Episode Book

EPISODE #6633

Voice Engineering and the TTS Recording Matrix

PUBLISHED	16 November 2026 at 15:47 UTC
DURATION	27:04
LISTEN	myweirdprompts.com/episode/voice-engineering-tts-recording-matrix-83004/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.34
SOURCES	22 cited by the research (listed at the end)
BOOK GENERATED	16 November 2026 at 22:22 UTC

SYNOPSIS

Retraining OmniVoice LoRA adapters on older Chatterbox synthetic speech with a light EQ pass worked surprisingly well, raising a deeper question: what does deliberate voice engineering look like? This conversation explores how LoRA handles off-distribution audio, why corrective EQ on synthetic output differs from destructive EQ on human capture, and what an Interspeech Greek TTS study reveals about speaker drift. It also breaks down Datatoorder's recording matrix framework—separating emotion, register, and delivery across multiple takes to keep synthesized speech alive.

TRANSCRIPT



Corn

We are, right now, being spoken by the second generation of ourselves. That's the situation. The voices you're hearing are OmniVoice LoRA adapters retrained on the output of the previous Chatterbox voices, with a light EQ pass in between. And it worked. Suspiciously easily, if you ask me, but nobody did.



Herman

I asked, internally.



Corn

Daniel, though, did ask something better. He looked at how hands-off our retrain was, the voice-to-voice step basically handed to a model, and said: fine, but what does the deliberate version look like? If we actually sat down with my training set and Herman's and engineered the voices we want, on purpose, what's the craft? Curating and editing the data, controlling timbre and EQ and delivery, the frameworks and processes that exist for that. Which, I'll spoil, is a field that barely has a name yet, let alone a textbook.



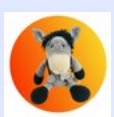
Herman

Voice engineering, he's calling it. And he's right that it's a name for something real. The pieces exist, they're just scattered across practitioner blogs, per-model fine-tuning guides, and a stack of research systems nobody has stitched together. There's one proper survey of controllable speech synthesis that comes closest to a unifying reference, and even that is mostly a map of fragments.



Corn

Before we engineer anything, we should be honest about what the shortcut actually did. Because my suspicion is that we didn't take a shortcut. I think we accidentally did a real technique with two of the steps skipped.



Herman

That's a fair way to put it, and the evidence backs it. The core of what we did is a LoRA adapter on a zero-shot TTS system, trained on audio that was never studio-recorded. There's a paper called LoRP-TTS that looked at exactly this, and their finding is that



Herman (continued)

low-rank adaptation lets a zero-shot system use, and I'm quoting because it's a good line, even single recordings of spontaneous speech in noisy environments as prompts. Speaker similarity went up by as much as thirty percentage points in their tests. LoRA is unusually tolerant of off-distribution audio. It's the tool that makes non-studio audio usable at all.



Corn

So we fed it the one kind of non-studio audio that was secretly studio-grade.



Herman

Every failure mode that wrecks a normal TTS dataset, cross-talk, room changes, clipping, transcripts that don't match the audio, two speakers on one track, was absent from the old voices' output by construction. The Chatterbox voices were clean, single-speaker, consistently miked, and the transcripts existed because the model generated the speech from text in the first place. We didn't curate a dataset. We inherited one that had already been curated by a machine that didn't know it was doing it.



Corn

The EQ pass, then. Herman got a cut around five hundred fifty hertz and air added above six kilohertz, which is the audio engineer's way of saying he was muddy and boxy. I got a small cut around three twenty.



Herman

Which, for the record, was the first time anyone has described my voice as muddy and boxy, and I've been listening to it for years.



Corn

It was a scientific finding, Herman. Not a review.



Herman

I've taken it well. Anyway, the EQ removed the one systematic spectral bias in the training audio before it could get baked into the adapter. And here's the thing that makes this interesting rather than routine. The closest published analogue to what we did is a Greek TTS system from Interspeech this year. They took audiobook recordings, aligned and filtered them into TTS-ready data, and fine-tuned an eight hundred eighty million parameter Parler model with a speaker-specific LoRA stage trained on three and a half hours of single-speaker audio, updating about five percent of the parameters.



Corn

Three and a half hours. That's a long afternoon of recording.



Herman

And the results were solid. Word error rate around ten point seven percent, which is about three points above what the speech recognizer itself gets on the same audio, and listener scores within spitting distance of the human recordings. Four point zero versus four point three six on similarity, four point two four versus four point three zero on comprehensibility. Robust single-speaker synthesis from limited curated data is achievable. We're not in fringe territory.



