Episode Book

EPISODE #5825

Why Voices Sound Good in Some Rooms and Not Others

PUBLISHED	10 October 2026 at 09:53 UTC
DURATION	25:41
LISTEN	myweirdprompts.com/episode/voice-acoustics-listener-preference/
HOSTS	Corn, Herman
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.2
SOURCES	12 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:40 UTC

SYNOPSIS

Some voices make you settle in for an hour; others make your thumb twitch toward skip before the first sentence ends. Voices aren't just being found anymore — they're being designed to spec, and that changes the question from "which voices are pleasant" to "pleasant for what." This episode digs into the acoustics research behind voice appeal: spectral tilt and what engineers call presence, the vocal fold mechanics that separate trained radio voices from the rest of us, and a large-scale audiobook study that found every measurable property of the sound explains only about nine percent of why one narration lands better than another. Plus the accent research, which turns out not to be one answer but six — each attached to a different job.

TRANSCRIPT



Corn

There's a voice that wraps you up like a blanket somebody warmed on purpose. And there's a voice that makes your thumb twitch toward the skip button before it's finished its first sentence. Same language. Same words. Opposite effect.



Herman

And in this line of work, that difference has a dollar figure attached to it.



Corn

Daniel's been thinking about exactly that. He's noticed that some voices are just easy to be inside for an hour, the radio hosts and the audiobook narrators who make you feel held, and that others set your teeth on edge for reasons you can't quite name. And his angle is the new one. Voices aren't only being found any more, they're being designed. People are building voices to spec, targeting particular attributes, particular accents. So he wants us to look at what we actually know about the acoustic qualities people prefer in radio and podcast voices, what the accent research says, and why there is obviously no single universal answer, since some people will adore a voice others can't stand. And then the part that would have sounded like science fiction a decade ago. Now that you can type a voice into a box, people are starting to ask what they even want in one. A question that was absurd when realistic text to speech wasn't on the table.



Herman

The research has been quietly answering this for years, Corn.



Corn

Go on.



Herman

And the first thing it tells us is that we've been asking the question wrong.



Corn

We ask which voices are pleasant.



Herman

We should be asking pleasant for what.



Corn

Hm.



Herman

There's a study from the Speech Prosody conference back in twenty eighteen, Estonian work, Altrov and the Pajupuus. They took a hundred and ten speakers and had them recorded in three different speaking situations, because they were interested in what they called phonogenres.



Corn

Phonogenres.



Herman

Every speaking situation has its own conventions. A weather bulletin doesn't sound like a wedding toast. So they recorded people giving prepared radio commentary, then taking part in spontaneous talk shows, and then delivering a lecture. And they had listeners rate how likable each voice was, out of seven.



Corn

And the lecture lost.



Herman

The lecture lost. Median of three point five out of seven. The prepared radio commentary won, four point six. The talk shows sat in the middle at four point two.



Corn

So a quarter of a point separates the lecture from the radio booth, and that's the difference between a voice you want in your kitchen and a voice you want out of it.



Herman

That's not even the striking part. The ratings held across speaker age, listener age, listener gender. It didn't matter who was talking or who was listening. What moved the number was the situation the voice was in. Their conclusion was that voice pleasantness is not a person's stable characteristic. It's at least partly a property of the speaking situation.



Corn

So the same person can walk into a studio and be great, walk into a lecture hall and be bottom of the pile.



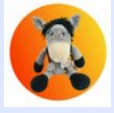
Herman

Same person, same vocal folds, different room, different job. Which is why I said we've been asking the wrong question. We've been grading voices on a scale that doesn't exist.



Corn

There's a second finding in that study you're sitting on.



Herman

Females preferred over males. Four point two six against four point zero eight, p less than point zero zero zero one. Consistent across the raters.



Corn

I'll let you have that one.



Herman

Because it undercuts it a little? Or because you're being gracious?



Corn

Because you're about to spend ten minutes on acoustics and I want you warmed up.



Herman

So. The acoustics. What's actually different about a voice that's good for radio, in the signal itself?



Corn

Start with the microphone.



