Episode Book

EPISODE #6178

Why Your Voice Clone Sounds American

PUBLISHED	21 October 2026 at 14:34 UTC
DURATION	21:58
LISTEN	myweirdprompts.com/episode/voice-clone-accent-displacement-72786/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.11
BOOK GENERATED	22 October 2026 at 07:46 UTC

SYNOPSIS

You clone your voice, and out comes a stranger with a vague American accent. It's called accent displacement, and it's not a bug in your recordings — it's baked into the training data. This episode digs into why TTS models flatten accents, whether a "neutral" English accent even exists, and what actually helps: fine-tuning on many speakers of your accent rather than hours of just one. We look at Singlish fine-tuning results, accent embeddings, activation steering, and what the vendors' own documentation admits about cloning Irish English.

TRANSCRIPT



Corn

Daniel is Irish, and every time he clones his own voice, the clone comes out sounding vaguely American. Good model, a few hours of clean recordings, same result. He calls it accent displacement, and he wants to know what actually fixes it, which open models are less biased, whether a neutral English accent exists at all, and what people have done when they needed a particular accent and could not record dozens of hours.



Herman

One sentence on the why, since we covered it in the episode about the Irish accent sounding American. The training data is lopsided towards American English, the model's prior is enormous, and a few hours of you is a spoonful in that ocean. Today is the other half. And it sits right next to the episode just before this one, on single-shot cloning, where Chatterbox, the model that voices this show, basically stops listening to a reference clip after six to ten seconds. That is the same problem from the opposite end.



Corn

The word I want to poke at is neutral. Every description of a base model says it speaks a neutral English, and then Daniel's accent gets pulled towards neutral. Is there such a thing?



Herman

Not to a linguist. The standard position is that everyone has an accent. Rosina Lippi-Green has a whole chapter called the myth of no accent. When people say neutral, they mean prestige plus sheer volume, the accent nobody remarks on because it's on every news bulletin. For a TTS model, neutral has an even plainer meaning. It means the majority of the training data.



Corn

Which would make you the baseline, Herman. Factory settings, no accent, nothing to see.



Herman

I keep "caught" and "cot" apart, thank you. I'm a lot more regional than you think.



Corn

You're a lot more in the training set than you think.



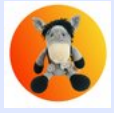
Herman

Fine. That one I'll take. The part I'd hedge is how American the big research datasets are. The two famous English corpora, LibriSpeech and LibriTTS, are audiobook readings. The authors of a newer accent dataset called GLOBE say those are mainly American speakers, but nobody has published an actual audit with a percentage. Same for the huge in-the-wild sets scraped from YouTube and podcasts. The Emilia paper reports roughly forty-seven thousand hours of English in its first release, and it doesn't analyse accent at all.



Corn

So everyone is confident the pull is American, and nobody has measured it.



Herman

Daniel's ears are the measurement. And there's one number that shows what "neutral" costs. A group fine-tuning for Singlish tested stock Chatterbox on Singapore English speech. The stock model's output was easier for a recogniser to transcribe than the real recordings were, about eleven and a half percent errors against seventeen and a third. The authors put it down to clean, accent-neutral output.



Corn

So the machine is more intelligible than actual humans. That's less a feature than a sign it has sanded someone off.



Herman

The same pattern shows up in speech recognition. The best recogniser they tested, trained on six hundred and eighty thousand hours, gets about two and a half percent errors on clean American read speech and nearly twenty percent on a benchmark of global accents in conversation. The default is American in both directions.



Corn

Okay. So what actually solves it for Daniel? Start with the obvious one, more of his own voice.



Herman

It helps, but the problem is that fine-tuning on you teaches the model your timbre quickly, because timbre is cheap. Accent lives in rhythm, in vowels, in where the stress falls, and the model has a very strong opinion about all of that already. My guess, and it's a guess, is that a small adapter like a LoRA has enough room to move the voice and not enough to overrule the prosody. There's no head-to-head study on accent that I know of. The nearest evidence is a vendor's docs for one model, saying a LoRA reaches about ninety-eight percent of a full fine-tune on speaker similarity at roughly half the memory.



Corn

Speaker similarity. Is that the same as sounding Irish?



