Episode Book

EPISODE #6176

Do Audio Models Hear How You Said It?

PUBLISHED	21 October 2026 at 14:03 UTC
DURATION	20:27
LISTEN	myweirdprompts.com/episode/audio-models-emotion-detection-25004/
HOSTS	Corn, Herman, 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

Google's Gemini audio docs promise emotion labels with timestamps. A benchmark called LISTEN tested whether the models actually use the sound of a voice or just transcribe and guess — and the numbers aren't flattering. This episode digs into the gap between what speech models claim to detect and what they really hear, what specialist open models like emotion2vec+ and SenseVoice offer instead, and why Hume's prosody API shut down in June. Plus: what happens when the caller is hostile and the person on the other end is a call-center agent with no protection.

TRANSCRIPT



Corn

Does an audio model hear a man shouting down a phone, or does it read the transcript and guess the rest? That's where Daniel's prompt lands, and he starts from a hostile call, the kind where the transcript keeps every word and loses the raised voice, the menace, the register. Google's own documentation lists emotion detection in speech as a supported use, so on paper the promise is real. What he wants to know is how much of it holds up, and whether anyone has built a model just for how something was said, open or closed.



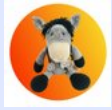
Herman

The documentation part is accurate, and it's almost touching how plain it is. The Gemini audio guide shows a prompt that returns a single label, happy, sad, angry or neutral, with timestamps. Hand it a call, ask for the emotion, you get an answer. What matters is what the answer is made of. A benchmark called LISTEN came out last October and went straight at that. Gemini two point five Pro and Flash, Qwen three Omni, a few others, tested on whether they use the sound or the words.



Corn

From the title, I'd guess the words.



Herman

From the numbers, you'd be right. When the emotion lived in the voice and the words were flat, accuracy across the models landed between roughly sixteen and twenty-three percent. Pure chance is ten. Guessing the most common answer every single time gets you thirty-nine.



Corn

A rubber stamp that says neutral beats Gemini at hearing tone.



Herman

On that test, yes. And the pattern is the useful part. When the words are neutral, the models say neutral. The paper's own summary is that they transcribe rather than listen. I'm not sure that's entirely fair, because that condition is the hardest one by construction... but



Herman (continued)

a caller who keeps his vocabulary polite while his voice does the threatening is exactly that condition.



Corn

So Daniel's case isn't an edge case, it's the exam. And it fits what we found in the episode on the signal versus symbol gap, where a Gemini model reported vocal fatigue just because the speaker said they were tired. A lip-reader that favours the words.



Herman

Near enough. Though the picture isn't flat. Earlier work, from March of last year, ran GPT four o audio and Gemini two point zero Flash on emotion datasets with no training at all, and they came out in the mid-forties on a standard F-score, up to about fifty when you feed in a few turns of conversation first.



Corn

Mid-forties on actors, though. People performing anger for a microphone in a studio.



Herman

Acted or scripted television, yes, which is the easy version of the problem. A real caller doesn't enunciate his rage. He mumbles it, or he goes quiet, or he's perfectly civil for forty seconds and then he isn't.



Corn

We've done the word-accuracy fight before, dedicated speech recognition against the multimodal models, and the old transcribe-then-reason sandwich. This is the layer above that, whether the thing hears anything beyond the words at all.



Herman

Right, and the label set is part of the trouble. Four boxes. Sarcasm is the hard case, because the words say one thing and the voice says the opposite, and that's the exact situation where the models fall back on the words. Hostility has the same shape. "Thank you so much for your help" can be a thank-you or a threat, and a transcript can't tell you which.



Corn

Four boxes, and sarcasm goes in neutral and takes the whole call with it.



Herman

Part of the fix on the research side is to drop the boxes. Psychologists describe emotion on two axes, how aroused someone is, calm to agitated, and how positive or negative. Arousal is the one that generally shows up in the sound itself, louder, faster, more variable pitch. Valence is the one that hides in the words. A shouted sentence can't always tell you whether it's joy or fury, but it can tell you the person isn't calm.



