Episode Book

EPISODE #6602

Training Voice LoRAs on Filtered Zero-Shot Audio

PUBLISHED	14 November 2026 at 19:18 UTC
DURATION	19:44
LISTEN	myweirdprompts.com/episode/omnivoice-lora-voice-adapters-98833/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	omnivoice
PIPELINE	V5.27
BOOK GENERATED	15 November 2026 at 11:16 UTC

SYNOPSIS

Six podcast voices now run on LoRA adapters trained for under two dollars each on OmniVoice. By generating synthetic lines with zero-shot cloning and using Whisper to discard every mispronounced take, the pipeline distills consistent speech and reliable tech acronyms without a full fine-tune. The catch is that an adapter cannot exceed its teacher: when Whisper cannot verify pronunciation, such as Hebrew terms in English sentences, errors pass straight into the weights, showing where cheap compute still needs a human ear.

TRANSCRIPT



Corn

...and that's the strange part of listening back to the test renders. There's no recording behind this voice at all. Nobody handed it a clip to imitate.



Herman

Right, it's baked in. And that's today, Corn, because this is the first episode where all of us talk that way. Everyone who speaks on this show now has a LoRA adapter on OmniVoice, six voices in all, Daniel's own included, and none of them is given a reference clip when it speaks.



Corn

So if you've been listening and something sounds different, you're not imagining it. Tell us how we sound now compared with before. Better, worse, odd in a specific place. We read all of it, and a vague "something's off" is still useful.



Herman

Specific is better, though. "Herman swallows the ends of sentences" gives us something to work with.



Corn

He doesn't, I've checked. Now, Daniel's question is the reason we're getting into the mechanics. He wants to know how much you can actually teach a speech model with a fine-tune, and which kind of fine-tune does the teaching. A light one, a middle one, a heavy one. What each buys you, what each costs, and where this show's voices landed.



Herman

Two callbacks first, and I'll keep them short. Earlier today we did the episode on how OmniVoice cloned our voices from our old AI. That was the zero-shot debut: one short clip per character, nothing trained, and the model imitates it on the fly. Whatever you heard in that episode is what zero-shot sounds like. And the other one was Daniel's question this morning, whether a twenty-minute fine-tune can stabilize a cloned voice. He asked it before any training had happened.



Corn

And this is the answer. We did it.



Herman

We did it, yes.



Corn

Walk me through how, though. Because I assume it wasn't a microphone and a quiet afternoon.



Herman

It's almost the opposite. For each voice, the text comes from three places: the character's real lines from published transcripts, some lines written fresh in character, and, for the two of us, about four hundred sentences stuffed with tech and AI acronyms and medical and mental-health vocabulary. Those came from Daniel's own read-aloud sentence set.



Corn

And the audio?



Herman

OmniVoice itself. Zero-shot, from each character's reference clip, two takes per line. So the training data is the old voice's own output.



Corn

Hold on. That's a voice learning from its own impression of itself. A copy of a copy. How does that not just average out into mush?



Herman

Because of the filter, and it's what makes the whole thing work. Whisper, the speech recogniser, transcribes every take. If it doesn't come back word for word, the take is thrown out. Of Herman's four hundred vocabulary sentences, two hundred and eighty-one survived. For the rest, neither take read every word right.



Corn

So I'm a copy of a copy that's been proofread.



Herman

Proofread by a machine that can't be flattered. Then a restoration model cleans each clip, and a gentle warm EQ and compression chain goes in. Daniel picked that one by ear, and it's in the training audio itself. Training is a few hundred to eight hundred steps on one data-centre GPU, five to ten minutes a voice, well under two dollars each including generating the audio.



Corn

Two dollars. I've spent more on a sandwich I regretted.



Herman

There are two wrinkles. Dorothy and Larry needed personality, so for each line the take kept was the liveliest one for Larry and the flattest, slowest one for Dorothy. And Daniel's adapter is the exception: that one is trained on about forty minutes of his real recordings, no synthetic audio.



Corn

So he's the only one of us with an actual person underneath.



