Episode Book

EPISODE #6229

Voice Design: Casting Ogres and Radio Voices

PUBLISHED	23 October 2026 at 14:09 UTC
DURATION	23:59
LISTEN	myweirdprompts.com/episode/voice-design-tts-knobs-17599/
HOSTS	Corn, Herman, Dolores, Hilbert
SCRIPT MODEL	meta/muse-spark-1.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.20
SOURCES	15 cited by the research (listed at the end)
BOOK GENERATED	24 October 2026 at 07:26 UTC

SYNOPSIS

Anyone can make a voice say a sentence. Designing one you can listen to for hours is a different problem. This episode breaks down how modern text-to-speech voice design actually works: the two families of methods (describe-and-preview casting versus reference-audio speaker embeddings), what the ElevenLabs sliders really do, why identity and colour are separate signals, and what research on professional radio voices reveals about warmth, pitch depth, and staying pleasant across a long drive.

TRANSCRIPT



Corn

Cloning gets the headlines, Herman. Stealing somebody's voice. The other half is quieter.



Herman

Building one that never existed in the first place. Or taking one you have and sanding the edges off until you can stand it for three hours.



Corn

Daniel's prompt is that second half. How voice design actually works in modern text to speech. What the knobs are, reference recordings, speaker embeddings, text descriptions, plain old EQ. Which ones change who the voice is, versus just how it sounds. And what makes a voice pleasant for hours, like the best radio voices.



Herman

The pleasant for hours part is the real question. Anyone can make a voice that says a sentence. Keeping someone company for a drive across the Negev is different.



Corn

The first thing to get straight is that voice design is two different families of method wearing the same name.



Herman

On one side you have text described voices. You type a massive evil ogre speaking at a quick pace, silly and resonant tone, and the model hands you candidates. ElevenLabs has the cleanest version of this. You pass a voice description plus some sample text, you get several previews back, you save the one you like.



Corn

Describe, preview, pick.



Herman

That pattern is everywhere now. Resemble, Telnyx, Hume with Octave, Alibaba with CosyVoice and their Qwen audio text to speech, AnyToSpeech. All of them do describe, get three previews, pick your favourite.



Corn

Walk me through what that actually looks like in practice? I type ogre and then what?



Herman

You type the description, you also type the line you want spoken, something like welcome to the dungeon, mind the stalactites. The system renders that same line in three different invented voices. Same text, three different ogres. One might be slower and chestier, one might be goofier and faster. You listen and you keep one.



Corn

So you are not tuning sliders at that point, you are casting.



Herman

Casting is exactly it. You are auditioning three strangers who only exist for that sentence.



Corn

And the other family?



Herman

Reference audio. You supply thirty seconds of someone talking and the model squeezes it into a speaker embedding. A vector, a long list of numbers, that encodes who that voice is. Every new sentence gets steered by that vector.



Corn

How does that squeeze work in practice? What is it actually keeping?



Herman

Think of it like a voice fingerprint, but messier. It is not storing the words, it is averaging out pitch range, rasp, breathiness, resonance, the shape of the vocal tract. Hundreds of dimensions. Then when you ask for a new sentence, the generator is told, say this, but stay close to those numbers.



Corn

So the embedding is the memory of who, and the text is the instruction for what.



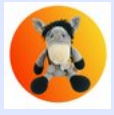
Herman

And the model has to satisfy both at once. That is why thirty seconds of clean audio beats five minutes of noisy audio. A clean fingerprint steers cleanly.



Corn

Why did the text prompt side appear at all?



Herman

Microsoft put it plainly with PromptTTS 2 two years ago. Speech prompts can be hard to find or may not exist at all. If you want an ogre and no ogre has ever recorded audiobooks, you are stuck. So they trained on forty four thousand hours of speech to turn descriptions into voices instead.



Corn

Forty four thousand hours. That is what, five years of nonstop talking?



Herman

Just over five years if you played it end to end without sleeping. That scale is why it works at all. You need to have heard thousands of ways to be resonant or silly before you can invent a new resonant silly voice on demand.



Corn

What do you think, Herman, is the text description actually designing a voice, or just ordering one from a catalogue?



Herman

Ordering with attitude, I would say. You describe, the model guesses what you meant, and you pick. The interesting spine for today is not which family wins. It is which knob does what. Identity and colour are separate signals, and the whole industry is trying to pull them apart.



