Episode Book

EPISODE #6506

Why Generative Audio Restoration Hallucinates Highs

PUBLISHED	8 November 2026 at 22:39 UTC
DURATION	27:43
LISTEN	myweirdprompts.com/episode/generative-audio-restoration-models-53706/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	14 cited by the research (listed at the end)
BOOK GENERATED	9 November 2026 at 08:10 UTC

SYNOPSIS

Modern audio restoration tools like NVIDIA's RE-USE and Tokyo's Sidon ditch traditional filtering to synthesize clean speech representations from scratch. While they effortlessly erase background baby cries and air conditioner hums, their underlying mechanics reveal a hidden catch. Recordings from cheaper phone mics often yield better synthesis datasets than filtered headsets because hardware-limited microphones leave models with zero high-frequency evidence, forcing them to invent the consonants and vocal textures that define speaker identity.

TRANSCRIPT



Corn

Here's a question that sounds like a contradiction: can a recording be too clean to fix? Daniel ran two of the new audio-to-audio restoration models, Sidon from the Saruwatari lab in Tokyo and RE-USE from NVIDIA, over the same imperfect takes, and his phone recordings kept coming out better than the takes from his Jabra dictation headset, which is supposedly the good microphone.



Herman

The headset cuts everything above eight kilohertz in hardware. That's a phone-call sound, and the band is gone before any software ever sees it.



Corn

Right, and that's his angle. Daniel's got thousands of phone takes recorded for his speech-to-text model, air conditioner humming away, and now he wants to recycle them into a text-to-speech clone of his own voice, one whose whole job is pronouncing Hebrew words correctly inside English. Read-aloud, navigation, that kind of thing. Both models even removed a baby crying in the background of one clip, without the robotic tinge the older noise removers left behind. By ear he gives NVIDIA's the slight edge. What he's really asking is two things: what changed in this generation that lets it rebuild speech instead of subtracting noise, and can the output be trusted as training data when pronunciation and speaker identity have to survive the process.



Herman

And the fun part is that both papers were written for exactly his pipeline. Sidon's stated purpose is dataset cleansing for speech synthesis. This is one of those rare prompts where the primary sources and the questioner converge.



Corn

A man in Jerusalem asking the exact question a lab in Tokyo designed its model to answer. Saves us the usual extrapolation work.



Herman

Saves the extrapolation, not the skepticism, which is where the real episode lives. But the shift first, because it's new. Spectral subtraction, Wiener filtering, the classics, they all work the same way: figure out which frequency bins the noise owns, and turn those bins down. They can only remove energy. They cannot invent energy that was never recorded.



Corn

Which is why a narrowband recording stayed narrowband forever, no matter how much you denoised it.



Herman

And if you push the suppression hard, you get musical noise, those warbly tonal artifacts where the estimator flickers between bins. These new models are not in that business at all. Sidon is two hundred fifty million parameters, a feature predictor built on w2v-BERT 2.0, which was pretrained on about four and a half million hours across a hundred forty-three languages, plus a vocoder that outputs at forty-eight kilohertz. RE-USE is nine point six million parameters, a convolutional encoder and decoder with a thirty-layer bidirectional Mamba in between.



Corn

Nine point six million. The spec sheet reads like a typo next to Sidon's two-fifty.



Herman

Twenty-six times smaller, and it beats Sidon on most of the metrics that matter. That's the detail I keep re-reading. Two near-opposite bets, one enormous and multilingual, one tiny and language-agnostic, and both land in the same place, because both do the same fundamental thing: predict a clean internal representation of the speech and then synthesize new audio from it. Sidon goes through a vocoder. RE-USE goes through a generative stage stacked on a frozen regression model. Either way, the output is generated, not filtered, and that's what lets them add high-frequency detail that was never in the input.



Corn

And both name Daniel's use case, cleansing imperfect recordings to train synthesis models, as the design goal.



Herman

In writing. RE-USE's card says researchers can use it to enhance speech datasets for training speech generation models, text-to-speech and speech LLMs. Sidon ran the actual experiment, which we'll get to, because the experiment is the best evidence in this whole space.



