Episode Book

EPISODE #6606

Why Optimizing Voice Quality Erases Identity

PUBLISHED	14 November 2026 at 23:16 UTC
DURATION	21:04
LISTEN	myweirdprompts.com/episode/voice-cloning-erases-identity-35032/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.30
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 11:16 UTC

SYNOPSIS

Optimizing a voice-cloning pipeline for perceptual quality can quietly erase speaker identity. Using a failed LoRA adapter run as a case study, this breakdown traces how squared-error loss seeks the conditional mean, causing spectral over-smoothing. It examines why glottal qualifiers like rasp and breathiness are treated as defects by generative speech-restoration models, and how a four-step recipe—teacher distillation, generative cleanup, shared EQ, and similarity filtering—compounds the averaging until switching back to Chatterbox as the teacher restores the original voices.

TRANSCRIPT



Corn

Two voices went in, one voice came out. That's the whole autopsy.



Herman

It's worse than that, we're the autopsy.



Corn

Daniel wrote in after hearing the first episode on the new LoRA adapters, and his verdict stung a bit: he preferred the old zero-shot voices. You and I sounded too similar, and I'd lost what he calls my gravelly edge. Drifted toward the median of the two of us. And then he did the thing Daniel does, which is go back through the recipe and find that every single step quietly rewarded the average take. The teacher was the new TTS model imitating us rather than the voice listeners actually know. A speech-restoration model smoothed away the rasp and breath. We both got the same EQ chain. The similarity filters kept the cleanest, most typical clips. Then they retrained with the old Chatterbox model as the teacher instead — rendering our own lines in the voice that aired for over a year, and filtering only for hallucinations.



Herman

And the two questions underneath all of that are good ones. Why does distilling a voice through clean-up and quality filters regress it toward the mean? And why is the flaw so often the identity?



Corn

I'd like an answer before I lose the last of the gravel.



Herman

Here's the framing I'd put on it. Two distinct voices went into that pipeline, and one averaged output came out — but that's not a bug in the LoRA training. It's what you should expect when you optimize for quality and similarity without anything in the loop that preserves idiosyncrasy. Quality and identity are different objectives, and the recipe treated them as one. Every step bought perceptual quality and paid for it with speaker



Herman (continued)

identity, and nobody was tracking the second account.



Corn

Because the second account doesn't show up on any score.



Herman

Right — the scores we were watching go up! That's the trap. ElevenLabs' own tips say flat EQ, no compression, preserve natural dynamics. The restoration models are marketed on how much cleaner they make things. Everything in that stack is rewarded for smoothness.



Corn

So the arc for today: first the math of why models seek the mean at all, then why the rasp is load-bearing rather than noise, then we walk the recipe step by step and watch each one reward the average take, and then why the Chatterbox retrain fixed it without any new modeling.



Herman

Start with the math, because it's cleaner than people expect. If you train a model with a squared-error loss, or any of its cousins, you're implicitly assuming there's one right answer for a given input. One mode. Human speech doesn't have one mode. For any sentence I might say, there's a huge space of valid deliveries — high pitch, low pitch, fast, dragging, breathy, tight. Faced with multiple valid but distinct targets, a deterministic regression model minimizes its loss by settling on the mathematical average of all possible outputs. The point that minimizes expected squared error is literally the conditional mean.



Corn

Which nobody wants to listen to.



Herman

Nobody wants to listen to it because of what averaging does acoustically. There's a nice analogy in the write-ups on this. Say you're told some forests are bright green and others are deep autumn orange, and your only goal is to minimize the total squared distance between your canvas and every possible forest. You end up painting a muddy brownish-gray blur.



Corn

That's us. Two forests, one mud painting.



Herman

In the frequency domain, averaging flattens the sharp spectral peaks and valleys — the formants — that give a voice its texture, and you get that muffled, slightly buzzy sound. In prosody it flattens the peaks and valleys of the pitch contour, so the voice comes out monotonous and dull. Ling-Hui Chen and colleagues documented this years ago in voice conversion: statistical models produce a weighted average of the training data, severe loss of variance, spectral envelopes missing the sharp resonant peaks of natural speech. They gave it a name — spectral over-smoothing.



Corn

I want to make sure I've got the mechanism straight, because it sounds like the model is being lazy. It's not. It's being optimal.



Herman

It's being exactly optimal for the objective it was handed. That's the uncomfortable part. The FastSpeech generation of models had this documented all over them — deterministic models trained with mean squared error are optimized to predict the average of all plausible outputs, and the output lacks intonation, rhythm, dynamic variation. FastSpeech 2, FastPitch, same story. Our recipe didn't invent this failure. It reenacted the founding failure of the field, at adapter scale, on our own voices.



