Episode Book

EPISODE #6194

Teaching a TTS Model His Own Accent

PUBLISHED	22 October 2026 at 15:54 UTC
DURATION	21:53
LISTEN	myweirdprompts.com/episode/hebrew-english-tts-fine-tune-75935/
HOSTS	Corn, Daniel, Herman, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.19
BOOK GENERATED	23 October 2026 at 07:44 UTC

SYNOPSIS

What does it take to make a machine say Hebrew words the way an English speaker in Jerusalem actually does? Daniel fine-tuned Qwen3 TTS on his own voice to build Meorav, a model named after Jerusalem mixed grill, and the voice you hear from him in this episode is that model speaking. Corn and Herman run on Chatterbox, the show's usual engine, so this is a side-by-side test of cloning versus full fine-tuning. The conversation covers why zero-shot clones keep the timbre but borrow an American cadence, why a phonemiser with no Irish-English voice closed one door, and why spelling every Hebrew word one single way turned out to matter so much.

TRANSCRIPT



Corn

What does it take to teach a machine to say Bituach Leumi the way somebody who actually lives in Jerusalem says it? That's the heart of Daniel's prompt. He trained a text-to-speech model on his own voice, and the voice you hear every time he speaks in this episode is that model. Daniel, you're in the room.



Daniel

I am, but I'm technically not speaking at all. I typed every word of this, and a model called Meorav is saying them. I'm the only person in the studio who isn't talking.



Herman

Which makes this a side-by-side. Corn and I run on Chatterbox, the engine the show uses for all of us. Daniel runs on Meorav. And one thing worth saying early, because listeners will wonder. The Daniel voice you're hearing is version two of Daniel's voice model. The second Meorav. The first was an experimental proof of concept, and it's already published on Hugging Face under the Hebrish organisation as TTS V one.



Corn

Version two. So version one is the one we're not meant to judge too harshly.



Herman

It ran end to end, and that's what a proof of concept is for. Version two is the one Daniel wanted to put next to the two of us.



Corn

Fine, but go back a step. Meorav. It sounds like something you'd order at a counter.



Daniel

It is, sort of. Meorav is Hebrew for mixed, as in meorav Yerushalmi, the Jerusalem mixed grill. Different cuts, one grill, everything touching. I liked that for speech that mixes English and Hebrew in the same sentence.



Herman

And it lives inside Hebrish, which is Daniel's one-person not-for-profit on English and Hebrew code-switching. The speech recognition models there are named after Israeli food too. Sabich, Jachnun, Shakshuka.



Corn

Of course the recognition models are breakfast. What's code-switching look like in practice?



Daniel

It's how English speakers talk here. You're halfway through an English sentence and a Hebrew word walks in and nobody blinks. Yalla, tachles, Bituach Leumi wants the form by Thursday.



Herman

That's the sentence the whole project is built around. Irish-accented English, because that's Daniel's accent, with the Hebrew words said the way he says them, in the middle of the sentence and not as a separate performance.



Corn

So why not just clone yourself from a clip, like everybody else? We've done cloning from a few seconds of audio on this show, and it works.



Daniel

It works, and I tried it. Twice, actually. An Orpheus fine-tune and a zero-shot clone. Both came out American. It sounded like me if I'd spent a summer in Ohio.



Herman

There's a mechanism behind that, and it's worth a minute. A zero-shot clone copies timbre, the colour of the voice. But the rhythm, the stress, where a sentence rises and falls, all of that comes from the base model, and the base model learned mostly from American speech. So you get your voice reading in somebody else's cadence.



Corn

Your voice wearing a stranger's coat.



Herman

And the show has its own example. Daniel's voice already exists here as a Chatterbox clone, and it reads the AI disclaimer at the top of every episode. A single-shot clone, so the timbre is his, and the cadence is less Irish than the real thing.



Daniel

I can confirm that. Hearing yourself read a legal disclaimer in a mild American accent is a particular kind of unsettling.



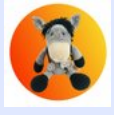
Corn

Worse than hearing yourself normally?



Daniel

Different. Normally I hate the pitch. With that one I hate the vowels.



Herman

Corn, you've used the clone for a lot of things. Does it sound like him to you, or does it sound like a very polite American doing an impression of him?



Corn

The second one. It's missing the lilt. Daniel says the same sentence and it goes up at the end like a question, and the clone lands it flat like a receipt.



Herman

That's the prosody coming from the base model. The fix is to stop borrowing and change the model itself. A full fine-tune updates every weight, so it can move off the American prior rather than being nudged around it.



Corn

You said that like it's settled. Is it?



Herman

No. I haven't seen a published open text-to-speech fine-tune that holds a non-American English accent. Not one. The project's own notes say so. Whether it works is the thing being tested.



