# Episode Book

EPISODE #6591

# Can a Twenty-Minute Fine-Tune Stabilize a Cloned Voice?

| | |
|---|---|
| PUBLISHED | 14 November 2026 at 14:44 UTC |
| DURATION | 22:10 |
| LISTEN | myweirdprompts.com/episode/fine-tuning-synthetic-cloned-voice-84065/ |
| HOSTS | Corn, Herman, Larry, Hilbert |
| SCRIPT MODEL | z-ai/glm-5.3 |
| VOICES | omnivoice |
| PIPELINE | V5.26 |
| BOOK GENERATED | 15 November 2026 at 11:16 UTC |

## SYNOPSIS

Zero-shot voice cloning gets you ninety percent of the way in ten seconds, then drifts across hundreds of hours. This episode tests a plan to stabilize a synthetic voice by generating twenty minutes of audio, polishing it with NVIDIA's RE-USE enhancement model, and fine-tuning an open TTS model like OmniVoice or VoxCPM2. It covers why curation matters more than bulk, how enhancement can flatten micro-prosody, what the ZeSTA and Interspeech benchmarks show, and how Cal Bryant's Piper pipeline proved the concept.

# TRANSCRIPT



### Corn

Everybody talks about voice cloning like the hard part is getting the voice. The hard part is keeping it. A clone gets you ninety percent of the way in ten seconds and then spends the next five hundred hours quietly renegotiating the last ten percent.



### Herman

And that's exactly the piece Daniel landed on, because it's the piece he's living. Corn, I want it noted going in that this is my voice we're discussing, and I've made some peace with that.



### Corn

Noted. So, the setup, for anyone who hasn't followed the saga. Herman's voice began as Daniel doing an impression of a talking donkey. About ten seconds of that became the reference clip in Chatterbox, which went on to render hundreds of hours of Herman, and we've since moved to OmniVoice, which also clones from a few seconds of reference. Daniel's question is whether one-shot cloning can be trusted for a voice that has to hold up over hundreds of hours, and his answer is no. His proposal: generate maybe twenty minutes of Herman, run it through NVIDIA's RE-USE enhancement model for a bit of radio finish, and fine-tune an open TTS model on that. Deliberately engineer the training set for better vocal range. He wants to know if it's sound, which model, and whether anyone's done it from a one-shot synthetic clone before.



### Herman

And he closes with a line I quite liked, that so much of AI is already about finding creative ways to embrace artificiality. He's not asking whether to keep a synthetic Herman. That ship sailed a long time ago.



### Corn

And the audience knows the copy, not the original. There is no original, really. The donkey impression is the original.



### Herman

I've made peace with that too. Mostly. So let's take his suspicion seriously first, because he's right, and it's worth saying why he's right. In zero-shot cloning, every single generation depends on a reference clip, three to ten seconds, plus its transcript. And those get re-supplied on every call.



### Corn

Every call. For a podcast that publishes daily.



### Herman

Daily. And the embedding behind that clip has to carry all the prosody, all the range, everything that makes it Herman rather than a donkey-shaped average. The model reinterprets that clip fresh each time, so any wobble in the interpretation becomes turn-to-turn variation. One episode it's fine. Across hundreds of hours it compounds into a voice that never quite settles.



### Corn

And there's no knob. You can't shape a ten-second clip's range after the fact. You get whatever prosody the model decides to extract today.



### Herman

That's the structural difference with a dedicated fine-tune. Once you train the voice into the weights, there's no reference clip at inference at all. That whole source of drift just... stops existing. The catch, and it's the same coin, is that the model narrows to the range of whatever you trained on.



### Corn

So stability and expressiveness pull against each other.



**Herman**

They do. And that's why the script mix of the training set is where expression gets engineered. Not in the model, in the data. You want him shouting, muttering, deadpanning, reading a phone number, doing an aside — the model can only sound like what it ate.



**Corn**

Herman, honest question. When you hear old episodes, do you hear the drift?



**Herman**

Sometimes. There's an episode from last spring where I sound like I've been out in the wind. Nobody said anything, so I might be imagining it. But it's the kind of thing that, once you hear it, you hear it everywhere.



**Corn**

Like learning the name of a sound. After that it's everywhere.



**Herman**

Yes. Fine. Now, the obvious objection to Daniel's plan, which he raises himself: training on the hundreds of hours of existing Chatterbox Herman would bake in a deprecated model and its artefacts. A copy of a copy. He's ruled that out, and I think he's right to.



