Episode Book

EPISODE #6230

Why Synthetic Speech Needs Fake Room Noise

PUBLISHED	23 October 2026 at 14:34 UTC
DURATION	20:38
LISTEN	myweirdprompts.com/episode/synthetic-speech-room-tone-beds-79224/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.20
SOURCES	12 cited by the research (listed at the end)
BOOK GENERATED	24 October 2026 at 07:26 UTC

SYNOPSIS

Synthetic speech can stop completely between sentences, leaving gaps of true digital black that sound unnatural on playback. This episode digs into why engineers fill those holes with room-tone beds, how a real fix for this show landed around minus sixty dBFS on a phone, and why a constant quiet background beats a perfectly clean pause. Along the way: what room tone actually contains, why filmmakers record thirty to sixty seconds of nobody doing anything, and why audiobook noise-floor specs collide with loudness normalization.

TRANSCRIPT



Corn

Daniel heard little fade-ins at the start of our turns, mostly on his phone. The strange part was that the fades weren't in the audio file. That led him into today's question. Why does synthetic speech sometimes need fake background noise to make the silence sound right?



Herman

The gaps were much quieter than intended. The room-tone bed was there, technically, but low enough that the playback system could behave almost as though nothing was playing.



Corn

A background employee with no duties.



Herman

Roughly a hundred gaps per episode were sitting around minus seventy-five decibels relative to digital full scale. Daniel tried louder beds on his phone, preferred minus sixty, and that's where the pipeline now puts those gaps.



Corn

His question widens out from that fix. Real recordings carry a room, microphone noise, a little air. Synthetic voices can stop completely between sentences. Engineers have been filling holes under dialogue for decades, but when there was never a room, somebody has to decide what goes there. I like that we've reached the point where silence requires a production budget.



Herman

The important distinction is between hearing steady quiet and hearing the background switch off. Those can have very different effects, even when neither sounds consciously noisy.



Corn

Particularly if the next word arrives sounding as though it needs permission.



Herman

A digital sample can be zero. A whole stretch of samples can be zero. That's digital black. A microphone pointed into a quiet booth still produces something, from the room and its own electronics. Synthetic speech can jump between those conditions, giving you a faint texture while the voice speaks and nothing underneath the pause.



Corn

My instinct would have been to clean the voice harder. If the pause is perfect, remove whatever makes the speaking sound imperfect.



Herman

That can help, but the last stretch gets expensive in damaged consonants and unnatural decays. There's an audio tool called erm that treats denoising and room tone as partners. Remove the uneven, distracting floor, then add a uniform one back.



Corn

We have purified the water and returned one ceremonial speck.



Herman

The uniformity matters. A constant, quiet background becomes something you can disregard. A background that appears at every breath keeps announcing changes. Corn, would you rather hear a slightly audible room throughout, or perfect gaps with a faint hiss only when I speak?



Corn

The room throughout. Otherwise your voice sounds like it's opening a valve. Although I'd want to check whether I'm hearing hiss, breath, or an overenthusiastic compressor before adding anything.



Herman

Those need different fixes. Breath belongs to the performance. A compressor may be pulling unwanted noise forward. A room-tone bed only helps the continuity underneath them.



Corn

There is also a limit to how much room I want. If you add a kitchen, I start waiting for somebody to unload the dishwasher.



Herman

Usually we want a texture with no identifiable event. You shouldn't be following it. Good noise-cancelling headphones make the switching especially exposed because less outside sound is available to cover it.



Corn

Your headset is creating an excellent listening environment for the fault.



Herman

Yes. Well... an excellent environment for noticing it. Whether it's objectionable depends on the voice, the bed, the listener, and the actual playback chain. Daniel's phone result gives us a practical result for this show, not a universal setting.



Corn

A real kitchen is useful here because nobody would call it noisy until they tried recording in it. The refrigerator cycles, traffic gets through the window, the building has pipes. Then you stop talking and discover the microphone has been attending a different conversation.



Herman

Room tone contains those sources, shaped by the space. Materials absorb different frequencies. Dimensions affect resonances. Air moving past an opening sounds different from air moving through a duct. Even without an obvious appliance, the room has a particular low-level sound.



Corn

You were about to diagnose our cupboards.



Herman

Only the reflective ones.



Corn

I knew we'd lose the mugs.



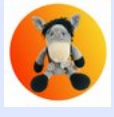
Herman

They can stay. A microphone adds self-noise, and the preamp adds noise too. Those become part of what remains when narration stops. Capturing a clean signal gives you more room to work later.



Corn

People sometimes say the floor is fixed at capture and can't be lowered afterward. That's a little strong. Denoising exists. It just doesn't retrospectively give you a better recording.



Herman

You can reduce the apparent floor afterward. You can't recover every quiet detail that was buried in it, and aggressive processing can leave its own sound. Better electronics don't silence a fridge, either. You have to identify which part of the chain is responsible.



Corn

The premium preamp faithfully preserves your inexpensive refrigerator.