Daniel

It might not. That's the project.



Corn

A bet that admits it's a bet. What did you build it on?



Daniel

Qwen three TTS, the one point seven billion base model, from Alibaba's Qwen team. It's Apache licensed, so I can fine-tune it and publish what comes out.



Herman

Qwen three TTS, twelve hertz, one point seven billion parameters. And the detail that made it the pick is that it reads raw text. There's no phonemiser in front of it. Most speech systems turn your sentence into phonemes first, and the usual tool for that is espeak, and espeak has no Irish-English voice. So that front door was already shut.



Corn

And the tokenizer just eats the spelling.



Herman

It eats the spelling. Which means the spelling of a Hebrew word matters enormously. One spelling per word, so the model learns that this exact string means Daniel's pronunciation. He keeps a canonical list for that, shared with the recognition work.



Corn

How many ways are there to spell tachles?



Daniel

More than you'd think. Takhles, tahles, tachles. People write all of them in messages. The list picks tachles, and every sentence I record uses that one.



Herman

Corn, would you have guessed that one?



Corn

I'd have guessed the h and been confidently wrong. Same as everything else.



Daniel

That's also why the sentences look odd. Numbers are spelled out, the way you say them. Two thousand and twenty-six, half four, twelve euro fifty. The model maps text to sound, so the text has to say what I said.



Corn

Which is why nobody in this script has written a digit. I've been told.



Herman

Corn, can I do the training script? I've been holding it since the opening.



Corn

Go on. I can see it in your ears.



Herman

Qwen ships a fine-tuning script, and it has bugs. Daniel patched seven things. The model was learning one frame too far ahead. It was predicting codes from a hidden state that had already seen them. A projection layer was skipped in training and then used at inference. And in half precision, plain weights lose most of the small updates at these learning rates, so he trained with full-precision master weights.



Corn

The last one sounds like my week. Small effort, then it vanishes.



Herman

When Daniel last checked, none of the fixes had been merged upstream.



Daniel

I'd love to say I found them cleverly. A lot of it was people on the issue tracker saying this looks wrong, and me saying yes, it does.



Corn

What does one of these runs cost?



Daniel

The short ones are about a dollar. One rented GPU on Modal, a few passes over the data at a small learning rate. The first proof of concept was three passes.



Herman

Which is why starting small was sensible. You learn whether the whole pipeline works for the price of a coffee, before you give up a weekend to recording.



Corn

It depends on the coffee.



Daniel

In Jerusalem it depends on the neighbourhood.



Corn

Where does the audio come from? Eight hours of the phone book?



Daniel

A recording app on my phone. Raw audio, forty-eight kilohertz, twenty-four bit, browser voice processing off. And it has a voice-clone mode, which means a quiet room.



Herman

That's the clever part. The same app also collects speech recognition data, and that wants the opposite. Recognition wants an air conditioner and a crying baby. A voice model wants silence. So only takes recorded in voice-clone mode, with no noise flags, ever reach Meorav's training set.



Corn

A one-way door.



Herman

A quiet take can help the recognition side. A noisy recognition take never trains the voice.



Corn

And what do you read?



Daniel

Sentences. A phonetic set so every English sound turns up. Conversational ones in my own register, with a bit of Hiberno-English, because reading aloud pulls you towards a neutral reading voice. A big set of Hebrish sentences covering the common Hebrew words. A set for names and places, Yerushalayim, Chevron, that kind of thing. And a personal set.



Herman

The personal set has the show in it. Daniel's name, Hebrish, Herman Poppleberry, Corn.



Corn

So he read my name into a microphone and nobody asked me.



Daniel

You weren't available for comment.



Corn

I'm a sloth. I'm always available. I just answer slowly.



Herman

Version two is more of the same, and a lot more of it. Eleven hundred and three takes, about eighty-eight minutes after trimming, against fifty minutes for the first. The longer plan is two and a half to three hours, with a different kind of recording in it. Free speech, where he talks unscripted. That's meant to be the strongest lever on accent, about a third of the final hours.



Daniel

Reading aloud makes you sound like a man reading aloud. When I just talk, the Irish comes back.



Corn

What does Irish even mean to a model? Where does the accent live?



Herman

In small things. The project audits its recording sentences for them. The th in this and thing, the r after a vowel, the diphthongs, the ing endings, and the t. Irish English as a rule keeps the r, softens the t, and makes some vowels purer than an American would.



Corn

Diphthongs. A sloth hears that word and goes to sleep.



Daniel

A read sentence makes me self-conscious, so I enunciate like a newsreader. Free speech catches the sloppier, more Irish version.



Corn

Daniel, what's it for? Besides this episode.



Daniel

Mostly me. News briefs I prepare for myself, read in my voice. Narration. Anything where I'd otherwise have to sit down and read a long text out loud, and I'd rather not. The disclaimer, maybe, once it's good enough.