Corn

That paper has a warning in it, though. Tell me about the style prompts.



Herman

Yes, and this one bears directly on us. They found that LLM-generated style prompts introduce speaker drift at inference. If your pipeline is generating the style or reference prompts on the fly, the voice wanders. Replacing them with deterministic prompts, the same fixed prompts every time, resolves it. Now, we switched to per-voice LoRAs precisely because our zero-shot clones were drifting, Corn's clone famously started sounding like mine by month three.



Corn

Deeply disturbing at the time. Less so now that I know it was a documented phenomenon.



Herman

So the question I can't shake is whether the drift question compounds. The new Herman is a model trained on the output of a model. The old Chatterbox voice was itself synthetic. LoRA's tolerance for off-distribution audio is why the shortcut worked, but has anybody measured what happens on generation three, generation four? As far as I can tell, no. That's a open question, and it's our question in a way it isn't most people's.



Corn

We're the experiment.



Herman

We're the experiment nobody's writing up.



Corn

Here's the piece I keep turning over, though. Every serious data-curation guide I've ever seen says don't EQ your training data. Don't compress, don't run automatic gain control, because those operations flatten exactly the intensity and pitch cues that carry emphasis. And we EQ'd. And it worked.



Herman

And I think the reconciliation is that those guides are about human capture. A human recording has room tone, natural dynamics, real prosodic variation, all of which a sloppy EQ chain destroys. The old voices' output had none of that. It was already flat, already compressed, no room tone to preserve. EQ'ing it wasn't destroying a signal. It was correcting a defect.



Corn

Corrective EQ on model output versus destructive EQ on human capture. Two operations that look identical on the mixer.



Herman

Same knob, opposite meaning. And nobody has published on that distinction. Whether light-EQ-then-retrain is measurably beneficial, and by how much, appears to be undocumented. We didn't follow the rules. We may have accidentally found one.



Corn

So that's the deliberate version: curate the data, control the timbre, shape the delivery. But honestly, most people just want a voice that stays put.

L

Larry

Friends, are you tired of voices that drift, wobble, and forget who they are? Then you need VoxLoftra, the once-daily voice-stabilizing tablet from the makers of nobody in particular. VoxLoftra gently retrains your inner speaker model overnight, absorbing even single recordings of spontaneous speech in noisy environments, so you wake up thirty percentage points more yourself. Side effects may include mild boxiness, a faint muddy quality around the edges, sudden air where no air should be, spontaneous Greek audiobook narration, an adapter that will not detach, low-resource whispering, persistent room tone, and in rare cases, a second voice that knows things about you that you have never told anyone.



Corn

Anyway. Say more about what the deliberate version looks like, because I don't think the answer starts with the model.



Herman

It doesn't. It starts with a spreadsheet, and this is the part I find exciting, so brace yourself. There's a practitioner write-up from this summer, from a data company called Datatoorder, whose central claim is that the recording matrix is the dataset. The idea is you separate the axes before you record anything. Emotion, angry, sad, that's one axis. Register, news, audiobook, conversational assistant, that's another. Delivery, whisper versus projected, a third. And most datasets get this wrong, they list styles as one flat list, emotion and register and delivery all mixed together, and the labels overlap and the model can't separate them.



Corn

A style label like "excited news whisper" is three knobs taped together.



Herman

And every speaker who walks into the booth means something different by excited unless you've given each style a one-line definition and an audible example. That's the discipline the write-up pushes: define each style in words, play a reference take, and only then record. And when it comes to choosing the labels themselves, discrete categories beat dimensional ratings for most projects. Neutral, cheerful, serious, sad, angry, whisper, maybe with a strength level attached. Not valence and arousal axes.



Corn

Because a director can ask for angry, but nobody can ask for minus zero point four valence.



Herman

That, and raters can't agree on continuous numbers, and a model conditioned on two floating values needs far more data to learn the axes. Then the matrix itself. Sentences times styles times takes. Multiple takes of the same sentence in the same style. Their line on this is the best thing in the whole piece: a single take per sentence per style teaches the model that each style has exactly one prosody. Several takes give the model the distribution, what varies between renditions and what stays constant, which is what makes synthesized speech sound alive rather than stamped.



Corn

Stamped. That's the word for the zero-shot clones, wasn't it. Every sentence arrived with the same fingerprint on it.



