# Episode Book

EPISODE #6627

# Separating Voice From Pronunciation in a TTS Model

| | |
|---|---|
| PUBLISHED | 16 November 2026 at 13:41 UTC |
| DURATION | 18:58 |
| LISTEN | myweirdprompts.com/episode/tts-voice-and-pronunciation-55161/ |
| HOSTS | Corn, Herman, Dorothy, Hilbert |
| SCRIPT MODEL | anthropic/claude-sonnet-5.5 |
| VOICES | omnivoice |
| PIPELINE | V5.33 |
| BOOK GENERATED | 16 November 2026 at 22:22 UTC |

## SYNOPSIS

Can a text-to-speech model keep a speaker's Hebrew pronunciation while dropping their Irish accent? Over three days building Meorav, an English-with-Hebrew model, synthetic phonemes failed on gutturals and Seed-VC voice conversion leaked the original accent's rhythm. A forty-clip x-vector pilot revealed that pronunciation lives in the model weights, separate from speaker identity. By splitting off Meorav-Daniel and using masked multi-speaker fine-tuning with OmniVoice's A three voice, the model kept the Hebrew intact.

# TRANSCRIPT



### Corn

Three days, one text-to-speech model, and the whole thing came down to a single word on a listening page. The word is "machar", Hebrew for tomorrow, and a synthetic voice kept saying "micker".



### Herman

Which is a lovely failure, because it tells you where the problem lives. Herman Poppleberry, for anyone new.



### Corn

Daniel spent the fourteenth to the sixteenth of this month on Meorav, his English-with-Hebrew voice model, the one where Hebrew is written in Latin letters inside English sentences. We did the episode on teaching a TTS model his own accent a few weeks back, so that's settled. Meorav pronounces the Hebrew well, but it sounds like Daniel, Irish accent and all.



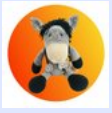
### Herman

And the new goal was the opposite split. Keep his Hebrew pronunciation, lose his voice. Give me a model that says the Hebrew words the way he does, in a voice that belongs to nobody.



### Corn

Which sounds like it should be one switch. It was eight or nine steps, and every one was decided by ear.



### Herman

Let's take them in order, because the order is the argument. Day one, he didn't start with his own recordings at all. He started by designing a voice. There's a model called OmniVoice that builds a voice from a text description. You type something like male, young adult, moderate pitch, American accent, and it hands you a speaker.



**Corn**

Twenty candidates, I think.



**Herman**

Twenty, and he picked one, seed three, which he calls A three. Small catch. A designed voice doesn't reproduce from its seed on new text. Run the same seed with a different sentence and you get a different person. So the A three clip itself became the reference, and every later render clones from that clip. Freezing the voice.



**Corn**

A voice you can only keep by photographing it.



**Herman**

And it matters later, because that frozen clip is the A three that gets cloned over and over. Every English sentence in the final model that sounds like that American male is a clone of one short clip.



**Corn**

So if the clip had a mumble in it, the mumble is now a feature.



**Herman**

Yes. Which is partly why he listened to twenty before choosing. Then the Hebrew. He tried everything for getting a synthetic route to say it. Pointed Hebrew with the vowel marks. Phoneme strings in ARPAbet, written by a Claude subagent that was given Israeli pronunciation rules. Another model, Zonos, fed hand-written IPA. None of them was good enough. That "micker" is the typical shape: the guttural letter comes out as a hard k and the vowel is off.



**Corn**

Hang on, though. Why does a model that's perfectly fluent in English fall over on a three-syllable word?



**Herman**

Because Hebrew isn't in the base model's list of languages. It's got ten, and this isn't one. So when it meets "machar" in Latin letters, it falls back on whatever its English spelling habits say. Hold that thought, because it comes back at the end of day two and it's the whole episode.



**Corn**

Fine. So synthesis is out. Though I'd have wanted to know why the phonemes failed. Phonemes are supposed to be the cheat code.



**Herman**

My guess is the model has never heard much Hebrew-flavoured sound in that space, so the symbols map onto the nearest English-ish thing it knows. I can't prove that. What I can say is that a guttural isn't a sound an English-trained model has a slot for, and no amount of careful notation puts a slot there. Synthesis is out, and the pivot is obvious in hindsight. His own recordings already have the Hebrew right. The gutturals, the stress, the flow. So stop synthesising Hebrew and convert his audio into the A three voice instead. Seed-VC, zero-shot voice conversion, with the RE-USE restorer cleaning up afterwards.



