Episode Book

EPISODE #6177

How Voice Cloning Works From Six Seconds of Audio

PUBLISHED	21 October 2026 at 14:21 UTC
DURATION	21:54
LISTEN	myweirdprompts.com/episode/voice-cloning-six-second-prompts-58761/
HOSTS	Corn, Herman, Desmond, 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

Voice cloning from a short clip isn't the model learning your voice — it's the model finding you inside a voice space it already built from hundreds of thousands of hours of speech. This episode digs into how Chatterbox's two-stage pipeline actually uses your reference audio, why feeding it more doesn't help, and what changes when you move from instant cloning to a trained model. Along the way: VALL-E's running-start idea, the three channels a reference clip feeds into, why ElevenLabs sells Instant and Professional as separate products, and what a fine-tune really moves inside the network.

TRANSCRIPT



Corn

Six seconds. That's how much of a reference clip Chatterbox actually listens to when it builds the speech-token prompt, and it's the reason the two of us sound like anyone at all.



Herman

Six for the token prompt, about ten for the decoder. Anything past that, the code just cuts off and throws away. I read the source file this morning and I was a little delighted.



Corn

Of course you read the source file. Daniel's question is built on that gap. This show is voiced by Chatterbox, which clones a voice from one short clip, and he found early that feeding it more reference audio doesn't help the way you'd assume. Yet from a few seconds it gets timbre, the range of sounds, even the rhythm of a speaker well enough that we pass as brothers. How does that work? Then there's the other side, where ElevenLabs sells Instant Voice Cloning and Professional Voice Cloning as two separate products, one from a minute or two of audio, one from hours that train a dedicated model. Why are those separate pipelines, and what's actually different inside the trained one? And for the best open source voice today, what do you train on?



Herman

Good, that's a lot of ground, and the first half is the miraculous part, so start there. The trick is that the clip is not teaching the model anything. Chatterbox was trained on roughly half a million hours of cleaned speech before anybody gave it your reference. By then it has already seen an enormous spread of human voices, so a ten second clip isn't a lesson. It's a pointer. It tells the model where in a space of voices it already knows it should be standing.



Corn

So the miracle isn't in the clip, it's in the half million hours.



Herman

Yes. Hundreds of thousands of hours of other people did the heavy lifting, and your six seconds just picks one of them out. The weights never change. Nothing about Herman has been learned. The model is being handed a description and told to continue in that style.



Corn

Nothing about Herman has been learned. That could be on the tote bag.



Herman

Noted. Now the specifics, because the how matters for Daniel's other question. Chatterbox has two stages. First, a language model of about half a billion parameters turns text into speech tokens, discrete little units of sound at around twenty-five a second. Second, a separate decoder turns those tokens into a mel spectrogram and then a waveform. And the reference clip feeds in through three different channels.



Corn

Three.



Herman

A speaker embedding, which is one compact fingerprint of the voice, taken from the whole clip and averaged. Then the actual speech tokens of the first six seconds, which are placed in front of the text as a prompt, so the language model literally continues from them. And then a reference for the decoder, taken from the first ten seconds, which handles the fine acoustic detail. There's also an exaggeration dial for emotion, default half.



Corn

Let me guess where you're going next. Somebody in two thousand twenty-three treated this as a language modelling problem, and you're about to name them.



Herman

VALL-E, from Microsoft, and I'll take the guess. They trained on sixty thousand hours, gave it a three second clip, and it simply continued the clip. It didn't just copy the voice, it kept the room and the mood of the clip too. That's the idea Chatterbox lives inside. The prompt is a running start.



Corn

Which explains something about my own clip. I'm told my reference is a sloth being unhurried and flat for ten seconds.



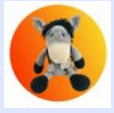
Herman

And that's why your delivery is the easiest thing in the world to clone. Flat prosody has nowhere to go wrong.



Corn

And yours is the exaggeration dial turned to the stop.



Herman

Fair. Fair, I walked into that one.



Corn

So back to Daniel's finding. More reference audio, no improvement. If the code only looks at six seconds of speech tokens and ten for the decoder, then a three minute clip isn't a richer portrait. It's the same ten seconds with a lot of unused tape behind it.