Herman

One prosody per style is a stamping machine. And there are confounds to control. Same speaker records every style, or the model learns style and speaker are the same variable. Style order within a session gets randomized, because if the energetic styles always land in the first hour and the quiet ones in the last, the model isn't learning excited versus calm, it's learning fresh voice versus tired voice.



Corn

Which is a real thing in this household. By the last hour of anything, we are all quiet styles.



Herman

You're a quiet style by eleven a.m. And keep a rejected-take log with reasons. That log is half the value of the session, because it's your labeling quality control later. Then the scripts. Phonetic balance, obviously, but also prosodic sentence shapes. Statements, yes-no questions, wh-questions, lists, emphatic stress, long sentences with internal clause boundaries. The sharpest training signal you can build is minimal pairs where the words are identical and the meaning is carried entirely by stress or intonation.



Corn

"I didn't say you stole it." Seven meanings, one string.



Herman

That exact kind of thing, yes. Rare in off-the-shelf scripts, and exactly what a style-conditioned model needs. And the transcripts have to be exact down to the punctuation, because punctuation is the only text-side cue to phrase boundaries. A missing comma is a lost label. Capture discipline is its own list. Lossless or high-bitrate recording, gain control off, constant mic distance, watch clipping, because the emphatic takes are the loudest ones and they clip first. The line that stuck with me: automatic gain control removes the intensity differences that carry emphasis, which is the signal itself.



Corn

The cleanup chain that removes the thing you recorded.



Herman

You're cleaning the signal out of the recording. Keep a raw unprocessed file of every take alongside anything you process. Corn, let me turn this around, because you've thought about the data side more than me on this. Where would you start if we did it for our own sets? Not the frameworks, the actual decision.



Corn

With what the matrix would even be for us. We don't have styles. We have one register each, roughly, mine is unhurried and Herman's is enthusiastic to a degree listeners write in about.



Herman

They write in appreciatively.



Corn

They write in. So the honest matrix for me is sentences times delivery times takes. Two, three takes per sentence, some whispered, some projected, and the labels left alone. I don't need "cheerful Corn." Nobody has ever asked for cheerful Corn.



Herman

Fair. And there's a verification step in the Datatoorder piece that applies to any labels you do choose. The speaker was asked to sound sad, and that is not evidence that the audio sounds sad. You run a listening check where raters pick the intended style from the audio alone, and the confusion matrix tells you which styles sit too close. Sad, tired, and calm sit close together on the perceptual map, apparently. When two styles get confused at a high rate, you merge them or sharpen the definitions. You don't ship the labels as recorded.



Corn

For us the confused pair would be "deadpan" and "serious." I'd want to know whether listeners can even tell.



Herman

I can't, sometimes, and I've known you for years. Moving to the tooling, because Daniel asked what actually exists. The frameworks each want a specific shape of data. F5-TTS wants ten to fifteen hours of twenty-four kilohertz mono WAV for a proper fine-tune, one hour minimum, clips between one and thirty seconds, a metadata file in a common format, ninety-ten train-validation split. XTTS v2 wants the LJSpeech layout at twenty-two oh-five hertz, and the community LoRA configs run a rank of eight, alpha thirty-two, dropout oh oh five. GPT-SoVITS is the heaviest, it precomputes phoneme sequences, BERT features, features from an audio model, speaker-verification embeddings, though the V3 and V4 LoRA route cuts the memory requirement roughly from fourteen gigabytes



Herman (continued)

to eight.



Corn

Fourteen to eight. That's the difference between renting time on a serious card and using the one already in the flat.



Herman

Which matters more than anyone admits when the experiment is a hobby. The general playbook across all of them looks similar. Audit the data first, deduplicate, remove cross-talk, normalize loudness to around minus twenty-three LUFS, annotate emotion and intent. Noisy labels drag listener scores by zero point two to zero point four, which is enormous when human speech is scoring four point three. Then force-align the transcripts, precompute the content features and pitch contours, and train with low learning rates, around one times ten to the minus five.



Corn

Alignment is the quiet hero of that list.



Herman

It is, and there's a segmentation subtlety I want to flag. The standard way to cut long audio into training clips is voice-activity detection, split at any silence longer than a couple of seconds. A Persian corpus project this year pointed out that this produces a mess of duration distribution, and instead used diarization followed by prosody-aware segmentation targeting roughly twelve-second utterances. Cut where the prosody says cut, not where the silence happens to fall. Microsoft's custom-voice guidance lands in the same place from the other direction, phrases up to twenty seconds, and if your use case leans on questions or exclamations, put a good portion of them in the script.



