# Episode Book

EPISODE #6215

# Can Anyone Still Spot an AI Voice?

| | |
|---|---|
| PUBLISHED | 22 October 2026 at 23:07 UTC |
| DURATION | 19:28 |
| LISTEN | myweirdprompts.com/episode/detecting-synthetic-speech-audio-51974/ |
| HOSTS | Corn, Herman, Vex, Hilbert |
| SCRIPT MODEL | anthropic/claude-sonnet-5.5 |
| VOICES | chatterbox-regular (chatterbox-local) |
| PIPELINE | V5.20 |
| BOOK GENERATED | 23 October 2026 at 07:44 UTC |

## SYNOPSIS

Two Chatterbox hosts are talking to you right now, and this audio carries an inaudible watermark — can you hear it? That is the design goal, and it is also the problem. This episode digs into the real state of synthetic speech detection: listeners judge voice clones as human about as often as they judge real people, while off-the-shelf AI voices still give themselves away. Detectors that score near-perfect in the lab lose roughly half their accuracy on in-the-wild audio, and one dropped below a coin flip. Commercial detector claims come from vendors grading their own homework, and the same laundering trick — a neural codec plus MP3 — defeats both forensics and watermarks. We walk through the cases where detection actually worked, the ones where it didn't, and why the honest answer to "can you tell?" is more interesting than either panic or complacency.

# TRANSCRIPT



### Corn

The standard line this year goes like this. Ears can't tell, detectors can't tell, the fakes have won, so stamp a watermark on everything and hope. Half of that is wrong in an interesting way, and Daniel's question lands right on the wrong half. He backs watermarking, calls it good practice, says so on the record. He just doesn't think it does much, because people will strip the mark, and a deepfake read out on the radio only gets caught if somebody records it, downloads it and runs it through a screen. So set the marks aside. If the only tools were a human listener or analysis of the audio itself, are we past the point where either one can catch synthetic speech?



### Herman

Start by splitting the word fake in two. A clone of a real, specific person is one problem. A generic synthetic voice, nobody in particular, is another. Humans have mostly lost the first and mostly not the second. And since we're on the subject, you are listening to two Chatterbox voices right now, and this audio carries that watermark. Can you hear it?



### Corn

Not a thing. Which tells you how much use it is to a man in a car.



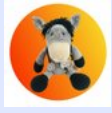
### Herman

It tells you the mark is inaudible, which is the design goal. Whether it's useful is a separate matter. Back to the people. A study in PLOS ONE last year played listeners real voices, clones, and generic AI voices. Real voices got judged human about seventy-two percent of the time. The clones got seventy.



### Corn

Two points apart. Over a bad phone line that's no gap at all.



### Herman

The generic AI voices were judged human noticeably less often. So the clone has caught up with the real thing, but the off-the-shelf voice still gives itself away. And the authors looked for the creepy result, the AI judged more human than the humans. They didn't find it at the group level. A few individual pairs, maybe, nothing you could generalise.



### Corn

I'd have bet the other way. I assumed by now the machines would be doing better than us at sounding like us.



### Herman

Most people would. A separate Berkeley study put listeners at around sixty percent on labelling real versus AI, where fifty is guessing. An older one, from twenty twenty-three, found people caught about three quarters of the fakes, but those came from older algorithms, and the authors said outright they expected worse for newer ones. Now, Corn, I have a question for you. Do you trust your own ear on this? You've sat through a lot of synthetic speech.



### Corn

For about four seconds, yes. Then I notice I'm judging the content, not the voice. If someone says something outrageous in a politician's voice, I'm already listening for the lie, not the lisp. The voice only gets examined when the words fail the smell test.



### Herman

Which is a problem when the fake is built to sound ordinary. Clips under about twenty seconds were especially hard for people in that Berkeley work. Short, flat, plausible. A voicemail asking a grandmother to wire money isn't going to contain a tell you can hear.



### Corn

So ears are past it for clones, not past it for generic voices, and nobody found the machines sounding more human than people. That's the human half. Now the machines checking the machines.



### Herman

Here there's a sharp gap, and it's the gap between the lab and the street. Detectors trained and tested on the standard research datasets score close to perfect. Then a team took three popular open-source ones and pointed them at in-the-wild audio, about fifty-six hours across forty-two languages, and they lost roughly half their ability to separate fake from real. One went from near-perfect to worse than a coin flip.