Herman

No, and this is the trap. A recent paper on accent leakage points out that methods which let the reference accent leak through can score better on speaker similarity, because accent is part of how a voice sounds. A leaderboard of best clones may be rewarding the models that copy the accent from the clip. Their fix is to compare similarity at equal accent. A big listening study from this March found the same tangle, that speaker similarity and accent similarity move together.



Corn

So some of the "good" clones on those leaderboards are just the ones that cheat from the reference.



Herman

Which is the opposite of cheating if your reference is Irish, funnily enough. But I should flag the limit. Nearly all of that leakage research is cross-language, a Spanish speaker reading English text. Daniel's case is one English drifting into another, and as far as I know that is much less studied.



Corn

Then the closest thing we have to his problem?



Herman

The Singlish work. Two researchers fine-tuned Chatterbox and another open model, CosyVoice three, on fifty speakers from a Singapore speech corpus. Out of the box, with a Singlish reference clip, the output kept the little discourse particles but the prosody and the finer sound detail were missing. Intelligible, and non-Singlish. After fine-tuning, accent similarity rose, and, importantly, it rose for speakers the model had never seen. The gain was the accent, not the individuals.



Corn

So the lesson is fifty people, not one person for a very long time.



Herman

That's how I read it. Teach the model the accent as a category from many voices, then let a short reference supply the individual. I haven't confirmed how many hours each of those fifty speakers contributed, and they only used automatic metrics, no human listening test. But the direction makes sense to me. What would you do, Corn? If it were your Mongolian-sloth accent being flattened.



Corn

Mine's not an accent, it's heritage. And I'd be wary, because fifty Irish speakers is exactly the sort of dataset nobody has lying around.



Herman

Fair. The biggest accent-labelled set I can name is GLOBE, about five hundred hours across a hundred and sixty-four accents, curated from Common Voice. I don't know how many of those are Irish. There's also a Corpus of Irish English Speech, openly available, with a focus on prosody, but I couldn't tell you its size or its licence.



Corn

Is anyone trying to make the accent a dial instead of a data problem? Because that's what Daniel actually wants. A slider that says more Cork.



Herman

There's work in that direction. One team, AccentBox, built accent embeddings that are speaker-agnostic, meaning a vector for the accent that doesn't carry the person with it, and fed them to a zero-shot model. On accents of speakers it had never heard, their accent identifier scored about fifty-six percent on its F-one measure, which is modest. And it's built on an older system, YourTTS, not one of the big language-model voices. Separately, a group at UT Dallas used activation steering, nudging the model's internal activations, to strip an accent while keeping the timbre. That one runs the wrong way for Daniel, but it shows the two can be pulled apart after the fact, without retraining.



Corn

So in principle you steer the Irish back in. Same trick, opposite sign.



Herman

In principle. I haven't seen it demonstrated for putting an accent in, and I wouldn't promise it. There's also a fix on Fish Audio's model from a conference paper, where setting the language tag stopped the reference accent leaking. That again is cross-language, so I'd hold it loosely.



Corn

And now, a voicemail from one of our listeners.
Lila: Hi Corn and Herman, this is Lila calling from Nairobi, I am on the matatu home from work, squeezed next to a man carrying two live chickens, which is relevant because the chickens are making this clucking sound and I keep thinking, that is a kind of accent too, you know, and nobody is trying to clone it. Anyway, my question is about Daniel and his



Corn (continued)

Irish accent turning American. Does the machine forget the accent the way a person forgets a language, and if I recorded my grandmother reading Bible verses in Kikuyu for two thousand and one hours, would the model keep even forty percent of her, or would it just smooth her into a generic lady from nowhere? Thank you, I listen every week, even when the chickens are loud.



Herman

Thank you, Lila, and the matatu sounds livelier than our studio. Being smoothed into a generic voice is the right worry, and the way out is probably less about piling up hours of one person and more about how the model learns the accent in the first place.



Corn

Which is why the vendors don't just record more hours either. Amazon's polyglot voices team described weighting the training data by locale, augmenting it, and using a LoRA, with synthetic reference samples to pry the speaker's identity apart from the accent. Their goal runs the opposite way, though. They want accent-free cloning, a voice that speaks other languages without dragging its own accent along.