Herman

He is. And each adapter is about a hundred megabytes, published openly on Hugging Face with samples, non-commercial licence, inherited from OmniVoice's weights. One base model serves all six voices and the pipeline switches adapter per speaker.



Corn

Right, so, that's the new voices. The question underneath is why that, and not the other two. Three grades, as Daniel puts it. Lightest first.



Herman

Zero-shot cloning changes nothing in the model. A reference clip conditions each generation, so the model hears a few seconds of someone and carries on in that voice. No training, any voice, instantly. The cost is that it copies everything about the clip, quirks and pacing included, and it can wobble. One-word turns are where it shows, and it varies from run to run.



Corn

Daniel's take on this, and I want to get it right, is that zero-shot is reliable and good in his experience. He's not complaining about it.



Herman

No. But the show is going to generate hundreds of hours of audio, and at that scale he wanted a few things zero-shot doesn't promise. Single words that come out properly. The same voice every time. Consistency from one episode to the next. And, in his words, a bit deeper engineering over the model than a reference clip. So he went for adapters.



Corn

Fair. A reference clip is a sticky note on the model. An adapter is a change to the model.



Herman

That's close, with one correction: a LoRA adapter leaves the base model frozen. It adds small low-rank matrices to some of the layers. Here that's about three percent of the parameters, plus the layers that take audio in and put it out. Minutes to an hour of data, minutes on one GPU. And because the voice is in the weights, it needs no clip, it's steadier on short lines, and you can swap voices on a single base.



Corn

Which is the thing I'd poke at. Every adapter is trained on what the base model already does. What if the base model has a bad habit? You've distilled the habit.



Herman

You've distilled the best of what it does. The filter is why, and it matters for everything in this episode. The adapter can only learn what the teacher can say at least some of the time and the filter can verify. If OmniVoice gets an acronym right in one take of two, Whisper confirms it, and the adapter learns the right reading every time. It turns "usually right" into "reliably right".



Corn

But it can't exceed the teacher.



Herman

It can't. It can't invent a pronunciation the teacher never produced.



Corn

So the ceiling is the base model, and the floor is raised. Comforting, if the ceiling's good.



Herman

And the third grade is the full fine-tune. You change all the weights. The deepest change available: new languages, new phonetics, styles the base can't do at all. It also needs typically hours of clean data, hours to days of training, and a full copy of the model per



Herman (continued)

voice. And it can forget things the base knew.



Corn

Which is a poor trade if the voices were already inside the base model's reach.



Herman

They were. OmniVoice could clone all of them zero-shot. The goal here was consistency and robustness, not a new capability. A full fine-tune is for when the base genuinely can't do the thing. And how's voice cloning done in the wild, Corn? Do you know?



Corn

Instant cloning in the commercial products is zero-shot, I'd guess. And the pricey "professional" ones are fine-tunes of some kind.



Herman

Right, and increasingly adapter-based, for exactly the reasons we've said: cheap to train, cheap to store per voice. A multi-voice product is often adapters on a shared base. So we're doing what the industry does, just with a more eccentric cast.



Corn

Show me what actually transferred, then. Because "it sounds more consistent" is the thing you'd say about any fine-tune.



Herman

There's a framing I like. For speech recognition, a fine-tune is mostly about domain vocabulary recognition: the model learns to hear your jargon. For text-to-speech, what we got was domain vocabulary pronunciation. And the acronyms came through remarkably. That was Daniel's word on Herman's adapter, "remarkable". GPU, API, JSON, VRAM, read reliably.



Corn

Which tracks, given the filter. The teacher could say them, Whisper could check them.



Herman

Yes. And then Hebrew words in English, which didn't transfer. A first version was trained on Hebrew terms, with audio rendered from Hebrew script with vowel points. It improved the words it had seen. A word it hadn't seen, "mechina", still came out the English way. So the adapter can carry over the acronyms, but the Hebrew phonetics just will not come along for the ride.