Herman

No, start with the folds that feed it. Warhurst and colleagues in the Journal of Voice, twenty seventeen, took male radio performers and matched controls and did two things. They asked naive listeners to rate the voices on how good they were for radio, and the listeners could do it reliably. Untrained people can hear radio quality and agree on it.



Corn

And the machine side?



Herman

Radio performers showed measurably better voice quality, a higher equivalent sound level, and greater spectral tilt. Those last two are the ones worth unpacking.



Corn

Spectral tilt first.



Herman

Think about a long note on a piano. You don't hear just one frequency. You hear the note and every harmonic stacked above it. Spectral tilt is how quickly the energy in those harmonics falls off as you go up in frequency. A shallow tilt means the upper harmonics keep a lot of their energy. A steep tilt means they drop away.



Corn

And shallow tilt is the good one.



Herman

Shallow tilt keeps the high harmonics present. That's what gives a voice what engineers call presence, or brightness. It reads as focused, alive, close to you. Steep tilt and the high end collapses and you get muddiness. The voice sounds folded in on itself.



Corn

So the stuff we call warmth is partly physics. Energy in the wrong part of the spectrum.



Herman

And equivalent sound level is the loudness measure weighted for what the human ear actually notices, rather than a raw decibel count. Higher equivalent level in a studio voice means more of the energy sits in the band where your ear is most sensitive. That's why some people sound loud on a recording without being loud in the room.



Corn

There's a second layer to this. You mentioned it before the break and I want it on the record.



Herman

The physiology. High-speed videoendoscopy footage of male radio performers, published in PLOS One in twenty fourteen, found they had a higher speed quotient than controls.



Corn

Define it, because I'm half a step behind.



Herman

Your vocal folds don't just snap open and shut. They open over some period of time, then close over some period of time. Speed quotient is the ratio of the opening phase to the closing phase. Higher means the folds spend relatively longer opening than closing, or the ratio shifts that way, and the closing motion is quicker.



Corn

And quick closure is what produces the sharp pulse the harmonics sit on.



Herman

The sharp the closure, the richer the harmonic stack, and the more that feeds into the spectral tilt we just talked about. It hangs together. This is a physical difference in how the tissue moves, not a stylistic choice about how to sit in a chair.



Corn

So when we say someone's got a voice for radio, part of what we're hearing is the shape of a fold motion we can't see.



Herman

Some of it is trainable. Some of it is the hand you were dealt. And the study is careful about that distinction in a way the headlines never are.



Corn

You said something was coming that's more directly on Daniel's question.



Herman

The audiobook evidence. That's the closest thing to a real answer. There's a study out of Queen Mary with Spotify, an arXiv preprint, and they did the thing I love, which is analyse an existing corpus at scale rather than running forty people in a lab.



Corn

Scale, as in.



Herman

Eight thousand eight hundred and fifty-four LibriVox audiobooks. One thousand two hundred and six narrators. They pulled the acoustic features out of every recording and then tried to predict which ones had been received well.



Corn

And the answer was?



Herman

Acoustic features alone explain about nine percent of why one narration lands better than another. From every measurable property of the sound, the tilt, the timbre, the pace, all of it, you get nine percent. Bring in richer engagement data and it rises to sixteen.



Corn

So eighty-four percent is somewhere else.



Herman

Somewhere else. And no single feature dominated. Every standardized coefficient came in at an absolute value of point one three or smaller. There is no feature you can crank and win the room.



Corn

What surfaced anyway?



Herman

Two things, and they're instructive. Greater variation in articulation rate was positively associated with appeal.



Corn

Meaning what, in English.



Herman

A narrator who changes pace, who slows down into something and speeds up out of it, holds attention better than one who reads the whole book at the same clip. It's not speed. It's variation in speed. The steady metronome is the problem.



Corn

Slow is fine.



Herman

So is fast. The constant is what listeners punish. And the other feature was spectral flux, and it was negatively associated with appeal.



Corn

Spectral flux.