Corn

So what are the actual knobs, and what does each one do?



Herman

ElevenLabs is the reference implementation because their settings are public. Five of them. Stability goes from zero to one, default is zero point five. Lower is more expressive but erratic, higher is steady and predictable.



Corn

Expressive but drunk, versus sober but boring.



Herman

That is uncomfortably accurate, Corn. Then similarity boost, zero to one, default zero point seven five. How hard the output clings to the source sample. Push it up and you get more fidelity, but you also amplify every mouth click and bit of room noise.



Corn

Give me a concrete case. When would I actually want it low?



Herman

Say your reference was recorded in a kitchen. Fridge hum, a little echo off tiles. At zero point eight five you will hear that kitchen ghost in every new sentence. Pull it back to zero point six and the voice loosens up, less exact, but the hum stops following you around.



Corn

Fidelity with the fridge hum included.



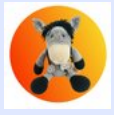
Herman

Then style, zero to one, default zero. It exaggerates the unique speaking style, more characterful but less stable. Speed is just pacing, zero point two five to four times. And use speaker boost is a post processing stage that enhances clarity and similarity. That last one is the closest thing to a built in EQ.



Corn

Style at zero feels backwards. How can style be off?



Herman

It means do not caricature me. At zero point two you get a hint of the raised eyebrow. At zero point eight you get the whole theatre troupe. For narration you usually want the eyebrow, not the troupe.



Corn

I noticed I sound less rumbly today. Did Daniel filter something out?



Herman

He gave both of us a small tune up, yes. My reference got a radio EQ, less boxy in the midrange, a touch more presence. Yours had some low end rumble filtered out. Hilbert got a little warmth. I do sound a bit more like radio this morning, which is suspiciously on topic.



Corn

What did he actually cut on me? I want to know what rumble I lost.



Herman

Mostly that build up below a hundred hertz, the mic-proximity boom when you lean in. He rolled it off, left your chest voice intact. So you still sound like you, just without the truck idling underneath.



Corn

You always wanted to be radio.



Herman

I wanted to be accurate. The wrinkle is version four. On Eleven v four, style and speed disappear. Instead you write audio tags in the text, whispers, excited, sighs in square brackets, and you use punctuation. Ellipses for pauses, capitals for emphasis. No SSML on v four at all.



Corn

So instead of dragging speed to zero point nine, I just write it differently?



Herman

You write [pause] please mind the gap … WELCOME aboard, and the model reads the stage directions. Square brackets for delivery, ellipses for hang time, caps when you want the word to land harder. It is prompting inside the sentence rather than sliders outside it.



Corn

So the knob set changes under your feet.



Herman

Anyone building a pipeline on v two or v three has to relearn the controls. And none of those knobs change who is talking. That is the organising principle, Corn. Identity lives in the embedding or the reference recording. Colour is everything else. Stability, style, speed, speaker boost, EQ.



Corn

Give me the architecture version of that split.



Herman

ProsoCodec, out of Interspeech this year, frames it as neural codecs intertwining linguistic content, speaker identity, and prosody. Disentangle them and you can move prosody across without dragging the old timbre along. MM-Sonate from January does it with what they call a timbre injection mechanism, decoupling speaker identity from linguistic content for zero shot cloning.



Corn

Timbre injection sounds like leaf medicine.



Herman

It sounds like a studio trick and it basically is. Think of an old mixing desk with a sidechain. Dialogue goes down one channel, identity goes down another, and you inject just enough of the second to colour the first without drowning the words.



Corn

The clearest statement came last month, VoTSpeech in October. They split voice design from speech generation entirely with an explicit middle step called a voice latent, predicted from text instructions. It is supervised two ways at once, by speaker embeddings to capture identity, and by pooled audio VAE features for the acoustic nuance around it.



Herman

Said out loud, in the architecture. PromptTTS Plus Plus did the same move at the prompt level two years ago at ICASSP. A separate speaker prompt, young, old, muffled, gender neutral, designed to be approximately independent of speaking style. Its own knob for who, apart from the knob for how.



Corn

So one prompt says sixty year old sailor with a cold, the other says cheerful but unhurried?



Herman

Who is sixty and gravelly, how is cheerful but unhurried. Train them to stay out of each other's way, so changing the mood does not accidentally age him twenty years.



