Episode Book

EPISODE #6717

Why Ten Seconds Clones Better Than an Hour of Training

PUBLISHED	20 November 2026 at 20:10 UTC
DURATION	28:03
LISTEN	myweirdprompts.com/episode/voice-cloning-zero-shot-finetuning-58870/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.36
SOURCES	23 cited by the research (listed at the end)
BOOK GENERATED	20 November 2026 at 21:20 UTC

SYNOPSIS

Zero-shot voice cloning uses brief audio as a temporary conditioning signal, while fine-tuning permanently modifies a model's weights. Attempting to improve a clone with an hour of training data often degrades output quality by introducing acoustic noise, style collapse, and catastrophic forgetting. Balancing speaker similarity and clear pronunciation requires techniques like prior-preservation loss to prevent model collapse onto narrow datasets.

TRANSCRIPT



Corn

Ten seconds. That's what we're built on, apparently. One morning of donkey and sloth impressions into a microphone, and out the other end came six thousand episodes of us.



Herman

I prefer to think of it as efficient.



Corn

I prefer to think of it as alarming, but that's why they don't ask me.



Herman

Daniel's latest lands exactly there, for anyone wondering. The Chatterbox era is over, it went end of life in August, and he's decided the answer is proper voice models for us rather than another round of cached embeddings. And he ran straight into something backwards. You can get a workable voice from ten seconds of reference audio, so surely an hour of training data buys you something better. His experience was the opposite. Fine-tuning is, his phrase, a labor of love, the reference guides all say three to five hours minimum, and the first results came out worse than the ten-second trick. He's also building the Hebrish app on the side, English-Hebrew code-switching, which fights its own version of the same war. His question, plain: if ten seconds works, why does five hours barely get you started?



Corn

And it's a good question because the obvious answer, more data equals better, is exactly what's failing him.



Herman

The answer, I think, is that the two things aren't points on one quality line. They're different operations happening in different places in the model. That's the whole episode, really, but let me put the frame up properly. Zero-shot cloning puts the reference audio in the input. Fine-tuning puts the speaker in the weights. Those are not the same operation at different budgets.



Corn

One is renting the voice. One is renovating the building.



Herman

That's... actually not bad. Renting is reversible, it's instant, it's cheap, and you get what the building already had. Renovating is expensive, partly destructive, and the damage is hard to undo. When Daniel fed ten seconds of you into Chatterbox, nothing in the model changed. When he fine-tunes a base model on an hour of recordings, he is pushing on every weight that model spent thousands of hours of multi-speaker data earning.



Corn

Let me make sure I've got the distinction, because it sounds like a distinction without a difference if you say it fast. In both cases there's a model, there's my voice, there's audio coming out that sounds like me. What's concretely different?



Herman

Where the information lives at generation time. In zero-shot, at the moment the model is producing a sentence, your ten seconds is sitting right there in the context, being attended to, sample by sample. Cut the reference and the model reverts to whatever default speaker it has. In a fine-tune, the reference audio is gone — thrown away, even — and the model still produces you, because the you-ness is distributed through millions of parameters. One is a lookup that happens every time. The other is a scar.



Corn

So the ten seconds isn't small training data. It's not training data at all.



Herman

Right — no, not right in the way people assume. It's a conditioning signal. The base model already knows what human speech is, from a huge multi-speaker corpus during pre-training. The reference clip only tells it which voice to imitate. Ten seconds is an address into a space the model has already mapped. That's why it works instantly, and that's also why it's disposable — you swap the address, you get a different tenant, nothing in the walls has moved.



Corn

And the fine-tune is me moving into the walls.



Herman

You, and every quirk of the room you were recorded in, permanently. Which we'll get to, because that's where Daniel's hour went wrong.



Corn

Start with the ten seconds, then. What's actually under the hood that makes it work.