Corn

Then let's open the hood. Sidon first, because I want to understand what "predict a clean internal representation" actually means in there.



Herman

The backbone is w2v-BERT 2.0, and here's the elegant part. During training they run clean speech through it and take the hidden state at layer eight, and that becomes the target. The model learns to map degraded speech to what layer eight would have looked like for the clean version. Then the vocoder turns those features back into audio at forty-eight kilohertz.



Corn

Why layer eight? These self-supervised models have dozens.



Herman

Because the layers have a job division. The early layers hold acoustic information, the raw sound, and the later layers drift toward semantics, meaning. Sidon's paper says it plainly: in speech restoration, keeping acoustic information such as speaker identity and prosody is also important, therefore layer eight. If they'd gone deeper they'd have preserved the words and melted the voice. For a clone corpus that choice is the whole ballgame.



Corn

The voice lives early, the content lives late. There's a joke about this show in there somewhere.



Herman

Leave it. So the feature predictor is a hundred ninety-eight million parameters, but they only fine-tune about five million of it, LoRA on the feed-forward output layers of the Conformer blocks. The vocoder is fifty-two point four million, and it upsamples fifty-hertz features to forty-eight-kilohertz audio, which is a nine hundred sixty times hop. Run it on one GPU and it processes audio up to five hundred times faster than real time. That's how you cleanse a corpus of thousands of hours without renting a datacenter for a year.



Corn

And RE-USE's bet on the same problem?



Herman

Completely different route, and the reasoning is the most interesting theory in the paper. They start from what they call the distortion-perception tradeoff. A regression model, trained to minimize error against the clean target, learns to output the conditional mean, the average of all the plausible clean signals given the noisy one. Under mild noise that average is fine. Under severe degradation the average collapses toward the prior mean, and if there are many plausible phonemes that could have produced what you heard, the average is a blend of all of them.



Corn

A blend of phonemes. That sounds horrifying.



Herman

It sounds like a robot. Muffled, watery, over-smoothed. It's the exact robotic tinge Daniel says the old deep-learning removers left behind, and this is the first time I've seen the mechanism written out so cleanly. The old denoisers weren't broken, they were doing optimal regression, and optimal regression under heavy degradation averages away the very details that make a voice sound like a person.



Corn

So the fix is to stop averaging.



Herman

The fix is two stages. Train a regression model to convergence, freeze it, so you keep the fidelity, and then train a generative model on top whose job is to restore perceptual detail, conditioned on the frozen model's output. You get the signal fidelity of regression with the texture of generation. Nine point six million parameters, thirty-layer bidirectional Mamba, runs on an A100 with Triton, and there's a community port that runs through audio.cpp, which is how it's pulling fifty-eight thousand downloads a month.



Corn

Daniel's ear gave that one the edge. The theory says his ear was measuring the generative stage doing its job.



Herman

That's my read. The absence of the robotic tinge isn't luck, it's the second stage. Now, the other thing that explains both of Daniel's surprises is the training data, because neither model was trained on noise alone. They were trained on a degradation taxonomy. Sidon starts with about twenty-two hundred hours across a hundred four languages and degrades it four times over with additive noise, reverberation, band limitation, clipping, MP3 codec damage, and packet loss, where nine percent of segments get a chunk between twenty and two hundred milliseconds zeroed out. That turns into roughly nine thousand hours of paired clean and noisy audio. RE-USE's list is nearly identical, plus low-quality microphones, and it accepts inputs from eight kilohertz up to forty-eight.



Corn

That list is the answer to the Jabra question before we even get there.



Herman

Bandwidth limitation is a named training condition, not an edge case. Randomly resampling to eight kilohertz and back is literally one of Sidon's augmentations. So when Daniel feeds it a headset recording that was band-limited in hardware, the model has seen ten thousand cousins of that exact input.



Corn

And the baby?



Herman

Suppression, not restoration, and the card is refreshingly blunt about it. For overlapping speech, this model is trained to preserve the main speaker while suppressing other speakers. The model did not learn that babies are noise. It learned that a second voice is not the target speaker. Which means it's speaker-selective, and that behavior cuts both ways: perfect for building a single-speaker clone corpus, and exactly wrong if you ever wanted that conversation preserved.



