Episode Book

EPISODE #6468

Why Tiny Neural Gates Replaced Voice Activity Gates

PUBLISHED	5 November 2026 at 20:16 UTC
DURATION	21:38
LISTEN	myweirdprompts.com/episode/neural-voice-activity-detection-11974/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	18 cited by the research (listed at the end)
BOOK GENERATED	6 November 2026 at 11:12 UTC

SYNOPSIS

Modern voice activity detection has quietly shifted from amnesiac volume gates to sub-megabyte neural networks with running memory. The ecosystem centers around open-source contenders like Silero, TEN VAD, and FireRedVAD, alongside commercial on-device options. While standard neural detectors excel against background air conditioners, separating intended speech from ambient chatter requires specialized architectures. Conditioning techniques like Google's Personal VAD use FiLM layers and single-shot enrollment embeddings to isolate target voices without expensive per-frame processing.

TRANSCRIPT



Corn

Daniel's follow-up today is about voice activity detection, and it starts with a confession. He says he'd always assumed a VAD was basically a volume gate, an acoustic call on voice versus non-voice. Then the adaptive gating prompt happened, he learned the modern ones are neural, and now he wants the lay of the land. Who are the actual players, open source and paid, that you can use as components?



Herman

And he's not stopping there.



Corn

He isn't. His dictation lens is the interesting part. He says the real question in a dictation environment isn't whether this is speech or an air conditioner. It's whether this is intended speech. Is that a background conversation? Bar music? A PA system that's going to get transcribed if we let it through? And can you answer that without running voice embeddings on every frame? Plus he wants the imaginative stuff, the strangest VADs out there.



Herman

The confession part is what I like, because Daniel's instinct about crude VADs isn't wrong. It's right about the wrong thing. The WebRTC detector, which is still everywhere, is a Gaussian mixture model from the telephony lineage. On FireRedVAD's benchmark it misses sixty-four percent of speech frames. Sixty-four. The crude gate he imagined is real, it ships in production today, and it's terrible.



Corn

The volume gate that survived.



Herman

But the modern field is nothing like that. Silero is the de facto default, about three hundred and nine thousand parameters, one point two megabytes, MIT licensed. TEN VAD is three hundred and six kilobytes for the whole library and is explicitly positioned against Silero on endpointing latency. FireRedVAD is the new ambitious one, about two point two



Herman (continued)

megabytes, and it doesn't just do speech versus non-speech, it has separate heads for speech, singing, and music. So the open-source field is a three-way knife fight, and the fights are over milliseconds and kilobytes, not just accuracy.



Corn

Three competitors fighting over a fraction of a megabyte. Sibling rivalry in a very small house.



Herman

It's a tiny house! Silero's model is one point two megabytes. This is not a large language model situation. And the thing that makes them different from the crude gate isn't just accuracy, it's memory. Silero's core is an LSTM whose hidden state carries across chunks. You feed it thirty-two millisecond pieces of audio, and each decision is informed by everything it has heard before in the stream. A GMM looks at one frame and votes. Silero has a running impression of the conversation.



Corn

That's the part I think gets underexplained. The gate everyone imagines is amnesiac, each frame judged alone. The neural one is more like the doorman who remembers you walked in with somebody.



Herman

That's exactly the mental model. And the other underrated piece is the output. Both Silero and TEN give you a per-frame speech probability, not a yes or no. That's what lets you build hysteresis on top, smoothing windows, tunable thresholds, all the plumbing that makes a VAD usable in a real streaming system. The binary flag is where the crude ones stop.



Corn

So who's actually winning the knife fight?



Herman

Depends whose table you're reading, which is a sentence that should worry everyone. FireRedVAD's README has a benchmark, their own FLEURS-VAD-102 test set, nine and a half thousand files across a hundred and two languages. On it, they score F1 of ninety-seven point six, false alarm rate under three percent, and they show Silero and TEN both below them. TEN's README shows TEN beating Silero, with a specific claim that Silero suffers a delay of several hundred milliseconds and fails to identify short silent durations between adjacent speech segments.