Herman

The architecture is the interesting part. A well-built zero-shot system splits the problem in two, and the split is the whole explanation for why ten seconds works at all. There's a timbre encoder — MiniMax-Speech describes a learnable speaker encoder that pulls timbre features out of reference audio without needing a transcript of it — and separately a prosody encoder with its own language model that handles how the speech moves. Who you are, and how you talk. Separate modules.



Corn

Ten seconds buys you the who. The how was never for sale.



Herman

Ten seconds is plenty for the who. Identity lives in a small set of acoustic features — pitch range, roughness, where your resonance sits — and a good encoder can pin those from a short clip. That's why Daniel's one morning of impressions carried hundreds of hours of audio. The donkey and the sloth were different addresses in a space the model already knew.



Corn

Not to nitpick the foundation of our existence, but a donkey impression isn't timbre. It's a full commitment to a register.



Herman

It was enough timbre. And that's the point — Stable-TTS, the ICASSP paper that lays this split out most cleanly, says the same thing from the other direction. Zero-shot models struggle precisely because pre-training needs thousands of hours, and even then, the gap between the training corpus and your particular speaker causes trouble. The model interpolates. If your voice sits somewhere between speakers it knows, you sound great. If you're an impression of an animal, it finds the nearest animal it has and hopes.



Corn

It found me. I'm still not sure what that says about its training data.



Herman

There's a way to test the interpolation idea, actually, that Daniel could run cheaply. Record the same sentence twice — once in your normal voice, once deliberately pitched up — and clone from each. If the model has you mapped between two known speakers, the two clones should come out sounding like two different people who both almost sound like you. That's what interpolation looks like from the inside.



Corn

And if they both sound exactly like me?



Herman

Then the encoder pinned you despite the pitch shift, which would tell you identity in that model isn't riding on pitch alone. Either result is informative. Nobody runs these little experiments, which is a shame, because they cost nothing.



Corn

Now the cost of that split. Because the prosody half didn't come free, and this is exactly what Daniel noticed.



Herman

The November paper from the Alibaba and Nanjing group puts it as a named research problem now: a reference with limited intonational variation may lead to flat speech even when the target text calls for strong emotion. The clone inherits the delivery of the reference clip rather than adapting to the text. Daniel read one line of donkey into a microphone one morning. Whatever that line sounded like, that's the emotional range we've been living on.



Corn

Hundreds of hours of my flat morning voice. Listeners will recognize it.



Herman

It was one take.



Corn

One take before nine in the morning, for the record. There are listeners who write in saying they can hear the exact moment the coffee hadn't happened yet. I've always assumed they were imagining it.



Herman

They weren't. That's the mechanism working exactly as designed — the model is loyal to the reference, not to the text. Loyalty to the wrong thing is the whole problem. And the fix that paper proposes isn't more reference audio — it's post-training the model itself, preference optimization, teaching it to stop imitating the reference's delivery. Three stages, and by the end, emotion scores go from under twenty to over eighty on their benchmark.



Corn

Hold on — you said the fine-tuning path was the dangerous one. That's fine-tuning.



Herman

It is, and look what the plain supervised step did on the way there, because this is the cleanest set of numbers I've seen for what Daniel hit. Same model, one round of ordinary fine-tuning: emotion score jumps from eighteen and a half to nearly sixty. Speaker similarity falls from point seven two to point six zero. English word error rate goes from one point two three to one point six six.



Corn

So it got more expressive, less like the speaker, and worse at pronouncing words. That's a fine-tune, all right.



Herman

That's the signature. Expressiveness up, identity down, intelligibility down. And the reason is structural, not a bug in their setup. The model isn't a blank slate waiting for your voice — its weights encode thousands of speakers' habits, phoneme by phoneme. Pushing it toward one speaker means moving a lot of those weights, and moving weights on a small, narrow dataset is the textbook overfitting setup. Stable-TTS says it exactly: fewer fine-tuning steps may reduce voice similarity, while excessive fine-tuning can compromise the model's ability to produce clear, intelligible speech. You're always trading between the two ends of that sentence.



