Episode Book

EPISODE #5678

Why Voices Are Harder to Design For Than Music

PUBLISHED	2 October 2026 at 14:17 UTC
DURATION	24:56
LISTEN	myweirdprompts.com/episode/speech-audio-listening-effort/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.2
BOOK GENERATED	19 October 2026 at 17:38 UTC

SYNOPSIS

Daniel listens to podcasts with headphones while minding Ezra, and the show has become the soundtrack of his apartment — which got him wondering why it's easy to describe what makes music enjoyable and so hard to describe what makes a voice enjoyable. This episode digs into the audiology research on listening effort: why a podcast can be perfectly intelligible and still tiring, why the effort curve rises before the accuracy curve falls, and how pupil dilation and EEG alpha power let researchers measure the cost of listening. Then it turns to the production side — the consonant band, high-pass filtering, controlled warmth, compression, loudness normalization, de-essing, noise reduction, and mono compatibility — and why speech is judged on the absence of friction rather than on anything you can point at.

TRANSCRIPT



Corn

Okay, so the thing about a mudroom is that nobody ever actually builds it for the mud. They build it for the shoes.



Herman

Which is the same mistake as building a listening room for the furniture.



Corn

Right. Anyway. Daniel wrote this one in, and it's less a prompt than a confession with a question stapled to the end. He's listened to the show with headphones most of his life, but he started making it because he needed something to keep his brain occupied while he's minding Ezra, and it's turned into his DIY companion. Hannah listens too. Sometimes they're both doing their own thing and the show becomes the shared soundtrack of the apartment, and he says fairly plainly that it made unpacking the new place less miserable.



Corn

From that he pulls two threads. First, the EQ and digital side. What actually makes a podcast easier and more enjoyable to hear, from the production end. Second, speakers, where he wants the unlimited-money treatment: not brands, because products churn, but the characteristics a speaker needs to shine in the scenario he describes, which is two people, one soundtrack, moving around a living room for a few hours with a drill in one hand.



Corn

And then the line that stuck with me, and I think is the whole episode. He says it's easy to describe what makes music enjoyable. It's much harder to describe what makes a voice enjoyable. People default to "it's just a voice, it doesn't matter how you play it." He doesn't think that's true.



Herman

He's right, and there's a reason he can't put his finger on it. So where do we start?



Corn

With the obvious question underneath all of it. Why does a voice need designing for at all?



Herman

Because the brain does two completely different jobs. Music is designed to be heard as a whole. When you're listening to music you're parsing melody, harmony, rhythm, timbre, all at once, and you have enormous tolerance for coloration. You can hear the same song on a car radio, a tinny laptop, a good system, and it's still the song. The brain is doing an aesthetic job.



Herman

Speech is a carrier of semantic information. The brain isn't evaluating the sound, it's decoding phonemes, tracking syntax, building meaning. Which means the failure modes are different. Music fails as "that doesn't sound good." That's a judgment about tone. Speech fails as "I have to work to understand this." That's a judgment about cognitive load.



Corn

And that second one is the whole concept. There's a term for it in the audiology literature. Listening effort.



Herman

And here's the thing that makes it worth an episode. A podcast can be a hundred percent intelligible on a phone speaker, every single word arriving correctly, and still be exhausting. The words got through. The cost of getting them through is the problem.



Corn

Which is exactly what he's describing when he says easy audio makes it easier to get into flow. He's not complaining about not hearing it. He's complaining about the tax.



Herman

Yes. And that tax is what kills the immersion. So what is listening effort, actually, and why does it matter more than just being intelligible?



Corn

Start with the review, because it's the cleanest statement of the idea. Peelle, twenty eighteen, in Ear and Hearing. The argument is that when the acoustic signal is degraded, comprehension doesn't just risk word errors. It consumes executive resources. Resources that would otherwise be going to language processing and to memory.



Herman

So the brain has a budget. And if it's spending part of that budget on fighting the signal, it's not spending it on remembering what was said.



Corn

Which is the line I keep coming back to. Bad audio doesn't just sound worse. It makes you remember less.