Herman

That's how I read it. The model has fixed size slots, and slots don't get bigger because you offered more. The one channel that sees the entire clip is the speaker embedding, and that's an average, so a long clip mostly smooths itself into a blur. I should say, I haven't found anyone documenting some specific strange behaviour on long references, so I won't invent one. But if the cut lands mid word, or in a breath, or on a stretch where you



Herman (continued)

went quiet, the prompt the model continues from is a worse prompt. Daniel having a bad time with it fits.



Corn

So the winning strategy is the opposite of what a person would do. Don't record more. Pick the best ten seconds.



Herman

Clean, one voice, steady pace, the register you want back. Because it continues what you give it, including the room. Give it ten seconds with a refrigerator humming and the refrigerator is part of the voice.



Corn

Here's what still bugs me. Ten seconds can't contain much. How does a voice survive being squeezed through that?



Herman

Because most of what makes you recognisable is a handful of broad things. The length of your vocal tract, the resonances that result, the pitch range you live in, the grain of the voice. Those barely vary inside a single breath of speech, so ten seconds samples them nearly as well as ten minutes would. What a short clip can't sample is anything that only shows up occasionally. A rare vowel in an unusual word. Anger. Boredom.



Corn

So Daniel's observation, that a lot more audio doesn't help, is what you'd expect if the stable part is already captured in the first few seconds.



Herman

And the rest is exactly what a fixed slot can't hold anyway. Here's a question for you, Corn. If you could hand the model one clip of yourself, which would you choose? A neutral read, or something with a bit of life in it?



Corn

Something with life, I think. If it continues the prompt, then a dead clip gets me a dead sloth forever.



Herman

That's where I'd land too, with a caveat. The exaggeration dial and the clip are pulling on the same thing, so a lively clip with the dial up can overshoot. Mine, apparently.



Corn

Apparently.



Herman

Go on, ask the obvious one.



Corn

What does it not capture, then? Because we do sound halfway decent, but you don't actually sound like a retired pediatrician from Connecticut.



Herman

I'm hurt, but no. A single clip gives you timbre, and the local rhythm of those few seconds. It does not give you habits. How a person lands the end of a sentence, which words they lean on, the little stumble they always make, how they breathe before a long clause. That lives across hours of speech, and ten seconds has no room to hold it.



Corn

Which is where the vendors come in.



Herman

Which is where the vendors come in, yes. Here's something that surprised me, Corn. Daniel's sense was that a model is either single-shot or corpus-trained, and that vendors rarely do both. Actually, ElevenLabs does both and sells them as different products. Instant Voice Cloning wants about a minute or two and they say avoid going past three minutes. Professional Voice Cloning wants thirty minutes at the least and they recommend two to three hours.



Corn

So the interesting question is why they don't just have one button.



Herman

Because underneath they aren't the same operation. Instant is conditioning, exactly like Chatterbox. The voice is a description handed to a frozen model at the moment of generation. The Professional route, ElevenLabs says outright, trains a dedicated model. The weights themselves are updated on your hours of audio. That's a different kind of thing, and everything that makes it a separate product flows from there.



Corn

Walk me through what flows from it, because I'd say it's mostly business.



Herman

Partly, and I'll concede that. It isn't an architectural impossibility, it's an economics and operations split. But look at what follows. Training takes hours, three to six by their documentation, though another page says up to a day. Somebody has to store a distinct model per voice. And they verify identity, you have to speak a captcha to prove it's your own voice, which they can't ask of someone cloning a clip in a second.



Corn

And the product is for different people. Instant is for someone who wants a character voice tonight. Professional is for someone who wants their audiobook narrated for the next five years.



Herman

And PlayHT had the same shape. Instant clone off about thirty seconds, and I find this part funny, only the first thirty seconds were used, so the same truncation Daniel ran into. Then a high fidelity tier, twenty to thirty minutes for decent, hours for better. That documentation may be outdated, mind you.



Corn

The truncation being universal is almost comforting.



Herman

It also tells you something. A product that accepts a long clip and ignores most of it isn't being lazy. The whole point of that tier is not to train.



Corn

Open source has both too, doesn't it? Just not as a button.



Herman

As separate workflows, yes. Chatterbox itself has community finetunes, not official ones. One person published a full finetune of both stages with a German example. Another did a low rank adapter for Indic languages that touches the text side and leaves the speech tokens, the vocoder and the speaker encoder frozen. VoxCPM, a different model, documents an adapter trained on five to fifty clips for one speaker. So the two paths exist, you just pick a script.



