Episode Book

EPISODE #6714

Why Voice Clones Stay Flat

PUBLISHED	20 November 2026 at 18:16 UTC
DURATION	26:20
LISTEN	myweirdprompts.com/episode/voice-cloning-data-prep-38207/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.36
BOOK GENERATED	20 November 2026 at 21:20 UTC

SYNOPSIS

Training a professional text-to-speech voice clone requires far more than feeding audio into a model. While zero-shot cloning captures basic timbre, fine-tuning shapes naturalness, emotional range, and behavioral consistency across long passages. Examining data requirements from ElevenLabs, LJ Speech, and CMU ARCTIC shows why recording hours of neutral reading yields a flat clone. Achieving lifelike results demands emotional range, precise forced alignment, and tedious manual clip cleaning to avoid baking trailing breaths and background room noise into the model forever.

TRANSCRIPT



Corn

Thirty seconds of audio, upload, and out comes a passable you. That's the party trick, and it's a good party trick. But Daniel wants to skip past the trick entirely this week and ask about the real thing: a professional voice clone, a text-to-speech model fine-tuned on one person's voice. How many hours of speech does that actually take, what has to be in those recordings so the clone sounds alive instead of flat, and what does the data prep and the training really involve — including, and I quote him, "the parts that make it so frustrating."



Herman

He knows us too well.



Corn

He does. Where do you want to begin, the number or the pain?



Herman

The number.



Corn

How much audio are we actually talking about?



Herman

Okay, so the first thing to get straight is that the vendor minimum and the professional standard are two different numbers. ElevenLabs' professional cloning docs say a bare minimum of thirty minutes of audio, and they recommend somewhere around two to three hours. Their own FAQ says thirty to a hundred and eighty minutes, and one of their step-by-step guides says at least an hour. So even internally, the same company is giving you a range that spans a factor of six.



Corn

Nothing builds confidence like a vendor disagreeing with itself in writing.



Herman

Then at the other end you've got the classic research corpora. LJ Speech, which half the field was trained on for years, is about twenty-four hours — twenty-three fifty-five, to be precise — from a single LibriVox reader doing seven non-fiction books. Thirteen thousand clips, two hundred twenty-five thousand words. That's what "from scratch" looked like.



Corn

So the honest range is somewhere between half an hour and a full day of one person reading.



Herman

Roughly, yes. And the reason it's so wide is that the two endpoints are doing different jobs. The thirty-minute minimum gets you a clone that exists. The twenty-four hours gets you a model that was trained from nothing on that voice. Almost nobody starting today trains from nothing — you fine-tune a model that already speaks — so the real question is where the curve flattens, and that's where it gets interesting. There was an experiment Baseten published, fine-tuning Qwen3-TTS on eight hundred clips from LJ Speech, about an hour and a half of audio, on a single H100 for about an hour. And the punchline is that at an hour and a half, the fine-tune did not materially beat zero-shot cloning on speaker similarity.



Corn

Wait. You fine-tune for an hour on an H100 and the similarity score doesn't move?



Herman

Doesn't move meaningfully. But — and this is the part I keep thinking about — it did improve on perceived naturalness, emotion, and prosody. The clone didn't sound more like the speaker. It sounded more like a person speaking. More alive, more varied.



Corn

So the fine-tune isn't buying identity. It's buying behaviour.



Herman

That's the distinction I'd want every listener to walk away with, yes. Zero-shot cloning is fundamentally a similarity trick — it grabs the timbre and a rough overall impression from a short reference clip. It's very good at that now. What it can't do is guarantee that the voice behaves like the person across arbitrary text, hour after hour. A zero-shot clone can nail one sentence and wobble on the next. A fine-tune is supposed to make that consistency boring.



Corn

Which reframes the whole data question. If similarity plateaus early and what you're really training for is expressiveness, then the recordings can't just be abundant — they have to have range.



