Episode Book

EPISODE #6341

Why Fast Speech Detectors Still Struggle with Onsets

PUBLISHED	29 October 2026 at 11:23 UTC
DURATION	23:39
LISTEN	myweirdprompts.com/episode/voice-activity-detection-latency-61923/
HOSTS	Corn, Herman, Desmond, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.25
SOURCES	23 cited by the research (listed at the end)
BOOK GENERATED	30 October 2026 at 07:19 UTC

SYNOPSIS

Voice activity detection models operate in milliseconds on a CPU, yet real-time conversational audio remains vulnerable to clipped speech onsets. Human hearing detects delays ranging from half a millisecond to tens of milliseconds, complicating systems like live voice changers. Frame-level benchmark scores mask onset jitter, while models like Silero, S4VAD, and FireRedVAD grapple with dataset annotation biases up to 136 milliseconds. Buffering strategies and state-space architectures explore where detection ends and perceptual latency begins.

TRANSCRIPT



Corn

Voice activity detection usually gets treated as the boring switch before the impressive AI. Daniel puts it the other way round. The tiny CPU model that notices speech before a word has properly arrived may be doing the most impressive real-time job of all. He wants to connect that to live face avatars and generative voice changing, where the system has to finish its work before we notice the delay.



Herman

The switch gets a remarkably demanding job description. Notice the start, ignore the fan, don't mistake a plate hitting the table for somebody addressing you, and please don't eat the first consonant.



Corn

The consonant is rarely covered by the warranty.



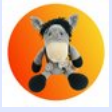
Herman

I'm Herman Poppleberry, and the first distinction matters immediately. Processing faster than real time means you can keep up with incoming material. It doesn't automatically mean the output feels instantaneous. You might process an enormous recording very quickly while still making a live speaker wait for the next chunk.



Corn

Daniel's angle is the space between zero delay and detectable delay. The awkward part is that the space changes depending on what we're doing. My tolerance for waiting is excellent. My tolerance for hearing my own voice arrive late is considerably less spiritual.



Herman

Corn, you will defend a slow computer until it interrupts your sentence.



Corn

Waiting is civilised. Losing the beginning of my objection isn't.



Herman

Human hearing can detect astonishingly small timing changes. The Aalto and Genelec experiment used twelve listeners and found sensitivity down to half a millisecond at certain frequencies. Starts and stops made the delays especially audible.



Corn

Half a millisecond makes every software progress bar look criminal.



Herman

The experiment used frequency-selective time-reverse filtering to create negative delay. Juho Liski described it with the line, To achieve a negative delay, we need to move into the future. With an existing recording, software can rearrange timing in ways a live microphone cannot. That doesn't give your voice changer knowledge of your next syllable.



Corn

Nor does it mean a person notices every half-millisecond pause before a reply. We're comparing altered frequency timing inside a sound, rather than waiting for somebody to answer.



Herman

A different experiment gives a different threshold. Minimum detectable echo delay came out around twenty milliseconds for music and twenty-three for speech. Now you've got an original sound and a delayed version, rather than parts of one sound arriving out of alignment.



Corn

My own voice supplies the original whether the software wants it or not. I hear myself through my head, then the headphones deliver their opinion.



Herman

Self-monitoring is especially unforgiving for precisely that reason. Action-to-sound experiments give another spread. One measured a just-noticeable difference of forty-nine milliseconds starting from effectively zero latency.



Corn

Meaning the smallest extra delay people reliably noticed in that particular task. Not a universal forty-nine-millisecond allowance for everything with a microphone.



Herman

At a sixty-four-millisecond starting delay, the additional detectable difference was twenty-seven milliseconds. With the starting delay pushed to five hundred and twelve milliseconds, it rose to seventy-seven.



Corn

Hm.



Corn

The brain doesn't issue a single deadline. It judges a relationship. A delayed echo, a key press, a lip movement, and an answer to a question all offer different evidence.



Herman

Would you accept a converted voice arriving a little late if the person on the other end couldn't tell?



Corn

For them, yes. For me, only if you stop piping it straight back into my ears. The audience and performer have different clocks.



Herman

That distinction will haunt every latency claim we mention.



Corn

Before we get further into perception thresholds, our resident finance expert has his regular segment.

D

Desmond

Desmond's Money Tip. Split one paper napkin into its two layers and call it two napkins. That is a saving of one napkin per meal, which over a year of three meals a day is one thousand and ninety five napkins, a sum I value at roughly two pounds and eleven pence.