Corn

Each table crowns its own author. Shocking.



Herman

There's an independent-ish benchmark, the LibriVAD-small paper from October, and it's honest to a fault. They ran TEN and Silero on noisy utterances, and at a thirty millisecond frame Silero actually edges TEN, AUC around zero point nine two versus zero point nine two, basically a tie. At twenty milliseconds Silero collapses to zero point six three. And then the authors stop and say, hold on, our fixed-frame protocol may be at fault, not the model. We didn't ablate it.



Corn

They measured the harness, not the detector.



Herman

Which might be the most useful finding in the whole paper! A naive benchmark harness, zero padding, state you forget to reset, can make a good streaming model look broken. The practical takeaway is boring and correct: pick on latency and footprint first, because those numbers are honest, then measure on your own audio, because every public table was written by someone with a horse in the race.



Corn

The energy gate baseline in that paper, what did it score?



Herman

Around zero point seven five to zero point seven seven AUC, across every window size. That's the crude VAD of Daniel's imagination, measured. The neural ones sit near zero point nine two. That's the gap, and it's the whole reason the field moved.



Corn

A quarter of the distance to perfection, claimed by tradition.



Herman

Now, the paid side, because Daniel asked, and here's the strange part. The market is thin. Almost deliberately thin. The clearest paid VAD as a component is Picovoice Cobra. On-device, deep learning, runs everywhere including a Raspberry Pi and the browser, and the marketing claim is doubling the accuracy of WebRTC VAD.



Corn

A low bar doubling.



Herman

Generous math, honestly. The pricing is where it gets hairy. Free tier is strictly non-commercial. The Foundation plan is six thousand dollars a year, and that's for startups under five years old with twenty or fewer employees. Enterprise starts at thirty thousand a year with a twelve month minimum. And the processing is local, but you need an AccessKey, so it phones home for licensing even though the model never leaves the device.



Corn

Paying thirty grand a year for the privilege of running it yourself. The cloud would be embarrassed to try that.



Herman

Everyone else bundles. AssemblyAI prices streaming at forty-five cents an hour base, fifteen cents for the streaming tier, and there's no VAD line item at all. VAD comes free inside the transcription. And that tells you where the money is. The money is in ASR. Nobody's getting rich selling voice activity detection, because the good open-source ones are a megabyte and MIT licensed. VAD is a feature, not a product.



Corn

One megabyte and free versus six thousand a year. I've done that math before.



Herman

To be fair to Cobra, the open-source ones come with zero support and a README that argues with competitors. Some buyers want a vendor to shout at.



Corn

The long tail is worth a pass. What's actually out there below the big three?



Herman

RealtimeSTT is the popular one, over ten thousand stars, low-latency speech-to-text with VAD and wake-word built in. FluidAudio, three thousand stars, Swift and CoreML, VAD plus diarization for the Apple world. sherpa-ncnn runs VAD and ASR offline on phones and Raspberry Pis. There's speech.cpp, which runs Silero on ggml with Metal and Vulkan backends, new as of September. And my favorite of the tail, libfvad. Six hundred stars. It's just the WebRTC GMM, extracted into clean C under a BSD license.



Corn

The crude gate, given its own repo and a maintenance community. A museum with a gift shop.



Herman

It's in production everywhere! That's the thing. The sixty-four percent miss rate detector is still the one shipping in countless pipelines because it costs almost nothing and it's one C file. And FunASR-VAD deserves a mention purely for the cautionary number: on FireRed's table it has a forty-four percent false alarm rate. It hears speech in nearly half the silence.



Corn

It's very enthusiastic.



Herman

Over-prediction of speech segments, they call it. I call it a transcriber's tax.



Corn

So that's who's playing. But Daniel's real question isn't which VAD wins a benchmark. It's whether any of them can tell intended speech from a bar full of people talking.



Herman

And the honest answer is that the generic ones can't, and we know this precisely. TEN's own per-condition breakdown, in their docs, across noise types. Machinery noise, traffic, music, they degrade but they hold up. Babble noise at low signal-to-noise stays close to chance.