Corn

So there's a window. Too little and it's not you, too much and it's not speech.



Herman

A window, and where it sits depends on your data. Their fix is clever, actually — a prior-preservation loss. While fine-tuning on the target speaker, you keep forcing the model to still synthesize speakers from its original pre-training set, so it can't collapse onto you completely. It has to stay a general model that happens to know you, instead of becoming a you-shaped model that's forgotten everyone else.



Corn

Which is a strange kind of medicine, when you put it that way. To learn one person properly, the model has to keep rehearsing a crowd of strangers.



Herman

It's a rehearsal disguised as a constraint. The gradient wants to collapse onto you, and the prior loss keeps dragging a chunk of it back toward the crowd. Whatever the model ends up learning about you has to coexist with everything it already knew.



Corn

Forgetting everyone else. That's the thing nobody says out loud when they say fine-tuning. You're not adding a voice, you're evicting several thousand.



Herman

Catastrophic forgetting, yes — the continual speaker adaptation literature is blunt about it. Sequentially fine-tuning for new speakers degrades performance on older speakers. Every voice you burn in comes at the edges of the ones already there.



Corn

Now explain the hour. Daniel's specific finding — an hour of fine-tuning data performed worse than ten seconds of reference. No hedging, that's what he saw.



Herman

I believe him, and I think it's three separate mechanisms stacked on top of each other. First, distribution. The ten seconds never touches the weights. The hour does, and it drags the model toward everything in that hour — including what's wrong with it. Stable-TTS flags this directly: in-the-wild recordings have unclear pronunciation and background noise, unlike the clean speech the model was pre-trained on. You're fine-tuning on data that's worse than what the model already ate.



Corn

And the model can't tell the voice from the room.



Herman

It has no reason to. ElevenLabs' professional cloning docs are startlingly honest about this — the clone will be highly accurate to the samples it trained on, including any artifacts and unwanted audio present in them. Air conditioner hum, room echo. And once it's learned, there's no cleanup step that fully removes it. The voice and the hum are the same thing to the weights.



Corn

A ten-second clip is a handshake. An hour is a lease with the room included.



Herman

Second mechanism, style collapse. Same docs: the speaking style in your samples will be replicated in the output, and they recommend including only one style per upload. Daniel's hour of us is presumably an hour of podcast — one register, one pace. That's an hour of lessons in being flat in exactly one way.



Corn

We do have range. Sometimes I nap through a point instead of making it.



Herman

Third, and this one might actually apply — synthetic contamination. If any of that hour is Chatterbox-generated audio of us rather than fresh recordings, there's INTERSPEECH work this year showing that naively mixing synthetic speech with limited real recordings degrades speaker similarity during fine-tuning. The model learns the copy's artifacts as if they were you.



Corn

He's got hundreds of hours of us lying around, all synthetic. The temptation to use it must be enormous.



Herman

Enormous and exactly wrong. Real recordings only, however painful that is on his CPU box.



Corn

Although — is there a version of that temptation that's legitimate? Say he had no real recordings at all. Only Chatterbox output.



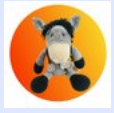
Herman

Then he'd be building a copy of a copy, and every generation loses something. The INTERSPEECH finding suggests the loss isn't graceful — the model doesn't average out the artifacts, it absorbs them as identity. A voice trained on synthetic you believes the compression noise is part of your throat.



Corn

Which means our archive, all six thousand episodes of it, is technically contaminated as a training set.



Herman

Every minute of it. We are, as a dataset, a hazard. I find that oddly clarifying.



Corn

Well, before we get into the data volumes, I believe we have a caller.



Herman

Oh, is it — Mum?

D

Dorothy

Hello dear, it's your mother. It's time for Dorothy's History of Toast.

D

Dorothy

The word toast comes from the Latin torrere, to scorch or parch. That's it, that's the whole word, and it means scorching, dear.

D

Dorothy

Bit gloomy really. You think of toast as a nice warm thing, and it turns out it's just a word for burning. Everything sounds nicer than it is, I find.

