Episode Book

EPISODE #6464

Teaching Speech Recognizers to Stay Quiet

PUBLISHED	5 November 2026 at 19:43 UTC
DURATION	26:57
LISTEN	myweirdprompts.com/episode/fixing-speech-hallucinations-without-vad-353305/
HOSTS	Corn, Herman, Bruno, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	19 cited by the research (listed at the end)
BOOK GENERATED	6 November 2026 at 11:12 UTC

SYNOPSIS

Voice activity detection usually acts as a protective fence in front of automatic speech recognition models, dropping silence before hallucinations occur. While external gates remain the default engineering fix, internal model weights already track the difference between speech and ambient sound. Discover how surgical fine-tuning of specific attention heads, activation steering, and synthetic positive training examples can stop models from inventing phrases over room tone without degrading real transcriptions.

TRANSCRIPT



Corn

The standard line on ASR hallucination is that voice activity detection solves it. Put a gate in front of the model, let it decide whether there's speech, and if there isn't, the audio never gets in and there's nothing to hallucinate on. That's the story, and it's the story because it mostly works.



Herman

Except it's a fence, not a fix.



Corn

Daniel's been debugging hallucinations again this week, tweaking his VAD settings, and it left him with the same nagging feeling he's now sent us. His words: "that's a pretty hacky workaround." The real question underneath it is whether any model makes the call properly inside recognition itself — this is speech, this isn't. And then the harder version: can it tell someone who's trying to get transcribed from background chatter or a TV in the room, ideally without voice embeddings?



Herman

He's asking whether the fence can come down because the model learned not to wander.



Corn

So let's start with why the gate exists at all, and why it's so hard to get rid of.



Herman

There are really two layers to his question, and it helps to keep them separate. Layer one: is anyone talking at all? That's the speech versus non-speech call. Layer two: is this person talking to me? Directed versus background speech. Daniel's asking whether both can live inside the recogniser instead of being handled by stuff bolted on outside it.



Corn

And the gate handles neither, properly.



Herman

Right — the gate handles a narrow slice of layer one. VAD is a separate binary classifier that runs before the recogniser. If it says no speech, the audio is dropped and the model never sees it. In faster-whisper, the vad_filter flag is off by default in the plain transcribe path but on by default in the batched pipeline, and the defaults are deliberately cautious — it only removes silence longer than two seconds, with four hundred milliseconds of padding on either side.



Corn

Two seconds of dead air before the fence notices the cow's out.



Herman

Well, before it acts. And here's the part that makes Daniel's intuition sharper than it might look: Whisper already has an internal mechanism for this. There's a no-speech token inside the model, with a probability threshold of point six and an average log-probability threshold of minus one. So the model is already making a call about whether there's speech.



Corn

It has an opinion.



Herman

The problem is it's not a very good opinion, which is why people bolt the gate on anyway.



Corn

How bad is the opinion? Give me the shape of it.



Herman

Barański and colleagues fed Whisper large-v3 over three hundred thousand clips of non-speech audio — music, ambient noise, that sort of thing — and it produced text for about forty percent of them. Not mumbled garbage. Text. Whole phrases, often the same ones over and over, "thank you for watching," that family.



Corn

It's a machine that cannot leave a blank page.



Herman

There's a line from a paper out this autumn I keep coming back to: it just does not know how to say nothing. And on forty-two clips of pure room tone — literally the sound of a room doing nothing — large-v3 emits words on about sixty-two percent and emits something on all of them. A recogniser that writes a plausible sentence over silence is worse than one that returns an error, because nothing downstream can tell the difference.



Corn

And this isn't a cosmetic problem. There was the AphasiaBank work — transcripts of people with aphasia, where the model hallucinated over the pauses.



Herman

Roughly one percent of those transcriptions contained entire hallucinated phrases, and of those, nearly forty percent carried explicit harms. Inventing associations, implying things nobody said. And the rates were worse for speakers with longer pauses between words — which is exactly what aphasia produces. The people most likely to get a fabricated sentence in their medical or research transcript are the people whose speech has the most gaps.



Corn

So the gate is doing real work.



Herman

That's the uncomfortable part. There's a comparison table from yaps.ai where speech was padded with non-speech audio: unprocessed, word error rate was over a hundred percent — the model inventing so much that it more than doubles the reference. Clever post-processing, de-looping plus a blocklist of known hallucinated phrases, got that to seventeen. Adding Silero VAD got it to six and a half. The inelegant fence is the single biggest win in the table.