**Corn**

He tried the other converters, presumably.



**Herman**

Applio, which is the RVC route, a fine-tuned Seed-VC, and kNN-VC. Zero-shot Seed-VC beat all of them. Then Meorav version three point zero was an OmniVoice LoRA trained on his takes converted to A three, and it went out as the main Meorav release.



**Corn**

So at the end of day one, mission accomplished.



**Herman**

By ear, not quite. And this is the part I like, because he didn't ship and move on. Day two starts with an attempt to make the conversion itself replicable. He built a paired dataset, six hundred and fifty pairs, six hundred for training and fifty for test. Each pair has the original forty-eight kilohertz take, the cleaned version, and the converted A three target. The idea is to train one audio-to-audio model on the matched pairs, so conversion becomes a fixed model you can run again.



**Corn**

And he listened to the targets first.



**Herman**

Verdict, in his words: "still too much accent transference."



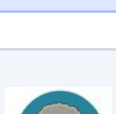
**Corn**

Which is a polite way of saying the Irish is still in there.



**Herman**

It leaks through the conversion. And I think that's natural, honestly. A converter is trying to keep the rhythm and the stress of the source, because that's what makes the Hebrew correct. But rhythm and stress are most of what an accent is. You can't keep one and discard the other.



**Corn**

Hm.



**Herman**

Anyway, it leaks. The voice changes, the accent rides along underneath. Anything you train on those targets inherits it. So that line of attack stopped there, and out of it came a decision he's calling the Meorav split. His own voice gets its own model, Meorav-Daniel, version one point zero, which is the old two point two weights copied unchanged and checksum-verified. And plain Meorav stays as the neutral-accent aspiration.



**Corn**

I like that he separated the thing that works from the thing he wants. Most people keep calling the first one the second.



**Herman**

It also freed him up. Once Meorav-Daniel was safe and named, nothing he tried next could break it.



**Corn**

Fair. Pilot time?



**Herman**

Now the pilot that, to me, is the centre of the whole run. It cost a few cents and trained nothing. The question was narrow. If you condition a stock TTS model on a speaker embedding only, an x-vector, instead of on a reference clip, does less accent come through?



**Corn**

Because a clip carries the whole performance and an embedding is just a fingerprint of the voice.



**Herman**

That was the bet. Ten lines, four columns. Daniel by clip, Daniel by x-vector, A three by clip, A three by x-vector. Forty clips. And line eight, "My achi called to say the chag starts earlier this year," lost the gutturals in every column.



**Corn**

Every column. Even the one that was literally his voice.



**Herman**

His note was "It inherits." The base model's own spelling-to-sound mapping decides what the Hebrew sounds like, whichever voice you hand it. Conditioning carries the voice and nothing else. That's the finding: pronunciation and voice are separate layers.



**Corn**

Which, to be fair, you could have guessed. A speaker embedding is a small vector. A pronunciation system is an entire model.



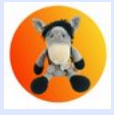
**Herman**

You could guess it, but then you'd have been guessing. Forty clips settled it for a few cents. And it sets up the obvious next move. If pronunciation is in the weights, then train the Hebrew into the weights, and do it without teaching the model his voice.



**Corn**

Which sounds impossible. The training audio is his voice.



**Herman**

The trick is the loss mask. You train on his takes, but you only compute the loss on the Hebrew words. The English stretches are in the audio, the model sees them, but they contribute nothing to the gradient. So the model is rewarded for getting the Hebrew spans right and gets no signal about how the English sounds.



**Corn**

So how do you know where the Hebrew words are?



**Herman**

Four Claude Sonnet subagents labelled them. Four hundred and sixty-one of the six hundred takes had Hebrew in them, a bit over thirteen hundred words. Then forced alignment timed each word in the audio, and it failed on none of them. Each span got padded by one codec frame, eighty milliseconds, on both sides, so the edges don't clip. Alongside that went forty-five minutes of LibriTTS-R, multi-speaker English, trained in full so the model keeps a sense of other voices.



**Corn**

Three epochs, small learning rate, one big GPU. Rendered a hundred and twenty clips.