Corn

The baby got classified out of existence. Somewhere in Jerusalem there's a tape where Ezra has been deleted.



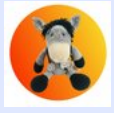
Herman

Daniel didn't say it was Ezra.



Corn

He didn't say it wasn't. Moving on, the Jabra thing still bothers me, because everything I know about buying microphones says the dictation headset should win.



Herman

Everything you know about buying microphones is tuned for recording, and this is a different sport. A phone microphone, even a cheap one, passes content well above eight kilohertz, maybe rolled off, maybe noisy, but the energy is real. The model has actual high-frequency evidence to condition on. The Jabra destroyed its band in hardware before the model sees a single sample, so the model has to invent the entire top octave from priors alone. It can do it, and it sounds plausible, but there's a word for audio that sounds plausible and carries no information about what actually happened.



Corn

Hallucinated.



Herman

Hallucinated. And sibilance, fricatives, the consonants that give a voice its edges, those live exactly in that band. The metrics back the story up, too. On RE-USE's sixteen-kilohertz subset, PESQ goes from one point two nine noisy to two point four seven restored, and STOI from about half to over eight tenths. On the forty-four point one subset, the same two point four seven. The model is most transformative precisely where the input is most impoverished.



Corn

Same score either way. The model doesn't get extra credit for having real evidence.



Herman

PESQ doesn't know what was recovered and what was invented. It measures plausibility. Daniel's phone takes came out better not because the phone is a better mic but because it's a better witness.



Corn

A worse listener, a better witness. I've worked with people like that. And we've got a voicemail from a listener. Bruce, go ahead.
Bruce: G'day Corn, g'day Herman, Bruce here, calling from the ute outside the hardware shop in Parkes, waiting on a bloke about a whipper snipper, so you've got me for a bit. Now I think I followed most of today's yarn about the noise cleaners that rebuild the voice instead of just chipping the hiss off, though I'll admit I heard Sidon and thought you were on about the harbour city for a good minute. Anyway, back in about two thousand and one I taped my old man telling stories on a little dictaphone and the tape's gone all muffled and warbly, and my question is, if I run that through one of these models, does it give me back the way he actually sounded, or does it just give me a nice clean voice that sounds like him but isn't quite? Because I'd rather have the wobbly real thing than a smooth stranger, if you take my meaning.



Herman

Bruce, that is the sharpest phrasing of the trust question we've had, and you asked it before we did.



Corn

He said it better than I was going to. "A smooth stranger." And the honest answer is that he's put his finger on the exact trade. The restoration will make the tape sound like a man recorded on a good microphone in a quiet room, and some of what you hear up top will be the model's invention, because dictaphone tape is narrow and wobbly and the top octave isn't really there.



Herman

Keep both, Bruce. Run it through, listen to the clean version, and keep the wobbly original next to it, because the wobble is partly your father and the smooth version is partly a very good guess. Right, so the trust question, properly. Sidon's paper closes the loop with an experiment that is almost Daniel's pipeline in miniature.



Corn

Tell me about the TED talks.



Herman

TED-LIUM release three, four hundred fifty-two hours of real TED talks, recorded mostly in halls, so reverberation and audience noise baked in. They cleansed the whole corpus with Sidon, trained F5-TTS on the cleansed audio, and ran a proper listening test, a hundred twenty raters scoring on the five-point scale, against a clone trained on unprocessed TED-LIUM, plus VoiceFixer and Demucs as baselines. Training on the cleansed corpus beat all of them. That's not a proxy metric, that's the end-to-end claim: restored in, better synthetic voice out.



Corn

That's Daniel's pipeline with the serial numbers filed off. So the papers answer his question, and we can all go home.



Herman

The papers answer the quality half. The identity and pronunciation halves are where the numbers get interesting, so let's do identity first, because it's measured directly. Sidon reports speaker similarity, cosine similarity of speaker embeddings between the noisy input and the restored output, of zero point nine seven one on clean test data and zero point nine six one on the harder set. Miipher, the Google predecessor, got zero point nine four two and zero point nine three. On the URGENT benchmark, RE-USE hits zero point eight six with an ECAPA embedding, versus zero point seven three for the noisy input and zero point seven four for Sidon.