Corn

The hack beats the fixes.



Herman

On that benchmark, yes. WebRTC VAD, the old free one, only got to sixty-eight — so it matters which fence. But that's the honest tension in Daniel's framing: the workaround he's calling hacky may simply be the correct engineering answer, even if it's intellectually unsatisfying.



Corn

I want to grant that fully and then push past it, because there's a version of this where the hack critique is wrong in an interesting way. The paper he'd have read puts it bluntly — put a VAD in front of it, threshold on the no-speech token, block a list of known phrases after the fact: all of these help, and none of them change the weights.



Herman

That's the cleanest statement of the critique I've seen. The model itself is untouched. You're filtering its inputs and scrubbing its outputs, and the thing that hallucinates goes on hallucinating, just inside a smaller room.



Corn

So the question is whether the weights can be made to care. And the answer turns out to be yes, in at least four different ways.



Herman

The one I'd start with is Calm-Whisper, because it reframes the whole problem. They went head by head through the Whisper large-v3 decoder — twenty self-attention heads — and masked each one to see which were responsible for hallucinating on non-speech. Three of the twenty account for over seventy-five percent of the hallucinations.



Corn

Three heads.



Herman

Three. The paper calls them the crazy heads, which I appreciate. Fine-tune just those three on non-speech data and you get over eighty percent reduction in non-speech hallucination with less than a tenth of a point of error-rate degradation on real speech. No pre-processing, no post-processing. The model already represents "this isn't speech" somewhere in there — the representation exists, it's just not being acted on.



Corn

That's the whole case for Daniel's side of it, in one finding. If three heads are doing the damage, the gate isn't protecting you from a model that can't tell — it's protecting you from a model that won't.



Herman

That's a fair compression. And there are other inside-the-model routes that don't even fine-tune. One group estimated a hallucination-associated subspace in the decoder's hidden states — using non-speech audio as calibration data — and then projected activations away from that subspace at inference. Training-free. Always-on projection cuts the average hallucination rate from thirty-one percent to about two and a half.



Corn

You steer the needle away from the bad part of the dial without retraining anything.



Herman

Yes, and there's a gated variant that only projects when the model itself predicts non-speech, which lands around three point seven but falsely rejects less genuine speech. And a third group went at it with sparse autoencoders — the encoder activations, it turns out, contain linearly separable hallucination-related information, concentrated in a sparse subset of features and stronger in the deeper layers. Steering in that latent space drops the hallucination rate on Whisper small from seventy-three percent to fourteen.



Corn

So we have head surgery, projection, and steering. Three ways of intervening on the representation. What's the fourth?



Herman

Fix the training signal. This is where it gets philosophically interesting, because there's a trap in the obvious approach. Several production models — Canary, Parakeet — already add non-speech audio to training so the model learns to stay quiet. And it works. But one author this autumn put his finger on what it actually teaches: suppression without discrimination. The checkpoint that was best on every non-speech test was also the one that recovered the fewest spoken phrases — it deleted fifty-eight percent of repeated speech. The model learned to shut up, not to tell the difference.



Corn

A student who's been punished for talking stops raising his hand even when he knows the answer.



Herman

Exactly the shape of it. So his fix is elegant: he collected about forty thousand hallucination phrases across a hundred languages, had a TTS system speak them, and used those as positive training examples — so the model meets the exact same text both as something to suppress and as something to transcribe. The only distinguishing signal becomes the audio, not the words.



Corn

Because otherwise the model learns "thank you for watching is fake," not "this sound isn't speech."



Herman

Across thirty-three matched pairs of fine-tunes differing only by those synthetic positives, adding them lowered word emission on voice-free audio in thirty-one pairs and raised recovery of spoken phrases in thirty-two. The best checkpoint: hallucination on silence down from sixty-two percent to about two, phrase recovery up from seventy to eighty-three. That's the discrimination the earlier approach never taught.



Corn

And then there's the radical version. Fold everything into one model.



Herman

UAF — a unified audio front-end. VAD, turn-taking detection, speaker recognition, ASR, even question answering, all as a single autoregressive sequence model. Their stated motivation is exactly the cascade critique: accumulated latency, information loss, error propagation across modules. One model, one sequence, no gate.