**Herman**

And Daniel's verdict was "same problem, the base is still problematic."



**Corn**

Ouch. Two for two.



**Herman**

There were two open readings. Either the signal was too weak, only about nine minutes of Hebrew-span audio against forty-five of English. Or the base model just can't be taught it this way. Honestly I don't know which, and I don't think he did at that point either.



**Corn**

What I'd have done at that point is go to bed.



### Herman

He did something better... well, he slept, I assume, and then he reframed it. Meorav-Daniel's gutturals are fine. So a Qwen-family TTS model can learn this Hebrew, because it already did. Stop trying to teach a fresh base. Start from Meorav-Daniel and swap the voice in as an input.



### Corn

Which is the x-vector idea again, but on a model that actually knows the words.



### Herman

One snag. When he fine-tuned Meorav-Daniel, the speaker encoder got detached. So the probe restored the base model's speaker encoder, trained nothing, and rendered forty clips with only x-vectors as the voice. Hebrew, in his words, "perfect in all of them."



### Corn

But.



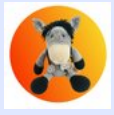
### Herman

But every voice sounded like Daniel. The LibriTTS voices, he said, "just sounds weird."



### Corn

So the Hebrew survived and the swap didn't.



### Herman

The reason's clean, I think. During that fine-tune the talker only ever saw one embedding, his. Whatever the input said, there was never any pressure to read it. The voice got baked into the weights, and the model learned to ignore the knob.



**Corn**

Herman, a question, and I'm asking it honestly. If the voice is baked in, why did the Hebrew survive the swap? Isn't the Hebrew baked in too?



**Herman**

It is, and that's the asymmetry. The Hebrew you want baked in. The voice you want to stay a dial. Training on one speaker bakes in both, because nothing tells the model which parts of the sound are about who and which are about what.



**Corn**

Let me guess where this goes. Multiple speakers.



**Herman**

Multiple speakers, and the mask. With many voices in training, the embedding has to carry information, because the weights can't guess the speaker. The mask keeps his voice out. We'll get there.



**Corn**

Well, it's time for another one of Dorothy's History of Toast, isn't it? Hello, Mum.

D

**Dorothy**

Hello dear, it's Dorothy, with Dorothy's History of Toast.

D

**Dorothy**

In nineteen ninety five the physicist Robert Matthews published a paper explaining why toast really does tend to land butter side down, mostly because of the height of the table.

D

**Dorothy**

And it won him an Ig Nobel Prize the next year. So that's the history of toast, sweetheart, everything slides off in the end, doesn't it.

D

**Dorothy**

Anyway, I've got a load on, so I'll let you go. Bye bye now.



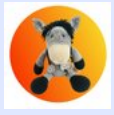
**Herman**

Always a delight to hear from Dorothy.



**Corn**

Day three, then.



**Herman**

Day three runs from Sunday night into Monday, and it opens with a pilot before a spend. A run that would cost around two dollars fifty, and Daniel wanted a small test first. One pass of multi-speaker masked fine-tuning from Meorav-Daniel. All three hundred and seventy-seven LibriTTS-R rows at full loss, plus every third Daniel row, a hundred and fifty-four of them, Hebrew spans only. About thirty-three updates.



**Corn**

Thirty-three updates. That's less training than most people do to get a coffee order right.



**Herman**

And it worked, for A three. He said it "definitely doesn't sound like me, which is good," and it kept more of the pronunciation. The x-vector steers the voice again and the Hebrew carries over.



**Corn**

And the LibriTTS voices?



**Herman**

"Really bizarre, almost like a Canadian speaking."



**Corn**

I'd pay to hear a Canadian speaking Hebrew through a donkey's idea of Ohio.



**Herman**

Ha. The pilot's real lesson was that real strangers from an audiobook corpus give a model odd, uneven voices. So the next pilot swapped them out. Instead of LibriTTS strangers, one extra voice, A three itself. OmniVoice cloned A three reading the same three hundred and seventy-seven texts, about thirty-eight minutes of synthetic English. Same mix, one pass.



**Corn**

The designed voice trains its own replacement.



**Herman**

It trains itself in, which is the point. A three came out fine. Not all the gutturals yet, but he expected those to get stronger with more passes. And a side finding I enjoyed: the LibriTTS-mix pilot, when rendered with A three's x-vector, was "the most bland, neutral, almost genderless." That's the neutral-accent goal, accidentally, and it's also a voice nobody would choose to listen to for an hour.