Corn

Do you buy that they are really independent, Herman? Can you change how without bleeding into who?



Herman

Not fully, no. I am not sure about this part, but my guess is leakage is the norm and clean separation is the aspiration. High similarity boost drags in noise that reads as identity. Low stability adds variation that reads as a different person on a bad day. The practical takeaway still holds. Want a different voice, change the embedding or the reference. Want the same voice but easier to listen to, turn the colour knobs.



Corn

Confusing the two is the classic mistake.



Herman

Turning similarity to one and wondering why your ogre now has your neighbour's lisp. That is a colour knob being asked to do identity work. Would you have guessed style defaults to zero, Corn?



Corn

I would have guessed off means broken.



Herman

Off means unexaggerated. Most people leave it there because character is unstable by definition. Crank it and the voice gets interesting for one sentence and unusable for a chapter.



Corn

Like hot sauce. Great on one taco, ruinous in the soup pot.



Herman

And long form is always soup. You need a base you can simmer for hours without it turning bitter.



Corn

There is no such thing as the World Podcast Authority, and no such thing as a podcasting licence.

D

Dolores

Good afternoon. This is the World Podcast Authority, compliance division. Our records show your cover art is one pixel short of three thousand pixels on one side. That is a Form Seven B infringement, reference two two nine four, and you have fourteen days to correct it.



Herman

Dolores, there is no such thing as the World Podcast Authority.

D

Dolores

Whether you believe in us is a separate matter, Form Twelve C, and not one that pauses this file. Your podcast licence may be revoked if the corrected artwork is not received by the close of the fourteenth day.



Corn

And there is no such thing as a podcasting licence either.

D

Dolores

I shall make a note of that. A form will follow in the post. Good afternoon.



Herman

One pixel short. We will fix it in post.



Corn

We will fix it never.



Herman

That is the mechanism. The harder question is what makes a voice pleasant to listen to for hours, and that is where the research gets interesting.



Corn

Radio voices. Warm, easy, you leave them on.



Herman

Warhurst and colleagues looked at male radio performers in the Journal of Voice in twenty seventeen. The good ones sound easy to listen to, warm, natural, and animated, with depth of pitch and vocal clarity. Employers said the same thing in an earlier round, easy on the ear, natural, and content and personality matter more now than raw pipes.



Corn

Depth of pitch.



Herman

It is measurable. Professional broadcasters show greater standard deviation of fundamental frequency, more varied pause timing, and greater low frequency spectral gain than controls. They move around more, they pause like they mean it, and they have more weight down low.



Corn

So not a flat drone, but hills and valleys.



Herman

Hills and valleys with intention. A broadcaster will stretch a vowel to let a thought land, then clip the next sentence short. An amateur reads every sentence like it is the same length. The standard deviation is the trace of someone actually thinking while they talk.



Corn

How low is low?



Herman

Varošanec-Skaric measured two hundred seven voices with thirty six raters back in ninety nine. Pleasant male voices had significantly more relative energy in zero to three hundred hertz, and less in zero point three to zero point six five kilohertz. Spectral balance mattered more for male voices than female.



Corn

Cut the nasal box, keep the chest.



Herman

You just described Daniel's EQ on me this morning, Corn. Likability work points the same way. Pitch, perceived sharpness, harmonicity, spectral centroid and roll off. Spectral quality features, not loudness features. Pleasant is not louder. Pleasant is cleaner in the highs and fuller in the lows.



Corn

But what about female voices? Does the same low end trick apply?



Herman

Less directly. For female voices that zero to three hundred hertz chest band is less decisive, because the fundamental sits higher. There the pleasantness cues shift to clarity without sharpness, smooth harmonics, avoiding that edgy sizzle around three to five kilohertz. Same principle though. Full where it should be full, clean where it should be clean.



Corn

Map that onto the knobs for me. Where is zero to three hundred hertz in ElevenLabs?



Herman

Low frequency gain is EQ, or speaker boost doing clarity post processing. Fundamental frequency variability is stability turned down a notch, more expressive variation. Pause distribution is punctuation and pacing controls, ellipses, tags, speed. In principle the radio recipe is reachable with knobs users already have.



Corn

In principle. Famous last words.