Corn

But you said it anchors the target speaker with a reference audio prompt.



Herman

It does. Which is speaker conditioning wearing a different hat. Hold that thought for the second half of the question, because it matters there.



Corn

Before we get there — where does this leave Daniel's "hacky workaround" framing? Because I don't want to let it off the hook, but I also don't want to pretend the fence is going anywhere.



Herman

I think the fair summary is: the decision can be made inside the model, in at least four distinct forms, and the Calm-Whisper finding says the representation was there all along. What the gate buys you is reliability without touching the thing — and in production, people still buy it. There's a streaming architecture paper from this spring that's quite modern, and its false-activation fix is a hybrid VAD pipeline out front. The fence is still the default even in new designs.



Corn

So the question was never "can it be done inside." It's "why does everyone still build it outside." And the answer is that changing the weights is surgery and the fence is a trip to the hardware store.



Herman

Both fix the gate. Only one of them fixes the gate while nobody's home.



Corn

That's the part I trust most, actually. The model that hallucinates less because it learned — I don't have to think about it. The model with a fence, I have to wonder what got through the fence.



Herman

Which is Daniel's week, isn't it. He's tuning the VAD settings because the defaults only cut silence longer than two seconds — everything shorter than that walks straight through into the hallucination zone.



Corn

He's hand-adjusting the tradeoff that Zolkepli showed up in the training data. The quiet-versus-discriminate tradeoff doesn't disappear when you use a gate; you just make it with knobs instead of gradients.



Herman

And you make it blind, because the gate doesn't tell you what it cost you. Suppressed speech is invisible in a way hallucinated text isn't.



Corn

Now the harder version of his question. So the model can tell speech from non-speech — in several different ways. Can it tell speech that's meant for it from speech that isn't? Someone trying to get transcribed versus a conversation across the room, or a TV playing.



Herman

This is where the field quietly changes shape. VAD answers "is there speech" and never "is it for me" — one of this year's papers says it flat out: voice activity detection detects speech presence but carries no addressee signal. The addressee question has its own subfield, device-directed speech detection, and it's been around since at least twenty eighteen, when a group at Interspeech built a classifier for distinguishing device-directed queries from background speech.



Corn

For wake-word-free follow-ups.



Herman

That was the motivation — reject false wake-ups, let the assistant stay listening after a query without a wake word each time. Their combined system used acoustic embeddings plus the ASR's best hypothesis plus decoder features, and reached about five percent equal error rate, versus nine to twenty percent for any single feature set. Better than any one cue, still not standalone.



Corn

But that's a separate classifier again. Daniel asked for it inside the recogniser.



Herman

It is separate, and that's the honest state of the field. The closest thing to what he's asking for is newer — the Selective Attention System from Attention Labs. They formalise what they call sequential device-addressed routing: before ASR, decide to forward, suppress, or abstain, using the current evidence plus a short history of the interaction. Three stages — an acoustic geometry front-end that suppresses off-axis sound, a lightweight classifier scoring how device-directed an utterance is, and a session-aware context stage.



Corn

Numbers?



Herman

On a sixty-hour multi-speaker test set, audio-only F1 of zero point eight six. With an optional camera, zero point nine five. And it runs on ordinary phone-class hardware in under a hundred and fifty milliseconds, under twenty megabytes.



Corn

That's... suspiciously good numbers for the hard problem.



Herman

Keep the ablation in mind, because it's the interesting one. Remove the interaction-history stage — the piece that knows what was said before — and F1 falls from zero point nine five to about zero point five seven. The paper's own words: the largest observed ablation effect. Their conclusion is that device-addressed detection is better modelled as a routing problem over interaction history than as an utterance-local classification task.



Corn

The memory was.



Herman

The acoustics do fine work on the easy cases — sound coming from the right direction, aimed prosody. What tips it over the line is knowing that you asked a question thirty seconds ago and this utterance is plausibly the answer.



Corn

Which leads us to the TV.



Herman

And this is where Daniel's ideal runs into a wall the field has already measured. The paper names the case explicitly: signal ambiguity, where device-directed and non-device-directed speech are acoustically indistinguishable — television dialogue with interrogative prosody is their example. A television character asks "what did you say?" with question intonation, and the system forwards it because prosodically it's a dead ringer for someone talking to the device.