Herman

That's the finding. People who heard degraded speech remembered less of it afterwards, performed worse on concurrent tasks, had more trouble with linguistically complex sentences. And the imaging is the part I find striking. Degraded speech doesn't just light up the usual language networks. It recruits prefrontal regions. The brain is pulling in executive control to help with what should be automatic.



Corn

That's a brain doing work it doesn't want to be doing.



Herman

Right. Now the second piece, and this one is clever. Kadem and colleagues, twenty twenty, Trends in Hearing. They used pupil dilation as a direct physiological index of cognitive demand.



Corn

Pupils.



Herman

Pupils. They dilate with mental effort. You can watch the cost of listening happen in someone's eyes. And they found that even without any added noise, semantically ambiguous words dilated pupils. So just the ambiguity of language itself costs something. Add noise and the masking effect dominates the whole signal.



Corn

So anything that makes the signal cleaner is measurably reducing load. That's the practical takeaway.



Herman

It is. And then there's the one that I think is the heart of the episode for Daniel's purposes. McHaney and colleagues, twenty twenty-four, in Scientific Reports. Listening effort rises before intelligibility falls.



Corn

Say that again slowly, because I think it's the most useful sentence in the whole episode.



Herman

At moderate difficulty, you're still understanding everything. You're not missing words. But you're working harder to get them. The effort curve climbs well before the accuracy curve drops.



Corn

So the phone speaker. He can hear every word. He's not confused. And he's tired, and he doesn't know why.



Herman

That's the mechanism. It's not that the phone is failing to deliver the audio. It's that the phone is delivering it in a form that costs more to unpack.



Corn

And then the proof that this is a real engineering target and not a vibe. Valderrama, twenty twenty-five, also in Scientific Reports. They took beamforming, the technique that improves speech in noise, and they asked whether it also reduced listening effort. It did. Faster reaction times on a dual task, lower pre-stimulus alpha power.



Herman

Alpha power being a marker of how much the cortex is gearing up to do work.



Corn

So effort is a separate axis. You can measure it independently of whether the words arrived. Which means "easy to listen to" is a spec you can design for, not a feeling you either get or you don't.



Herman

That's the license for the whole back half of this episode. If it were just a vibe, the only advice would be "buy something that sounds nice." Because it's measurable and separable, we can actually say what to do.



Corn

So let's do the production side first, because that's the half that's directly relevant to us and to anyone else making audio. The EQ and digital choices. Start with the band that carries the information.



Herman

Consonants. The intelligibility band is roughly one to four kilohertz, and that's where consonants live. Vowels are loud and low and carry a lot of the energy, but the information, the difference between "bat" and "back," is up in the consonants. A gentle presence lift in that region is the standard move. Not a spike. A lift.



Corn

And at the bottom, the high-pass. Eighty to a hundred hertz, cut the rumble.



Herman

Because it's eating headroom and adding no information. A truck going past, the building's air conditioning, the thump of someone walking upstairs. None of that is your voice, and all of it is using up the dynamic range you'd rather spend on the consonants, and on the level. Get rid of it before it costs you anything.



Corn

Then the interesting one. Warmth.



Herman

Warmth is the most misunderstood word in audio. People hear "warm" and think "bass." It isn't bass. For speech, warmth lives around two hundred to five hundred hertz, and it's a double-edged sword. Too much and the voice goes muddy and boxy, and intelligibility actually drops, because you're smearing the low end over the consonants. Too little and it sounds thin, and thin is fatiguing, because the brain is hearing a signal instead of a person.



Corn

Controlled warmth, then. Enough body to feel human. Not so much that you're burying the information.



Herman

And then the lever that I'd argue matters more than all the others combined. Compression.



Corn

Dynamic range compression, for anyone whose first thought was the file format.



Herman

Speech has enormous peak-to-average variation. Plosives, breaths, laughter, someone leaning away from the microphone, someone leaning in. If you don't compress, you're asking the listener to ride the volume for you. The quiet passages vanish in a room with any ambient noise, and the loud ones spike.



Corn

And riding the volume knob is itself listening effort. It's a task.



Herman

It is. Broadcast-style compression and limiting raise the average level so the quiet parts stay intelligible, in a noisy room, at low volume. For a background-listening context this is the single biggest thing you can do. It directly reduces the amount of work the listener is doing.