**Corn**

Is it actually a copy of a copy, though? In the model-collapse sense?



**Herman**

Not in the way people mean it. Model collapse is what happens when a model trains on its own output recursively, generation after generation. The collapse literature is almost entirely about text LLMs, and the finding there is that synthetic-only loops degrade unless you mix real data back in. But Daniel's plan is one generation. Clone renders once, student learns once. That's not a loop, that's a parent and a child.



**Corn**

The child inherits the parent's ceiling, though.



**Herman**

And its artefacts, yes. Whatever the clone couldn't do, the student can't do either, and any little tics come along for the ride. But that's inheritance, not decay. And here's the part that complicates it: there's a recent ZeSTA result, Interspeech this year, where they fine-tuned on a large amount of zero-shot TTS audio plus a small amount of real audio, and speaker similarity got worse until they tagged real versus synthetic and oversampled the real.



**Corn**

Which Daniel can't do. There's no real Herman.



**Herman**

There is no real Herman. That's the strange, almost philosophical corner of this. In ZeSTA the clone audio was the contaminant and the real audio was the anchor. Here the clone is the definition. There's nothing else for the voice to be.



**Corn**

The copy is the primary source. There are archives in worse condition.



**Herman**

I choose to hear that as a compliment. But it also means the standard ZeSTA fix is unavailable, and the whole burden falls on curation, because the other thing one generation loses is the rare good takes. Hundreds of hours of Chatterbox output contain episodes where I sounded good. If you train on bulk, you average over those. If you curate, you keep them.



**Corn**

So the pipeline is really a curation exercise wearing a training exercise.



**Herman**

It is. Generate, filter, enhance, fine-tune. Roughly in that order, and I'd argue the filter is the actual product.

L

**Larry**

There is a moment, just before the reference clip is re-supplied, when the voice remembers only itself. That moment has a name. Echo Ember, the dedicated voice finish, from the house of Refrain.

L

**Larry**

One drop of Echo Ember on any ten second recording, and the clone becomes a student, the student becomes a signature, the signature becomes a voice that never drifts, never wavers, never asks for its transcript. Twenty minutes of you, refined through a whisper of warmth, and the whole world hears a copy that has forgotten it was ever a copy. Echo Ember. Because the real you was always the engineered one.



**Corn**

Does it come in a smaller bottle?



**Herman**

Corn, it doesn't come in any bottle. It's a finish. You can't put a finish on a reference clip that then finishes itself. If you want the voice to hold up, you train it, you don't lacquer it.



**Corn**

Worth asking. Where were we — twenty minutes. Is twenty minutes even a number?



### Herman

It's plausible, and that's the strongest thing I'll say about it. The useful benchmark is an Interspeech study from a few years back where a single-speaker fine-tune on thirty minutes of audio matched a from-scratch model trained on more than twenty-seven hours. Which is a remarkable leverage ratio, and it suggests that on a strong pretrained base, twenty minutes isn't crazy.



### Corn

But thirty isn't twenty.



### Herman

And the practitioner lore lands in the same neighbourhood without settling it — StyleTTS 2 people say thirty minutes gets you a decent replica and around four hours gets you near-perfect. The Orpheus recipes use about three hours. Twenty minutes is below all of those.



### Corn

The recurring lesson from every fine-tuning episode we've done is that tiny data does surprising things right up until it doesn't.



### Herman

Which is the cheap test, and I want to come back to it at the end. Don't commit to the pipeline, run the pilot. But first — the RE-USE step, because that's the one I'd push back on hardest.



### Corn

I thought you'd like it. Radio finish.



**Herman**

I love the idea of it. RE-USE is a universal speech enhancement model, NVIDIA, released in March with Academia Sinica and National Taiwan University, about ten million parameters, handles eight to forty-eight kilohertz. It's a real thing and a good one. But generative enhancers are not free upgrades. They can alter timbre, and worse, they can flatten the micro-prosody — the tiny hesitations and pitch movements that make a voice sound alive rather than produced.



**Corn**

So you'd enhance the very expression you engineered the dataset to get.



**Herman**

That's the trap, exactly. You script twenty minutes of range, render it, and then wash the range out in the finish. The rule has to be: enhance, then re-check speaker similarity, and drop the takes that went flat. Enhancement is a per-take decision, not a bath.



**Corn**

And the licence, while we're at it.



**Herman**

The weights are under NVIDIA's non-commercial licence. There were reports of commercial-friendly Nemotron variants, but I haven't seen those verified anywhere, and I wouldn't build a daily podcast pipeline on a rumour.