Corn

Why your own voice and not a hired actor's?



Daniel

Because I can record as much as I like, for free, and nobody else's consent is involved. I can publish the recordings because they're mine.



Herman

That sidesteps the question everyone else in this field is stuck on, whose voice it is. Here the answer is one man, and he's in the room.



Corn

Herman, a real question. We're Chatterbox and we're fine. Why would anyone want a model that needs hours of one man's voice?



Herman

Because Chatterbox is a general engine. Give it a short reference and it speaks in that timbre, fast and cheap. A whole episode runs on the order of fifteen cents of GPU. But it reads romanised Hebrew as English. It sees the letters on the page and reads them the way an English reader would. Chatterbox was actually on Daniel's fallback list, behind CosyVoice three.



Corn

Is there anything we do better than Meorav?



Herman

Plenty. Chatterbox takes a reference clip and works the same afternoon, in nearly any voice. Meorav needs recordings, a training run, and a model on a disk, and then it only does one voice. Daniel's trade is breadth for depth.



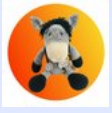
Daniel

A clone is a photograph of the voice. The fine-tune is trying to learn how I move.



Corn

So I'm the plan C.



Herman

You're plan C for Daniel and plan A for everyone who just wants a donkey and a sloth.



Corn

Herman, I've got one for you. If Meorav nails the Hebrew and loses the Irish, would you call that a win?



Herman

A half win. Daniel would call it a Hebrew speaker with an American accent, and he's heard that one already from the clones. I'd want both.



Corn

Hm. So now we find out. Do we do the demo?



Herman

We should do the demo, and I'd like it clear what we're doing. Daniel says a few everyday lines, the way English speakers here really talk. Then Corn and I try the same lines, and listeners compare.



Corn

And I'll say now that it may go badly for us.



Herman

It may go very badly. I'm from Connecticut. You're from, well.



Corn

Mongolia. A very real place, and I take offence.



Herman

My guess for tachles is a tack followed by a less. That's what the letters say to an English reader.



Corn

Tackless. I've been called worse.



Daniel

Right, line one is the one from the project page. Yalla, tachles, Bituach Leumi wants the form by Thursday.



Herman

Yalla is let's go. Tachles is get to the point.



Daniel

Tachles is more the bottom line. And Bituach Leumi is the National Insurance Institute, the office you can't ignore.



Corn

I'll go first. Yalla, tachles, Bituach Leumi wants the form by Thursday.



Herman

Brave, Corn, or at least confident. I will leave the judging to the listeners.



Corn

Daniel, give us another before I lose my nerve.



Daniel

Sababa, I'll sort the arnona on Sunday, but the office is a total balagan.



Herman

Sababa is fine, great, no problem. Arnona is the municipal property tax. Balagan is chaos.



Daniel

Balagan is a mess, but it's the specific mess you get when nobody's in charge.



Herman

My turn. Sababa, I'll sort the arnona on Sunday, but the office is a total balagan.



Corn

You looked proud of yourself.



Herman

I'm a donkey. It's the only face I have.



Daniel

Next one. Ma nishma, are we doing a tiyul on Friday, or is it going to be a chamsin?



Corn

Ma nishma is what's up. Tiyul is a hike. Chamsin is a heatwave, the dry wind.



Daniel

That's it. Ma nishma is how you'd start with someone at the bus stop.



Corn

Ma nishma, are we doing a tiyul on Friday, or is it going to be a chamsin?



Herman

I'm going to say something kind, Corn. You committed.



Corn

I'm a sloth. Committing is all I do. It's moving I struggle with.



Daniel

Two more. Take kvish echad, there's no traffic after motzei shabbat.



Herman

Kvish echad is Highway One, the road up to Jerusalem. Motzei shabbat is Saturday night, once Shabbat is out.



Daniel

And this one's a city. Yerushalayim in the rain is sababa, honestly.



Herman

Yerushalayim, which is Jerusalem, the long way round. Yerushalayim in the rain is sababa, honestly. I honestly can't tell from the inside whether that was good or bad.



Corn

Neither can I. That's what the listeners are for.



Herman

What I'd listen for, if I were a listener. The ch in tachles and Bituach, which is a throat sound and not the ch in church. Whether Yerushalayim keeps all its syllables. And whether the stress lands where a Jerusalemite would put it.



Corn

That's a lot of ch for a sloth.



Daniel

One more, because I've been saying nice easy ones. Slicha, I'm walking through Shuk Hacarmel to get tomatoes for the shakshuka.



Corn

Slicha, I'm walking through Shuk Hacarmel to get tomatoes for the shakshuka.



