real movement is done with arrays, several microphones, where the arrival time difference gives direction.



Corn

So my dictated shopping list, one phone, can tell you roughly that I walked away. Not which way.



Herman

Roughly, in good conditions. And the thing that stuck with me on this is the spatial benchmark, STAR-Bench, nineteen models. If you give the model a caption of the audio instead of the audio, accuracy falls by about thirty-one percent on timing questions and thirty-five percent on spatial ones. Meaning the cues aren't in the words, and a pipeline that transcribes and then reasons has already lost them. Most audio language models are mono input anyway, no direction, no elevation, no distance. There's a binaural one, BAT, that was built for exactly this.



Corn

Herman, you said gait too. Is that people being identified by their steps?



Herman

Research stage. A group at NTU and Kentucky reported identifying individuals from footsteps at around ninety percent from short clips, on a microphone array. There's also work on detecting footsteps at home to monitor how elderly people move. Promising. Not something I'd say is deployed.



Corn

Ninety percent at a nursing home, and the other ten is a different resident getting the wrong lunch.



Herman

Which is why the third tier makes me uneasy. Here the claims run well ahead of the evidence. Corn, how do you feel about a phone call being checked for depression?



Corn

Uneasy. But the idea isn't silly. Speech changes when you're depressed. Slower, flatter.



Herman

It does change, and that's the trap, because the science can be real while the product fails. The best known voice depression company was Kintsugi, out of Berkeley. In February it shut down and put its models, methods and research out as open source. The chief executive, Grace Chang, said they were choosing "the integrity of the science over the limitations of a distressed market." They'd raised about thirty million dollars. One source says it ran out of time on an FDA route called De Novo. The main article I read doesn't mention the FDA at all, so I wouldn't say it was turned down.



Corn

Which is it, then? Was it cleared, rejected, abandoned?



Herman

I couldn't find a cleared voice or cough biomarker anywhere. Recent AI authorizations I found were imaging tools. Noah Labs has what's called Breakthrough Device designation for its heart failure voice product, but that's a designation, not a clearance. I should be careful, that's from searching around, not from checking the FDA's own database, so "we couldn't find one" is the right phrasing and "none exists" is not.



Corn

And Google? They had the cough one.



Herman

HeAR, announced in August twenty twenty-four. Trained on about three hundred million audio clips, with the cough model drawing on something like a hundred million coughs. Access is for researchers. An Indian company, Salcit, with a product called Swaasa, is using it for tuberculosis screening research. It's a research tool. It isn't a cleared diagnostic. And vendors like Canary Speech claim they can flag Parkinson's or mild cognitive impairment from speech, with no regulatory status I could find.



Corn

So health from voice sits somewhere between a good paper and a pitch deck.



Herman

And lie detection is on the far side of that. Voice stress analysis. A test of a commercial device by Washington University in two thousand four found it performed at or below chance, and the reviews from around two thousand two said the same. Stress changes the voice. Stress is not lying. A nervous honest person and a calm liar produce the opposite of what the device assumes. I haven't found a newer study that overturns it. Intoxication from voice is similar territory for me, I haven't seen anything deployed with real validation, though I'd put less weight on my search there.



Corn

A product that's right half the time is a coin with a subscription.



Herman

The reason it sticks around is that it feels right. A voice does crack when you're frightened, and nearly everyone has a story where they knew someone was lying from the sound of them. The stories are real. The sample of the times we were wrong is the one nobody keeps.



Corn

I'd bet I'm excellent at it, personally. In my own head.



Herman

Everyone is. Do you think a good interviewer can do it, though? Sit across from someone, listening, for years?



Corn

I think a good interviewer asks the same question three ways and checks whether the story moved. That's not the voice. That's the content.



Herman

There's one more, which is the reverse: detecting that a voice is synthetic. The ASVspoof five challenge had fifty-three teams, and the organizers say the systems consistently struggle to generalize, and degrade under adversarial attacks, codecs and out-of-domain data. Earlier work showed error rates rising by two hundred to a thousand percent on in-the-wild audio compared with the test set. That one isn't a settled task, it's an arms race.



Corn

Which means the hearing that matters, whether the voice is a person at all, is the one the machines are worst at.



Herman

For now. And that's a good place to see the whole scale. Where the cue is physical and narrow, a bird, a bearing, a machine, it works. Where the label is a person, it gets looser. Where the label is an inner state or an intention, it starts to be pseudoscience.



Hilbert

Sanyo M twenty-four hundred. Nineteen dollars and ninety-five cents.



Corn

A cassette dictaphone, if I have the era right.



Hilbert

Port Gannet Harbour Licensing Office bought it so I could count the gulls. By ear. They had a register of forty-one gulls, Herman, because forty-one was how many had been invoiced. One dollar forty a quarter. A Mr Osric Plumb sent the invoices to the pier.



Herman

To the pier, or to the gulls?



Hilbert

Third bollard from the lighthouse. Payment came back in sardines, nine to the dollar, and Mr Plumb logged every one.



Corn

Sardines go off. There has to be a spoilage column somewhere in that ledger.



Hilbert

Column F. Three hundred gulls on the first tape. Three hundred and twelve, I don't count the ones on the roof of the licensing office, they have a contract. Any gull I couldn't prove was licensed was legally a pigeon, so I kept a second tally and carried it across the road to a Mrs Ilse Frakes at Parks, who stamped it on Thursdays only. The eject button on the Sanyo is not built for claws, so I used a pencil.



Herman

The overlap is the technical problem there. A single gull makes a lot of calls, and ten calls on a tape can be one bird or ten. The bird models today work on exactly that, pulling overlapping calls apart and scoring each against a reference, which is a long way from a pencil.



Hilbert

An hour of tape took me eleven hours. Mr Plumb wanted it in two. Herman, you said roughly four times in the middle there. He said roughly too, and then he'd bill the lighthouse.



Corn

He wanted the count in two hours, and the gulls were paying in sardines. Someone in that office was always going to lose.



Hilbert

The lighthouse keeper lost. He was invoiced for eleven gulls that had never landed on him. He paid in a barometer.



Herman

Which is a spatial problem, in fairness. If a gull was never within range of the lighthouse, the localisation was wrong, and an array of microphones would have caught that where a single cassette recorder cannot.



Hilbert

Mr Plumb had one microphone and a theory that gulls came in from the east. They came in from the east on Tuesdays. The rest of the week it was the car park.



Corn

A harbour office that cannot tell a pier from a gull has bigger problems than the tape.



Hilbert

Then he stopped invoicing the gulls and started invoicing the trawler owners for them. Two dollars a gull, a quarter. That ran six years. The trawlers paid in sardines.



Herman

And the three hundred and twelve may have been calls rather than birds. If a gull can make seven calls, you were counting forty-odd birds and the register was closer than anyone thought.



Hilbert

One bird, seven calls. Yes. That would be the eleven hours.



Corn

If you take one thing from this, it's that hearing a lot of what's in a recording is easy now, and being allowed to conclude something about the person who made it is a separate matter.



Herman

Sharper than that, I think. The things that work are the ones where the answer is in the physics of the sound. Birds, rooms, machines. Everything that asks the audio to confess something about a human, their health, their honesty, still needs a person and a proper trial. And the count in a recording is rarely the first number you'd guess.



Corn

I'd like to see what happens when someone builds a proper open benchmark for accent, one that returns a spread, not a label. That would tell us how much of the middle tier is real.



Herman

Thanks to our producer, Hilbert Flumingtop, for keeping all of this in one piece.



Corn

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Herman

See you tomorrow.