Corn

Which connects straight to the loudness normalization question. Minus sixteen LUFS, the podcast standard.



Herman

Consistent perceived level across episodes and platforms. If your show is leveled properly, a listener can set the volume once and forget it for the rest of the hour. If it isn't, they're adjusting every time an episode starts, and every adjustment is a small reintroduction of effort.



Corn

De-essing. Sibilance control.



Herman

Matters more for speech than for music, for two reasons. One, sibilance is fatiguing over a long listen. Two, at high levels it's a distortion risk on small speakers, which is exactly what a lot of background listening is happening on. Harsh esses are the thing that makes someone turn a podcast down, and turning it down makes the low-mids vanish, and now it sounds thin. It's a cascade.



Corn

And noise reduction and de-reverb. There's a hearing-aid finding here that I think generalizes cleanly. Slugocki, twenty twenty.



Herman

Noise reduction that improves speech in noise also changes the cortical and subcortical evoked responses. The brain's response to the sound itself changes when you clean it up. Which is the general principle stated as evidence: removing competing noise and reverb reduces the cognitive cost of listening. Reverb especially. Reverb is just the room smearing the signal across time, and the brain has to un-smear it.



Corn

And then one that nobody thinks about until it breaks. Mono compatibility.



Herman

Background listening happens on a single kitchen speaker, a portable, a phone in a dock. If your mix relies on stereo width for intelligibility, or if it collapses badly to mono, then in exactly the scenario Daniel's describing, your show falls apart. Check the mono fold-down. It's the least glamorous production note in this entire episode and it probably matters as much as any of them.



Corn

So all of that is the production half. Now the question that's been sitting there the whole time. Why is it so much easier to describe what makes music enjoyable than what makes speech enjoyable?



Herman

Because the qualities are inverted. Music is judged on presence. You point at what's there. Warmth, brightness, soundstage, punch, air. Those are all things you can hear added to a recording. The vocabulary exists because the goal is aesthetic and the listener is consciously attending to the sound itself.



Corn

And speech is judged on the absence of friction. You don't notice good speech reproduction. You notice that you got to the end of the episode and weren't tired, and you noticed the content, and you never once noticed the audio.



Herman

Effortlessness, intelligibility at low volume, non-fatiguing over hours. All of those are negative. They're defined by what isn't happening. Which is why Daniel can hear the difference and can't articulate it. There's nothing to point at. The thing you'd point at is the absence of a thing.



Corn

Music you notice. Speech you only notice when it's wrong.



Herman

And that's a framing, not a finding. But I think it's the right frame for his question. So if listening effort is the cost, what does a phone speaker actually do to that cost, and what would a speaker that reduces it look like?



Corn

The phone is the floor, so start there. What is a phone actually doing?



Herman

Physically, it can't do the thing. The low-mid body that gives a voice chest and presence lives around a hundred and fifty to four hundred hertz. A phone speaker is tiny, it's sealed into a chassis with no volume behind it, and it produces almost nothing below five hundred hertz. It's not that they tuned it badly. The physics won't allow it.



Corn

So everything below five hundred is just gone.



Herman

Gone. And what's left is a thin, telephone-like band that the brain has to work harder to parse, because it's been stripped of the cues that make a voice sound like a person. High listening effort, low immersion. Which is the physical explanation for his intuition. He's not imagining that the nice speaker is better.



Corn

So now define the thing the nice speaker is doing.



Herman

Warmth is the first one. A slightly elevated, well-controlled low-mid response. It's the difference between hearing a person in the room and hearing a signal that happens to be shaped like a person. That's the quality he's intuiting when he says warm sound and good resonance.



Corn

Resonance being the cabinet.



Herman

The cabinet, the enclosure, the port, the transmission line. A properly tuned enclosure produces fuller, more extended low-mids without boom. That's the mechanical difference between a nice speaker and a phone. The phone has nowhere for the air to go. A real speaker has a designed volume of air that resonates at the right frequency and extends the low end down where the voice actually lives.



Corn

So the production half and the playback half are aiming at the same target from opposite ends. The production side is trying to keep two hundred to five hundred clean and controlled. The playback side is trying to actually reproduce it.