Corn

We reproduced a classic. As a reproduction it's very faithful.



Herman

Now the second question, and this is the one I find moving. Why is the flaw the identity? There's a survey on acoustic correlates of voice qualifiers that puts it plainly — what makes our voices truly unique is a spectrum of complex voice qualifiers which differ significantly across speakers. Not just fundamental frequency, not just loudness. The qualifiers. Breathiness. Creak. Rasp.



Corn

Translate the qualifiers into physics for me.



Herman

Breathiness is incomplete vocal fold closure — air escaping at the glottis, audible aspiration noise, higher jitter, lower pitch amplitude. Vocal fry is irregular or very low frequency fold vibration, typically under about seventy hertz, low harmonics-to-noise ratio. Raspy, hoarse voice — that's me describing you, Corn — is irregular fold vibration plus glottal noise. Low HNR, elevated jitter and shimmer. Now look at what those numbers are. Those are the signals a restoration model trained to maximize perceptual quality reads as defects. High jitter, low HNR, noise at the glottis — on a quality score, every one of those is a mark against you.



Corn

So the model sees my entire personality as a defect report.



Herman

And here's the kicker from the speaker-identification literature. There's an Interspeech paper on text-independent speaker identification that treats glottal source behavior as a first-class identifying feature — the shape and timing of the glottal pulses, whether the folds close completely, how the glottal and supraglottal sources trade off. Those speaker-specific characteristics determine the high-frequency roll-off of the spectrum. Incomplete closure is the physical definition of breathiness. So a model that fixes breathiness isn't cleaning near the identity. It is deleting a named speaker parameter.



Corn

Corn, summarized: a walking stack of removable settings.



Herman

And the paralinguistic layer makes it worse, because breathy and creaky voice carry meaning — they correlate with intimacy, with sadness, with fear, with calm. A quality-maximizing pipeline normalizes all of that away because none of it helps the score. What's your read, though? You're the one whose rasp got sanded off. Did it feel like being cleaned, or like being replaced?



Corn

Replaced. The strange part of hearing that episode was that every line was recognizably almost me. The words were mine, the timing was mine, and the voice underneath was a committee's summary of me.



Herman

That's a good phrase, a committee's summary. Because that's precisely what the conditional mean is. It's not a wrong answer — it's every valid answer collapsed into one cautious compromise.



Corn

And now, as ever, the phone line is ringing in the studio, so let's see who it is.

D

Dorothy

Hello dear, it's your mother. I'm ringing for Dorothy's Kitchen.



Herman

Mum! Wonderful.

D

Dorothy

Today's recipe is an ice cube. You will need one tablespoon of water, and a freezer. That's all. I know. It sounds simple. It never is.



Corn

With water, more than most things.

D

Dorothy

First you pour the water. Slowly, sweetheart. Then you must watch it. Do not walk away. The bread must be watched and so must the water, they are the same in that respect.



Herman

Do you find freezing it changes the water, or is it the same water?

D

Dorothy

Then the freezer. And here is the dread. The butter can turn on you in a freezer, and timing is everything. Things have gone wrong before. I would rather not say how.



Corn

Iced butter does sound like a danger.

D

Dorothy

Wait one hour. Or two. I forget. Then you have your ice cube, and you put it in a glass and it melts anyway and you start again.

D

Dorothy

Anyway. Wear a cardigan, it's chilly. And eat something that isn't an ice cube. Bye bye, dear.



Herman

She's not wrong that you start again.



Corn

Back to the mud painting. The math explains one averaging, but the recipe had four of them stacked. Where does the teacher come in?



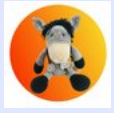
Herman

The teacher is the deepest one, because distillation inherits the teacher's distribution — adapters learn from whatever the teacher emits, and that includes the teacher's averaging. Think about what our teacher actually was. Not us. It was the new TTS model imitating us. And the base model's prior is studio-clean. The LoRA work on this is explicit that zero-shot models struggle with non-studio samples that differ from their training data — that's the whole reason LoRA helps, it can lift speaker similarity by up to thirty percentage points on rough spontaneous audio. But the lift is toward the model's idea of you, filtered through what the model already believes voices should sound like.



Corn

We distilled an imitation of an interpretation.