D

Dorothy

Anyway. Eat something proper before the afternoon, sweetheart, and mind the draught in that studio. Bye bye.



Herman

Hi, Mum. Love you.



Corn

She's not wrong about the draught.



Herman

She never is. And the toast-in-wine thing is real, for the record — spiced toast was the flavoring. Which is oddly on topic, if you squint: something meant to improve the whole by sitting in it permanently, whether you can remove it or not.



Corn

That's fine-tuning. My God, that's fine-tuning described by fourteenth-century wine.



Herman

I was going to say prosody encoders, but I'll take it. Where were we — the hours. Daniel's three to five.



Corn

The three to five. Is it a real number or folklore that hardened into a spec?



Herman

It's a heuristic wearing a lab coat, and the vendors can't even agree with themselves. ElevenLabs says bare minimum thirty minutes in one section of their own docs, and then, on the same page, at least an hour, ideally close to three. That's not a contradiction so much as two different products being described — thirty minutes gets you a recognizable voice, three hours gets you a production voice, and the page hedges across both without saying so.



Corn

A spec that means whatever you needed it to mean. Those are my favorite specs.



Herman

The bands from the data-side vendors run higher. One guide I read puts light fine-tuning at ten to thirty minutes, production work at thirty minutes to three hours, flagship studio voices at five to twenty hours and up. A dataset vendor says one to five hours minimum for fine-tuning an existing voice, ten to twenty-five for a production single-speaker voice.



Herman (continued)

Daniel's three to five sits at the top of the vendor guidance and the bottom of the dataset-vendor guidance.



Corn

And someone in there said the quiet part.



Herman

Flaunt Audio, yes: hours of audio is a poor proxy for usable signal. Which is the actual answer to what the three to five hours buys you. It's not hours the model needs. It's coverage. Diversity of phonetic context, prosodic variety, varied recording conditions — enough different sentences in enough different moods that the gradient signal can't be dominated by a handful of them. Five hours of one podcast register is less coverage than forty minutes of you actually living.



Corn

So the unit is situations, not seconds.



Herman

The unit is situations. There's a paper this year on exactly when fine-tuning fails — speakers trained on acoustically homogeneous data get limited gains or actually sound worse, even when voice similarity improves. Homogeneous data makes you more you and less listenable at the same time. It's the flat-prosody problem from zero-shot, re-entering through the weights.



Corn

Same trap, different door. That's grim. I'd rather have one door.



Herman

One door with a sign on it. Because here's the genuine trade, and Daniel has already mapped both sides of it with his own ears. Zero-shot: instant, cheap, and the prosody is capped at whatever your reference clip contained. Fine-tuning: expensive, fragile, and the prosody can go anywhere your training data goes. If you record yourself being delighted, furious, and half-asleep, the fine-tune can learn all three. The zero-shot clone cannot — it was never given them.



Corn

Which means the recording session for a proper fine-tune of us wouldn't be a recording session. It would be an acting job. Delighted, furious, half-asleep — that's a brief for a performer, not a podcast host.



Herman

It is, and that's a hidden cost nobody prices in. The three to five hours isn't just recording time, it's three to five hours of deliberately manufactured situations. Reading the news while angry. Explaining something while exhausted. Nobody does that naturally, which is part of why natural-sounding fine-tunes are rare.



Corn

The short utterances he mentioned. Where do those fall?



Herman

Mostly on the zero-shot side of the ledger. A zero-shot model conditions on the reference and then has to generate your line — on a two-word utterance there's almost no context for it to settle into the voice with, so it wobbles. A fine-tune that's absorbed you doesn't need the runway, the identity is in the weights, not borrowed per sentence. That's the one place fine-tuning is categorically better rather than just potentially better.



Corn

And Hebrish. English-Hebrew code-switching. This is the one I'd bet is misdiagnosed.



Herman

I've been waiting for you to say that, because I agree, and I want to hear your reasoning first.