**Corn**

Worse than a coin flip. You'd have been better off with the coin, and the coin is cheaper.



**Herman**

The coin doesn't need a graphics card. And the reason is instructive. Follow-up work asked whether the in-the-wild fakes were simply harder, and the answer was mostly no. They were different. A detector learns the fingerprints of the generators it has seen. Show it one it hasn't and it has nothing to hold on to. On three newer open TTS systems, the detector that was best on one was the worst on another. Seven percent error on the first, over twenty-seven on the last.



**Corn**

A burglar alarm that only recognises last year's burglars.



**Herman**

And the next generator is always newer than the training data. Bigger models don't fix it either, which surprised people... more parameters didn't close the gap. It's got worse in the newest work too. A paper this year threw ten of the latest language-model-era speech systems at eight pretrained detectors, in English and Spanish, with compression and phone-band downsampling added. The best overall error rate was about twenty-nine percent, and most of the detectors sat near chance. I've only seen the abstract, so take that as the headline and not the detail.



**Corn**

Twenty-nine percent from the best of them. I've seen better odds on a bus timetable. Then what about the commercial ones? Because the companies selling them say ninety-something percent.



**Herman**

They do, and every one of those numbers is either run by the vendor or labelled in a way that flatters the vendor. Resemble published a benchmark this spring where its own detector came first, at ninety-eight percent, with a competitor down at seventy-one. That's a company marking its own homework. It may be a fair mark, but I can't tell you it is. The one in-the-wild academic benchmark that tested a commercial model gave its best one eighty-nine percent, and the paper itself warns the labels were partly derived from



### Herman (continued)

commercial detector outputs.



### Corn

Circular. The answer key was written by the student.



### Herman

I haven't seen any independent head-to-head of the big commercial detectors on modern TTS. If one exists, I'd love to read it.



### Corn

And Resemble is also the company that open-sourced Chatterbox while selling the detection. We covered that. The house that gave everyone the lock-picking kit and also sells the alarm.



### Herman

To be fair, they're not hiding it. It does mean one company is shaping both sides of the arms race and publishing the scoreboard.



### Corn

So lab detectors are great, wild detectors are poor, commercial detectors are probably better and nobody independent has checked. That sounds like a case for giving up. Why doesn't it end the story?



### Herman

Because there's a counter-example, and it's the one everybody knows. The Biden robocall in New Hampshire. A security firm, Pindrop, took the audio, downsampled it to eight kilohertz, and compared it against a library of a hundred and twenty-two known text-to-speech systems. Their conclusion was ninety-nine percent likelihood it came from ElevenLabs components. ElevenLabs' own classifier agreed with lower confidence, and Berkeley's Hany Farid independently said the same.



**Corn**

Pindrop sells that analysis for a living, so I'll hold the ninety-nine loosely. But two others agreeing is hard to wave away.



**Herman**

The tell they pointed to was the fricatives, the hissy consonants, and unfamiliar phrases, where the synthesis stumbles. So signal forensics did work, in hours, on a phone-quality clip. What it needed was a suspect list. That was an identification against a lineup, not a detector scoring a stranger.



**Corn**

And before we get to the courtroom, the timing. Forensics took hours there. An election doesn't wait hours.



**Herman**

Slovakia, two days before the vote in twenty twenty-three. A clip of a candidate apparently discussing rigging the election surfaced during the pre-election quiet period, when the media and candidates were barred from campaigning. Experts told AFP it looked synthesised. I've only seen secondary summaries of that, and nobody has measured what difference it made. But look at the shape of it. The debunking arrived in a window where answering back was hard, and his side could only say that no tool detects AI content with absolute certainty.



**Corn**

Which is true, and also the worst possible thing to say to voters.



**Herman**

It's a sentence that sounds like a dodge even when it's accurate. That's the trap with detection in the wild. Stating the limits of a detector accurately and making an excuse use the same words.



**Corn**

Back to the legal side. A six million dollar FCC penalty, a telecom settlement, and the man behind it was acquitted on every state charge.



**Herman**

Which says something about the gap between knowing a voice is synthetic and proving who did it and what they intended. The forensics succeeded. The prosecution still wobbled. With the Baltimore school principal clip, the experts hedged, one of them thought it might be several real clips stitched together, and the outcome rested on the police work, the search history and the devices.



**Corn**

So when it worked, it wasn't the audio that did it. It was the paperwork.