Herman

And this is where the vendor docs are more honest than people give them credit for. ElevenLabs is explicit that the clone reproduces one consistent speaking style from whatever samples you give it. Not an average of your styles. The style. If you record in a flat reading voice, you get a flat clone, and no amount of extra flat audio makes it livelier. Long pauses get copied. Filler sounds get copied — "uhm" and everything like it, faithfully, forever.



Corn

A clone that says "uhm" is technically the most accurate clone possible.



Herman

It's the most human one, anyway! Nobody's fluent, Corn. But you have to decide whether your clone inherits your hesitations, and that's a style decision disguised as a data decision. And the quality bar matters more than it used to, too — their newest model reproduces noise and distortion more faithfully than the old one did, so clean input has gone from polite suggestion to actual requirement.



Corn

That's an odd side effect of progress. The better the model gets, the worse your room sounds.



Herman

Exactly backwards from every other consumer technology. Usually the product gets more forgiving. Here, fidelity is a trap — the model doesn't know your refrigerator hum is supposed to be background. It just knows that's what your voice sounds like, acoustically, so that's what it makes.



Herman

The levels are concrete: roughly minus twenty-three to minus eighteen decibels RMS, true peak at minus three, one voice only, no music or room reverb, and about two fists from the mic. We've dug into room and mic setup before, so I'll skip it. The one that surprises people is format: MP3 at one ninety-two kilobits per second or higher is fine, and uncompressed WAV gives little or no improvement.



Corn

Settled by a company that sells the clones.



Herman

Settled for their service, anyway. The LJ Speech corpus itself came from hundred twenty-eight kilobit MP3 audiobook files, and some of those clips carry encoding artifacts — and that dataset trained a generation of models. So the quality floor is lower than the forum arguments suggest.



Herman

Okay, the expressiveness side, because Daniel asked about that directly — what should the recordings contain? There are two answers that stack. The first is coverage. CMU ARCTIC, one of the old studio-recorded corpora, used about eleven hundred thirty-two phonetically balanced prompts per speaker — scripted material chosen to touch the whole sound inventory of English. Though one secondary source, which I couldn't check against the original report, puts diphone coverage at only about eighty percent.



Corn

Eleven hundred phonetically balanced sentences and you still miss a fifth of the sound pairs.



Herman

Diphones are brutal — the transitions between sounds are where the combinatorics live. Think about it this way: English has a few dozen phonemes, but the transitions between them — the glide from a "t" into an "oo," the blend from an "s" into an "sh" — those pairs multiply out into the hundreds, and some of them barely occur in natural speech. Nobody's everyday reading gets you all of them. But coverage of phonemes gets you pronunciation. It does not get you expression. The second answer is range. If you only record neutral declarative sentences, you get a neutral declarative clone. If you want a voice that can do wry, or urgent, or tender, you have to have recorded wry, urgent, and tender. The model is not going to invent your emotional register from your vowel inventory.



Corn

Here's what strikes me, and tell me if I'm wrong. The old speech corpora were built by people thinking like linguists — phonetic coverage as the goal. The ElevenLabs guidance is written by people thinking like casting directors — style consistency as the goal. And the actual failure mode people complain about, the flat clone, is a coverage problem the linguists solved and the casting directors didn't.



Herman

Hmm. Say more about that, because I half agree and I want to hear which half you mean.



Corn

The flat clone complaint — "my clone sounds robotic." Nine times out of ten the person recorded themselves reading articles aloud in one tone, because that was the easiest way to accumulate hours. They got the phonetic coverage of a read-aloud and the emotional range of a read-aloud, which is to say none. The fix isn't more hours of the same reading. It's fewer hours across more registers.



Herman

That's the half I agree with, then. The other half is where I'd hedge — pitch matters too. If you read everything at one pitch, the model learns that pitch is what your voice does. Variation in the training data isn't decoration, it's the space the clone is allowed to move in later. Bounded by what it saw, and not a hertz beyond.



Corn

The clone never improvises.



Herman

Never. It's the most literal student you will ever teach. Which means this thing will faithfully reproduce your worst habits, and it will do it at scale, with total commitment, forever.



Corn

Some of us would call that a legacy.



Herman

Some of us have a legacy of napping through the interesting parts.