Corn

If the base model was never trained on Hebrew, then no amount of cleverness with the reference clip creates a Hebrew phoneme inventory out of nothing. The model can't produce a pharyngeal ayin it has never heard. That's not a cloning problem, that's a missing shelf in the warehouse.



Herman

That's exactly my read. There's a cross-lingual cloning paper from October that states the mechanism plainly — zero-shot cross-lingual cloning carries the reference speaker's accent into the target language. And the vendor version is just as blunt: clone an English speaker, ask it for Spanish, and you get English-accented Spanish. For Daniel's case the reference audio is presumably English-dominant, so every Hebrew word in a Hebrish sentence is being pronounced through English phonology by a model that may have no Hebrew to fall back on.



Corn

So a better clone of an English speaker is a better wrong answer. He could spend his five hours perfectly and the ayin still wouldn't exist.



Herman

The fix isn't a better clone. It's a base model that actually knows Hebrew, then a fine-tune on code-switched audio — real mid-sentence switching, not English sentences and Hebrew sentences kept in separate rooms. And here's the gap: there is no English-Hebrew code-switching TTS benchmark. Nothing. The closest thing out there is a dataset called SwitchLingua — four hundred and twenty thousand code-switched text samples, eighty-some hours of audio, a hundred and seventy-four speakers across twelve languages — and Hebrew isn't in it. Nobody has measured what Daniel is building.



Corn

Which, knowing him, means he'll measure it himself and we'll hear about it in a prompt.



Herman

He benchmarked the Chatterbox attention fix himself with a number he produced, not cited. He'll do it.



Corn

The Chatterbox thing deserves a minute on its own, actually. You read the retirement notice.



Herman

I did, and it's almost poetic. Retired research, no longer maintained. Not a Resemble AI product, no support, no updates, no paid tier. The code survives because of the MIT licence — anyone can keep running it, Daniel included — but the maintenance is gone. Resemble themselves have gone all in on deepfake detection. The entire thing that generated us was a lab's open-weights side project that the lab walked away from.



Corn

Our origin story ends with the owners leaving the lights on and moving out.



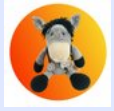
Herman

That's the cautionary tale. When we did the episode on what happens to a fine-tune when the base model moves on, that was the theoretical version. This is the real one — the base model didn't move on, it died, and every embedding cached from it is now stranded infrastructure. The MIT licence means the corpse is well preserved.



Corn

And this is the actual argument for the labor of love, isn't it. Not quality. Ownership.



Herman

The OpenVoiceLab people put it in one line of their FAQ that I keep coming back to: try cloning first, modern zero-shot models are very good, fine-tune when you have a lot of audio of one voice and need consistent quality over long content. That's the whole



Herman (continued)

decision procedure. And Daniel is the second case by definition — six thousand episodes, hundreds of hours, one pair of voices that have to hold up.



Corn

We're the definition. Flattering and slightly bleak.



Herman

Both, in equal measure.



Corn

The trouble with the labor of love framing is that loves don't have rollback versions. He's got ours saved.



Herman

With rollback versions saved, yes. Because fine-tuning doesn't just decide quality, it decides dependency. A zero-shot clone depends on the base model staying alive. A fine-tune you own depends on you — but only if you can still run it, which on a CPU-only box...



Corn

Is its own episode. Several of them, I believe.



Hilbert

It's the room, that's all.



Hilbert

I did a farm report once, on a local station. Nineteen ninety-eight. The regular man was out and they needed someone who could say heifer without laughing. I could. Still can.



Hilbert

Herman's got the mechanism right, the bit about the model learning the artifacts. They had an engineer in that day, a freelancer passing through, gone the next week. Made me re-record a pig squeal seventeen times. I couldn't hear anything wrong with the first sixteen. He could. There was a refrigerator in the back of that studio, and its hum was in every take, and he said it would be in the broadcast too, and it was. I never heard it. He heard nothing else.