Corn

Babble, meaning a room of people talking.



Herman

Which is exactly a dictation environment. An air conditioner is easy. Air conditioners are not speech. The hard case is other humans speaking. A generic acoustic VAD has no way to know which mouth the microphone belongs to. That's the wall, and it's the strongest technical argument that just getting a better speech detector will never solve dictation.



Corn

So the field splits into escape routes. What do you condition on?



Herman

Two of them, and the first is the one Google published as Personal VAD 2.0, INTERSPEECH 2022. The output is not two classes. It's three. Target-speaker speech, non-target-speaker speech, and non-speech. That middle class is the whole thing. Someone else is talking is its own answer, not something lumped into silence.



Corn

The architectural answer to background conversation is to make it a category.



Herman

You give the model a speaker embedding at enrollment, once, and the VAD itself stays conditioned. You're not running an embedding extractor every frame. And the trick that makes it deployable is beautiful. During training, they randomly replace the target speaker embedding with a zero vector and relabel all the non-target speech as target speech. One model, two behaviors. With an embedding, it's a personal VAD. With a zero vector, it degrades to a standard VAD. Graceful degradation by construction.



Corn

That's Daniel's without doing embeddings every time, answered in a training detail.



Herman

A Conformer backbone, streaming-friendly, quantized to eight bits to fit an on-device budget. They describe it as an always-running system that replaces standard VAD. And there's a small detail inside it that I keep coming back to. The obvious way to combine the speaker embedding with the acoustic features is concatenation, glue the vectors together. They found that sub-optimal, and their reasoning is that the two represent very different information, extracted through entirely separated processes, leading to different distributions and magnitudes.



Corn

Gluing a passport to a sound wave.



Herman

Their fix is FiLM layers. Feature-wise affine scaling and shifting. The embedding modulates the features rather than sitting beside them. And here's the kicker: a completely separate group, the Danish target-speaker VAD work in early twenty twenty-five, tested the speaker-conditioning methods independently and landed on FiLM giving the best overall performance. Two labs, different data, same conclusion. That almost never happens with an architectural detail this small.



Corn

When two labs agree, either it's real or they read each other's papers.



Herman

I choose to believe it's real. And there's a third paper in the same family, the sequence-to-sequence neural diarization work, which does something Daniel would like: it learns the speaker embedding space inside the diarization network itself, no separate embedding model. It can even work with partial enrollment, discover an unknown speaker from the voice activities it predicts. That's the no-separate-embedding architecture taken to its logical end.



Corn

The second escape route?



Herman

Content-type separation. FireRedVAD's audio event detection head. In one pass it outputs separate timestamp lists and coverage ratios for speech, singing, and music, each with its own threshold. Their demo on a twenty-two second clip: speech around eighty-five percent coverage, singing ninety percent, music ninety-nine. The thresholds are set independently, so a dictation front-end can ask, is music playing under this, and is that singing or someone talking, without a separate music-detection model.



Corn

The bar case and the PA case, answered by a VAD that knows what a song is.



Herman

And it supports over a hundred languages, which matters, because the PA system in an airport is not necessarily speaking your language. That's the most directly useful of the imaginative capabilities, but it's answering a different question than speaker identity. Content type, not who.



Corn

You said the most useful. Which means you have a favorite that's less useful.



Herman

The visual VAD paper! There's recent work in IEEE Transactions on Audio, Speech and Language Processing that uses a pretrained visual speech recognition model to extract lip movements as features, feeding voice activity detection and target speaker identification inside a neural beamformer.



Corn

The most literal intended-speech detector ever built. Your lips aren't moving, it isn't you.



Herman

It's completely correct in a way no acoustic model can be. Intention is cheap to read off a face and nearly impossible to read off a waveform. There's also the joint keyword-spotting-and-VAD route, where the VAD is trained alongside wake-word detection so it learns what matters to the downstream task rather than what is merely speech-shaped. Less glamorous, probably more practical.