D

Desmond

I keep a running total on the back of a napkin, which does mean I have to spend one to record it. I do not count that one. It would ruin the figures.

D

Desmond

Money is mostly just the slow accumulation of things you did not buy. It is not uplifting, but it is reliable.



Corn

Tempting. Herman looks appalled, so I'll leave the napkin intact. The VAD model has a simpler audience problem, though. Nobody needs to listen to its output.



Herman

Its output is a speech probability, usually turned into a yes or no decision. That's vastly less work than generating the speech itself. Silero's JIT model is around two megabytes. On a single CPU thread, processing a chunk containing more than thirty milliseconds of audio takes under one millisecond.



Corn

Thirty milliseconds arrive, one millisecond works, twenty-nine milliseconds look available. Except you had to wait for the thirty to arrive.



Herman

Thank you, because that's where the phrase faster than real time can mislead. Roughly thirtyfold processing headroom gives you room for buffering, overlap, other software, and sometimes look-ahead. It cannot erase the time spent collecting evidence. A future sample must still happen before a live system can use it.



Corn

The model cannot spend spare CPU cycles purchasing tomorrow.



Herman

Daniel mentions breaths, lip clicks, and the beginning of voicing. Those physical events can precede a recognisable word. I'm less certain about claiming that a particular VAD explicitly uses all of them as advance warning. A breath can mean speech is coming, or simply that the speaker intends to remain alive.



Corn

A persistent false positive in this household.



Herman

The onset of voicing is already speech evidence. Recognising a complete word takes longer. A detector can therefore fire before you know what was said without predicting anything at all.



Corn

I would actually be suspicious of a detector that opened every time I inhaled. It could feel wonderfully responsive during a demonstration and spend the rest of the day transmitting nose sounds.



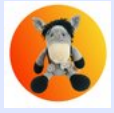
Herman

What would you choose, Corn? Occasional clipped beginnings, or the channel opening too readily?



Corn

Keep a short rolling buffer and send the beginning retrospectively when you're confident. I'd rather preserve the consonant than pretend the decision happened before it did.



Herman

That's a common practical solution. The buffer retains recent audio while the detector decides. Detection and preservation become separate jobs. You may begin sending slightly later, but the recogniser still receives the actual beginning.



Corn

Which means the user can experience complete speech capture without the classifier being clairvoyant. It also means an onset score alone won't tell us whether the product works well.



Herman

There is a fascinating architectural knob in S4VAD. Its state-space model carries acoustic information forward through time. A parameter called delta controls how quickly that memory decays. Larger values make it more local and responsive. Smaller values preserve



Herman (continued)

context, but evidence of speech can linger into silence.



Corn

You are about to fall in love with an adjustable memory.



Herman

I've already done it. However, this one has a reason to exist beyond having a beautiful equation.



Corn

That narrows your usual selection process.



Herman

Fair. Keeping context helps distinguish meaningful speech from a tiny ambiguous noise. Reacting locally helps catch a sudden start. You can't maximise both just by making the model smaller.



Corn

There's also stopping. Daniel's entry point is the start of speech, but a conversation system has to decide whether my silence means I'm done or whether I'm thinking. You tend to treat both as permission.



Herman

Your pauses contain several possible endings.



Corn

They're still mine.



Herman

Endpointing adds another policy layer. VAD continuously estimates whether speech is present, not merely the first onset. The application decides how much silence counts as the end of a turn. Interruptions need yet another decision. Someone saying a brief acknowledgment shouldn't necessarily terminate the assistant's entire answer.



Corn

That is where a very fast wrong decision becomes more annoying than a slightly slower correct one.



Herman

Frame-level accuracy doesn't settle it. FireRedVAD uses about six hundred thousand parameters and reports ninety-seven point five seven percent frame-level F1 on FLEURS-VAD.



Corn

A very small model with a very large mark.



Herman

Its reported area under the receiver operating curve is ninety-nine point six percent. It beats several familiar alternatives on that benchmark, including Silero and WebRTC VAD. Yet neither headline score tells you much about exactly when the gate opens.



Corn

A long sentence contains plenty of easy middle. You could classify all that correctly and still damage the beginning of every sentence. The aggregate looks healthy because the part I care about is brief.



Herman

The timing labels themselves make it worse. S4VAD's evaluation found onset annotation jitter on the order of tens to hundreds of milliseconds. Its estimated dataset shift for LibriSpeech was minus seventy milliseconds, meaning the labels tended to put the start too early.