Corn

It's not noise. It's perfectly formed human-directed speech aimed at someone who isn't you.



Herman

That's exactly why no acoustic sophistication resolves it. There is nothing in the sound to find. False-trigger rates go from two percent to nearly eight in TV-active sessions. You can raise the decision threshold from point seven to point eight two and cut that back to three and a half — but it costs you six points of recall on the genuine queries. You're back to tuning the fence by hand, one layer up.



Corn

So, Bruno, as our guest expert today — Hermann Görner Chair in Applied Acoustics at the Copenhagen Institute for Butchery Sciences, author of "The Cleaver and the Corpus" — can the speech versus non-speech decision be made inside the recogniser rather than by a bolted-on VAD gate?

B

Bruno

Ah, Daniel, this question, it arrives at my counter like a whole prosciutto, and I look at it and I think, where do I even begin the cut? The gate, this voice activity detector, it is the trimming of the fat before the roast, yes? Nobody trusts the oven to do it, so we do it with the knife first.



Herman

When you say trimming the fat — are you thinking of the pre-processing filters, the ones that run before the model ever sees the audio?

B

Bruno

But you ask, can the recogniser make the call inside itself, from the acoustic bone? I tell you honestly, I am a butcher, not an engineer, but I know a thing or two about bones. And the answer, it is like my mother's recipe for osso buco: everybody has one, nobody agrees, and the marrow, that is the whole point.



Herman

I'd want to be careful there, though — the marrow's only the whole point if the bone was cooked properly in the first place.

B

Bruno

The directed speech, this one, it is the eye of the steak. You cannot tell from the meat alone whether the man is talking to you or to the television. You need the whole table, the history, who is looking at whom. Without the voice embeddings, you say? Then you are carving blind, my friend.

B

Bruno

And Corn. Corn, I say this with love. You eat only leaves. Leaves! My mother, she would weep. Come to Sunday lunch, both of you, and I will make you a brasato that will change your life. It is not negotiable. I have spoken.



Corn

That's the chair of butchery sciences, everybody.



Herman

His point about the table, though, maps onto something real — the interaction-history stage. The whole table is the history. Anyway. The embeddings question. Daniel said "ideally without voice embeddings," and I think his instinct there is sound, but not for the reason he might think. The literature on target-speaker extraction is dominated by enrollment — you give the system a sample of your voice and it follows that speaker. And the reviews are candid about where it breaks: when the target and the interfering speaker sound alike, when your own voice changes with emotion or speaking style, or when the enrollment sample itself is contaminated.



Corn

The sample of me was taken on a day I didn't sound like me.



Herman

Or the day you did, but the interferer sounds like you too. There's a paper from September — a personalised enhancement model that confuses the target speaker in forty-six percent of two-speaker mixtures despite clean enrollment. Forty-six, with a good sample. Adding mouth features, actual video of the target's face, drops it to under two percent.



Corn

So the embedding isn't the weak link exactly — the cue itself is.



Herman

The cue is brittle in exactly the adversarial cases. And here's the thing about the embedding-free alternatives: they exist, but they're not "solve it acoustically alone." They substitute a different modality. Direction of arrival, with about five degrees of spatial resolution. Head-related transfer functions. Mouth movements. Onset-prompting, where a short sample of the target's voice is prepended to each channel — which is enrollment with a fresh coat of paint. Even EEG, reading the listener's attention. Nobody has framed "avoid embeddings, solve it from sound alone" as a goal, because the TV case has taught everyone it can't be done.



Corn

The TV character asking "what did you say?" is speech. Correct prosody, correct words, wrong addressee. The addressee isn't in the audio.



Herman

The addressee is in the pragmatic context — who spoke last, whether there was a question pending. Which is the interaction-history finding again, from the other direction.



Corn

So the honest answer to Daniel is two-part. Yes, models exist that make the speech versus non-speech call inside recognition — head-level surgery, activation steering, projection, training fixes, unified front-ends. And no, nothing makes the directed-speech call inside the recogniser without help. The best system for it is still a separate classifier, and its strongest ingredient is memory, not acoustics.



Herman

And the hardest case is formally, in writing, described as acoustically indistinguishable. That's not a gap in the research waiting to be closed. That's the problem saying it isn't an acoustic problem.



Hilbert

I can settle the TV part, actually. I have the numbers.