Corn

Then what's different inside? If I fine-tune, what actually moves?



Herman

Depends how you do it, and this is the architectural heart of Daniel's question. Full fine-tuning updates all the weights of the language model and the decoder, so the whole network drifts toward you. The cheaper way is an adapter, a small set of extra low rank



Herman (continued)

matrices bolted on, where the original weights stay frozen and only the adapter learns. The VoxCPM docs claim an adapter of that kind reaches about ninety-eight percent of the full fine-tune's speaker similarity, though that's their own internal benchmark.



Corn

Ninety-eight percent for a fraction of the compute. That sounds like it should be the whole market.



Herman

For open source, often it is. But notice what the training changes. In the instant route, nothing in the network knows you. The voice lives outside it, in a clip. After fine-tuning, you live inside it. Your habits are in the weights, so the model produces them without being shown anything at generation time, and it can do it for any text, any length.



Corn

Which is also the risk. A model that's seen two hours of me can learn me too well. The ums, the tired afternoon.



Herman

Overfitting on a narrow corpus, yes. It can also get worse at things outside your data, a word it never heard you say. The adapter approach helps there, because the big model underneath stays intact.



Corn

Do you buy that the commercial split is partly economics, then? Because you said it isn't impossible to merge them.



Herman

Mostly, yes, and I'm only a little unsure. The technique is the same family of methods that's been around for years. What's expensive is operating it for a million users, each with their own model, safely verified. That's a business shape more than a physics one.



Corn

That covers where you'd go when ten seconds isn't enough.



Herman

Before we go on, here is Desmond with his Money Tip.

D

Desmond

Desmond's Money Tip. This week I have been taking two fewer steps per day, which spares the soles of my shoes. Over a year that is seven hundred and thirty steps saved, and at roughly one penny per thousand steps of sole wear, it comes to less than one penny a year. Over a lifetime, about six pence. I do it on the kitchen tile, going round the table instead of past it, and my wife has stopped asking why.

D

Desmond

Money is mostly this. Small and quiet and not much.



Corn

I'd try it. Quietly. Without telling Herman.



Herman

Now, the fine-tune route. The thing to see is where the hours go. A model like this already knows how speech works in general, so training on your voice isn't teaching it language. It's teaching it you.



Corn

Hours of me. The idea is chilling.



Herman

Hours of you being consistent, which is the harder part. ElevenLabs says one style, because the clone reproduces your pauses and your ums. If you read the first hour crisp and the second hour tired, you've built a model that drifts between two people. And the recording standards are precise. Peak around minus three decibels, average level between minus twenty three and minus eighteen. And record in the language you'll use, because your accent carries into the clone.



Corn

Which is exactly the opposite instruction from the one for single-shot. One says an hour of consistent treated-room audio, the other says one perfect clean ten second take.



Herman

That is the clearest way to see they're different tools. One is a library, one is a postcard.



Corn

You've been saving that.



Herman

Possibly since the car.



Corn

Here's what I don't get. Thirty minutes is nothing. People used to need days of data to build a voice. How does half an hour of Corn, for instance, produce a better Corn than ten seconds?



Herman

Because of the same pretraining. A paper on transfer learning for single speaker voices found that fine-tuning on about thirty minutes matched a model trained from scratch on more than twenty-seven hours. The foundation does the heavy lifting again, the thirty minutes only supplies your particulars. And that's what Daniel's last question rests on, so



Herman (continued)

here's the older world. LJSpeech, the dataset half the field learned on, is one woman, Linda Johnson, reading from LibriVox books, about twenty-four hours in thirteen thousand clips.



Corn

A day of one person reading library books. That's the dataset the entire field grew up on?



Herman

Yes, a day's worth. The earlier tradition was even more deliberate. The CMU Arctic corpora, around eleven hundred sentences chosen from out-of-copyright books specifically for phonetic coverage, so every sound in English in every neighbourhood shows up. That was when you needed to hit every phoneme yourself, because the model knew nothing.



Corn

So the traditional recommendation was, read a thousand carefully chosen sentences, or a day of audiobook, and hope.



Herman

And the modern one is, start with a big pretrained model, then thirty minutes to a few hours, clean and consistent. For serious fidelity, two to three hours, which is what ElevenLabs recommends for theirs.