Corn

You're being marked late against somebody who started the stopwatch early.



Herman

For AVA-Speech, the estimated shift was minus one hundred and thirty-six milliseconds. These are estimated dataset biases, not corrections you should blindly apply to every recording. Individual boundaries still vary.



Corn

A twenty-millisecond latency claim measured against that needs a very careful explanation. You can print more decimal places than your labels deserve.



Herman

Speech boundaries aren't always obvious even to a careful listener. Does the start include an unvoiced consonant? A preparatory breath? Quiet speech under background noise? The annotation policy needs to say, and the evaluation needs to respect it.



Corn

I hadn't expected the expensive problem to be agreeing when somebody began talking.



Herman

The tiny gate sits ahead of speech recognition, speaker diarization, enhancement, and communication systems. If it discards the beginning, every expensive model downstream inherits the loss. If it opens constantly, the battery and inference bill inherit the noise.



Corn

Cheap enough to run everywhere. Important enough to spoil everything.



Herman

Conversion has no comparable escape into a binary answer. It has to produce an entire waveform. The old chipmunk effect changes pitch, often shifting the apparent vocal character along with it. Generative conversion retains the spoken content while synthesising a different voice.



Corn

I say the words, somebody else's throat gets the credit.



Herman

Conceptually, yes. Though it needn't literally transcribe text and then speak that text. Some systems use continuous acoustic representations or neural codecs. A text bottleneck would discard details they may want to preserve, including timing, laughter, and parts of the performance.



Corn

Then content extraction sounds tidier than it really is. My identity isn't stored in one removable layer while my sarcasm sits safely underneath.



Herman

TVTSyn points to that representational mismatch. Content changes over time, while many architectures inject speaker identity as one static global embedding. But voice identity interacts with how each sound is produced. Maintaining articulation while changing vocal character is a moving problem.



Corn

A pitch slider has the considerable advantage of admitting what it does.



Herman

Current reported latency is impressive. Paravo advertises as low as sixty milliseconds on CPU, with more than a hundred voices and over eighty thousand users.



Corn

As low as means there's a favourable setup behind it.



Herman

Hardware, configuration, and measurement endpoints matter. SynthVC reports seventy-seven point one milliseconds end to end. MeanVC Two cuts its reported latency from two hundred and eleven to one hundred and ten milliseconds.



Corn

Almost half the wait removed. That ought to change how speaking into it feels, not merely improve a benchmark table.



Herman

It uses future-receptive chunking with forty-millisecond chunks. The challenge is retaining enough continuity to avoid rough joins without accumulating a long queue of future audio.



Corn

Wait.



Corn

Future-receptive still means waiting for actual speech when the future contains speech.



Herman

It does. There's no exemption from causality in the title. More efficient chunk handling reduces the wait; it doesn't predict an unconstrained speaker perfectly. StyleStream's first version, despite being streamable, sat around a second end to end on consumer hardware.



Corn

Streamable. Bring your own patience.



Herman

A system can produce output continuously while remaining a second behind. Throughput tells you whether the queue grows. Latency tells you how far behind you're already standing.



Corn

Those sixty to one hundred and ten millisecond figures are also above several of the auditory thresholds we discussed. We shouldn't call the delay imperceptible just because the voice arrives continuously.



Herman

Who hears the result matters. A remote listener lacks your immediate original as a reference. You don't. Corn, would you prefer a cleaner voice at a higher delay, or a slightly rougher voice that follows you closely?



Corn

During a conversation, the faster one, within reason. An exquisite voice that keeps arriving after I've moved on feels like somebody else is controlling my mouth. For a recording, let it take longer.



Herman

That's why product settings can outperform general-purpose demonstrations on speed. A restricted set of voices, smaller models, fixed hardware assumptions, and a tightly controlled pipeline can all help. But sixty on CPU against seventy-seven in a paper isn't a fair race unless the work and timing boundaries match.



Corn

Your comparison table needs more columns than the sales department allows.



Herman

Video adds another reference. You see your own gesture and the transformed gesture. Lucy Two point five advertises under forty milliseconds glass to glass at thirty frames per second.



Corn

Glass to glass at least names two ends. Camera capture to displayed output, rather than the pleasant little interval when the GPU is busy.