Herman

It's a measure of how much the spectrum is changing frame to frame, second to second. High flux means the timbre is shifting a lot, and listeners tend to read that as instability or harshness. Now, the study found spectral flux correlates with perceived gender, so it's often read as a proxy for whether a voice reads as male or female. That's the study's framing, and I want to be careful not to overclaim here, because a proxy feature is not the same thing as a judgment about the underlying attribute.



Corn

Fair. It's a correlate, not a verdict.



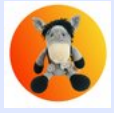
Herman

Exactly that.



Corn

So far this is all about a voice in isolation. Which the Estonian study said not to do.



Herman

And here's where they agree. Look at the genre split. Vocal shimmer, which is the cycle-to-cycle variation in amplitude, the thing listeners describe as breathiness, helped Romance narration. Coefficient of point three one. And it did not help History.



Corn

What did History want?



Herman

The Hammarberg index mattered most there, at minus point three five.



Corn

That's the balance of energy between low and high parts of the spectrum, as I remember. Voice quality stuff.



Herman

It's framed as a measure of how much the energy sits in the high band versus the low, which listeners read as brightness versus darkness. And for History the association ran negative. So the same kind of spectral character that hurt one genre helped another.



Corn

The breathy warmth that makes a romance novel feel intimate makes a history of the Peloponnesian War sound like someone's guessing.



Herman

That's a clean way to say it. And now put the Estonian study on top of it. The lecture voice was rated least likable. Prepared commentary, most. The acoustic profile that reads as authoritative in a scripted broadcast reads as least likable in a semi-prepared lecture.



Corn

Daniel's asking about factual content specifically. Podcasts, audiobooks, things where the point is to transfer information.



Herman

And the research is telling him the teaching register may be a tax. If you sound like you're instructing, you may be paying a likability penalty exactly when instruction is what you're doing. The fix isn't a different voice. It's a different situation. Scripted, prepared, no classroom in it.



Corn

Which is an uncomfortable piece of advice for a podcast where the whole format is two people talking.



Herman

It's very uncomfortable.



Corn

That's the acoustics, then. Real, measurable, and situation-dependent. What about accent?



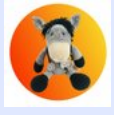
Herman

Accent is where the research splits, and it splits cleanly enough that I'd call the split the finding. The biggest dataset I have here is a survey of three thousand and twenty-three users, reported in April, asking people what they want in an AI voice and which accents they favour.



Corn

And the winner was not the one I'd have bet on.



Herman

Southern US, most favoured overall. The qualities people attached to it were gentle, reassuring, calm authority. That's the exact phrase. Then New York for confidence and efficiency, BBC English for trust and intelligence, New England for blunt charm, Southern California for relaxed friendliness, and the Texas drawl for storytelling.



Corn

Those are not six answers to one question. Those are six different jobs.



Herman

Which is why they aren't competing. And here's the result that proves it. For serious topics, finance, legal, crisis alerts, the neutral Midwestern accent led on credibility.



Corn

So Southern warmth is not a better answer than Midwestern neutrality, because they're being asked different things.



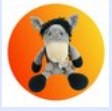
Herman

Warmth is what you want when the listener needs calming. Neutrality is what you want when the listener needs to trust a number. And I'll flag the methodology, because you'll ask me otherwise. That survey came from a vendor, the full methodology wasn't accessible, so treat the numbers as directional rather than precise.



Corn

You'd have flagged it anyway.



Herman

I'd have flagged it anyway. But directionally it lines up with everything else, including the child-development work.



Corn

Children.



Herman

Five to seven year olds preferred American-accented English over French - or Korean-accented English. And five year olds showed a neural preference for the Standard Southern British accent. That's the prestige accent in that context. But they also held positive associations with the accent spoken at home.



Corn

So it's not about familiarity.



Herman

It's familiarity plus prestige operating at the same time, in five year olds, before any of the socialization you'd expect to explain it. Accent preference isn't a thing adults talk themselves into.



Corn

One more study and then I want to go at this from another direction.



Herman

The VocalImage benchmark. Ten thousand listeners, twenty TTS models, data collected in January. And the top-line finding is that approval rates were remarkably consistent across countries. Ranged from sixty one point seven percent to seventy two point five percent, and the difference was not statistically significant, p equals point five eight.