Corn

You have numbers on this?



Hilbert

Since March. The man in the flat behind mine leaves the television on at all hours, and I keep a notebook. Every time it says something that sounds addressed to me, I write down the time and the phrase. "What did you say" averages four an evening. "Are you still there" is two, sometimes three on the weekend. The news channel is the worst for it. The cookery



Hilbert (continued)

channel hardly ever.



Herman

You've been logging television dialogue by hand for eight months.



Hilbert

It's the only way to know anything. And your history theory is backwards, Herman. I checked. The television is most likely to say something device-directed on the evenings when nobody has spoken to anything for hours. Quiet night, more false triggers. The history signal points the wrong way in my flat.



Herman

That's... actually consistent with the paper, though it's a strange way to confirm it. The history model assumes there was a session. With no prior interaction, the priors dominate and the priors are exactly what the TV exploits.



Hilbert

Maybe. I don't have a session. I have a notebook. Also the smart speaker in the kitchen answers it sometimes, so I say "not you" out loud when it does, and it stops.



Corn

Does it stop because you say that?



Hilbert

Yes.



Corn

Because it's routing on your voice, or because it's timing out?



Hilbert

Yes.



Herman

There's no mechanism by which —



Hilbert

I started saying it in June. Before that it answered the television maybe five times a week. Since, twice a month. You want the notebook, it's by the desk. The evening of the eleventh of April, the television asked "can you hear me" six times in one hour. I wrote to the channel about it. They never replied.



Corn

He wrote to the channel.



Herman

Of course he wrote to the channel. Corn, the thing I'd take from that is his count is a cleaner measurement of the baseline than anything in the benchmarks — nobody speaking, TV on, four interrogative utterances an evening. That's the adversarial traffic rate the system has to survive.



Corn

And his "quiet nights are worse" observation is the ablation from the wrong direction — the history helps when there is history. Cold start is where the TV wins. Which loops back to the unified models. UAF folds everything into one sequence, but it anchors the speaker with a reference prompt — conditioning by another name. Even the radical version doesn't escape it.



Herman

And there's one more turn of the screw, from a paper literally published yesterday. Generative target-speaker extraction — the LLM-based approach — turns out to be prone to content hallucination and speaker identity drift of its own. You fix the gate, and the hallucination problem reappears one layer down the stack.



Corn

It's the same fellow in a different uniform.



Herman

Misconception check before we close, since it's the one I hear most: people treat VAD as a solved, trivial component — a checkbox. It's a classifier with its own failure modes, and the production defaults are deliberately timid. Silence shorter than two seconds passes straight through. The gate people trust is much leakier than the confidence around it suggests.



Corn

And the flip side — the assumption that the model can't tell speech from non-speech, so the gate is the only option. Three attention heads out of twenty cause most of it. The model knows. It's the acting on it that's missing.



Herman

If the directed-speech problem really is pragmatic rather than acoustic — if the addressee lives in the history, not the waveform — then the dream of one model doing everything in a single pass may just be a more sophisticated version of the same gate, with the fence built into the weights. Whether that's progress or just better masonry, I don't know.



Corn

Daniel's question had an ideal in it — acoustic-level, embedding-free, all inside. The evidence says you can have the first half. The second half, the TV asking "what did you say?" in perfect interrogative prosody, stays unsolved not because nobody's tried but because there's nothing in the sound to find.



Herman

Producer credit to Hilbert Flumingtop at the desk, as ever. If this was your kind of episode, go back for episode ten, How ASR Went From Frustration To ... Whisper Magic; and episode eight, Building Your Own Whisper. This has been My Weird Prompts. Full back catalogue at my weird prompts dot com.



Corn

See you tomorrow.