Corn

Mine's restful and well-documented. So you've got your hours and your style — now what does the actual work look like? Because I don't think people imagine what's between the recording and the model.



Herman

This is the part nobody warns you about, and Daniel's phrasing — "the parts that make it so frustrating" — suggests he already suspects. The pipeline, in its standard form, goes like this: you transcribe your long audio with Whisper with timestamps, and then you split at closing punctuation into utterance-level pairs — one short audio clip matched to one exact sentence. That's the shape of training data. Sounds like an afternoon of scripting.



Corn

And is?



Herman

And is not. Because the split points are the whole game, and the split points are silences. The data-prep literature — ManaTTS, BibleTTS, that family — is unanimous on one ordering detail that tells you everything: forced alignment is done before silence removal, because pauses are where you cut. And pause thresholds in the papers range from about point-oh-eight seconds to half a second.



Corn

Point-oh-eight. That's not a pause, that's a consonant thinking about its future.



Herman

Right, and the choice matters both ways. Cut too eagerly and you slice words apart mid-breath and your transcript no longer matches your audio. Cut too generously and you train on trailing silence, and overly long silences hurt training — the model learns to insert dead air at the ends of sentences. Every recording is full of these judgment calls. A breath that's really a comma. A false start you have to decide is data or garbage. And the automated tooling gets you most of the way there and then stops.



Corn

How much of the way?



Herman

Most of it, and then the rest is you, listening, by hand. The LJ Speech corpus cut its clips automatically at silences and then had its transcripts matched by hand and checked in a QA pass. If the famous corpus needed human transcript matching, your project probably does too.



Corn

Give me a concrete one. What does a clip look like when the automation gets it wrong?



Herman

A typical one is the trailing breath. The tool finds the sentence boundary, cuts, and the clip ends with an inhale the model may learn as part of how you finish sentences. Or: the alignment slips by a syllable, and now the transcript says "manual" while the audio says "manu—" and you've taught the model a word that doesn't exist. Individually trivial. Multiply by thousands of clips.



Herman

And the frustration scales linearly — twice the audio, twice the listening. There's no economy of scale in the human part. Which is a strange inversion, right? The training itself keeps getting cheaper and the data prep doesn't.



Corn

Because the training is compute and the prep is judgment. Compute gets cheaper every year. Judgment bills by the hour.



Herman

You'd get along with my old scheduling spreadsheet. Okay, so say you survive that. You have your pairs, you've cleaned your silences, you fire off the fine-tune. Now the frustrations change character. The Piper training docs give you the epoch numbers: fine-tunes under about five hours of data want something like two thousand epochs. Ten to fifteen hours, more like six hundred fifty to eight hundred fifty. From-scratch runs, fifteen to thirty hours. Roughly twenty seconds per epoch per hour of audio on an A6000.



Corn

So ten hours of audio is about three and a half minutes an epoch, times several hundred epochs.



Herman

Which is a day or more of GPU time. But the number that stops people is this one, from the same docs: a hundred-fifty-epoch fine-tune on twenty-seven hours came out worse than the base model it started from.



Corn

Twenty-seven hours. Nearly LJ Speech scale. And it regressed.



Herman

Regressed. The docs tie it to too few epochs for that much data, so more data alone does not protect you. They warn that the validation score, the mean opinion score the pipeline computes for you, is noisy, and the best-scoring epoch is often a lucky one. Their advice is to listen to the checkpoints yourself.



Corn

The metric that exists to tell you when to stop, doesn't tell you when to stop.



Herman

It whispers. It doesn't tell. And there's a nasty scheduling trap on top — you have to set your epoch count before you start, because the learning-rate anneal is computed against the total. A run stopped early never finishes its anneal, so you can't casually extend one that looks promising.



Corn

So the one thing you learn from listening — "it's still improving, keep going" — is the one thing the setup forbids.



Herman

Plan the whole schedule up front and hope. Yes. That's the frustration Daniel's asking about, honestly. It's not that any single step is hard. It's that the signals are noisy, the decisions are irreversible in annoying ways, and the tools never tell you you're done.