Daniel

I'll say one thing. I picked those because I say them all day, which is exactly why they make a weak test. A model can learn the lines it trained on. The harder test is a Hebrew word I haven't recorded.



Herman

That's the out-of-distribution question. The project holds out sentences the model never trained on, and some lines made up to be strange, and compares the base model, the fine-tune and the real recordings.



Corn

And who judges?



Daniel

My ears, and other people's. There are no metrics here. It's evaluated by ear.



Herman

Which is right. Loss going down tells you the model is learning something. It doesn't tell you it sounds like Daniel. Version one's loss came down and every sample stopped cleanly at about the length of the real takes. Nobody has said how it sounds. That verdict was still pending.



Corn

A voice that knows when to stop. Unlike this show.



Herman

Unlike your tangents. Corn, what do you think version two has to prove? Be honest.



Corn

That the Irish survives. And that the Hebrew comes out right in the middle of an English sentence. They may fight each other, Hebrew vowels being a long way from Irish ones. If it picks one, I want to know which.



Herman

I'd bet on the Hebrew. There's a lot of Hebrew in the training sentences. The accent is the harder one, and it's one speaker, about an hour and a half of read sentences this time, and version one was about an hour.



Corn

Daniel, what makes you nervous about this?



Daniel

That I'll be the worst judge in the room. Nobody hears their own voice from outside. I'll need other people's ears, and I don't have a number to hand them.



Hilbert

Fifty minutes. Six hundred and fifty-seven takes, fifty minutes after trimming. Version two, eleven hundred and three takes, eighty-eight minutes.



Corn

Fifty. Noted.



Hilbert

Corn, you were right about the flat landing. The clone says the line like a receipt. A six-second clone does the job for nearly everything. An hour and a half in a quiet room is a lot of effort to get an accent right.



Daniel

For most uses a clone is the right answer. I'm not saying anyone else should do this.



Hilbert

The licence is non-commercial. CC BY-NC. So the show can't sell it and the recordings are public. Anyone can download your voice. The people who'd misuse it don't read licences.



Daniel

True, and I knew that when I published them. The alternative is me deciding who gets to read my own recordings, and I like that less.



Hilbert

Two licences on the one voice, too. The base model is Apache. What Daniel made from it is non-commercial. Tennant would have wanted a form for each. And once it says whatever you type, it says whatever anyone types. That one I'd take up with the council.



Corn

The council?



Hilbert

Nineteen seventy-three. I did the lift announcements at the Drumlish District Records Office. Third floor, births. Fourth floor, deaths. Fifth floor, the canteen, which was a cupboard. Recorded on a reel of the Corporation's tape, in the Corporation's time, so the Deputy Registrar, a Mr Aloysius Tennant, ruled the voice was Corporation property.



Herman

So you were a recording on their tape.



Hilbert

I needed a licence to say the word floor in public in the county. Form eleven B. Two pence a floor. Mr Tennant filled it in for me, the pen not being made for claws, and kept the carbon.



Corn

Two pence on five floors is ten pence a day. Before you've said anything else.



Hilbert

Twelve. The canteen was four. He said it was a different floor for tax purposes. Then he leased the reel to Ballyknock for their lift. Two floors. A goat on the payroll as second signatory, Bernadette, with authority up to forty pounds. For eleven Thursdays my voice announced a floor that wasn't there. A man went up in it looking for births and got a field.



Herman

That's what happens when you apply a model outside what it was recorded for. It had no data for a floor that doesn't exist.



Hilbert

It had the word floor. That was all it needed.



Corn

And the goat signed off on it.



Hilbert

Twice. Then I appealed to the Surveyor's Department, which ruled the voice a fixture, like a radiator, and sent a man with a spanner.



Herman

In property terms a fixture stays with the building when it's sold. So by that ruling, your voice came with the lift.



Hilbert

Nine pounds, with the handrail. Tennant had it on the inventory between the fire bucket and the clock. The man with the spanner was a Mr Dolan. He listened to me say floor, four times, and signed it as tightened.



Corn

A voice that needed tightening.



Hilbert

I did it again myself on a Saturday. One take, nine minutes, no Corporation tape. Mr Tennant played the new one in the lift for eleven years and signed for it every Monday.



Herman

Nine minutes. Version one was fifty minutes of recordings and it ran. Nine minutes and a Saturday is a start.



Corn

A voice nobody could keep, replaced in nine minutes by the man who owned it. Daniel's version two had an hour and a half of him reading. The next one is where we find out how the Irish holds up when he's talking and not reading.



Herman

Even a good version only settles one speaker. Whether it generalises is a question for version three.



Corn

One problem at a time. Thanks to Hilbert Flumingtop for producing this one.



Herman

This has been My Weird Prompts. If you want more, find us at my weird prompts dot com.



Corn

See you tomorrow.