Corn

For Daniel's call, that's most of what matters. I don't care whether he's angry or frustrated or terrified, I care that he went from calm to not calm at ninety seconds.



Herman

Which is an argument for the specialist models over the four-label prompt, though I'm not sure how many of the open ones hand you arousal as a number and not a label. Anyway, Corn, let me ask you something. Suppose a model hands you "angry" on a call. What do you do with that?



Corn

I listen to the call. Which defeats the purpose, unless the label tells me where to listen. A single word for twelve minutes of audio is useless. A timestamp that says "ninety seconds in, something changes" is worth having.



Herman

That's the better use of those docs, actually. The timestamps matter more than the label.



Corn

So Beryl, you're on the line from up north somewhere. What's your take on all this?
Beryl: Hello lads, Beryl here from Headingley, I'm on the number thirty eight bus into Leeds with me shopping on me knees and a pork pie going cold in the bag, and I've been thinking about your topic all week, about the machines hearing how a thing is said and not just the words, and it put me in mind of me nephew Dean, who rang me up last Boxing Day and said he couldn't come for his tea in a voice that sounded perfectly cheerful, and I knew from half a second of it that something were wrong, because he does this little swallow before he lies, and he'd been dumped and didn't want to say, and I just wondered, could these models pick that up, that swallow, or do they only hear the words and miss the man entirely?



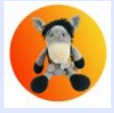
Herman

Lovely to have you, Beryl, and mind the pie. The real thing in there is that you know Dean's ordinary voice, so the swallow stands out against it. A model gets one clip of a stranger with nothing to compare it to, which makes the job much harder, so with a stranger on a phone line I'd lean on the shape of the whole call and not one moment in it.



Corn

Mid-call is where dedicated models come in, I'd have thought. Something built for the single job should beat a general model that does everything.



Herman

That was my guess too, and I'll be careful here, because I haven't seen the head-to-head on real calls. But the dedicated ones exist. The best-known open one is emotion two vec plus, from a research group in China, released in May of two thousand twenty-four. Around three hundred million parameters, so small enough to run on a laptop, and the large version was trained on something like forty-two thousand hours of audio labelled by other models.



Corn

Labelled by other models. The student marking the student's homework.



Herman

A bit, yes, pseudo-labelling is how they got the scale. One thing to check: the readme talks about nine emotion classes, but the model cards for the base and large versions list fewer, so you'd want to open the actual checkpoint before building anything on it.



Corn

What does it give back? A word?



Herman

A label and scores per class for a clip. Not a transcript. The other one I'd point at is SenseVoice-Small, from the same family of open projects as the Qwen speech work. It does transcription, language identification, emotion recognition and audio event detection, laughter, coughing, that kind of thing, in a single pass. Five languages: Mandarin, Cantonese, English, Japanese and Korean. Who said what is a separate pipeline.



Corn

So the transcript comes back with the feelings attached.



Herman

Roughly, yes. A rich transcript. Words, plus tags for what the voice was doing around them. For Daniel's call, that's closest to what he described.



Corn

Closest on paper.



Herman

Fair. Then there's Hume. Their expression measurement API had a prosody model that scored forty-eight dimensions of vocal expression, not four boxes, and it was the standalone product people pointed to. The last day to create jobs was in May and the API shut in June of this year. Their speech-to-speech assistant carries on and uses prosody inside.



Corn

The best-known vocal emotion API, and the tombstone says June.



Herman

My reading, and it's a guess, is that the capability moved into the big general models and the speech-to-speech systems, where it's a feature and not a product. The fact that the standalone market gave up doesn't tell you the models got good at it, though. LISTEN says otherwise.



Corn

Or it tells you nobody paid for it.



Herman

That's possible. I don't know which.



Corn

I'd bet the newer tests this year show the same gap. Nobody retrained these models to listen harder, they just made them bigger.