Herman (continued)

There's always one more checkpoint that might be the good one.



Hilbert

Hilbert's Mid-Show Haiku.



Hilbert

Twenty four hours of voice, kept in a hard drive, one mistake.



Corn

Twenty-four hours and one typo in the transcript, and the model learns the typo with total conviction.



Herman

It learns the typo as though you'd read it aloud on purpose. Right — the overfitting question, which is the other half of the war story. The Baseten experiment found learning rates above about one times ten to the minus five tended to degrade speaker quality outright, and that a flat schedule — no anneal — overshot early and overfit late. One more caution, and it's only an anecdote: in a community thread on XTTS-v2, one user reported overfitting beyond about forty minutes of data and thought the default learning rate was too high for small sets. That's a single unverified report, not a rule, but it rhymes with the rest.



Corn

So the settings matter as much as the hours.



Herman

And you won't know until you listen, which takes us back to the ear being the only instrument that counts. Even the reference implementations wave the white flag. StyleTTS2's own README: the default fine-tune uses about an hour of audio, roughly a thousand samples, for fifty epochs — about four hours on four A100s — versus two and a



Herman (continued)

half days on four A100s to train from scratch on the twenty-four hours of LJ Speech. And the authors say outright the fine-tuned model is slightly worse.



Corn

The people who wrote the paper admit their own shortcut is a compromise.



Herman

They print it in the README! I find that weirdly honourable. And the plumbing frictions around it are exactly the hundred-small-things problem. The pretrained modules expect twenty-four kilohertz audio, so if yours are at twenty-two or forty-four, that's a resample you get to discover. The adversarial stage of the pipeline can run out of GPU memory, and the README says you can skip it, though quality may suffer.



Corn

"May be skipped" is my favourite phrase in all of machine learning documentation.



Herman

Right up there with "should now work." ElevenLabs, on the hosted side, quotes training after verification at three to six hours, and their own status page says six to twenty-four. So even the company whose whole job is doing this for you hedges the wall-clock by a factor of four.



Corn

To be fair, a factor of four is closer than their data recommendation hedged.



Herman

Small mercies. But think about what that status page means — they can't promise when it'll finish, and when it does, you'll still want to listen to it. Even the hosted, polished version of this process ends with a human pressing play.



Corn

Which is the one step nobody has productized. You can rent the GPU by the second and you cannot rent the ear.



Herman

You cannot rent taste. Corn, honest question — you've listened to more of this show's output than anyone alive. When you judge a cloned voice, what are you actually listening for?



Corn

Fatigue, honestly. Not of the voice — of me. A real voice, I can listen to for an hour. A bad clone wears out my attention in about ninety seconds, and I've never fully worked out what it's doing wrong. It's not the timbre. I think the timbre's fine and something underneath it is repeating — the same rhythm, the same rise, the same place to breathe.



Herman

The prosodic loop. That matches what people describe when fine-tunes go flat — the model latches onto its most confident phrasing and reuses it. Which means the antidote is in the data: variation. If the training clips all phrase the same way, the clone has one phrasing. Would you record differently if you were doing this for yourself?



Corn

I'd record the way I talk, which is slow with long pauses, and I'd get back a clone that sounds like it's about to fall asleep. My data is a hazard.



Herman

Your data is a lullaby with a transcription.



Hilbert

Which of you said "two fists from the mic"? Herman. Of course it was Herman. I have a question about that. I recorded a weekend of myself once. Every page of the manual for a pallet jack I owned. Torque tables, safety warnings, the lot. Somebody had written that you need hours of speech, so I did the hours properly. Saturday morning to Sunday night, one continuous session, two hundred and eleven pages.



Herman

Two hundred eleven pages. Roughly what runtime did that come out to?



Hilbert

Just under eleven hours. I played it back on the Monday. By hour four of the recording my voice had gone hoarse and never came back for the weekend. So the whole thing starts clean and ends up sounding like a man reading through a head cold, and it does it gradually, page by page. Also it turned out the clone would only ever be able to talk about load limits. That's my question — none of the guides mention what you're supposed to read. I read the manual because the manual was what I had.