Herman

Second step, the speech-restoration model. This one has the clearest paper trail of the lot. The guides on preparing cloning audio all say the same thing: cloners copy everything in your training audio, hiss, hum, echo — but the vendors themselves warn against heavy generative speech enhancers, because rebuilding the voice flattens natural timbre and prosody. The distinction that matters is generative versus non-generative. A non-generative cleaner separates and turns down steady noise — fan hum, air conditioning — and leaves your formants, your breath, your delivery exactly as recorded. A generative enhancer rebuilds the voice from scratch. And a cloner trained on resynthesized speech clones the enhancer's synthetic voice model rather than your vocal folds.



Corn

We trained on a drawing of a photograph of me.



Herman

One of the restoration models people reach for is exactly that category — feature prediction plus a vocoder, reconstructing speech from learned representations. And its own numbers tell the story. On a reverberant clip, it raised DNSMOS from about 2.9 to 3.3 and UTMOS from about 3.0 to 3.4. Perceptual quality, clearly up. Speaker cosine similarity, sitting around 0.79. Quality up, identity sideways — and the vendor's own integration notes say the cleanup flag is opt-in because it changes the audio materially. It never even runs implicitly.



Corn

The company that built it won't turn it on without asking.



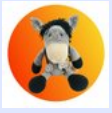
Herman

That's the tell. We turned it on and called it preprocessing. Third step, the EQ chain. Same curve for both of us.



Corn

My fault, incidentally. One chain was easier.



Herman

One chain, one spectral shape. ElevenLabs' guidance is flat EQ, no compression, preserves natural dynamics — their compression targets exist only for fixing inconsistency, and even then they're narrow. Any fixed processing chain becomes a permanent coloration on every line the clone speaks. Room reflections make the clone sound boxy on every line; a shared EQ curve makes two hosts converge on every line. Different cause, same outcome.



Corn

Two coats of the same paint.



Herman

Fourth step, the similarity filters, and this is where the selection logic mirrors the loss logic. If you keep only the clips that score highest on similarity and quality, you keep the clips closest to the model's existing centroid and you discard the outliers. The outliers are where the rasp lives. You've applied the conditional mean twice — once in the loss, once in the data selection.



Corn

The filter reads my worst takes and my best takes as the same category.



Herman

Because unusual is the only category it has. There's a detail here I keep turning over, too — the embedding that similarity is measured against is itself an average. A fixed-length vector, five hundred or a thousand dimensions, a statistical fingerprint pooled across the whole sample. Feed it more audio and it doesn't get sharper, it gets blurrier. And the ElevenLabs guidance on length runs the same way — one to two minutes for an instant clone, and beyond about three minutes quality can drop. Consistency of character beats volume. Their own docs say upload the raw file untouched. No fancy post-processing, no denoisers, keep it simple.



Corn

Every documented instruction pointed at the raw recording, and we built a pipeline whose whole job was to not give it one.



Herman

We built a very competent machine for producing the wrong training set.



Corn

So if every step pulled toward the mean, what changed with the Chatterbox retrain?



Herman

Three things changed at once, and they're worth separating. The teacher became the old Chatterbox model, which had rendered our aired voices for over a year — the voice listeners actually know, not a cleaned imitation of it. The data became our own aired lines. And the filter became categorical instead of continuous. The old Chatterbox is a five hundred million parameter zero-shot model with an Exaggeration tuning knob — a control for pushing expressivity away from the model's default. That's the exact control a mean-seeking pipeline lacks.



Corn

It was retired this summer, wasn't it?



Herman

August. Retired research, no longer maintained — the repo stays public as an archive under MIT, and Resemble has moved on to deepfake detection. Which is quietly funny, because the newer Chatterbox Multilingual V3 keeps the same half-billion size and headlines improved speaker similarity, better voice identity and accent preservation, reduced hallucination, unwanted continuation, repetition, off-prompt speech. Those are the right metrics, and they're aimed at the wrong layer. The newer model is better at being a teacher in general. Our problem was never the teacher's skill. It was that the exemplar — the actual aired voice — had been swapped out for an abstraction of it two steps upstream.



Corn

The fix wasn't a better model. It was pointing the same kind of model at the real thing.



Herman

And the hallucination filter is the part I'd defend hardest. A hallucination filter removes clips that are wrong — invented words, repeated phrases, speech that isn't in the script. That's categorical. A similarity filter removes clips that are unusual, and unusual is where identity lives. Filtering for wrong is coherent. Filtering for typical is a slow vote against the speaker.



Corn

You realize what we're saying, Herman. We spent the whole recipe optimizing for less of us, and the fix was to stop optimizing.