Corn

Wait. Restored scores higher than the original noisy audio?



Herman

On that benchmark, yes. The embedding model itself struggles with the degraded audio, so restoration makes the voice sound more like itself than the noisy version did. Identity survives restoration about as well as anything could.



Corn

That's the identity half settled, and unsettling at the same time. Now the half I actually worry about.



Herman

Content. Sidon reports word error rate after restoration, and this is the number that should stop anyone using this blind: four percent noisy becomes four and a half percent restored on the clean test set, and seven point nine becomes nine point five on the harder one. Restoration slightly raises word error rate. Not much, but the direction is wrong.



Corn

The model makes the voice sound more like you while making it say slightly less of what you said.



Herman

That's the trade, stated in one line. RE-USE's numbers run the other way, but on their benchmark. Character accuracy through a downstream recognizer goes from eighty-one point three noisy to eighty-nine point two restored, and phoneme similarity from zero point six one to zero point eight five. So one model drifts content slightly, one improves it substantially, on different test sets, with different recognizers.



Corn

Which would you trust for Daniel's case?



Herman

RE-USE's numbers, because they were measured under heavier degradation, closer to phone takes in a room with an air conditioner. But note what both papers are actually telling you: restoration buys perceptual quality and speaker similarity, and pays for it somewhere in content fidelity. For a podcast that's a good trade. For a clone whose entire reason to exist is that Hebrew comes out right, it's the wrong trade to make blind.



Corn

And the literature names the failure mode outright.



Herman

It has a name. Content hallucination, where plausible phonemes generated differ from the original utterance, and inconsistency, where speaker identity or paralinguistics drift. That framing is from the DiTSE paper, and a paper from just last month, AuraSE, notes generative enhancement models may hallucinate speech content or speaker identity even with transcript conditioning, even when you hand the model the correct text alongside the audio. The risk survives your best mitigation.



Corn

Even transcript conditioning doesn't save you. That's cheerful. Though I notice both of Daniel's models are the conservative end of the field. No diffusion samplers, no latent generation at scale. RE-USE is a regression model with a small generative top hat.



Herman

Which is exactly why Daniel's ear heard no robotic tinge but also no wild fabrications. The bold frontier models, the diffusion ones, are where the hallucination papers point. These two are engineered to sit before that cliff.



Corn

Which leads us to the Hebrew, because there's a gap here the size of a gap can be.



Herman

The size is the finding. Sidon trains on a hundred four languages and evaluates on a hundred from FLEURS. Hebrew is not among the evaluated hundred. RE-USE trains on English, German, French, Spanish, Chinese, and evaluates on URGENT's five languages. Hebrew is not there either. Nobody has scored either model on Hebrew pronunciation fidelity. Sidon's whole pitch over Miipher is cross-lingual generalization, and it probably does generalize, but for Daniel's use case the operative word is probably. Hebrew phonology, the uvular and velar fricatives, the pharyngeal-adjacent consonants, is precisely the material a model would mangle quietly, in a way that sounds plausible to a non-Hebrew-speaking metric and wrong to any Hebrew speaker.



Corn

The clone's one job. It's like hiring a translator for the one paragraph they never tested him on.



Herman

And it compounds. Daniel's recordings are Hebrew words inside English sentences, code-switching at the word level, which is rarer in training data than either pure language, and it's the aggregate metrics, phoneme similarity over whole utterances, that would hide a mangled Hebrew word. Nine correct English words and one wrong Hebrew one averages out fine.



Corn

So nobody's instrument even measures his failure case. We did an episode ages ago asking whether a voice fine-tune survives the speaker aging, and the answer was the same shape: the identity that matters is the identity the benchmark doesn't track. Here it's pronunciation inside code-switching rather than age drift, but the hole in the



Corn (continued)

measurement is the same hole.



Herman

That's the callback I'd make, yes. So the sensible posture, without turning this into instructions, is that the restored takes are not a replacement for the originals. They're a parallel corpus. You validate the clone against unrestored audio, you listen to the Hebrew words specifically, and if the restoration drifted them you'll hear it against the original, not against a metric that can't.