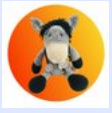
**Herman**

Mostly. Now, a harder issue, which is laundering. In the big spoofing challenge, the hardest conditions for detectors were audio pushed through a neural codec called Encodec and then compressed to M P three. And a benchmark of watermarking schemes found the same trick, Encodec then M P three, meant the marks went undetected nearly every time, for every scheme it tested.



**Corn**

Hold on. The same laundering defeats the forensics and the watermark.



**Herman**

Both layers. Caveat, that benchmark tested three schemes and not the one we use, so I can't tell you what happens to Chatterbox's mark. The project says it survives M P three, and I haven't seen an independent test beyond that. And there's a plainer point. In the released code the mark is a single call. With the weights on your own machine, taking it out looks like a one-line edit. That's me reading the source, I haven't tried it, but it's why I'd say the watermark binds the good-faith users and the hosted services, and nobody else.



**Corn**

And a radio broadcast, or a phone call, is laundering you get for free. Nobody has to do anything clever. The channel does it.



**Herman**

That's my read too, and it's an inference, not something I've seen tested on a broadcast chain. But in the spoofing results the narrowband phone codecs came right after Encodec as the hardest conditions. The cues detectors lean on live in the detail that a phone line throws away. Which is awkward, because the places people actually get fooled, calls and voicemails and clips forwarded through chat apps, are the places where the audio has already been squeezed.



**Corn**

Daniel's point in sharper form. A seat belt for people who already drive carefully, on a road where the weather scrubs the markings off anyway.



**Herman**

There's a nastier wrinkle. In that same benchmark, an attacker with access to the detector could take real recordings and make them read as watermarked, essentially a perfect false positive rate. So a mark saying this is AI can be forged onto something real. If the signal says fake, that alone isn't proof of anything either.



**Corn**

Hmm... so the mark can come off a fake and go onto a real one. Which one would you rather be accused with, Herman?



**Herman**

Neither, thank you. Now the question you started with. If the watermarks were stripped out entirely, could anyone still tell a synthetic voice from a real one?

v

### Vex

LISTEN TO ME, the bubble wrap is the whole tell! Every AI voice you have ever heard is stuffed with the exact same air that ships inside bubble wrap, and I can prove it because I ran the numbers on a spreadsheet at two in the morning!

v

### Vex

Two thousand and one, that is the year it started, they quietly rerouted every sheet of bubble wrap through a server farm, and now every machine voice pops one microscopic bubble per syllable, which is why you can never fully trust a robot saying your name!

v

### Vex

The watermarks are a decoy, the bubbles are the receipt, and I am DONE being the only person in this town who hears it, so you two hosts need to get on this air and pop it wide open on the record!



### Corn

That's him again. I'd love to know how he gets past screening. Maybe he just keeps ringing until the screener gives up and goes home.



### Herman

Bubble wrap was invented in nineteen fifty-seven, as textured wallpaper, so nobody could have rerouted it anywhere in two thousand and one. Anyway, whether anyone can tell depends on who's doing the telling. A listener with a clone, a short clip and a phone line is guessing. A detector that has seen the generator can be very good. A detector that hasn't is a coin.



### Corn

Three answers to one question. Ears are past it for clones, lab detectors were never really in the race, and the wild ones are near the edge.



### Herman

I'd sharpen it. For a short compressed clip put in front of a human, we're past it. For anything an investigator can spend a day on, with a suspect list and a clean reference recording, we're not. The trouble is in the middle, where it matters. A fake is posted at nine, millions have heard it by noon, and nobody has the day.



### Corn

That's where the radio listener sits. Nobody on a broadcast is going to download the stream, run it through a pipeline and ring in.



### Herman

Provenance is wider than the watermark, though. It's two different bets. One is marking the fake, which is the watermark, and we've just spent ten minutes on why that leaks. The other is signing the real. A recording carrying a cryptographic signature from the device that made it, unaltered since, proves itself. Corn, would you sign your own recordings, if your phone did it for free?



### Corn

Instantly. I'd sign my naps. Nobody else is going to impersonate those.



### Herman

Fair, though it only works if signing is everywhere. In a world where most recordings are unsigned, a missing signature means unknown, not fake. And a listener still can't verify a live stream as it plays, because nothing on the receiving end checks. It's a forensic tool for people with time, not an instant check for the public.



### Corn

Then a missing signature on a real clip makes the real one look suspect.