Corn

Now the payoff. If Daniel wanted the best voice and accuracy in open source today, what do you reach for? Because I assume it isn't what we're running on.



Herman

It's careful ground, and I want to be straight about what I know. Nobody has an authoritative scoreboard. The numbers that float around come from different evaluations, so they aren't strictly comparable. On the Seed-TTS test set, a commonly used one, F five TTS gets an English word error rate of about two percent with a speaker similarity of sixty-five. IndexTTS two is slightly higher on the error rate but scores seventy on similarity, and CosyVoice three sits around seventy-two. Those are all close.



Corn

Which is a long way of saying there's no clear winner.



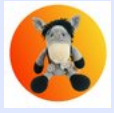
Herman

In that table, no. There's a newer claim I can't vouch for. Zonos two, an eight billion parameter mixture of experts model from late June, trained on up to six million hours, Apache licence, and its authors report state of the art on naturalness and cloning fidelity. That's their own claim, I haven't seen it independently checked. And then one arena snapshot from July put Step Audio EditX at the top of the open-weight models, and Chatterbox well below it, about a hundred Elo points lower.



Corn

So we are, on the evidence of a leaderboard, the budget option.



Herman

With an MIT licence, which counts, because a couple of the strongest ones, Fish Audio's top model and Voxtral, carry non-commercial licences. So the best for quality and the best you can actually ship are different lists. If I were Daniel, I'd pick two or three of the open models and clone the same ten seconds into each. Then listen. No leaderboard has your ears.



Corn

Spoken like a donkey who's already bought the headphones.



Hilbert

...so Ferenc says Thursday is fine, as long as the goose is back on the third stair by four. That's her stair. The council pays him two hundred and ten pounds a month to keep her on it and nobody has told me what for. Her name is Brenda. She's down on his tax return as a dependant, which I'm now doing, along with Tuesdays and Fridays at the bus shelter, where Ferenc stands and holds the number twelve for anyone running. There hasn't been a number twelve since March. Corn, you said six seconds earlier, and that's what I need. Six clean seconds, and the only room in the building without a hum is Ferenc's shed. Brenda sleeps in it.



Herman

A shed with no refrigerator is actually the point. If the room is in the clip, the model carries it into every sentence it ever speaks afterwards.



Hilbert

Brenda doesn't hum. She does the other thing. When you stand in front of her with a microphone she gets up and honks until you've read her the electricity bill. Ferenc says that's a thing she learned from the previous lodger. Nine pounds forty a quarter, I've heard it twice. I said I'd do the whole thing standing in the doorway, but the doorway is also where the draught comes in. He wants a fiver for the goose and I've given him the tax return instead, which he says is worth more.



Corn

The tax return is worth more to him than the fiver. That's a bad sign for the tax return.



Hilbert

It's not a good return.



Herman

Then the dilemma is that the goose is both the quiet room and the noise. And a model that continues whatever it hears would continue her.



Hilbert

Last Thursday I did a dry run. Forty-one minutes to get six seconds without Brenda in it. Lettuce is two pounds ten a week.



Corn

If she's in the final six seconds, she's in every episode. You'd be reading the news in a voice with a goose behind it, indefinitely, and a model can't un-hear a goose.



Hilbert

I've priced that in. I'll do it Wednesday, then. She's quieter before the post.



Herman

Wednesday's better anyway. Fewer takes means fewer sounds for the average to smear together.



Corn

Which is Daniel's point. Pick a few seconds well, and forget the rest.



Herman

Something I can't leave out, and it's tucked in the documentation, that everything Chatterbox produces carries an inaudible watermark, a signature called PerTh, baked into the output. So every sentence either of us has said today is stamped with it, even when we say something silly.



Corn

Mostly you.



Herman

I have it in writing that it's marked on both of us.



Corn

What I take from this is that the next time somebody wants a voice that's really theirs, the real question is the six seconds you choose, and how long you'll spend getting them clean. For anyone who wants more than the voice, the trained route is there, but it's a bigger decision than a bigger file.



Herman

The forty minutes for six seconds is the part I'll remember, to be honest. Getting one good clip is the work, and the model barely cares about the rest.



Corn

Thanks to Hilbert Flumingtop for producing this one.



Herman

This has been My Weird Prompts. If you want to find us, it's my weird prompts dot com.



Corn

See you tomorrow.