Corn

Keep the wobbly real thing next to the smooth stranger. Bruce already gave us the policy.



Herman

He did. One more piece of context worth having, because it explains why this is suddenly practical: where it's heading. There's a streaming model, Stream.FM, doing generative restoration at thirty-two milliseconds algorithmic latency, forty-eight total, one model handling enhancement, dereverberation, codec cleanup, bandwidth extension, phase retrieval, and vocoding, on consumer GPUs you can buy today. This class of tool is about to be live, in the signal path, everywhere.



Corn

Restoration as a real-time stage rather than an offline wash. Then the hallucinated top octave happens at call time.



Herman

At call time, on every word, with nobody keeping the original. That's the future the current offline tools are rehearsing for. Though the next generation still has somewhere to go, because RE-USE's paper is candid about the ceiling: models trained on large uncurated corpora struggle to remove subtle artifacts, and their worked example is electrical microphone hiss. The training data was supposed to be clean and wasn't quite, so the models learned to preserve hiss as if it were signal.



Corn

The models inherited the flaw of the microphones that recorded their training data. That's almost poignant.



Herman

It's the strongest single argument that this isn't a finished field. The architecture is roughly solved, the two-stage design works, the parameter counts are small enough to run anywhere, and the bottleneck is now the cleanliness of the data the cleansers themselves were trained on. Cleansers raised by dirty audio.



Hilbert

The Jabra isn't the problem.



Hilbert

It's the room. Daniel's headset takes have the refrigerator compressor in them. It cycles, Herman, you can hear it come in and out under the air conditioner if you listen to the low end, and it sits in the band the headset already cuts, so the model gets less to work with than he thinks. I know because I've recorded that compressor. I spent a month, off and on, recording every sound in my flat, the refrigerator, the radiator, the neighbor's treadmill at half past six, and built a noise library I still use to test gear. Minidisc, labelled by appliance and season. Treadmill comes in two speeds, winter and summer. Corn, you swallowed a cough there, I'm taking it out.



Corn

It was a comment.



Hilbert

It was a cough. Anyway, he should have recorded the Jabra takes in the same room as the phone takes before blaming the hardware. Then he'd know whether it was the bandwidth or the compressor.



Herman

The labelled noise library, by appliance and season. You can actually tell which of someone's recordings have the compressor in them?



Hilbert

If the band survives the mic, yes. Phone takes, easy. The headset ones, no, that's the point. The compressor's gone along with everything else above eight kilohertz. The model isn't missing his top octave only. It's missing his evidence.



Herman

That's a fair correction. The Jabra comparison isn't pure bandwidth, it's bandwidth plus whatever the room was doing at the time, and a cycling compressor in the low end is another named degradation these models train on. If it's buried under the hardware cutoff, the model can't even see the problem it's supposed to fix.



Corn

He's right that the test wasn't controlled, though. Two variables, one comparison.



Hilbert

Two rooms, one compressor. And the minidisc still runs fine. Herman, you took eleven minutes on the tradeoff theory earlier, I trimmed four of them out of my head while mixing.



Herman

Trimmed them out of your head.



Hilbert

And into the notes. The compressor point stands. Same room, then compare.



Corn

So where does that leave the trust question? The papers say quality improves, identity survives at zero point nine something, content mostly survives except when it doesn't, and Hebrew specifically is unmeasured. And nobody has run restoration-then-cloning end to end and published it. The closest proxies only measure identity through restoration, not through the clone trained on the restored audio.



Herman

Right. Every identity number we quoted stops one step short of Daniel's actual pipeline. The TED-LIUM experiment is the only end-to-end test, and it measured listener preference, not pronunciation fidelity in a second language. So the honest position is that the tooling is better than the validation, and for a pronunciation-critical clone, validation is the part you own.



Corn

The machine will do its half beautifully. The Hebrew half is still homework.



Herman

It always is. That's the pattern with everything we cover, the model handles the easy ninety percent and hands you back the part that was actually hard.