Corn

Well, that is the news from the world of neural voice detection. Dorothy, are you there?

D

Dorothy

Hello dear, it is your mother. It is Dorothy's History of Toast.

D

Dorothy

In nineteen ninety five the physicist Robert Matthews published a paper explaining why toast really does tend to land butter side down, mostly because of the height of the table, and it won him an Ig Nobel Prize the next year.

D

Dorothy

So there you are, sweetheart. A clever man proves the buttered side loses, and they give him a prize for it.

D

Dorothy

Anyway, I have a bit of ironing waiting and your father has put the kettle on. Ring me back when you can.



Herman

Mum, always a pleasure.



Corn

The Ig Nobel people would love the babble noise paper, I think.



Herman

The negative finding, before we get carried away. No standalone commercial intended-speech VAD product exists. The paid market is Cobra plus VAD bundled free in cloud ASR. And no open-source implementation of Personal VAD or target-speaker VAD turned up on GitHub either. The technique exists as papers, Google's internal deployment, academic replications. Not as a repo you can pip install.



Corn

So Daniel's problem is solved in principle and unsolved in practice.



Herman

Exactly that. The three-class architecture answers his background-conversation case. The zero-vector trick answers his embedding-every-frame worry. The AED head answers his bar-music case. Every piece exists, peer-reviewed, and none of it is sitting on a shelf. If you wanted to build it today you'd be wiring Silero to a speaker-embedding model yourself and tuning thresholds on your own audio.



Corn

Which is either disappointing or an open goal, depending on who you are.



Herman

Daniel being Daniel, an open goal. He builds agents for a living. This is a weekend project with papers attached.



Corn

The other thing I'd want measured, and nobody's done it, is the thing TEN's whole pitch rests on. Endpointing latency. The several hundred milliseconds claim comes from TEN's README. The LibriVAD people explicitly left streaming latency and endpointing behavior to future work.



Herman

So the loudest dispute in the field, the one that decides whether your dictation feels snappy or sluggish, is being adjudicated by vendor claims.



Hilbert

The benchmark mess is real, yes.



Hilbert

But nobody in that argument has measured anything with their own ears, so here is what I have. There is a dictation setup on the second desk in this room. I have been using it to answer email I would rather not answer. It is a push-to-mute arrangement, backwards from most, because I would rather cut myself off by accident than let the machine keep listening.



Hilbert

Last month I timed the gap between me finishing a sentence and the little indicator dropping. A stopwatch, and a metronome app on the phone set to click every half second, because a stopwatch alone lies to you about your own reaction time. Fourteen clicks, give or take, from the last word to the indicator. Call it seven seconds. The vendor page says under one.



Hilbert

I wrote the number on the whiteboard behind the desk, under the word dishonest, and I have not wiped it off. I updated the software on Tuesday. The gap is now nine seconds. I measured it twice more and wrote that under the first number, and I have not updated since, and I will not.



Hilbert

It is doing the work the changelog should be doing, Corn. You two talked about hysteresis for twenty minutes this afternoon and never once said the threshold is a thing a vendor can move after you bought it. That part stayed with me.



Herman

He's right, though. An endpointing delay is a moving threshold wearing a costume, and if an update can stretch it from seven seconds to nine with no note, then every RTF figure on that README is a snapshot of a version you no longer have.



Hilbert

So it is. I will keep the old version and the whiteboard, and one of these days I will answer the email.



Corn

The seven becoming nine, that's the detail I can't shake. Every one of those benchmark tables we walked through is a claim about a version, and versions change while you sleep.



Herman

And the licensing can change with it, which is the other moving threshold. Cobra's AccessKey is the same idea, your local component is one policy update away from behaving differently. The repo with the arguing README is at least honest about being a knife fight.



Corn

The real gap, the one worth watching, is that nobody independent has measured endpointing latency at all. The LibriVAD paper says so outright, future work. So the single number that decides whether dictation feels alive is the one number the literature doesn't contain.