Herman

Same band, same reason. The voice has to read as a person, and the person lives in the low-mids.



Corn

Now the scenario. This is where it gets different, because he's not describing a listening chair. Two people, one soundtrack, moving around a room for a few hours with tools in their hands. Every assumption a normal speaker review makes is wrong here.



Herman

Every one. A critical listening setup assumes one listener, seated, in a fixed position, attending to the sound. Every one of those four is false in his case. So walk the characteristics.



Corn

One. Wide, even dispersion. You're not in a sweet spot. You're bent over a box, or up a ladder, or in the next room. A speaker with narrow directivity sounds wonderful in one chair and hollow two meters away, and you will spend the afternoon drifting in and out of the good zone without realizing why the show keeps sounding different.



Herman

Wide, even dispersion means consistent tonality wherever you actually are. In practice that argues for multi-driver arrays, coaxial designs where the drivers are physically aligned, or speakers deliberately designed with broad radiation patterns. Which is the opposite of what a studio monitor is for.



Corn

Two. Full, warm low-mid response at low volume. And there's a real psychoacoustic reason this matters.



Herman

Equal-loudness contours. Human hearing is less sensitive to bass and treble at low levels. So at background volume, a speaker with a flat response will sound thin, because your ears are the thing rolling off the low end, not the speaker. It's why old stereos had a loudness button. The problem it was solving is real.



Corn

And the trap is that a thin speaker at background volume pushes you to turn it up. Which fixes the thinness and makes it intrusive, and now you're not doing background listening anymore.



Herman

A good low-level performance, or a proper loudness compensation, is the thing that lets the show stay quiet and still sound like people. That's the enabler for the whole use case.



Corn

Three. Room-filling rather than directional. You want the sound to fill the space, not beam at a chair.



Herman

Look at an open-plan office. There's real work on sound masking. Mikulski, twenty twenty-two, describes using four loudspeaker columns with directional radiation patterns to distribute sound evenly across a whole office. The engineering problem there is literally "make it sound the same everywhere," and it's the same problem in a living room. The office-acoustics world has thought about it harder than the consumer audio world has.



Corn

Four. Multi-room and multi-source. Two people, different tasks, one soundtrack is a distributed audio problem.



Herman

It's the category that actually matches what he described, better than any single speaker does. Grouped speakers, or a portable battery unit that follows you between the kitchen and the room you're actually unpacking in. The scenario is inherently distributed.



Corn

Five. Effortless control. Hands-free.



Herman

Which sounds trivial until both your hands are holding a flat-pack shelf and you need to skip a chapter. Voice control, or one physical control you can hit with an elbow. A touchscreen across the room is not a control interface, it's a decision.



Corn

Six. Durability and placement flexibility. Dust, knocks, awkward spots.



Herman

A workshop context eats speakers. Rugged portables, wall and ceiling mounts, anything that can be placed badly and still work. This is not a category where you want something precious.



Corn

Seven. Low fatigue over hours. Smooth treble, low distortion, no aggressive presence peak.



Herman

The distortion part is the one people underestimate. Over a multi-hour session, a large chunk of the fatigue is coming from distortion, not from the frequency response. A speaker that measures flat but distorts at moderate volume will wear you down in a way that a smoother, less "revealing" speaker won't. The audiophile tuning is the wrong tuning for this job. You don't want to hear the recording. You want to stop hearing the speaker.



Corn

Which brings us to money being unlimited. What are the categories a budget-driven buyer never looks at?



Herman

Line arrays, or column speakers. They're designed for even coverage across a wide area rather than a sweet spot. That's the office-sound-masking problem solved in a different package. For a room where people move, a column is doing exactly the right job.



Corn

Omnidirectional, three-sixty speakers. Radiate in all directions, so tonality is consistent as you orbit the room. Which is literally described in his prompt. Moving about a living room.



Herman

Electrostatics and planar-magnetics. Dipole radiation, extremely low distortion, famously warm and non-fatiguing in the midrange. These are the classic "voice sounds like a real person" speakers. There's a whole subculture refurbishing old Quad electrostatics, and once you hear a voice through one of those, it's hard to go back. The midrange is doing almost no work, and you can hear that it isn't.