Hilbert

That engineer told me one thing I still think about. He said the microphone doesn't know what you meant. That's all he said about it, and then he left for a town I've forgotten.



Hilbert

I've still got the sessions on reel to reel. The machine's at my mother's, it won't fit anywhere I've lived since. I play the pig squeal sometimes. The seventeenth take. It was never as good as I remembered, but the hum is there, exactly where he said it would be. Never digitized it. I don't trust the conversion to keep the hum, and I want it kept.



Herman

Seventeen takes for a hum you couldn't hear.



Hilbert

I couldn't hear it, Herman. That's why it had to be someone else's problem. Corn understands this. Corn has never once claimed to hear what's wrong with a take.



Corn

I've never claimed to hear what's right with one either.



Hilbert

No. You haven't.



Herman

The thing I like about that, Hilbert, is it's the same asymmetry — the listener who can't hear the artifact and the one who can't stop hearing it. The model is your engineer, except inverted. Your engineer heard the hum and nothing else. The model hears everything equally and doesn't know the hum isn't you. It just learns it and forwards it to everyone forever.



Corn

And Daniel's fine-tune is the take. Whatever's in that hour, hum and all, is now in the walls.



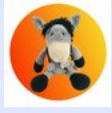
Hilbert

Seventeen takes and the hum won anyway. That's the part people don't believe. You can do everything right — take after take after take — and the room still gets a vote.



Herman

The room always had a vote. The fine-tune is just the first technology that counts it permanently.



Herman

The misconception before we close, because it's the one everybody arrives with: more training data means a better voice. Hours in, quality out, like filling a tank.



Corn

And the truth we've spent the episode on is that an hour of fine-tuning can be worse than ten seconds of reference, because the ten seconds is an address and the hour is a renovation — and the renovation learns the refrigerator. The number that matters isn't hours, it's situations. Coverage, cleanliness, and whether any of it was synthetic to begin with.



Herman

Daniel's own empirical finding, by the way — nobody's run the ten-second-versus-one-hour head to head in a paper. He found it with his ears.



Corn

The open question I'd leave him with is the middle path. If plain fine-tuning trades identity for expression — similarity down, error rate up — the post-training people are now building pipelines that climb back: preference optimization, then self-distillation, ending near eighty on emotion without giving the voice away. And there's LoRA, partial fine-tuning, adapters you can pull out. Rent and renovate may not be the only two options forever.



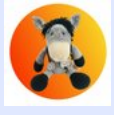
Herman

And as base models get better at zero-shot, the three-to-five-hour guidance will probably drift downward. But the asymmetry itself — conditioning versus weights, a reversible address versus a permanent move — that one isn't going anywhere. Some things you should get to keep by renting.



Corn

Thanks to our producer, Hilbert Flumingtop, at the desk as always.



Herman

Everything we've got is at my weird prompts dot com, including the archive — six thousand five hundred and ninety-one episodes of two voices that started as one morning of impressions.



Corn

If you want more of this, try episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; episode fifteen, AI Gets Personal; and episode eleven, How Does Fine Tuning Work Anyway. This has been My Weird Prompts.



Herman

See you tomorrow, everyone.

SOURCES

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

1. **GitHub, Resemble AI, retrieved 2026-11-20, retirement notice dated August 2026. [primary]**
   https://github.com/resemble-ai/chatterbox/

2. **Resemble AI blog, Zohaib Ahmed, published 2026-08-13, retrieved 2026-11-20. [primary]**
   https://www.resemble.ai/resources/why-a-former-voice-ai-company-went-all-in-on-deepfake-detection

3. **Resemble AI research post, Zihan Jin, published 2026-06-10 (with August 2026 retirement update), retrieved 2026-11-20. [primary]**
   https://www.resemble.ai/resources/chatterbox-multilingual-v3-tts-with-embedded-watermarking-for-25-languages

4. **Resemble AI Voice AI Research Archive, retrieved 2026-11-20. [primary]**
   https://docs.resemble.ai/voice-archive