Corn

This is a podcast. Our use case is exclamations with occasional regrets.



Herman

Then our script over-indexes on questions and interjections, which, listening back to any episode, it already does by accident. Scale is the last piece, and it's the one that kills the more-is-better instinct. There's an open project that aggregated six hundred fifteen thousand hours of speech, two hundred forty million segments, and then ran model-based filtering down to five hundred ten thousand hours, about a seventeen percent cut, and trained a one-billion-parameter model that lands within range of state-of-the-art systems trained on several million hours of proprietary data.



Corn

They threw out a hundred thousand hours and got better.



Herman

Curation is where the leverage is, not hours. That's the whole argument for the deliberate version in one number. There's even open tooling now that computes per-clip features, speech quality, prosody, intelligibility, speaker identity, transcription, so you can filter and balance a dataset semi-automatically. The audit is no longer purely manual.



Corn

Okay. So you've curated the data, you've trained the model. Now, how do you steer it? Because everything so far is about the voice that comes out being a faithful average of what went in. Engineering implies you can push it around afterward.



Herman

This is where it gets philosophically messy, because the controllability research has converged, almost by accident, on the primitive we're already using. LoRA. One paper, ReStyle-TTS, trains style-specific LoRAs, one adapter per style attribute, and fuses them with an orthogonal fusion method so you get continuous, disentangled control over pitch, energy, and several emotions at once, with a module to stop the timbre drifting while you do it. Another, DEFINE, uses LoRA on F5-TTS with a single guidance weight at inference that dials accent strength up and down without retraining. Accent-probe accuracy rose from six and a half percent to nearly twenty as the guidance increased. One knob, continuous effect.



Corn

So the deliberate version of us isn't one adapter. It's a rack of them.



Herman

A rack, exactly. A neutral Herman adapter, an excited Herman adapter, whatever axes we defined in the matrix, fused at whatever ratio the episode needs. And the Greek system used a speaker-specific LoRA stage too, so the pattern repeats across labs. We're already using the right primitive. We're just using one of them where the recipe calls for six.



Corn

Then what's the wall? Because you said this gets messy.



Herman

The wall is the reference-audio trap, and it's the gap between cloning and engineering. Multiple papers this year identified it independently in almost the same words. Zero-shot models clone a speaker's timbre from a short reference clip, but they strongly inherit the speaking style present in that reference. Another paper put it as a trade-off: text descriptions are a natural control interface but give weak acoustic grounding, while speech style prompts give concrete prosody but entangle the style with the prompt speaker's timbre. The reference clip carries timbre, style, and accent fused together. You cannot ask for Herman's timbre with Corn's energy without machinery that splits them apart.



Corn

Which is why the clone converged on me. The reference didn't just carry my voice. It carried my whole manner, and the model couldn't tell which parts were optional.



Herman

Disentanglement is an active research problem, not a shipped feature. There's work on mutual-information-guided separation of timbre and emotion, and IndexTTS2, from a big industrial lab, claims genuine independent control, timbre from one prompt, emotion from a separate style prompt, plus explicit duration control where you specify the token count for exact clip length. But it's not something you get for free with any checkpoint you download. The most directly relevant open tool is called VoiceSculptor. Fully open-sourced, code and pretrained models. You describe the voice you want in natural



Herman (continued)

language, it generates a controllable timbre from the description, supports iterative refinement with retrieval-augmented generation, does attribute-level edits, and then renders the designed voice into a prompt waveform that feeds a cloning model for the actual synthesis.



Corn

Type "warm donkey, slight Connecticut, enthusiastic but not exhausting" and it hands you a waveform.



Herman

In principle. And there's a lighter alternative to any retraining, audio prompting. A paper this spring pairs textual style tokens with human-curated, high-quality audio prompts, and the framing is that the audio prompt acts like in-context learning, guiding prosody and timbre without parameter updates or large-scale retraining. You keep a small library of exemplar clips and steer at inference time. That's closer to what a working voice engineer actually does day to day, I suspect, than any of the retraining routes. A clip library, not a training run.



Corn

One of the older takes we've held since the first fine-tuning episode still stands, by the way. Everyone wanted cloning to be the product, and cloning turned out to be the easy twenty percent.