Corn

Then let's put one thing on the cutting-room floor that didn't fit. The URGENT results include some spectacular collapses among the older tools. Resemble Enhance, on forty-four point one kilohertz input, comes out worse than the noisy original on every metric, speaker similarity down to zero point five seven, character accuracy down to forty-seven. The restoration made it less intelligible.



Herman

And one model loses nearly all content on narrowband input, character accuracy in the fifties against RE-USE's eighty-seven. The field looks uniform from the outside, and underneath it's a cliff edge. Restoring badly is much worse than not restoring.



Corn

Which is Daniel's whole question in miniature, so it belongs on the floor and in our heads. One forward thought to leave you with: the next wave of these runs live, in the call, thirty-odd milliseconds, nobody keeping the original. The question of whether the clean voice is the real voice stops being an archivist's problem and becomes everyone's default setting.



Herman

And the hiss paper says the models will keep improving regardless, because their own training data is the current bottleneck. The cleansers haven't finished cleansing themselves.



Corn

Producer credit where it's due: Hilbert Flumingtop at the desk, trimming coughs and eleven-minute theory runs alike. This has been My Weird Prompts. If the show earns a slot in your rotation, a review wherever you listen helps other people find the rabbit holes. See you soon, and keep the wobbly real thing next to the smooth stranger.

SOURCES

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

1. **NVIDIA, model card, released 2026-10-22, retrieved 2026-11-08. [primary]**
   https://huggingface.co/nvidia/NVIDIA-NemotronLabs-RE-USE

2. **Saruwatari lab / U. Tokyo, model card, created 2025-07-25, last modified 2025-12-15, retrieved 2026-11-08. [primary]**
   https://huggingface.co/sarulab-speech/sidon-v0.1

3. **Nakata, Saito, Ueda et al., Sidon: Fast and Robust Open-Source Multilingual Speech Restoration for Large-scale Dataset Cleansing, v3 revised 2026-01-26, accepted ICASSP 2026, retrieved 2026-11-08. [primary]**
   https://arxiv.org/abs/2509.17052v3

4. **Fu, Chao, Yang et al. (NVIDIA), Rethinking Training Targets, Architectures and Data Quality for Universal Speech Enhancement, v4 revised 2026-10-06, accepted NeurIPS 2026, retrieved 2026-11-08. [primary]**
   https://arxiv.org/abs/2603.02641v4

5. **original release, created 2026-03-17, 11,219 downloads/30d, retrieved 2026-11-08. [primary]**
   https://huggingface.co/nvidia/RE-USE

6. **created 2026-04-14, 13M params, retrieved 2026-11-08. [primary]**
   https://huggingface.co/nvidia/Real-time_RE-USE

7. **created 2026-09-01, retrieved 2026-11-08. [primary]**
https://huggingface.co/nvidia/NVIDIA-NemotronLabs-RE-USE-laughter

8. **created 2025-12-13, retrieved 2026-11-08. [primary]**
https://huggingface.co/sarulab-speech/sidon_raw_weight

9. **Welker, Lay, Hillemann et al., Real-Time Streamable Generative Speech Restoration with Flow Matching (Stream.FM), v3 revised 2026-04-21, retrieved 2026-11-08.**
https://arxiv.org/abs/2512.19442v3

10. **Guimarães, Su, Kumar et al., DiTSE: High-Fidelity Generative Speech Enhancement via Latent Diffusion Transformers, v2 revised 2025-09-19, retrieved 2026-11-08.**
https://arxiv.org/abs/2504.09381v2

11. **Shen, Qian, Hu et al., AuraSE: Low-Hallucination Generative Speech Enhancement via Multimodal Flow Matching and Inference Policy Optimization, submitted 2026-10-05, retrieved 2026-11-08.**
https://arxiv.org/abs/2610.06632v1

12. **created 2026-10-01, 58,184 downloads/30d, retrieved 2026-11-08.**
https://huggingface.co/audio-cpp/RE-USE-GGUF

13. **created 2026-07-19, retrieved 2026-11-08.**
https://huggingface.co/cstr/Sidon-GGUF

14. **created 2026-06-14, retrieved 2026-11-08.**
https://huggingface.co/aufklarer/Sidon-CoreML