Corn

Active speakers with room correction. Measure the room, compensate automatically.



Herman

Solves the "sounds different in every corner" problem without you having to think about placement. Which for a living room that's mid-unpacking is a genuine gift.



Corn

Portable high-fidelity. And there's a nice case study here. Teenage Engineering's OB-four. Deliberately unusual portable hi-fi speaker, has a mode they call orthodynamic, and the whole product is built around the joy of listening rather than around a spec sheet.



Herman

It's a good example of exactly what he asked for. A product a budget-driven buyer would never think to consider, because it's not competing on price or on measured performance. It's competing on what it feels like to have it in the room.



Corn

Sound-masking and ambient systems, from the office world. Because that's the discipline that has thought hardest about pleasant sound that fills a space without demanding attention.



Herman

DIY and kit speakers, which is worth a mention given he's building things anyway. Kit and open-baffle designs get you into the electrostat-and-omni character space at a fraction of the cost, because you're supplying the labor instead of paying for the finish.



Corn

Every one of those categories is one move. They're all ways of reducing the cognitive cost of listening over hours.



Hilbert

The word's wrong.



Corn

Which word.



Hilbert

Effortless. You've both said it about six times. The speaker isn't effortless. The speaker is one of three things in the room, and you keep talking like it's the only one.



Hilbert

I did acoustic treatment for a small chain of hearing clinics for a stretch. Not the audiology side. The room side. Panels, corners, the boring part. And the thing I learned there, which took me three clinics to believe, is that the single best predictor of whether a patient would keep wearing their hearing aids wasn't the fitting and wasn't the price.



Corn

What was it.



Hilbert

Whether the room they talked to the audiologist in had been treated. That was it. The hardware was fine. The room was undoing it.



Hilbert

There was one consultation room, and I remember it because we went back to it twice. Blank wall behind the patient's chair. So the audiologist's voice leaves her mouth, hits that wall, comes back, and arrives at the patient's ears a few milliseconds after the direct sound. Nothing you'd notice. Just a smear. And the patient would sit there nodding, with a hearing aid in each ear, and then get out to the car park and tell their wife they'd caught about half of it.



Corn

A few milliseconds.



Hilbert

A few. We put one absorptive panel on that wall. One panel. It cost less than the coffee machine in the waiting room. And the complaints from that room stopped.



Herman

That's the room as a third term.



Hilbert

Nobody in the consumer audio conversation wants to talk about it, because you can't buy your way out of it. You can spend a fortune on a system and put it in a room that fights it and lose. Or you can put a modest system in a room someone thought about for an afternoon and win. I've watched both. The second one wins every time for what he's describing.



Corn

The room is a variable, not a constant.



Hilbert

It's the cheap variable. That's the part I'd want him to hear. He asked what to buy. Half the answer is where you put it. He's already moving boxes around in there. He's already got his hands on the furniture.



Corn

The room is the third term. And that changes the question from what to buy to how to think about the space you're listening in.



Herman

It also explains something we skipped. The office-sound-masking thing isn't really about the speakers. It's about how the room distributes and absorbs what the speakers put out. The columns matter, but the ceiling tiles matter too. The room was always in the equation. We just filed it under "speakers" because speakers are the thing you can buy.



Corn

The one thing. If you take one thing from this, it's that the currency of a podcast is not whether the words arrive. It's what they cost to receive. A show can be perfectly intelligible and still be draining you the whole time.



Herman

The fix is rarely one thing. It's the signal, and the speaker, and the room, and the room is the one nobody sells you.



Corn

Daniel's scenario is two people, one soundtrack, moving about a room for hours. Almost every consumer audio product is built for one person, seated, in a sweet spot, paying attention. The moving, shared, background case is treated as an afterthought. There's a real question about what gets built if you design for that case first.



Herman

As spoken-word audio keeps growing, that question gets more important, not less. The places people listen are getting noisier and more mobile, not quieter and more seated. And the listening-effort frame scales with that.



Corn

If this episode made you think differently about how you listen, a review helps other people find the show. Thanks to producer Hilbert Flumingtop. This has been My Weird Prompts.



Herman

Email us at show at my weird prompts dot com. We'll be back soon.