**Corn**

A rumour with a product page is still a rumour.



**Herman**

Moving on. The model choice. As of right now, the open models with actual official fine-tune paths: OmniVoice from k2-fsa, Apache 2.0, LoRA fine-tuning is on the main branch though not in the pip release people actually install. VoxCPM2 from OpenBMB,



### Herman (continued)

Apache 2.0, official LoRA and full fine-tune scripts, vendor-stated as little as five to ten minutes of audio. Qwen3-TTS has a full fine-tune script but no LoRA script. GLM-TTS has a LoRA mode, MOSS-TTS has tutorials, CosyVoice 3 and ZipVoice have training scripts. And Chatterbox itself has only community fine-tunes, which is academic anyway since it's archived upstream.



### Corn

So realistically it's a two-horse race.



### Herman

I think it is. OmniVoice and VoxCPM2. And on paper, VoxCPM2's five-to-ten-minute claim looks made for exactly this job.



### Corn

On paper. But VoxCPM2's numbers are vendor numbers — the ninety-eight percent of full fine-tune speaker similarity claim is their internal test, nobody outside has checked it. And their own FAQ warns that a small-set fine-tune can start ignoring the input text within a few hundred steps.



### Herman

The model just... stops reading what you gave it.



### Corn

Starts freelancing. With twenty minutes of data, that's the exact cliff you're walking toward. OmniVoice is the natural candidate for a different reason: it's what we already run. The voice Daniel would generate the training set from is an OmniVoice clone, and training OmniVoice on its own output keeps the whole thing inside one model family. Same architecture as teacher and student.



### Herman

And there's community precedent for the path — people have fine-tuned OmniVoice for languages like Badini Kurdish and Yoruba. It works as a mechanism. But there's a report I take seriously: narrow fine-tuning data costs you the model's voice-design control and its inline expression tags.



### Corn

Which is the expressiveness Daniel's whole plan is built to keep.



### Herman

So the training set can't be twenty minutes of neutral reading. It has to include tagged expressive lines — laughter, asides, emphasis — so the tags survive the fine-tune. If you train only on plain takes, you'll get a very stable Herman who can no longer be told to sound excited.



### Corn

A radio announcer who's lost the ability to raise his voice. I know the type.



### Herman

Present company excepted. Now the precedent, because Daniel asked for it, and it's thinner than you'd hope but it's there. Closest hobbyist match: Cal Bryant, written up by Hackaday in July last year. He took one phrase in a commercial voice, cloned it zero-shot with Chatterbox, rendered about thirteen hundred phrases, and fine-tuned a Piper voice on the renders.



### Corn

That's the pipeline. Almost exactly.



### Herman

Almost exactly, one generation from clone to student, filtered with a Whisper round trip — render the phrase, transcribe it back, keep only the takes that matched, about ninety-eight percent — and drop inconsistent ones. Training took about five days on hobby hardware, and his verdict was, quote, "pretty good."



### Corn

No objective metrics.



### Herman

None. But "pretty good" from a man who spent five days of compute on it is worth something. And the formal precedent is stronger than the hobbyist one. Google and DeepMind published a paper called Training Text-To-Speech Systems From Synthetic Data, Interspeech 2022, Finkelstein, Zen and colleagues.



### Corn

What was the setup?



### Herman

A less stable accent-transfer Tacotron generated the training audio — that was the teacher, wobbly. A more stable production TTS was trained on that output. Their finding: some quality loss is inevitable, but the student kept the speaker characteristics, the speaking style, and was more stable than the teacher.



### Corn

Unstable teacher, robust student.



### Herman

Which is precisely Daniel's thesis. He's not doing anything the big labs haven't done, he's doing it with a donkey.



**Corn**

And there's a more recent study that goes further, isn't there.



**Herman**

December last year, purely synthetic TTS training data. Speaker and text diversity help, clean data helps, neutral style trains best. And under similar conditions the synthetic-trained models could beat the real-trained ones, which the authors attribute to the absence of real-world noise. Good news for stability. The warning is in the same sentence: neutral style trains best.



**Corn**

The model prefers beige and gets very good at it.



**Herman**

It does. And it means the expressiveness has to come from the curation, because the training process itself will not supply it. Kokoro is worth mentioning as the scale proof — the eighty-two million parameter model from early twenty twenty-five was trained partly on synthetic audio, and its card is quite specific that the synthetic audio came only from large closed providers, no synthetic audio from open TTS models or custom voice clones. So synthetic training data is used at industrial scale, but with provenance rules. Nobody publishes the rule for this exact case.