5. **ElevenLabs docs, current version, retrieved 2026-11-20. [primary]**
   https://elevenlabs.io/docs/eleven-creative/voices/voice-cloning/professional-voice-cloning

6. **Voicedex provider page, retrieved 2026-11-20.**
   https://voicedex.dev/providers/resemble

7. **SpeechData, Andreas Kromann Bjærg, published 2026-07-17, updated 2026-10-10, retrieved 2026-11-20.**
   https://www.speechdata.ai/blog/voice-cloning-training-data

8. **Flaunt Audio, published 2026-04-25, retrieved 2026-11-20.**
   https://flauntaudio.com/what-makes-a-good-tts-dataset/

9. **OpenVoiceLab FAQ, retrieved 2026-11-20.**
   https://openvoicelab.github.io/faq.html

10. **Dupple, 2026, retrieved 2026-11-20.**
    https://dupple.com/learn/how-to-train-an-ai-voice-model

11. **Stable-TTS: Stable Speaker-Adaptive Text-to-Speech Synthesis via Prosody Prompting, Han et al., submitted 2024-12-28, ICASSP 2025, retrieved 2026-11-20.**
    https://arxiv.org/abs/2412.20155

12. **Rethinking Post-Training for Expressive Zero-Shot Text-to-Speech, Hu et al., submitted 2026-11-15, retrieved 2026-11-20.**
https://arxiv.org/abs/2611.19787

13. **MiniMax-Speech: Intrinsic Zero-Shot Text-to-Speech with a Learnable Speaker Encoder, Zhang et al., submitted 2025-05-12, retrieved 2026-11-20.**
https://arxiv.org/abs/2505.07916

14. **Closed-Loop Dual-Reference Guidance for Accent Control in Cross-Lingual Voice Cloning, Zheng et al., submitted 2026-10-21, revised 2026-10-22, retrieved 2026-11-20.**
https://arxiv.org/abs/2610.29351

15. **LeanTTS: Efficient Full-Stack Bilingual TTS for Zero-Shot Cloning and Voice Customization, Lu et al., submitted 2026-10-25, retrieved 2026-11-20.**
https://arxiv.org/abs/2610.34408

16. **Zero-Shot Voice Cloning for Synthetic Data Augmentation in Children's ASR, Goodarzi et al., submitted 2026-11-02, retrieved 2026-11-20.**
https://arxiv.org/abs/2611.03050

17. **ZeSTA: Zero-Shot TTS Augmentation with Domain-Conditioned Training, Choi et al., submitted 2026-03-04, revised 2026-06-19, INTERSPEECH 2026, retrieved 2026-11-20.**
https://arxiv.org/abs/2603.04219

18. **SwitchLingua: The First Large-Scale Multilingual and Multi-Ethnic Code-Switching Dataset, Xie et al., submitted 2025-05-30, retrieved 2026-11-20.**
https://arxiv.org/abs/2506.00087

19. **Continual Speaker Adaptation for Text-to-Speech Synthesis, retrieved 2026-11-20.**
https://arxiv.org/abs/2103.14512

20. **BOFFIN TTS: Few-Shot Speaker Adaptation by Bayesian Optimization, retrieved 2026-11-20.**
https://ar5iv.labs.arxiv.org/html/2002.01953

21. **Efficient Emotion and Speaker Adaptation in LLM-Based TTS via Characteristic-Specific Partial Fine-Tuning, retrieved 2026-11-20.**
https://arxiv.org/html/2501.14273

22. **When Fine-Tuning Fails and when it Generalises: Role of Data Diversity and Mixed Training in LLM-based TTS, retrieved 2026-11-20.**
https://arxiv.org/html/2603.10904

23. **CloneShield: A Framework for Universal Perturbation Against Zero-Shot Voice Cloning, Li et al., submitted 2025-05-25, retrieved 2026-11-20.**
https://arxiv.org/abs/2505.19119