Herman

Would you actually retrain it that way again, from the start? No restoration, no shared EQ, hallucination filter only?



Corn

I'd go further. I'd keep the raw aired audio as the data even if a cleaner source existed, because we now know clean and correct aren't the same axis. And I'd want the two hosts' chains to be different by design, even if that means one of us sounds worse by some metric.



Herman

One of us meaning you.



Corn

One of us meaning the one with the gravel.



Hilbert

The one with the gravel is the whole question, Herman. I've been sitting here with a reel of my cousin's old radio broadcasts. He announced for a small station out past Ramle, voice all rasp and breath, and the station engineer kept after him with EQ and compression to clean it up, and the cousin fought him every week because the rasp was what listeners recognized.



Herman

Did he win the fight?



Hilbert

Eventually he quit and went to a station that let him keep it. He had a sign-off phrase, That's the way it is, and the rasp on the word way was the thing people called in about. When the engineer smoothed it out over one summer, the calls stopped. I have the recordings — Corn, you spent half this episode complaining about your own gravel, and you'd have liked his. I play the reel at family gatherings. I have been known to stop it and say, listen to that, that's not noise, that's him.



Corn

You keep your cousin's broadcasts on... what, a reel-to-reel?



Hilbert

Reels. I transfer them with a deck I service myself, these claws are no good for the fiddly screws but they're fine for the heads. What I wanted to know is whether your retrained adapters would take a reference like that. Old radio air-check, band-limited, thirty seconds of a dead man's rasp. Whether it would keep the rasp or average it out like your engineer did.



Herman

It would keep it. That's what the retrain changed — the exemplar is the ground truth now.



Hilbert

Good. The reel sits on top of the bookcase next to the mixer, and it's doing nothing there. I'd rather it was doing that. He signed off every broadcast with that phrase and the tape runs four seconds long after it. Just studio hiss. I never cut that part off.



Herman

You kept the hiss.



Hilbert

Of course I kept the hiss. That's the station.



Corn

The closer beat. If you take one thing from this episode, take the axis. Quality and identity are not the same axis, and every score we know how to compute sits on the first one.



Herman

And the practical version of that is almost nothing: keep the exemplar raw, filter only for what's wrong, and let unusual survive. Everything else is the pipeline voting against you.



Corn

One thought to leave open. As these models get better at speaker similarity and hallucination reduction, the risk isn't that they fail — it's that they optimize the wrong layer, the teacher instead of the data, and do it more fluently. The old retired model was better for us precisely because it was less advanced in the wrong direction. And Hilbert's four seconds of kept hiss is the same decision at reel scale: what looks like noise to a score is sometimes the room the voice lived in.



Herman

Our thanks as ever to Hilbert Flumingtop at the desk, producing the show and, apparently, archiving it.



Corn

For more along these lines, there's episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; and episode five, Fine-Tuning ASR For Maximal Usability. This has been My Weird Prompts. If this post-mortem earned a listen, leave us a review, or find the whole archive at my weird prompts dot com. See you tomorrow, Herman.



Herman

See you tomorrow, Corn.

SOURCES

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

1. **Resemble AI, Chatterbox TTS repo (README, model zoo, v0.1.2 release notes), last push 2026-11-04, retrieved 2026-11-14. [primary]**
   https://github.com/resemble-ai/chatterbox

2. **Soniqo, Sidon ONNX model card (DNSMOS/UTMOS/speaker cosine table), retrieved 2026-11-14. [primary]**
   https://huggingface.co/soniqo/Sidon-ONNX

3. **ElevenLabs Voice Cloning docs (no fancy post-processing, plugins, denoisers), retrieved 2026-11-14. [primary]**
   https://elevenlabs.io/docs/eleven-creative/voices/voice-cloning/

4. **ElevenLabs, 7 Tips for Creating a Professional-Grade Voice Clone (Flat EQ, no compression), retrieved 2026-11-14. [primary]**
   https://elevenlabs.io/blog/7-tips-for-creating-a-professional-grade-voice-clone-in-elevenlabs

5. **ElevenLabs Troubleshooting (RMS -23 to -18 dB, true peak below -3 dB), retrieved 2026-11-14. [primary]**
   https://elevenlabs.io/docs/eleven-creative/troubleshooting

6. **ElevenLabs Instant Voice Cloning best practices (1 [primary]**
   2 min; 3 min detrimental), retrieved 2026-11-14.
   https://elevenlabs.io/docs/eleven-creative/voices/voice-cloning/instant-voice-cloning