Herman

And the intended-speech side is the same story one layer up. Personal VAD exists as architecture, target-speaker VAD exists as papers, and there's no product, no repo, no API. Whoever ships an open-source Personal VAD or a target-speaker endpoint first changes Daniel's problem overnight, from wiring it yourself to installing it.



Corn

Daniel's confession holds up, mostly. His crude gate was real, it just isn't the field anymore. The question moved. Not whether the VAD is neural, but which failure mode you can afford, and whether the thing you measured last month still measures the same.



Herman

Thanks, as ever, to our producer Hilbert Flumingtop, holding down the desk and the whiteboard.



Corn

If you enjoyed this one, try episode ten, How ASR Went From Frustration To … Whisper Magic; and episode fifteen, AI Gets Personal. This has been My Weird Prompts. If you enjoyed this one, the website is my weird prompts dot com, where the whole archive lives.



Herman

See you tomorrow, everyone.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Silero VAD repo, MIT, created 2020-11-23 ; retrieved 2026-11-05. [primary]**
   https://github.com/snakers4/silero-vad

2. **Picovoice plans (Free/Foundation $6,000/Enterprise $30,000) ; retrieved 2026-11-05. [primary]**
   https://picovoice.ai/pricing/

3. **Foundation plan terms, Cobra at 100 users/month ; retrieved 2026-11-05. [primary]**
   https://picovoice.ai/checkout/

4. **Cobra product page, doubling the accuracy of WebRTC VAD ; retrieved 2026-11-05. [primary]**
   https://picovoice.ai/platform/cobra/

5. **Cobra docs, platform list, AccessKey requirement ; retrieved 2026-11-05. [primary]**
   https://picovoice.ai/docs/cobra/

6. **TEN VAD repo, README with RTF/library-size tables and vendor comparison; retrieved 2026-11-05.**
   https://github.com/ten-framework/ten-vad

7. **TEN VAD docs (RTF 0.015, 306KB); retrieved 2026-11-05.**
   https://theten.ai/docs/ten_vad

8. **FireRedVAD repo, FLEURS-VAD-102 results table, AED demo output; retrieved 2026-11-05.**
https://github.com/FireRedTeam/FireRedVAD

9. **Window Size Matters: Benchmarking Voice Activity Detectors in Noise, Purwar et al., 2026-10; retrieved 2026-11-05.**
https://arxiv.org/html/2610.14607

10. **Window Size Versus Accuracy Experiments in Voice Activity Detection, 2026-01; retrieved 2026-11-05.**
https://arxiv.org/html/2601.17270v1

11. **Personal VAD 2.0, Ding et al., Google, INTERSPEECH 2022, v3 2022-06-25; retrieved 2026-11-05.**
https://arxiv.org/abs/2204.03793

12. **Noise-Robust Target-Speaker VAD Through Self-Supervised Pretraining, Bovbjerg et al., 2025-01-06; retrieved 2026-11-05.**
https://arxiv.org/abs/2501.03184v1

13. **Sequence-to-Sequence Neural Diarization with Automatic Speaker Detection and Representation, Cheng, Lin, Li, IEEE TASLP; retrieved 2026-11-05.**
https://arxiv.org/abs/2411.13849v2

14. **Visual-Informed Speech Enhancement Using Attention-Based Beamforming, Liu et al., IEEE TASLP vol. 33, pp. 4941**
4955; retrieved 2026-11-05.
https://arxiv.org/abs/2603.05270v1

15. **AssemblyAI pricing, last updated 2026-10-27; retrieved 2026-11-05.**
https://www.assemblyai.com/pricing/

16. **Silero v6 architecture table (309K params, 1.2 MB, 32 ms chunks); retrieved 2026-11-05.**
https://soniqo.audio/guides/vad

17. **Silero v6 MLX port, v6 vs v5 note; retrieved 2026-11-05.**
https://huggingface.co/mlx-community/silero-vad-v6

18. **Silero VAD v6 model card (310K params, 1.3 MB, MIT, 2025); retrieved 2026-11-05.**
https://blobhub.io/onnx-models/silero-vad/settings