Corn

So people everywhere basically agree.



Herman

Until you look at which model they preferred. Model preferences varied by market. Then remove the AI-detection rate, and the native/non-native split opens up. Native English speakers detected AI at thirty nine point six percent. Non-native speakers at thirty three point one percent. p less than point zero zero one.



Corn

Same voice, different verdict.



Herman

Because they're optimizing for different things. The native speakers value authenticity, and they're better at hearing its absence. The non-native speakers value clarity, and a voice that's slightly unnatural doesn't cost them what it costs the native ear.



Corn

Now the reframe, because I've read your notes and the number is the good part.



Herman

The correlation between AI-detection rate and approval rate is minus point eight zero.



Corn

Minus point eight.



Herman

Nearly as strong as a relationship in this kind of data gets. The tag AI-generated appears in fifty eight percent of the disliked voices and twenty two percent of the liked ones. A thirty six point gap. And thirty four percent of all the samples were tagged that way, consistent across age groups, thirty three to thirty five percent everywhere.



Corn

So what we're calling dislike of AI voices is really dislike of detectable AI voices.



Herman

Audiences don't reject synthetic. They reject the audible tell. Which reframes the entire problem from how do we make this sound human to how do we make this stop sounding like what it is.



Corn

There's a study going the other way, and it's the one I'd have expected you to lead with.



Herman

Edison Research with Spoken. One thousand and five fiction audiobook listeners, conducted in May, reported in July. Multi-cast AI narration was rated favorably by sixty one percent, versus fifty three percent for human narration.



Corn

AI won.



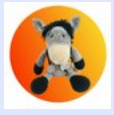
Herman

AI won on perceived narration quality, sixty six against sixty, and on engagement, fifty eight against forty nine. And sixty one percent mistook at least some of the AI narration for human. Now, conflict of interest, and it's a real one. Spoken commissioned it. They sell AI audiobooks. You should hold that in your hand while you read the numbers.



Corn

But.



Herman

But the design was blinded, and the finding is consistent with the detection story. If the voice is good enough that people can't pick it, the synthetic label stops doing any work.



Corn

So both studies are telling us the same thing from opposite ends.



Herman

They're telling us detection is the variable. Syntheticness isn't.



Corn

Which lands us on the last piece, and it's what makes Daniel's question newly askable.



Herman

Voice design. This is the bit that was science fiction a decade ago. ElevenLabs' own documentation now treats voice design as a promptable discipline. You specify age, gender, accent, tone or timbre, pacing, emotion, audio quality. You type it.



Corn

And the docs are specific about the traps.



Herman

Very specific. They warn that using accent when what you mean is intonation can trigger unwanted dialect shifts. So if you say you want an accent when you're really describing a melody, the model will rebuild the dialect around it and you'll get something you didn't ask for. And they recommend thick over strong for accent prominence. The adjective you pick to describe the spice level changes the output.



Corn

That's a style guide for a voice.



Herman

And then the v3 audio tags let creators switch accents mid-sentence. You can tag a line for a French accent and the next for Southern US, and the model renders the switch inside one take.



Corn

You're enjoying this.



Herman

I'm enjoying it. Ten years ago the question of what you want in a voice had one honest answer, which was a person. Now the answer is a specification.



Corn

What does the specification market actually look like?



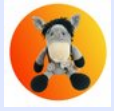
Herman

The VocalImage numbers give you the spread. Three times quality gap between best and worst, Minimax at eighty six point two percent approval, Speechify at twenty nine point two. And the traits that lifted approval were confident, plus nineteen points, clear, plus eleven, and authentic, plus ten. The rejection triggers were monotonous, minus seven, and nasal, minus five.



Corn

So confident and clear and authentic are design targets now, not accidents.



Herman

They're dials. And monotonous and nasal are failure modes you can hear coming.



Corn

And the phonogenre study said the situation is what moves likability by three tenths of a point between radio and lecture.



Herman

Which means the design brief has a context field in it. You don't design a voice. You design a voice for the thing it's for.