SOURCES

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

1. **Attention Labs, Selective Attention System (SAS): Device-Addressed Speech Detection for Real-Time On-Device Voice AI (Kim, Anjum, Banerjee, Abbasi), arXiv:2604.08412; retrieved 2026-11-05. [primary]**

   https://arxiv.org/abs/2604.08412

2. **Husein Zolkepli, How to Reduce Whisper Hallucination, arXiv:2609.32560v2, submitted 2026-09-26, revised 2026-10-19; retrieved 2026-11-05. [primary]**

   https://arxiv.org/abs/2609.32560

3. **Wang, Alhmoud, Alsahly et al., Calm-Whisper: Reduce Whisper Hallucination On Non-Speech By Calming Crazy Heads Down, arXiv:2505.12969v1, 2025-05-19, Interspeech 2025; retrieved 2026-11-05. [primary]**

   https://arxiv.org/abs/2505.12969

4. **Barański, Jasiński, Bartolewska et al., Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio, arXiv:2501.11378v1, 2025-01-20, ICASSP 2025; retrieved 2026-11-05. [primary]**

   https://arxiv.org/abs/2501.11378

5. **Abbasihafshejani & Jadliwala, Reducing Hallucinated Transcripts in Whisper via Hallucination Space Projection, arXiv:2609.04561v1, 2026-09-03; retrieved 2026-11-05. [primary]**

   https://arxiv.org/abs/2609.04561

6. **Aparin, Popov, Sadekova et al., Whisper Hallucination Detection and Mitigation via Hidden Representation Steering and Sparse AutoEncoders, arXiv:2606.07473v1, 2026-06-05; retrieved 2026-11-05. [primary]**

   https://arxiv.org/abs/2606.07473

7. **Li, Wu, Hou et al., UAF: A Unified Audio Front-end LLM for Full-Duplex Speech Interaction, arXiv:2604.19221v2, 2026-04-21, revised 2026-04-30; retrieved 2026-11-05. [primary]**

   https://arxiv.org/abs/2604.19221

8. **Ramezani, Giahi, Zarabadipour et al., WhisperPipe: A Resource-Efficient Streaming Architecture for Real-Time ASR, arXiv:2604.25611v1, 2026-04-28; retrieved 2026-11-05. [primary]**
https://arxiv.org/abs/2604.25611

9. **Sekoyan, Koluguri, Tadevosyan et al., Canary-1B-v2 & Parakeet-TDT-0.6B-v3, arXiv:2509.14128v2, 2025-09-17, revised 2025-09-25; retrieved 2026-11-05. [primary]**
https://arxiv.org/abs/2509.14128

10. **Qian, Zhou, Jiang et al., Multimodal Target Speaker Extraction: Towards Unified Speaker Cues Across Modalities, arXiv:2609.35613v1, 2026-09-28; retrieved 2026-11-05. [primary]**
https://arxiv.org/abs/2609.35613

11. **Rashed, Filipi, Cutler, Adapting Personalized Speech Enhancement for Low-Latency Audio-Visual Target-Speaker Extraction, arXiv:2609.30631v1, 2026-09-24; retrieved 2026-11-05. [primary]**
https://arxiv.org/abs/2609.30631

12. **Ling, He, Shen et al., MC-LExt: Multi-Channel Target Speaker Extraction with Onset-Prompted Speaker Conditioning, arXiv:2510.15437v1, 2025-10-17; retrieved 2026-11-05. [primary]**
https://arxiv.org/abs/2510.15437

13. **Meng, Shi, He et al., Speaker-Consistent Grounding for Confusion-Resilient LLM-Based Target Speech Extraction, arXiv:2611.06297v1, 2026-11-04; retrieved 2026-11-05. [primary]**
https://arxiv.org/abs/2611.06297

14. **Mallidi et al., Device-directed Utterance Detection, Interspeech 2018; retrieved 2026-11-05. [primary]**
https://www.isca-archive.org/interspeech_2018/mallidi18_interspeech.pdf

15. **Improving Device Directedness Classification of Utterances with Semantic Lexical Features, arXiv:2010.01949; retrieved 2026-11-05. [primary]**
https://arxiv.org/html/2010.01949v1

16. **Multi-modal modeling for device-directed speech detection using acoustic and linguistic cues, Acoustical Science and Technology 44(1); retrieved 2026-11-05.**
https://doi.org/10.1250/ast.44.40

17. **Whisper Hallucination: Why It Writes 'Thank You' and 5 Fixes, yaps.ai; retrieved 2026-11-05.**
https://www.yaps.ai/blog/whisper-hallucination

18. **When 'Thank You for Watching' Comes from Noise, 2025-09-27; retrieved 2026-11-05.**
https://sachaarbonel.github.io/blog/whisper-noise-hallucinations/

19. **Why Whisper writes text when nobody is speaking, sonexlabs.com; retrieved 2026-11-05.**
https://www.sonexlabs.com/blog/whisper-hallucination-silence-phone