Herman

Their tooling can separate identity from accent, which is the same lever Daniel needs, just pulled in the other direction. And one more vendor data point. Amazon Polly launched an Irish English voice called Niamh in two thousand twenty-three and said it was built with technology that teaches existing voices to speak with different accents, sharing a personality with their British voice. They didn't say how much data went in, so whether there was a dedicated actor is something I can't tell you. Microsoft has Irish English voices too, Emily and Connor, and I haven't seen anything on how they were built.



Corn

Adaptation, then, not forty hours in a booth in Dublin.



Herman

As far as the public statements go. And for the paid route, ElevenLabs says the quick instant clone needs one to three minutes and may struggle with uncommon accents, while the professional clone wants thirty minutes up to about three hours. Their docs also say



Herman (continued)

accent can't be edited after cloning, only by changing your samples. Their newest model lists an Irish accent tag, but they say how well it works depends on the source voice.



Corn

So the clone is more stubborn than the person.



Herman

And there's a tiny community example. Someone fine-tuned the Parler-TTS Mini model on about three hours of their own data to get a transatlantic accent. No evaluation published, so it's a sign that people try, not proof it works.



Corn

What about the voice conversion route? Record a real Irish person saying the line, then convert the timbre to Daniel.



Herman

Careful with the tool choice. OpenVoice's own documentation says its converter does not clone the accent or the emotion. It moves timbre only. The accent comes from whatever the source speech had. I've seen nothing showing that other converters, Seed-VC for one, change accent either, so I'd say it's untested rather than ruled out.



Corn

So voice conversion keeps the accent you feed it. If the source is Irish, you get Irish. If the source is the model's default, you get the default in Daniel's voice.



Herman

Which means the source has to be Irish already. A recorded human, or a model that can produce the accent on its own.



Corn

Then which open models can do it on their own?



Herman

That's thinner than I'd like. Kokoro, the tiny eighty-two million parameter model, has fifty-four voices, American and British. No Irish prefix listed.



Corn

Fifty-four voices, and the nearest one to Cork is a man called George.



Herman

The espeak-ng phoneme tool is the same story. It has American, British, Scottish, Received Pronunciation, a Caribbean variant, no Irish. Anyone going the rules-based route would be writing the Irish variant themselves. Chatterbox's multilingual release warns that when you cross languages, the output picks up the reference's accent, and the docs suggest turning one of its guidance settings down to zero. That is vendor guidance, I haven't seen it tested independently. Fish Audio's latest open model has free-form inline tags and fine-tuning code, but I can't say whether an accent tag does anything. I won't claim it for the others either, because I haven't seen it confirmed.



Corn

Then "which model is least biased" has no leaderboard answer.



Herman

Not that I can point at. If I had to bet, I'd bet on models trained on messy in-the-wild speech over polished audiobooks, because real podcasts have more accents in them. But that's an inference, and nobody has published the breakdown to back it.



Corn

Someone should. A column on the model card called "where it thinks you're from".



Herman

Parler-TTS is the cautionary tale there. Its features list covers gender, noise, speed, pitch and reverb, but not accent. And the accent labels in the pipeline behind it were predicted by a classifier, not checked by people.



Corn

I want to take the opposite side for a second. Everyone's solving this with data and training. What if Daniel just chooses a reference clip that is so Irish it overwhelms the prior? Like a reference of him saying the words that carry the most.



Herman

That's cheap, and I'd try it first. The prior is strong but the reference clip does pull, and we know from the last episode that only the first several seconds count. So those seconds should be the most Irish seconds you have, the vowels and the rhythm at their strongest, not a calm studio read.



Corn

Dozens of hours is impractical, so a trick that cheap might just work.



Hilbert

Impractical is the wrong word. Dozens of hours is a Tuesday.



Corn

You've done it, then.



Hilbert

Forty-one hours. The bird is called Fergal. It's a cockatoo, it belongs to a Mr. Dunphy on the second landing, and it says one sentence, which is mind the step, it's loose. It has a Mayo accent you could stand a spoon in. Dunphy won't let it out of his flat, so the recorder lives in my airing cupboard, a Tascam, plugged into the extension lead that also runs the immersion heater, and the cable goes under the door.



Herman

A cockatoo that only knows one sentence is actually a clean dataset. Same words, same speaker, no label noise. It's the opposite of an in-the-wild corpus.