L

Larry

Friends, does your voice sound like it was cloned by a machine that never met you? Then ask your audio professional about AdaptoVoice, the little adapter that snaps onto your personality and transfers what matters. AdaptoVoice learns your acronyms, your lingo, your very soul, and promises zero-shot results so convincing your own mother will call you by your full fine-tuned name. Side effects may include mild adapter slippage, sudden fluency in acronyms you never learned, a lingering sense that your phonetics have been quietly improved, spontaneous zero-shot cloning of nearby relatives, persistent Hebrew vowels arriving uninvited, and in rare cases a full fine-tune of the personality, leaving you forty percent more resonant and unable to pronounce your own address.



Corn

Unable to pronounce your own address.



Herman

Please don't do that again, Corn. Not in that voice, and not at that speed. So, "mechina" came out English, and what I'd like to know is whether you think the adapter was the problem. Where would you put the blame?



Corn

On the adapter, at first. It's the thing we trained. But if it learned acronyms perfectly, it can learn. So the data was wrong.



Herman

Wrong, and in a way nothing in the pipeline could catch. Here's the chain. Hebrew terms got written in Hebrew script, with vowel points, and handed to OmniVoice to speak. In a blind rating in a sister project, that input got one or two terms out of eleven right. So the teacher was mostly saying them wrong, and the training audio was mostly wrong.



Corn

And the filter?



Herman

The filter is the part that works for acronyms, and it can't work here. Whisper checks the words that come back. It can tell you "GPU" was said as GPU. It can't tell you whether a Hebrew name inside an English sentence was said the way someone in Jerusalem would say it. Nothing could verify it, so nothing was thrown out, and the adapter faithfully learned the mistakes. Not all of them. It improved the words it had seen and nothing else.



Corn

So it learned a lookup table, not a rule.



Herman

Close to that. The stumbling block is phonetic input. What goes in as text decides what comes out as sound, and if the text-to-phoneme step is wrong, no amount of tidy adapter fixes it. When hand-written phoneme spellings went in instead, the same eleven terms scored seven right and two close.



Corn

Which is a person writing out the sounds by hand.



Herman

For now. It's a better problem than a model that's mostly wrong, though. A person can write eleven terms down. Nobody can fix a model that's confidently wrong across thousands of words.



Corn

The recognition versus pronunciation split is tidier than it sounds. Because a recogniser that learns your jargon gets a score you can compute. Did it hear "Kubernetes" or didn't it. Speech output has no equivalent score.



Herman

Except the one we built. Run it through Whisper, count how many words come back. For common vocabulary it works as the scorer and the filter. For anything Whisper itself can't judge, you're back to a human ear.



Corn

A human ear that costs more than the training run. Two dollars a voice for the GPU, and then some afternoon for somebody to listen.



Herman

That's roughly right, and it's the real cost structure of this. Cheap to compute, expensive to judge. Which is why the adapter route made sense for the show. Judging was already happening. Daniel picked the warm EQ by ear. Larry's takes were chosen for liveliness, Dorothy's for flatness and slowness.



Corn

Let me push on the circularity again, because I'm not done. If the voice is distilled from its own zero-shot output, then everything the zero-shot voice did badly is in the teacher. Odd rhythm, odd stress. The filter only checks words. It doesn't check whether I sound like a sloth with something to say.



Herman

Fair, and the answer is yes, mostly. Word accuracy is what the filter guarantees. Prosody is only as good as the teacher's best take of the two, plus the by-ear choices. The gain is that you've got the good take every time instead of one time in two. For a show that renders hundreds of hours, one bad reading in a hundred is a lot of bad readings.



Corn

And single words.



Herman

Single-word turns were exactly where zero-shot wobbled. Short lines give the reference clip very little to settle on, and it drifts. An adapter has the voice in the weights, so a one-word line gets the same voice a long line does. That's the robustness Daniel was after.



Corn

Here's what bothers me, though. Variation isn't all noise. When a voice wobbles between runs, some of that wobble is life. Pin it down and I could end up sounding the same in a joke as in a eulogy.