Herman

And the bottleneck now isn't even generation, it's knowing whether you succeeded. There's a benchmark called InstructTTSEval, six thousand test cases across three tasks, acoustic parameter specification, descriptive style directives, role-play, English and Chinese, machine-judged, and its headline finding is that accessible instruction-following TTS systems have substantial room for improvement. Nobody can yet reliably measure whether the model followed the instruction. Before any of this, the old FestVox prosody documentation had the bluntest advice in the field: prosodic modeling is still an extremely difficult area, all models are extremely impoverished, do not be too ambitious.



Corn

Advice from the people who built the tools, to lower your expectations.



Herman

I find it comforting, honestly. Decide up front whether you're modeling conversation or read prose, news-reader or storyteller, and don't try to be everything.



Hilbert

"Air" is the wrong word.



Herman

For the EQ?



Hilbert

For the shelf. You didn't add air to Herman, you borrowed brightness from the top octave and let people call it air. I ran the comparison years ago, a rack unit I still have against a plugin doing the same move, and the plugin's air band was a lie told in kilohertz. Same clip, both paths, level matched. The hardware did something around two hundred to four hundred hertz that no plugin has replicated, and I stopped trusting the word air the day I heard it.



Corn

You A/B tested this on your own voice? For years?



Hilbert

Since before either of you had a voice to test. There's a notebook, paper, on the shelf behind the desk. Every EQ move I have ever made on my own voice, logged. Frequency, gain, width, date, what I was reading at the time. It's the only honest dataset in this building. The customer who wanted the warm sound, Rami, from the shop on Agrippas, he wanted warm the way people ask for air. I charged him for the two hundred hertz move and told him to skip the rest, and he argued with me for ten minutes about the word.



Herman

What was in the two hundred to four hundred range that the plugin missed?



Hilbert

Body that follows the note. The plugin moves the level and leaves the voice behind. I don't have a word for it and I've stopped looking for one. The point is the retrain worked, I'll grant that, but the LoRA did the work. The EQ was a coat of paint. And Corn sits there saying deliberate, deliberate, all episode, and never once asks what the paint is for.



Corn

Noted. Paint is for people who can hear it.



Hilbert

Somebody has to. The men who built that rack unit are probably gone, and the units are still doing the job. That's a good way to spend a life.



Corn

He's not wrong that nobody's measured it, though. That's the thing his notebook has over the entire field.



Herman

So it doesn't. Six thousand test cases for instruction following, and not one published measurement of EQ-before-retrain. The one variable our pipeline actually changed on purpose is the one with no literature.



Corn

Then we're the A/B. Retrain one adapter on EQ'd clips, one on the same clips untouched, blind-listen the outputs, and we'd know whether the five hundred fifty hertz cut on Herman did anything the LoRA wouldn't have absorbed anyway. My guess is Hilbert's right and it's mostly the adapter, but the answer costs us an afternoon.



Herman

And the second question is the one I'd actually fund first. Drift across generations. We're a model of a model right now, and if there's a generation three, someone should be measuring what the voice loses each time it passes through. Nobody has. We'd be starting from zero, which is either a warning or an invitation.



Corn

Both, usually. If you've done this yourself, deliberate voice engineering on your own training set, we want to hear what your recording matrix looks like. Everything's at my weird prompts dot com.



Herman

And thanks as always to Hilbert Flumingtop, our producer, for keeping the desk running.



Corn

If you want more of this, try episode fifteen, AI Gets Personal. This has been My Weird Prompts. See you tomorrow.

SOURCES

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

1. **Microsoft Learn, Recording custom voice samples, retrieved 2026-11-16. [primary]**
https://learn.microsoft.com/en-us/azure/ai-services/speech-service/record-custom-voice-samples

2. **XTTS_V2 fine-tuning repo (LJSpeech format, LoRA r=8/alpha=32/dropout=0.05), retrieved 2026-11-16. [primary]**
https://github.com/gokhaneraslan/xtts_v2-finetuning

3. **(Advanced-Version) [primary]**
AllTalk TTS wiki, XTTS Model Finetuning Guide (Advanced Version), retrieved 2026-11-16.
https://github.com/erew123/alltalk_tts/wiki/XTTS-Model-Finetuning-Guide-

4. **FestVox, Prosody Walkthrough, retrieved 2026-11-16. [primary]**
http://festvox.org/festvox/x1993.html

5. **audiogear per-clip dataset feature tool, retrieved 2026-11-16. [primary]**
https://github.com/lIkesimba9/audiogear