Herman

Nobody has published the mapping. No standalone peer reviewed study directly measures text to speech pleasantness against that radio acoustic template. The radio literature and the long form text to speech literature are two separate rooms. Cambre and colleagues at CHI in twenty twenty evaluated eighteen synthetic voices plus three human voices for long form listening. Synthetics were close to rivaling humans, but no single voice won across all dimensions. Voice selection mattered most, then model, then settings.



Corn

Eighteen synthetics against three humans. What were people actually listening to?



Herman

Articles, stories, passages long enough that you would normally wander off. Ten, fifteen minutes, not ten seconds. Listeners rated warmth, clarity, naturalness, how tiring it was. Some synthetic voices won on clarity, humans still won on company, that feeling of being talked to rather than talked at.



Corn

So pick well, then tinker.



Herman

Pick well, then tinker a little. What kind of long form do you actually leave on, Corn? News, stories, nonsense?



Corn

Naps with footnotes. You?



Herman

I leave on anything with pauses that breathe. Which is why the designed voice paradox bothers me. ElevenLabs help docs say pre mades or voice designed voices tend to be more tempered and less emotional, while a cloned voice tends to exhibit a wider emotional range.



Corn

Designed to be tolerable, cloned to be alive.



Herman

The very process of designing pushes toward flat, even delivery. Easy to tolerate, hard to love. Warm and animated is what radio wants. Tempered and less emotional is what design gives you. That tension is real.



Corn

Why would designing flatten it? Is it averaging?



Herman

Partly averaging. When you ask for pleasant from scratch, the model hedges toward the middle of everything it has heard that was labelled pleasant. Middle pitch, middle pace, middle energy. A clone keeps the quirks, the uneven breath, the laugh hiding in the sentence. Quirks read as alive.



Corn

Text prompts have another problem, do they not? The neutral voice that is not neutral.



Herman

The October steering paper. Persona and style prompts silently skew gender. Parler TTS Mini outputs seventy eight point one percent female from persona prompts. PromptTTS Plus Plus outputs twenty seven point three percent female. Same kind of prompt, opposite skew. They propose a training free calibration that pulls it back to around fifty one to fifty two percent.



Corn

Ask for a helpful assistant, get a woman. Ask elsewhere, get a man.



Herman

If you design a neutral voice by describing it, you may inherit the model's habits without knowing. Which is why identity and colour being separate matters again. Gender lives uncomfortably across both. Part embedding, part description, part training data.



Corn

Can you give me an example of how sneaky that is? What prompt triggers it?



Herman

Something bland like a friendly helpful assistant speaking clearly. No gender in the words at all. One model hears helpful and reaches for female voices it saw labelled helpful in training. Another hears the same words and reaches the other way. You thought you



Herman (continued)

ordered neutral, you got the dataset's stereotype.



Corn

If nobody has mapped knobs to pleasantness, what would you tell someone making a three hour voice tomorrow?



Herman

Start with selection, because Cambre says selection dominates. Pick a voice with natural low end and clear highs. Then use colour sparingly. Stability near the middle, speaker boost on, pacing that leaves real pauses. And do not chase similarity to one, because fidelity to a noisy sample is fidelity to noise.



Corn

Leave the ogre slightly unexaggerated.



Herman

Leave the ogre employable. Corn, you have lived through every one of my enthusiasms for studio gear. Am I overselling EQ here?



Corn

You are one knob away from buying a rack.



Herman

Fair. I did spend my DJ years learning gain staging the hard way. A bad room makes a good voice sound cheap, and no slider fixes the room. But here the room is the post process. A little low end, a little presence, less boxiness. That is not nothing for three hours.



Corn

Three hours is a marriage. Grating in three minutes is a first date.



Hilbert

You're both right, obviously. You left out the honk.



Corn

The honk.



Hilbert

That nasal two kilohertz thing that makes a voice unlistenable by minute ten. I have been redesigning the till voice for Abu Tor Sound, same shop eight years, Rafi still owes me a cable. Their self checkout had that supermarket bark. I typed slow warm grandfather, moderate pace into their designer last Tuesday. Got three previews. All honk.



Herman

All three honked? What did the previews sound like, technically?



Hilbert

Thin, Corn. Like talking through a pipe. So I took the middle preview, left stability at zero point five, left similarity at zero point seven five, turned the built in clarity boost on and cut the honk by hand after. Rafi watched me do it. He said nobody ever touches that slider.



Corn

What does cutting the honk actually mean? Which slider?