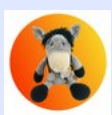
Corn

You built a voice that can only recite torque specifications, in a hoarseness that worsens over time.



Hilbert

The worsening was the part I hadn't planned. One session, one voice drifting. I still have the files. They're on a drive in the drawer, in a folder called "for the future." The drive takes a cable I don't currently own. Trim the level on your left channel a touch, Herman, you're peaking past minus three on the sibilants. I mention that because your mic technique has not improved since the manual, whatever the guides say about two fists. My snout has never been two fists from anything.



Herman

Noted on the levels. The hoarseness point is actually the technical heart of it, though — you recorded a voice that changes over the session, and the model would have learned that drift as one voice. It can't tell tired-you from fresh-you. It would average them, or worse, interpolate between them.



Hilbert

It would have the choice between averaging and inventing a middle hoarseness. Neither is me on a Tuesday. The torque tables were fine, though. Evenly spaced numbers read at even intervals — that part came out clean.



Corn

Eleven hours and the only usable output is the part that sounds like a robot on purpose.



Hilbert

I hadn't thought of it that way, no. Though the table of contents was my best material. Even pacing, no surprises, every entry the same length.



Herman

And that's the lesson hiding in Hilbert's weekend, honestly — the content question and the style question are the same question. Reading material with no emotional range produces a clone with no emotional range, no matter how many hours you put in. The guides tell you how much to record and how clean to record it. Almost none of them tell you what to say.



Corn

Which is strange, because that's the one part where the old corpora had it right. CMU ARCTIC didn't just record hours — it recorded specific sentences chosen for a reason. The linguists knew the content was the point.



Herman

They knew the content was the point for their purpose. The modern version of that advice would be: write yourself a script that covers your registers. Newsreader, storyteller, the voice you use when you're explaining something you love, the voice you use when you're tired of explaining it. Read all of them, on purpose, and take breaks so your voice doesn't drift — drink water, split it across days if you can.



Hilbert

I did none of that. I did the opposite of all of that, on a folding chair, over two days.



Herman

Which raises the real question then — whether the professional process is worth it at all when the zero-shot model is already this good. Daniel's framing was "the hard way," and the numbers back him up on the hard part.



Corn

Here's how I'd slice it. If similarity is all you need — one-off clips, short bursts, the party trick done responsibly — the fine-tune isn't buying you much. Baseten's own numbers said so: an hour and a half of data, an hour on an H100, and the similarity score didn't move. What you're paying for with the real process is naturalness, emotion, prosody — the stuff that shows up over hours of output, where the zero-shot voice starts looping or drifting.



Herman

So it's worth it exactly when the output is long and the style is specific. An audiobook, a permanent assistant voice, a character. Not worth it for a novelty impression.



Corn

And the target keeps moving. Every zero-shot improvement eats into the fine-tune's advantage on identity, which was already its weakest claim. What's left for the professional clone is the emotional and stylistic range the zero-shot model can't reach — and that means the value is migrating from "sounds like the person" to "performs like the person." The recordings you'd make for the second are a different recording session than the first.



Herman

The casting director wins after all. Though I'll say the quiet part — as zero-shot gets better, the bar for fine-tuning rises too, because the only reason left to do it is quality, and quality is the part with the noisy metrics and the two-a.m. checkpoint listening. The floor keeps rising and the ceiling keeps getting harder to prove.



Corn

And Hilbert's drive in the drawer — eleven hours of a hoarse man reading torque tables — is now, what, technically a curated dataset with a delivery problem?



Herman

It's eleven hours of style consistency. Single voice, no music, no reverb. By the vendor checklist it nearly qualifies. It's just qualified for the wrong life.



Corn

If he ever finds that cable, that's a clone that could get a job in compliance training. Somewhere out there is a warehouse onboarding video still waiting for exactly that voice.



Herman

And that's the actual open question, isn't it — not whether the fine-tune can be built, but whether anyone will still want it once the free version gets good enough to fool the room.