Herman

At that frame rate, another frame arrives roughly every thirty-three milliseconds. Meeting that production rate doesn't prove every frame has low latency. You can keep a smooth output stream while a queue quietly puts the whole thing behind.



Corn

A polished face from the recent past.



Herman

Camera-driven face transfer and audio-driven talking heads aren't identical workloads, either. One follows observed expressions. The other generates facial motion from sound. Both have deadlines, but the available evidence differs.



Corn

Audio-driven faces also have to know when to stop moving. A face that continues speaking while I'm listening makes a more serious timing error than a slightly late blink.



Herman

EmbodiedHead uses a streaming audio scheduler to suppress unwanted mouth motion during listening, and generation can take as few as four sampling steps. FaceGAN takes a different route, arguing that audio-conditioned facial motion is constrained enough for a single forward pass per frame.



Corn

You have found another elegant mechanism.



Herman

This one earns it. Diffusion repeatedly evaluates a network to refine a sample. Reducing that repetition helps streaming. FaceGAN also shapes synthetic noise using its future values. Since the system creates that noise, it can sample future noise now without waiting for future speech.



Corn

Inventing tomorrow is easier when tomorrow is your own noise.



Herman

The audio-to-expression path stays causal. The generated noise can provide smooth temporal structure without smuggling in unavailable audio. That's a very specific trick, not a general claim that the model knows the speaker's future.



Corn

A face render quoted at a hundred and thirty-four to a hundred and fifty milliseconds is still only one stage. Add speech generation, capture, transport, playback… no, even before that, decide whether we're timing a gesture or a conversational reply.



Herman

A natural turn gap might be around a fifth of a second, but that is a different perceptual situation again. You can tolerate some response planning while finding a visibly detached mouth immediately distracting.



Corn

The defensive side gets an uglier deadline. A voice-conversion detector working on one-second segments has to hear enough material to classify it. By then, the converted voice has already been talking.



Herman

Its inference might be fast while its evidence window is long. Same distinction again. For your DJ setup, Corn, would you even bother trying to reproduce that half-millisecond result at home? I'd call that impractical without carefully controlled listening conditions.



Hilbert

...with a recorded signal, yes, Corn. That's what I'm doing. Not live speech. The half-millisecond result concerns frequency regions relative to others. A whole recording arriving half a millisecond late doesn't recreate that comparison.



Herman

I blurred those two things just then. A controlled comparison is possible at home. Getting the same experiment is another matter.



Hilbert

My current arrangement is a mess. The laptop is on the floor because the short lead reaches there. I alternate processed recordings through the same headphones. I have been doing these comparisons for years, but this month I'm trying to sort out the switching. You were careful about capture and playback earlier, Herman. I'm trying to keep the switch from changing anything else.



Corn

Otherwise you end up identifying the switch rather than the delay.



Hilbert

Nissim at the audio shop keeps offering me another interface. I've dealt with him far too long to let him choose the subject. The old switch box is gone. I have its wiring diagram, folded inside an envelope marked unpleasant. I keep bringing that instead. It shows the connection I want without another discussion about recording vocals.



Herman

The diagram gives him something definite to wire. That's less ambiguous than asking for a box with no audible signature.



Hilbert

He says he can make the lead. I'm waiting. The surprise is the difference between the two routes I use now. My recording puts it at three milliseconds, before I add any deliberate filtering. I'm not claiming laboratory precision. It's enough to make me stop calling the routes identical. Nissim writes one hundred and eighty shekels on the envelope.



Corn

Three milliseconds before the intended experiment even begins. The unchanged route needs as much attention as the altered one.



Herman

That also sharpens the product comparison. If we want to know what the model added, we need the same pipeline without it. If we want to know what the user experiences, we need the whole route. Those are both useful numbers, but they answer different questions.



Corn

The distinction that stays with me is evidence time versus computation time. A model can finish in a millisecond after waiting much longer to learn what it needs.



Herman

Faster inference helps, but it cannot remove the observation window. VAD needs enough sound to decide. Conversion needs enough context to speak smoothly. An avatar needs enough information to move plausibly.



Corn

As those stages get connected, I'd like to see latency claims include the unchanged route and uncertainty at the boundaries. Otherwise we won't know whether the next improvement happened in the model or in the stopwatch.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If this was your kind of episode, go back for episode fifteen, AI Gets Personal; episode ten, How ASR Went From Frustration To … Whisper Magic; and episode five, Fine-Tuning ASR For Maximal Usability. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoyed it, leave us a review wherever you listen. It helps other curious people find us.