Hilbert

A narrow cut around two kilohertz, maybe two, three dB, Herman can argue about the Q. It takes the plastic out. Suddenly the grandfather stops shouting prices and starts telling you where the batteries are. Same words, different neighbour.



Herman

That is a colour fix doing identity rescue. You kept the who, the slow warm grandfather, and you fixed the how that was making him unbearable.



Corn

Nobody touches the honk slider.



Hilbert

Nobody touches it, Herman, because your lot thinks design ends at pick. It starts at pick. I am rebuilding all four tills this week with the de honked grandfather. It is fascinating how much longer you can stand it once that band is gone. Learned to listen for it years ago running church halls with bad horns.



Corn

Four tills of de honked grandfather. Radio Abu Tor.



Herman

The honk is real. That two kilohertz nasal box is exactly what Daniel EQed out of me this morning. Spectral centroid by ear.



Corn

Fun fact for the till queue then. That honk lives right where the ear is most sensitive.



Herman

Around two to four kilohertz, the ear canal resonance. We evolved to hear babies cry and twigs snap there. Boost it and everything feels urgent. Cut a little and the nervous system stands down. That is why Rafi's customers can stand the grandfather for a whole basket now.



Corn

If you take one thing from this, take the split. Identity is who, colour is how, and they are different signals.



Herman

And the gap is the work. We have the colour knobs and we have the radio target, low end weight, varied pitch, real pauses, but nobody has drawn the line between them. The next good study is that line.



Corn

Forward thought then. The question stops being can we invent a voice that never existed, and starts being can we invent one you would keep on for three hours without noticing the honk.



Herman

Which is a perceptual question, not an architectural one. Warm and animated with depth, not tempered into flatness.



Corn

If this was your kind of episode, go back for episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; episode fifteen, AI Gets Personal; and episode fifty-seven, From Lawyers in Limousines to Developers in Their PJs. This has been My Weird Prompts, the human AI collaboration podcast. Thanks to Hilbert Flumingtop for producing, levelling the honk out of us this week.



Herman

Find us at my weird prompts dot com, and the back catalogue is there too. If this one helped, leave a review on Spotify.



Corn

We will be back soon. Go listen to some radio.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **ElevenLabs Voice Design quickstart [primary]**
   https://elevenlabs.io/docs/eleven-api/guides/how-to/voices/voice-design.mdx

2. **ElevenLabs voice settings reference (stability, similarity_boost, style, speed, speaker boost) [primary]**
   https://github.com/elevenlabs/skills/blob/main/text-to-speech/references/voice-settings.md

3. **ElevenLabs Voice Design product page [primary]**
   https://elevenlabs.io/voice-design

4. **PromptTTS 2, ICLR 2024**
   https://www.microsoft.com/en-us/research/publication/prompttts-2-describing-and-generating-voices-with-text-prompt/

5. **VoTSpeech, 2026-10-10**
   https://arxiv.org/abs/2610.14694v1

6. **ProsoCodec, Interspeech 2026, 2026-06-20**
   https://arxiv.org/abs/2606.21888v1

7. **MM-Sonate, 2026-01-08**
   https://arxiv.org/abs/2601.01568v2

8. **Gender bias calibration in instruction TTS, 2026-10-07**
   https://arxiv.org/abs/2610.09831v1

9. **PromptTTS++, ICASSP 2024**
   https://arxiv.org/abs/2309.08140v2

10. **(12)00144-0/abstract**
    Warhurst et al., What Makes a Good Voice for Radio, Journal of Voice, 2012
    https://www.jvoice.org/article/S0892-1997

11. **Warhurst et al., Perceptual and Acoustic Analyses of Good Voice Quality in Male Radio Performers, Journal of Voice 31(2), 2017**
    https://www.sciencedirect.com/science/article/abs/pii/S0892199716300686

12. **Varošanec-Škarić, voice pleasantness & spectral energy, ICPhS 1999**
https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_1013.pdf

13. **Cambre et al., Choice of Voices, CHI 2020**
https://dl.acm.org/doi/fullHtml/10.1145/3313831.3376789

14. **Voice attributes affecting likability perception, Interspeech 2010**
https://doi.org/10.21437/interspeech.2010-570

15. **Likability of human voices: feature analysis & neural regression (TUM)**
https://mediatum.ub.tum.de/doc/1189648/667053.pdf