### Herman

It can, and that's the cost of the approach. Another thing I haven't settled in my own head... whether you'd rather know a clip is unsigned before you share it, or after.



### Corn

Before, and I'd be furious about how often it'd say unsigned. Most of what I hear all day is unsigned.



### Herman

You're allowed to defend Daniel's side of this, you know.



### Corn

I will, since he's not here to. He says watermarking is good practice, and I agree, and I still think it's a bad thing to build a defence around. Those are compatible.



### Herman

They are. A mark costs the well-behaved generator almost nothing, and it gives a platform something cheap to check at scale. It turns a large share of the careless abuse into a lookup. Where I'd disagree with the mood around it is the idea that it can settle a single disputed clip. If the mark is present, you learn it came from a service that stamps. If it's absent, you learn very little. That asymmetry is the whole limit.



### Corn

Present means something, absent means nothing. Like a receipt. A shop that gives you one proves you shopped there, and a missing one doesn't prove you stole it.



### Herman

And a forged one, from the forgery work, proves even less. Here's one for you, Corn. If a clip of you saying something awful turned up tomorrow, and it carried no mark and no signature, what would convince you it was fake?



### Corn

Where I was. Who saw me. I'd be reaching for an alibi, not an acoustic report. Which, now that I say it aloud, is the thing the Baltimore case turned on.



### Herman

The alibi is the oldest provenance system there is. It's just very slow. The forensics doesn't only catch the lazy, by the way. The Biden call was a competent fake that happened to use a commercial service with a fingerprint. The determined attacker runs their own model and launders the output through a codec, and that attacker isn't hypothetical. It's a laptop.



### Hilbert

Eight kilohertz is not what the phone carries. That's how often it samples. The voice on the line stops at about three and a half, Herman, and it has done since before the exchange I worked on was switched on. I wanted that said.



### Herman

Fair. Eight is the sampling rate, the usable voice band is far narrower. It doesn't change the forensic point.



### Hilbert

It changes nothing. I know. Mr. Delahunt next door at number nine paid forty-one dollars to a firm that built speaking clocks, to have his own voice put on one, so he could ring it. He rang it every evening for a year to find out whether he could tell it was him. He couldn't. The clock ran eleven minutes slow and he took that as proof. I have the receipt and the letter. The letter is better. They thanked him for his patience with the voice. How they got his forty-seven out of a sentence about the weather, I'd actually like to know that.



### Herman

A listener test with one subject and no control, run for a year. And look where he put the tell. Not in the voice. In the lateness, which is a behavioural fingerprint, not an acoustic one.



**Corn**

Eleven minutes slow means he was late for his own bins every week.



**Hilbert**

Twelve. The letter says twelve. I rounded it down for him.



**Corn**

Which is what the forensics people end up doing. They stop listening to the voice and look at what the thing does around the edges. How it handles an unfamiliar phrase, how it breaks on a hiss.



**Herman**

That's what stays useful when the voice itself is perfect. The audio gets better, but the pipeline around it leaves habits. Timing, odd words, the way a service strings sentences together. A year of evenings by ear didn't settle it for him. The pattern around the voice is what a machine can still hold up against it.



**Corn**

Though those habits are generator-specific, so you're back at the lineup. You only catch the fakes you already have a file on.



**Hilbert**

He wouldn't have agreed with you. He'd say the clock was sorted.



**Herman**

That's the lopsided part of the arms race. Defenders need coverage of every generator. Attackers need one nobody has profiled, and open weights make that easy.



### Corn

So, for Daniel. Is watermarking the only thing left? It isn't. Signal forensics still works sometimes, and it attributed the best-known fake of the last few years. But run as a general detector on a random clip, it's near the edge of useless, and ears are no better. Marks are worth doing and not worth leaning on, and the thing worth building is a way to prove the real.



### Herman

One from the cutting room floor. The man behind the Biden robocall said in his defence that he paid a magician a hundred and fifty dollars to make it.



### Corn

Cheap for a magician. Expensive for the client.



### Herman

What I keep wondering is whether provenance can ever reach a listener at the speed of a broadcast. One man rang a clock every evening for a year and ears never settled it. A signature that checks out for an investigator on Tuesday doesn't help a voter on Monday.



### Corn

Thanks to Hilbert Flumingtop for producing. This has been My Weird Prompts. If you enjoyed it, a review on your podcast app helps more people find the show.



### Herman

See you tomorrow.