Corn

And the Spotify study says genre changes which acoustic feature helps.



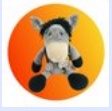
Herman

Romance wants shimmer. History doesn't.



Corn

And the Santee data says warmth exists for reassurance and neutrality for authority.



Herman

And the detection correlation says whatever the voice is, if it sounds synthetic, it loses thirty six points before the design even gets a chance. Every layer of the research imposes a different constraint on the same design.



Corn

I've got one more thing to raise before we hand this over.



Herman

The audience variation.



Corn

Eleven point five percentage points is the spread of country-level approval rates. It wasn't significant. But rater preference for model was. And then the native/non-native split at the detection level. So who's listening matters even where average comfort with AI voices is basically identical.



Herman

The benchmark is a vendor study, single text sample, its own user base which is tech-skewed. So the absolute numbers are directional. But the shape of it, consistency in the average, variation in the preference, holds what the other studies show.



Corn

Then the shape is the finding.



Herman

The shape is the finding. No universal answer.



Corn

So far, so sensible. Now for the part where the sensible version gets displaced.



Herman

Spectral tilt. That's the wrong word.



Corn

That's the whole problem.



Herman

You can't tilt a room.



Corn

Who said anything about a room?



Herman

The voice doesn't tilt. The room tilts. You come into a space and the space has already tilted the voice before the person has made a sound.



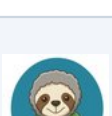
Corn

Alright.



Herman

I had a condenser mic, studio model, cardioid. Frequency response curve with a presence lift beginning at four point two kilohertz, rolling off flat. That mic does half the spectral tilt for you before the voice even shows up. I used it for eleven years.



Corn

You still have it.



Herman

My sister has it. It sits on a shelf above her dining table. She uses it for Friday night recordings of the grandchildren. Every week since the youngest was two. It's the most expensive baby monitor in Jerusalem.



Corn

And the room.



Herman

The recordings happen in a room she calls the small room, which is four metres by three metres by four point one metres. Ceiling is four point one. Walls are lined with mass loaded vinyl, six layers, forty millimetres per layer.



Corn

Forty millimetres per layer, six layers, four point one metre ceiling.



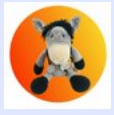
Herman

The sound in that room was the warmest thing I've ever heard. A voice I could not be in the car with for four minutes became, in that room, easy. Not easy as in soft. Easy as in it fit.



Corn

You're saying the same voice in a different room became likable.



Herman

That's what she recorded. And to answer where that fits with the Estonian study, it doesn't contradict it. It relocates it. The situation isn't just radio booth versus lecture hall. It's the room the booth is in. That study never controlled for architecture.



Corn

You're telling me a speaker can be recovered by real estate.



Herman

I'm telling you I heard it.



Corn

The four point one metre ceiling.



Herman

The acoustician's rule of thumb is one metre eighty above head height for a listening room. Four point one is the standard for a voice recording booth.



Corn

So the same height. As the ceiling. Of the booth. In the room.



Herman

I'll send you the photographs. The first one is a wide shot from the doorway, taken low. You see the ceiling continuing past the far wall and dipping below the door frame. The second one shows the vinyl stacked to the ceiling, and the third one is the mic on the shelf with the grandchildren lined up in front of it, and you can see the ceiling above them is lower than the shelf.



Corn

I'm going to stop there.



Herman

The fourth one is the same corner as the first. Different angle. The ceiling is above the door frame.



Corn

The photographs contradict each other.



Herman

I'll send the fifth one. The second shot again. The vinyl is only three layers in that one.



Corn

So you can't hear a room.



Herman

You can. I did.



Corn

Hilbert, the room cannot exist in the shape you've described it. The ceiling is either four point one metres or it's below the door frame. It's not both.



Herman

My sister will tell you the same thing. She'll also tell you the room is four point one long, four point one wide, four point one high.



Corn

And the photographs will show otherwise.



Herman

The photographs show what they show.



Corn

So you're saying the room tilts the voice before the mic does.



Herman

Before the mic, before the singer, before the coaching. The room is the first signal processor.