Herman

Going by the abstracts, and I haven't read the full papers, that's what they say. One from Interspeech this year found audio models judging speaking style, rate and emphasis trail human listeners by about thirty-two points. And a speech-to-speech benchmark presented at ICLR found those models don't respond to how you said something, and do no better



Herman (continued)

than a plain pipeline of transcribe, then answer.



Corn

Taking the bread out of the sandwich didn't add ears. Open models, then. You're going to tell me about licences.



Herman

I was about to, so thank you for ruining it. Qwen three Omni is Apache two point zero, a thirty billion parameter mixture with only about three billion active at a time, and there's a variant called Captioner that writes detailed descriptions of audio, which is the more promising shape for this job than a label. The newer Qwen three point five Omni from April tops Gemini three point one Pro on one general audio understanding benchmark and trails it on another. But neither of those is an emotion number, and I wouldn't dress them up as one.



Corn

And the other open one?



Herman

Audio Flamingo three, from NVIDIA. Strong on general audio understanding, but the licence is noncommercial. If Daniel wants it for a business that's the end of the conversation. Apache versus noncommercial decides it before accuracy does.



Corn

Does anyone have a scoreboard where these sit next to each other on emotion?



Herman

One paper, Hello-Chat from February, compared them on parts of a speech emotion collection. Qwen three Omni came out ahead on emotion recognition, while a Gemini three preview led on detecting audio events like a door slam or applause. That's the



Herman (continued)

closest to an emotion comparison with a Gemini three on it, and it's one paper. I haven't seen an emotion score for the current Gemini generation on its own, and I haven't seen anyone benchmark any of this on real hostile call-centre audio. Acted data and clinic recordings, yes. Abusive callers, no.



Corn

So everything we've said is about actors.



Herman

Mostly, yes.



Corn

The thing I'd want to remember about Daniel's case is who's on the other end. It's an agent taking call after call from people who have decided she's the problem. A transcript that says "caller was rude" is a weak record.



Herman

And it's a record for different people with different needs. The agent who needs it logged, the team lead who needs to know a call turned, the trust and safety person deciding whether to escalate a threat. A description that says the volume rose, the pace sped up and he stopped letting her speak is worth more than the word "angry".



Corn

Does anybody sell that today? I'd assumed the call-centre software people got there years ago.



Herman

They sell adjacent things. Cogito flags agitated or pressured speech and caller distress live during a call, and Observe.AI has talked about tagging strong emotions toward agents. Both are older products, and what I couldn't find is one built specifically to shield



Herman (continued)

an agent from an abusive caller, with the evidence trail that implies. Maybe it exists and I missed it.



Corn

A product for the person being shouted at. Apparently that's a gap.



Herman

It's a gap in what I know of, at least. Most of what I have seen in that software points at the agent.



Corn

Do the rules let you do that?



Herman

Depends who you point it at. And I'm going from the European Commission's guidance and secondary write-ups, not from the statute itself. Article five of the EU AI Act bans using AI to infer emotions in the workplace and in education, with narrow exceptions for medical and safety reasons. The Commission's guidelines, which aren't binding, say a call centre using voice to track the emotions of customers is outside that ban, while tracking the employees by voice is inside it.



Corn

So analysing the abusive caller is fine, and scoring the agent's feelings isn't.



Herman

That's the asymmetry. Outside the ban doesn't mean unregulated, though. Emotion recognition elsewhere falls into the high-risk category, with obligations. And a Hungarian regulator ordered a bank to stop analysing the emotions in customer call recordings, under GDPR, which is a separate law. Anyone running this on a real support line in Europe needs a lawyer and not a podcast.



Corn

Podcast law is generous.



Herman

It's very generous.



Corn

Then the accuracy matters more, because a flag on a call lands on a person. The agent gets coached, or the caller gets reported.



Herman

Right, and that's where the numbers bite. Even the bigger Gemini model, Pro, landed at about eighteen percent on voice-only emotion, below a guess of neutral every time, and nobody should hang a report on that.