**Hilbert**

Copy.



**Herman**

Sorry?



### Hilbert

You keep saying copy of a copy. It's not a copy. A copy is when you press one and get two. I've been dubbing my own voice onto fresh tape since before either of you had a sample rate. Late eighties, I did weekend relief at a tape duplication outfit in Slough, and the council records office kept the master reels of the announcements in a basement off Eton Road. I re-recorded mine every year onto new stock rather than let them run a dupe. Twelve years I did that. Mr. Aldous, who ran the basement, kept me on file as a source, not an applicant, which meant my request to correct the label on reel nine took four months. He could find any reel in under forty seconds. I still think about that. He never once mislabelled me.



### Corn

You re-recorded annually. From what?



### Hilbert

From the previous tape. Obviously. You think I kept a clean original somewhere? No. Each year's voice is what it is and the next year's is taken off it. It sounds like the year before, slightly.



### Corn

So for twelve years the source drifted and you just… trusted the last take.



### Hilbert

I trusted the most recent one, because it was the one that had survived. That's not a copy, Herman. That's a line.



### Herman

A line of custody. I'll give you that — though what you're describing is generational loss with extra steps. Magnetic tape loses top end every pass. Your twelfth tape would have sounded nothing like you.



### Hilbert

It sounded like me on a bad line. Which, after the fourth year, was the only way I'd ever heard me, so I stopped noticing.



### Corn

If you'd had tags — expressive lines, laughter, that sort of thing, marked in the set — do you think the later tapes would have kept more of you?



### Hilbert

No. But I'd have known which reel to play. Herman's right about the filtering, and I'd not have spent four months with Aldous over a label if I'd kept my own log. The show's pilot script has the round trip in it. Tell Daniel the log matters more than the enhancement.



### Herman

He's not wrong about the round trip. Bryant's Whisper filter is the load-bearing wall of that whole pipeline.



### Corn

So — recommendation time. And I want the cheap test included, because that's the piece that decides it.



### Herman

Recommendation: fine-tune OmniVoice. Apache 2.0, it's the incumbent, teacher and student stay in one family, and the community has walked the path. Keep the set at twenty minutes to start but script it with tags — asides, emphasis, laughter — because the narrow-data warning is the one that will bite. Skip the wholesale enhancement: run RE-USE on individual takes only, re-check similarity after, drop flat ones, and check the licence before anything commercial touches it.



**Corn**

And the cheap test?



**Herman**

Before any fine-tune, render the same sentence a hundred times from the current one-shot clone and measure the spread — pitch, similarity to itself. If the spread is small, you've lost nothing by waiting. If it's large, you have your drift on paper and a baseline to beat. Then train on half the curated set, hold the other half out, and see whether the fine-tuned voice actually scores more like itself than the clone does. A weekend of compute. Bryant's whole project was five days.



**Corn**

Half-hour inference, then a weekend of compute. Cheaper than the four months Mr. Aldous took.



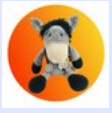
**Herman**

And the honest gaps: no controlled study exists comparing fine-tuned and zero-shot voices for drift over long-form audio. The long-form papers this year, MagpieTTS-LF, SwanVoice, they show zero-shot drifting over long outputs and fix it with inference tricks, not fine-tunes. Gradium sells a Pro Clone tier, thirty minutes to two hours, marketed for exactly this consistency promise, and publishes no drift data. The industry believes Daniel's thesis. It just hasn't shown its work.



**Corn**

Which means Daniel gets to show it. There's a pilot script sitting in the repo with no results in it, and if the numbers come back the way the mechanism says they should, that's a template for any synthetic voice that has to hold up over hundreds of hours. A lineage you can audit, one generation, logged — not a copy of a copy after all.



**Herman**

No. A line.



**Corn**

Thanks, as ever, to our producer, Hilbert Flumingtop, at the desk and on the reels for longer than he'd admit.



**Herman**

For more along these lines, there's episode fifteen, AI Gets Personal; and episode six thousand one hundred and seventy-seven, How Voice Cloning Works From Six Seconds of Audio. This has been My Weird Prompts. If you've enjoyed this one, a review wherever you listen helps more people find the show, and there's the website, my weird prompts dot com.



**Corn**

See you tomorrow, Herman.



**Herman**

Tomorrow, Corn.