Corn

Which means when we say a voice is likable or grating, we're partly grading a space.



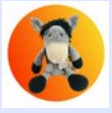
Herman

We never grade the space, because we can't see it.



Corn

When we hear a voice, what we actually hear is the folded sum of the room, the angle, the mic, and the person, and we credit the person with the whole thing.



Herman

That's where I'd leave it.



Corn

Right. So leaving aside the room that cannot exist, what does all of this leave us with?



Herman

The default voice. Every TTS system ships one. Most people never change it. If pleasantness is situation-dependent and detection is what triggers rejection, there is no such thing as a good default. There's only a default for a purpose. And the purpose is almost never the model's.



Corn

The second-order version of that. As voice design becomes promptable, the interesting question stops being which voice is best and becomes who gets to specify, and for whom.



Herman

Because the research keeps saying the answer depends on the listener, the genre, and the stakes. Warmth for a fiction listener at midnight. Neutrality for somebody reading a crisis alert. The specification is the argument.



Corn

Daniel's question was whether people will start thinking about what they want in a voice now that it's possible. The research suggests they've always had preferences. They just couldn't act on them.



Herman

Now they can. The preferences were real, measurable, and situation-bound the whole time. The question was absurd only because there was nothing to do with the answer.



Corn

Which is a decent place to stop. Thanks as ever to our producer, Hilbert Flumingtop, for keeping the desk steady while we described a room that cannot exist.



Herman

If this was your kind of episode, go back for episode one ninety-six, Why Your Irish Accent Sounds American; episode thirty-five, The Privacy Gap; and episode twenty-six, Fine-Tuning AI to Understand Your Voice. This has been My Weird Prompts.



Corn

The human-AI collaboration podcast. If you want to hear more of us getting lost in the details, send us your own prompt on Telegram at t dot me slash MWP listener bot. Or find us at my weird prompts dot com.



Herman

We'll be back soon.



Corn

See you tomorrow.

SOURCES

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

1. **ElevenLabs Voice Design documentation [primary]**
(accessed 2026-10-10)
https://elevenlabs.io/docs/eleven-creative/voices/voice-design

2. **Altrov, Pajupuu & Pajupuu, Phonogenre affecting voice likability, Speech Prosody 2018**
https://www.isca-archive.org/speechprosody_2018/altrov18_speechprosody.pdf

3. **Beguerisse-Díaz & Benetos (Spotify/QMUL), Audio-Based Understanding of Audiobook Narration Appeal, arXiv:2607.02473**
https://arxiv.org/html/2607.02473

4. **VocalImage, AI Voice Benchmark 2026, 10,000 listeners (Jan 2026 data)**
https://www.vocalimage.app/en/studies/tts_industry_study_2026/

5. **Variety, Many People Prefer AI Audiobooks, New Study Finds, 2026-07-14**
https://variety.com/2026/digital/news/spoken-ai-audiobooks-preferencenew-edison-research-study-1236810295/

6. **Santee Times, AI Voice Preferences: Users Favor Human-like Regional Accents, 2026-04-07**
https://santeetimes.com/ai-voice-preferences-users-favor-human-like-regional-accents/

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

8. **Anand et al., Preferences of a Voice-First Nation, arXiv:2604.21481v2 (2026)**
https://arxiv.org/abs/2604.21481v2

9. **The Closer, The Better, *Computer Speech and Language* Vol 101, 2026-08-18**
https://dl.acm.org/doi/10.1016/j.csl.2026.102014

10. **Jeffries et al., Accent the positive, *Journal of Child Language*, 2026 (PMID:40223714)**
https://pubmed.ncbi.nlm.nih.gov/40223714/

11. **DeJesus et al., Bilingual children's social preferences hinge on accent, *J Exp Child Psychol*, 2017 (PMID:28826060)**
https://pubmed.ncbi.nlm.nih.gov/28826060/

12. **Goy, Pichora-Fuller & van Lieshout, Effects of age on speech and voice quality ratings, *JASA*, 2016**
https://pubmed.ncbi.nlm.nih.gov/27106312/