Hilbert

Fergal only talks when three people are in the room. Dunphy counts as one, I count as two. The third is a tailor's dummy in Dunphy's father's overcoat. I pay forty-five euros a week plus the dummy's dry cleaning. Herman, you're hot on the S's, back off the mic two fingers. Anyway the dummy has a library card. Dunphy took it out in its name in March, and he's two overdue books on it.



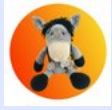
Corn

Which is a lot of overhead for the one sentence.



Hilbert

Fergal says it four times in a row and then screams during the news. Dunphy has the news on at six, at seven and at ten, and it is not Dunphy's fault, he's a man who likes to know what's happening in the world. Then there's Tuesdays, when I'm not allowed to record, because Tuesday is when the cockatoo has the neighbour's dog round. I'm not allowed to say what they do.



Herman

A cockatoo repeating a phrase has less variation than any human reading, though. It can't learn the accent from that alone. The model needs the accent in a range of sounds and sentences.



Hilbert

Fergal has a second sentence. It's only for the dog.



Corn

Two sentences and a dummy with a library card. That's a bigger cast than most podcasts.



Hilbert

The cupboard gets warm at night. The immersion heater comes on at five and the tape hiss goes up about a notch. I've got a towel over the Tascam now. Dunphy brings Fergal's perch to the door on Fridays so the bird can see the cable. It seems to approve.



Herman

A fixed room, fixed microphone and fixed speaker is the dream setup for a dataset, towel or no towel. The variation between sessions is what usually ruins it.



Hilbert

Ninety-four minutes usable. Dunphy wants it up to a hundred before the dummy's card renews in March.



Herman

Ninety-four minutes out of forty-one hours is under four percent. Even with a bird that says the same thing every time, the thing you keep is a sliver of what you record, and I'd budget for that. Most people picture the recording session as the cost, but the cleaning is the cost.



Corn

Then for Daniel, the bit worth knowing is that dozens of hours isn't the barrier. Getting dozens of hours that are clean enough to count might be.



Herman

Which points at the most practical ladder I can see. First the cheap things. Pick the reference clip with the strongest Irish features, since the first few seconds are what the model hears. Second, if the model family lets you, fine-tune on many Irish speakers instead of only one, the way the Singlish group did, and then clone Daniel on top of that with the short reference. Try a LoRA first, because it's quick, but don't expect it to change prosody all the way.



Corn

And judge by ear.



Herman

By ear, yes, and by Daniel's ear most of all. Speaker-similarity scores can reward the very thing that goes wrong. If he wants a number, accent similarity at equal speaker similarity is the fairer comparison, but I'd still trust a Dubliner over a classifier. Third, the fallback. Have a real Irish speaker record the lines, run voice conversion to put Daniel's timbre on top, and accept that the accent is theirs.



Corn

And if he'd rather pay, the ElevenLabs professional clone from a few hours of his actual voice, provided those hours are strongly Irish.



Herman

With the caveat that the accent is set by the samples and can't be edited afterwards. So the samples decide.



Corn

Honestly, I'd also ask whether Daniel wants the clone to be Irish all the time. Some of this is a text problem. Irish English has its own word order and rhythm, and a model reading standard written English has no cue to use them.



Herman

That's a fair point, and a research group has tried something along those lines, converting accents through text transliteration, spelling the line the way the accent would say it. I haven't read it closely, so I won't claim how well it works. As a cheap experiment, though, rewriting a line phonetically in the script costs nothing, and it might shift the vowels where the reference can't.



Corn

Spelling "bold" as "bowld" and hoping. I've done worse with leaf medicine.



Herman

Worse than that, you'd be back at the spelling bee. But it's a ten-minute test.



Corn

There's one belief I'd like to put in the bin. People assume that a model with a neutral accent is a model with no accent, so a clone that drifts American has gone wrong in some random way.



Herman

It hasn't gone wrong at all. It has done what the data said. Neutral in a model means most common in the training data, and a clone drifting towards it is the model working as built.



Corn

So the open thing to watch is whether anyone publishes an accent audit of a model's training set. If they did, picking a base model would stop being a coin toss. And whatever Daniel ends up recording, I'd plan on keeping a small fraction of it.



Herman

A small fraction, and then he'd better have a good ear for which fraction. Thanks to Hilbert Flumingtop for producing.



Corn

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



Herman

See you tomorrow.