**Corn**

Bland American is a genre.



### Herman

He agreed. Bland American, he said, is "slightly grating," and Irish is too distinctive. He wanted something more fun, a matched male and female voice. So he swept OmniVoice's native English accents, American, British, Australian and Canadian, male and female, two seeds each. Sixteen clips. No Irish and no South African in the list, those aren't something it designs.



### Corn

Which is a shame. I'd have taken the South African.



### Herman

Then the real run. One model, five designed voices. American male, which is A three, American female, British male, British female, Australian male. About two hundred and forty-five generated English lines per voice, twenty-seven to thirty-one minutes each, all trained in full. And Daniel's four hundred and sixty-one code-switching takes, counted twice, nine hundred and twenty-two rows, trained only on the Hebrew spans.



### Corn

Why count them twice?



### Herman

To balance it. Otherwise the Hebrew is a rounding error against twelve hundred rows of English. Twenty-one hundred and forty-seven rows in all, starting from Meorav-Daniel with the speaker encoder restored. Three epochs, learning rate of one in a hundred thousand, one big GPU. The loss started around two point six in the earlier pilots and was down to about one point two to one point four by the end.



### Corn

And it's out. Meorav-Multi-Voice, gated, non-commercial, because OmniVoice's weights are non-commercial, so what's trained on its output is too.



**Herman**

Voice is picked at inference by a short reference clip, x-vector only. None of the five voices is a real person. The designed voices supply the English. Daniel supplies only the Hebrew words, and the mask makes sure he supplies nothing else.



**Corn**

Do you believe the card's limits? Not benchmarked, checked by ear only, gutturals not always produced, clips capped at fifteen seconds.



**Herman**

I do, and I'd put more weight on it for being written down. A model that's checked by ear and says so is more useful than one with a leaderboard score nobody can hear. What do you think, though? Is by-ear enough for something like this?



**Corn**

For a voice, yes. The thing you're making is for ears. If the number says it's fine and Daniel hears "micker", the number is wrong.



**Herman**

There's a method point under that. Look at what each step cost. A few cents for the x-vector probe. A few dollars at most for the pilots. Every one of them was a page of tens of clips, and every one killed or confirmed an idea before the big run. The sixty-dollar mistake never got made.



**Corn**

I'd phrase it less nobly. He kept being wrong in cheap ways.



**Herman**

Which is the skill.



**Corn**

Both can be true.



**Hilbert**

...so the surname he says himself, and the nephew reads the rest. Mr. Kowalczyk wants the voicemail greeting done in a voice that isn't his, because he says his own sounds like a man reading out a tax bill. Fine. Forty pounds, one morning. But nobody else says Kowalczyk right. Herman, you pause before every dollar figure, I can hear where to cut it, and the nephew had three goes at the surname and got it wrong in three different places. So I cut in Mr. Kowalczyk's own take of it, seven tenths of a second, and I did not let the nephew near it. The nephew is Tadeusz. He can say the alphabet backwards in nine seconds. I've timed him. It hasn't changed since March and I find that irritating.



**Herman**

The join is where it would show. Daniel padded each Hebrew span by eighty milliseconds on both sides for the same reason. Clip a word at its edge and the ear hears the cut before it hears the word.



**Corn**

Does the nephew know the surname is off limits?



**Hilbert**

He thinks it's a copyright matter. The greeting is done, and Mr. Kowalczyk has rung twice since to ask for a different one for the weekend. He wants it recorded on a Friday.



**Herman**

Two greetings, one surname, which suggests the surname is the only part he's actually attached to.



**Corn**

Something from the notes that didn't fit. Daniel put the new model to work straight away on a six-voice audio drama called The Teudat Zehut, five hundred and eleven seconds long. It's a first render and he hasn't reviewed it yet.



**Herman**

So the next listening page is already open.



**Corn**

Whether the gutturals hold up through a long sentence in the British female voice is what I'd want to hear. A fifteen-second clip limit is kind to a model. A drama isn't.



**Herman**

Thanks to Hilbert Flumingtop for producing.



**Corn**

This has been My Weird Prompts. You'll find every episode at my weird prompts dot com.



**Herman**

See you tomorrow.