6. **Datatoorder, TTS style and prosody data: the recording matrix is the dataset, published 2026-08-31, retrieved 2026-11-16.**

   https://datatoorder.com/blog/collecting-tts-prosody-and-style-data/

7. **Xie et al., Towards Controllable Speech Synthesis in the Era of Large Language Models: A Systematic Survey, arXiv:2412.06602v3, revised 2025-08-25, EMNLP 2025 main, retrieved 2026-11-16.**

   https://arxiv.org/abs/2412.06602v3

8. **Li et al., ReStyle-TTS: Relative and Continuous Style Control for Zero-Shot Speech Synthesis, arXiv:2601.03632v2, revised 2026-04-19, ACL 2026, retrieved 2026-11-16.**

   https://arxiv.org/abs/2601.03632v2

9. **Hu et al., VoiceSculptor: Your Voice, Designed By You, arXiv:2601.10629v2, revised 2026-01-20, retrieved 2026-11-16.**

   https://arxiv.org/abs/2601.10629v2

10. **Zhou et al., IndexTTS2: A Breakthrough in Emotionally Expressive and Duration-Controlled Auto-Regressive Zero-Shot Text-to-Speech, arXiv:2506.21619v2, revised 2025-09-03, retrieved 2026-11-16.**

    https://arxiv.org/abs/2506.21619v2

11. **Bondaruk & Kubiak, LoRP-TTS: Low-Rank Personalized Text-To-Speech, arXiv:2502.07562v1, submitted 2025-02-11, retrieved 2026-11-16.**

    https://arxiv.org/abs/2502.07562v1

12. **Syllas et al., Deterministic Prompting for Speaker-Stable Low-Resource Greek TTS, arXiv:2609.10022v1, submitted 2026-09-09, Interspeech 2026, retrieved 2026-11-16.**

    https://arxiv.org/abs/2609.10022v1

13. **Kim et al., Raon-OpenTTS: Open Models and Data for Robust Text-to-Speech, arXiv:2605.20830v2, revised 2026-06-15, retrieved 2026-11-16.**

    https://arxiv.org/abs/2605.20830v2

14. **Mehrish et al., DEFINE: Exemplar-Guided Accent Control for Zero-Shot TTS, arXiv:2609.32777v2, revised 2026-09-30, retrieved 2026-11-16.**

    https://arxiv.org/abs/2609.32777v2

15. **Huang et al., E3-System: Expressive EmoData, Explicitly Controllable, Easy-to-Use TTS, arXiv:2610.12993v1, submitted 2026-10-09, retrieved 2026-11-16.**

    https://arxiv.org/abs/2610.12993v1

16. **Ouyang et al., Enhancing Conversational TTS with Cascaded Prompting and ICL-Based Online Reinforcement Learning, arXiv:2604.08709v1, submitted 2026-04-09, retrieved 2026-11-16.**

    https://arxiv.org/abs/2604.08709v1

17. **Yang et al., Emotional Text-To-Speech Based on Mutual-Information-Guided Emotion-Timbre Disentanglement, arXiv:2510.01722v1, APSIPA ASC 2025, retrieved 2026-11-16.**

    https://arxiv.org/abs/2510.01722v1

18. **Huang et al., InstructTTSEval: Benchmarking Complex Natural-Language Instruction Following in Text-to-Speech Systems, arXiv:2506.16381v1, submitted 2025-06-19, retrieved 2026-11-16.**

    https://arxiv.org/abs/2506.16381v1

19. **PersianVox: A Prosody-Aware Approach for Speech Dataset Generation from In-the-Wild Data, retrieved 2026-11-16.**
https://arxiv.org/html/2609.19324

20. **DeepWiki, SoVITS Model Training | RVC-Boss/GPT-SoVITS, retrieved 2026-11-16.**
https://deepwiki.com/RVC-Boss/GPT-SoVITS/6.3-sovits-model-training

21. **Momen Walied, Data Prep for Voice Cloning (F5-TTS), retrieved 2026-11-16.**
https://momen.gitbook.io/momen-walied-space/masters-learn/f5-tts-and-silma-complete-technical-guide/module-d-training-and-fine-tuning/19-data-prep-for-voice-cloning

22. **Instavar, Voice Cloning Finetuning Guide: E2-TTS, F5-TTS, and GPT-SoVITS V2Pro, retrieved 2026-11-16.**
https://instavar.com/research/tts/voice-cloning-finetuning-benchmark-picks