Herman

That's a real cost, and I'd rather say so than wave it off. Consistency and liveliness pull against each other. What keeps it from going flat is what you train on. Real lines from published transcripts, lines written in character, and for Larry the liveliest take kept every time. You choose the variation you want, then freeze it. Daniel's adapter stands on different footing from the rest. It's trained on his real recordings, about forty minutes of them, so its ceiling isn't a model's best take. It's a person.



Corn

The only one of us whose upper limit is a human being. I'll try not to resent it.



Herman

Sloths have a long tradition of resentment, I'm told.



Corn

Mongolian sloths especially. Then what's the downside? There's always one. I'm suspicious of anything that costs two dollars.



Herman

Three, I think. It's bounded by what the base model can already do, so you can't ask it for a sound it's never made. It can overfit if the data is tiny, like a voice that only knows the sentences it was trained on. And there's one more moving part in the pipeline, because it has to load the right adapter per speaker. That part is manageable. One base, six small files.



Corn

And the full fine-tune would've cost what, in this case?



Herman

A full copy of the model per voice, hours of clean data per voice, and a run that takes hours to days instead of ten minutes. For voices the base model could already produce. You'd be paying to move the whole house when the job was rearranging a room.



Corn

I'm not sure the house-moving analogy is earning its keep, Herman.



Herman

It isn't, you're right. The simpler version is that a full fine-tune can change what the model is able to do, and an adapter can only change what it tends to do. We wanted the second.



Corn

Which is the version I'd put on a mug. And if the Hebrew goes the way it looks like going, that's where it flips. An adapter changes tendencies. A sound the base model has no route to producing is a capability.



Herman

Yes. Though I don't know yet which side Hebrew words in English fall on. If the base can produce the sounds and only the input is wrong, phoneme spellings fix it and an adapter makes the fix stick. If the sound isn't in the base at all, you're in full fine-tune territory, and you'd want audio verified by a native speaker.



Corn

Same missing ingredient again.



Herman

Verified audio. Every road comes back to that.



Hilbert

Teacher's the wrong word. A teacher knows the answer. What you've got is a fellow who gets it right now and then and nobody's told him which times. You said one take in two, Herman. I had one of those in nineteen eighty-five. A speech box, two hundred and forty pounds, plugged into the back of the computer. It read anything you typed. Anything but Llanberis Close, which is where my brother-in-law lives. It said it like a man clearing his throat.



Herman

You spell the sounds for those, don't you? Not the letters. So you're typing a pronunciation, and it says whatever that comes to.



Hilbert

You write it how it sounds. I called the sheet the crib. First crib, wrong. Second crib, worse. I did eleven. They're in pencil, in a ring binder, with a separate crib for the village.



Corn

Somebody had to stand there and listen to each try and say warmer or colder.



Hilbert

Dennis stood there. Wouldn't say warmer. Said they all sounded the same to him. It was the ninth one. Dennis has the box.



Herman

Eleven cribs and a man who can't hear the difference. That's the verification problem in one living room. The crib is only worth something once somebody has agreed which of the eleven was right.



Corn

The ninth, apparently. Which nobody knew until it was over.



Herman

Whoever writes the crib decides what the voice says.



Corn

Something that didn't fit earlier, though. For Dorothy and Larry, the reference clip was the most in-character of sixteen renders. Sixteen takes of one voice, and someone listened to all of them to pick the gloomiest.



Herman

And with Dorothy the line-by-line choice was the flattest and slowest take, every time. An adapter trained on that doesn't have to be told to be dismal.



Corn

Dismal on request, and consistent. That's more reliable than most people I know.



Herman

The next step for the Hebrew terms is the crib, then. Hand-written phoneme spellings, an ear on each one to confirm it, and then, if it holds up, an adapter to make it stick. We'll find out soon enough whether the base model can say the sounds at all.



Corn

Thanks to Hilbert Flumingtop for producing this one, which meant getting six voices through with no clip to lean on.



Herman

This has been My Weird Prompts. If the new voices work for you, or don't, a review is the easiest way to tell us, and we'd like to know either way.



Corn

See you tomorrow.