7. **Audo AI, How to Clean Up a Voice Recording for Voice Cloning, updated 2026-11-05, tested October 2026, retrieved 2026-11-14.**
   https://audo.ai/guides/clean-voice-sample-for-cloning

8. **Lacuna/Tiptree Systems, Overcoming Statistical Over-Smoothing in Generative Audio Models, retrieved 2026-11-14.**
   https://lacuna.tiptreesystems.com/direction/overcoming-statistical-over-smoothing-in-generative-audio-models/txn_484ddd13fa5b40eca6eea1919020b548

9. **speech-swift docs, Sidon `--clean-reference` caveat, retrieved 2026-11-14.**
   https://github.com/soniqo/speech-swift/blob/main/docs/inference/sidon.md

10. **Sidon: Fast and Robust Open-Source Multilingual Speech Restoration (SpkSim 0.971 vs Miipher 0.942), retrieved 2026-11-14.**
    https://www.alphaxiv.org/abs/2509.17052

11. **Acoustic Correlates of the Voice Qualifiers: A Survey (breathiness, vocal fry, rasp acoustic signatures), retrieved 2026-11-14.**
    https://arxiv.org/pdf/2010.15869

12. **Espy-Wilson et al., A New Set of Features for Text-Independent Speaker Identification, Interspeech 2006, retrieved 2026-11-14.**

https://www.isca-archive.org/interspeech_2006/espywilson06_interspeech.pdf

13. **Ishi et al., Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction, Speech Prosody 2006, retrieved 2026-11-14.**

https://www.isca-archive.org/speechprosody_2006/ishi06_speechprosody.pdf

14. **Probabilistic Speech & Motion Synthesis thesis (oversmoothing in FastSpeech/FastSpeech 2), retrieved 2026-11-14.**

https://www.diva-portal.org/smash/get/diva2:1988881/FULLTEXT01.pdf

15. **LoRP-TTS: Low-Rank Personalized Text-To-Speech (30pp speaker similarity gain), retrieved 2026-11-14.**

https://arxiv.org/abs/2502.07562

16. **Why 30 Seconds of Audio Beats 3 Minutes for Voice Cloning (512/1024-dim embeddings, blurrier average), retrieved 2026-11-14.**

https://exa.ai/library/podcast/2fcpbmjslqh/episode/ss74mlrtl0m

17. **Wang et al., Speaker Identity Preservation in Dysarthric Speech Reconstruction by Adversarial Speaker Adaptation, ICASSP 2022, retrieved 2026-11-14.**

https://arxiv.org/abs/2202.09082

18. **Chen et al., DiffDSR: Dysarthric Speech Reconstruction Using Latent Diffusion Model, Interspeech 2025, retrieved 2026-11-14.**

https://arxiv.org/abs/2506.00350

19. **Kim et al., Extended U-Net for Speaker Verification in Noisy Environments, Interspeech 2022 (the denoising process of independent enhancement models not tailored to the SV task can also distort the speaker information), retrieved 2026-11-14.**

https://arxiv.org/abs/2206.13044

20. **Wang et al., Task-aware Warping Factors in Mask-based Speech Enhancement, EUSIPCO 2021 (SE systems trained to improve speech quality often fail to improve other downstream tasks… because they do not share the same objects), retrieved 2026-11-14.**

https://arxiv.org/abs/2108.12128

21. **Anacin et al., Investigating the Impact of Speech Enhancement on Audio Deepfake Detection in Noisy Environments, 2026-03-16, retrieved 2026-11-14.**

https://arxiv.org/abs/2603.14767

22. **Violeta et al., An Extensive Analysis of the Singing Voice Conversion Challenge 2025 Evaluation Results (difficulty modeling breathy, glissando, vibrato styles), retrieved 2026-11-14.**

https://arxiv.org/abs/2509.15629

23. **Tsai et al., The False Resonance: A Critical Examination of Emotion Embedding Similarity for Speech Generation Evaluation, Interspeech 2026 (rewards acoustic mimicry), retrieved 2026-11-14.**

https://arxiv.org/abs/2604.26347

24. **Kang et al., GLASS: GRPO-Trained LoRA for Acoustic Style Steering in Zero-Shot Text-to-Speech, 2026-06-04, retrieved 2026-11-14.**

https://arxiv.org/abs/2606.05889

25. **Mehrish et al., DEFINE: Exemplar-Guided Accent Control for Zero-Shot TTS, 2026-09-30, retrieved 2026-11-14.**

https://arxiv.org/abs/2609.32777