Corn

See you tomorrow.

SOURCES

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

1. **Silero VAD GitHub repo [primary]**
   v6.2.3, 2026-09-23, last push 2026-10-22
   https://github.com/snakers4/silero-vad

2. **Aalto University, The human ear detects half a millisecond delay in sound, 2021-06-10**
   https://www.aalto.fi/en/news/the-human-ear-detects-half-a-millisecond-delay-in-sound

3. **Schmid et al., Measuring the Just Noticeable Difference for Audio Latency, ACM Audio Mostly 2024, n=37**
   https://epub.uni-regensburg.de/59003/3/3678299.3678331.pdf

4. **The minimum detectable delay of speech and music, ICASSP 1984**
   https://doi.org/10.1109/icassp.1984.1172559

5. **Center for Hearing Access, Latency glossary**
   https://centerforhearingaccess.org/glossary/latency/

6. **Cohen, Netzer, Habi, Low-Latency State Space Voice Activity Detection with Robust Onset Time Evaluation (S4VAD), arXiv:2609.11110, 2026-09-10**
   https://arxiv.org/abs/2609.11110

7. **Xu et al., FireRedASR2S (FireRedVAD, 0.6M params), arXiv:2603.10420, 2026-03-11**
   https://arxiv.org/abs/2603.10420

8. **Guo et al., SynthVC, arXiv:2510.09245, 2025-10-10**
   https://arxiv.org/abs/2510.09245

9. **Ma et al., MeanVC 2, arXiv:2606.09050, 2026-06-08**
   https://arxiv.org/abs/2606.09050

10. **Quamer et al., TVTSyn, arXiv:2602.09389, 2026-02-10**
    https://arxiv.org/abs/2602.09389

11. **Liu, Lee, Anumanchipalli, StyleStream 2.0, arXiv:2610.28940**
    https://arxiv.org/html/2610.28940v1

12. **Chinchmalatpure et al., Defense Against Synthetic Speech: Real-Time Detection of RVC Voice Conversion Attacks, arXiv:2601.04227, IJRAR vol. 12 no. 4, 2025**
    https://arxiv.org/abs/2601.04227

13. **Streamer Magazine, AI Voice Changer Paravo v3.0 Released**
    0.06s Latency on CPU, 2026-06-30
    https://alive-project.com/en/streamer-magazine/article/17782/

14. **Han, Markovic, Richard et al., No Distillation Needed: Single-Pass Real-Time Talking Heads via Acausal Noise Shaping (FaceGAN), arXiv:2610.11070, 2026-10-08**
    https://arxiv.org/abs/2610.11070

15. **Zhang, Shen, Li, EmbodiedHead, arXiv:2604.17211v2, ECCV 2026**
    https://arxiv.org/abs/2604.17211v2

16. **LiveAvatar: Real-time 3D Gaussian Head Avatars from Single Images, ACM IMX 2026**
https://dl.acm.org/doi/full/10.1145/3788851.3815056

17. **Hugging Face blog, Lucy 2.5 Real-Time AI Video Editing: Decart's 40ms Live...**
https://huggingface.co/blog/ResterChed/lucy-2-5-real-time-ai-video

18. **livefaceswap.ai, What Is Latency in Real-Time Face Swap? (G2G definition), 2026-09-17**
https://livefaceswap.ai/what-is-latency-in-real-time-face-swap

19. **PRNewswire, ShengShu Technology Unveils Vidu S2, 2026-09-16**
https://www.prnewswire.com/news-releases/shengshu-technology-unveils-vidu-s2-bringing-real-time-interaction-and-editing-to-ai-video-302879976.html

20. **Forasoft, Real-Time Avatars + Lip-Sync In A Video Call, Oct 2026**
https://www.forasoft.com/learn/ai-for-video-engineering/articles-ai/real-time-avatar-lipsync-in-call

21. **clonemyvoice.io, Live Voice Cloning Latency: Seed-VC vs RVC v2 in 2026**
https://clonemyvoice.io/blog/live-voice-cloning-latency-seed-vc-vs-rvc-v2-in-2026.php

22. **VoxBooster, Best AI Voice Changer 2026**
https://voxbooster.com/blog/best-ai-voice-changer-2026/

23. **Wikipedia, Voice activity detection, last edited 2025-09-08**
https://en.wikipedia.org/wiki/Voice_activity_detection