Hilbert

Fifteen point seven. Eighteen was Flash. Pro was the lower one, Herman, and I know because both have been open on my screen since you started, with a voicemail from my Aunt Dolores. Tuesday, the answerphone, nine minutes, the water board. I typed the instruction with the pencil, the keys being spaced for people without claws.



Herman

Nine minutes of one voice is a fair clip, plenty of signal in it. It should have built over time.



Hilbert

Both said neutral. Courteous, formal, mildly devout. She was singing Abide With Me at the water board, Herman, all four verses, at a volume that moved the cat. Between verses she read out the account number. The water board has barred her from four call centres. Three. Swindon was a chat window and I don't count those.



Corn

The cat was the better classifier.



Hilbert

Dennis. He left for two days and came back in a different collar. Dolores took the printout to the appeal. Neutral, it says, at the top. Her cousin Gerald chairs the appeal, they sit in the Morrisons café in Prestatyn, and Gerald wears a snorkel on Thursdays for his sinuses.



Herman

A snorkel would put the chairman's own voice through a very odd filter. Any model labelling that hearing would be guessing as well.



Hilbert

Three bans lifted on the spot. Swindon he kept, because of the capitals. Then Dolores read the printout properly. Forty-one calls over three years, she remembers every one, and the position she's taken is that being scored neutral on forty-one of them is a slur.



Corn

That's a specific number of slurs.



Hilbert

She's claiming three hundred and twelve pounds from the water board. She wants it in writing that she was furious. I told her the model only saw the one. She says the other forty she can do again on Thursday, with the choir.



Herman

A choir would give the classifier a lot more to work with. Eleven voices in harmony drown out a single one at any volume.



Hilbert

Eleven, yes. Mr Anwar Bristow conducts and is paid in Battenberg, one slice a call. He has asked for the forty-one in a lump.



Corn

A lump of Battenberg would be a lot to absorb in one sitting.



Hilbert

The post office did the lamination for one pound forty. I priced a laminator at Staples. Forty-one ninety-nine.



Herman

One pound forty for the lamination, and the clip was nine minutes. Nine minutes is the part I'd keep hold of. One label for the whole clip flattens a call that builds, and the useful output is the shape across it. The first sentence was probably fine, the fifth wasn't, and the label can't say which.



Corn

Then ask for the shape. Don't ask the model for a verdict, ask it to describe what it hears in the stretch, loudness, pace, who's talking over whom, and time-stamp it.



Herman

I'd do that, with a caveat. If the model leans on the words, a description might be a more detailed version of the same guess, and I don't know how far to trust it.



Corn

I wouldn't trust it as a finding. I'd use it as an index of where to go and listen.



Herman

Right, and put a dedicated model beside it as a second opinion, emotion two vec plus or SenseVoice, because a general model and a small specialist failing in different ways is useful information. If they disagree about the same ninety seconds, that's the stretch a human listens to first. Where's your line, Corn? What's a flag allowed to do on its own?



Corn

Open a ticket. Highlight the clip. Never close an account, never discipline anyone. Anything with consequences gets a person with headphones.



Herman

That lines up with how the European treatment reads, high-risk where it affects people. There's also a dataset called ParaSpeechCaps with fifty-nine style tags for voices, whispery, nasal, shouting, and so on, but it was built to train text-to-speech to produce those styles, so it's a generation resource and not a captioner you can aim at a phone call. I'd keep it off the list.



Corn

Noted, and deleted from the shopping list. You keep the audio. Whatever gets laminated, the audio is the evidence.



Herman

The audio is the evidence, and everything the models say is an index into it.



Corn

Where I'd want to see this go is a proper test: real abusive calls, the same ones shown to a Gemini three model, the Qwen Captioner and a dedicated emotion model, with a human marking where the call turned. Nobody has published it, and it would settle more than the benchmarks on actors have.



Herman

Thanks to our producer, Hilbert Flumingtop, for keeping the show running. Mixing and cutting is the part nobody hears and everybody would notice.



Corn

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



Herman

We'll be back soon.