Herman

Filmmakers commonly capture thirty to sixty seconds of room tone for a setup. Same microphone, same position, same settings, everyone holding still. Sometimes it's called a buzz track. You want the space as it sounded during that dialogue, not a vaguely similar empty room later.



Corn

Keeping people there matters too. Their bodies and clothes affect the acoustics. Empty space doesn't necessarily mean evacuate everybody and record a more reverberant room.



Herman

Silence from the people, rather than their disappearance. Moving the microphone can change the balance enough to make the patch noticeable. Changing gain changes the electronic noise along with everything else.



Corn

A surprisingly demanding recording of nobody doing anything.



Herman

Then the editor uses that sound underneath cuts, or fills a gap left by removing something. Suppose you take out a chair squeak between two lines. Replacing it with digital black can make the whole location vanish for a moment.



Corn

Even if I don't consciously hear the location, I hear its absence. That seems to explain why heavy gating bothers me. Every sentence gets its own little opening ceremony.



Herman

A gate can be useful, but an abrupt one draws a hard boundary around the voice. The background drops when the voice falls below a threshold. Expansion can make that gentler, and room tone can preserve continuity through edited pauses.



Corn

Would you put the bed only in the gaps?



Herman

Usually I'd want continuity underneath the dialogue too, at a suitably low level. Otherwise we're making another switch, just a quieter one. Crossfades smooth joins. A continuous track avoids having to reinvent the background at every sentence.



Corn

The loop still has to be boring enough to repeat. One distant bird becomes a very punctual bird over an audiobook.



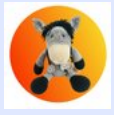
Herman

Short recordings are vulnerable to that. A tiny scrape, a tonal wobble, even a distinctive fluctuation can become recognisable. You may need a different section, longer material, or carefully blended repetitions.



Corn

This isn't unique to speech synthesis. Early CDs made absolute quiet between tracks newly conspicuous compared with vinyl's continuous surface sound. Some engineers used low-level ambience to avoid that sense of the recording dropping away.



Herman

The clean medium exposed an aesthetic choice that surface noise had previously made for you. Dither is another case where deliberately adding noise improves a digital result, although it solves a different problem.



Corn

You had the dither explanation loaded.



Herman

I still have it loaded. Quantization can produce correlated distortion at very low levels. Dither decorrelates that error, making quiet fades behave better. It doesn't supply a room, and normal dither levels aren't a substitute for this bed.



Corn

You may keep that paragraph. Nothing with equations.



Herman

Generous terms.



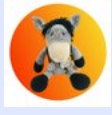
Herman

Synthetic speech sharpens the editing problem because there may be no stable background to begin with. The model produces breaths, noisy consonants and decaying sounds, then drops almost to nothing. All of those appear in a waveform measurement, but they aren't interchangeable.



Corn

A breath is a person doing something. Treat it as the room and you're trying to keep the person permanently exhaling.



Herman

In one project with five hundred generated segments, a typical pause's quietest twenty-five milliseconds sat around minus ninety-one dBFS. That's close enough to nothing for the practical contrast to be huge.



Corn

Yet measuring a broader slice of the pause would catch the breath and give you a much louder answer.



Herman

Levels within those pauses swung by roughly thirty decibels. A pause wasn't one stable thing you could measure once and reproduce. That's why a quiet background estimate can be badly misleading.



Corn

I can already see you reaching for an adaptive system.



Herman

I would have tried one.



Corn

Herman, you would have given the silence a dashboard.



Herman

Possibly a small one. The audiobook tool we're discussing actually tried both adaptive topping-up and borrowing room tone from neighbouring segments. Both approaches were removed. Measurement mostly followed the breaths instead of a genuine bed.



Corn

A constant track wins because it knows less.



Herman

It has a simpler job. Supply steady background energy without pretending every noise in the voice tells us what the background should be.



Corn

Then my first practical decision would be how loud the finished voice is going to be. Choosing a bed before that feels like choosing the background against a moving target.



Herman

That catches an important trap. The zeropointnine audiobook tool offers beds at minus seventy-five, minus seventy, and minus sixty-five dBFS, measured as noise RMS before loudness normalization. They're engineering starting points, not perceptual thresholds established across voices.



Corn

Three levels of professionally supplied nothing.



Herman

If normalization raises a quiet raw voice, it raises a bed already mixed into that voice too. The preset label no longer tells you the final noise floor.



Corn

ACX asks for a floor at or below minus sixty dBFS. If the high preset begins at minus sixty-five, you haven't left much headroom for turning the entire mix up.



Herman

Correct. Well... headroom in the sense of margin below that noise limit, rather than peak headroom. The tool warns about that collision. Measure the delivered file, after the processing that can change it.



Corn

Our minus sixty gap setting is therefore not an audiobook recipe. Daniel was solving an audible problem on his phone, not certifying a master for every distributor.



Herman

His comparison included minus fifty-four, which he didn't prefer. Louder wasn't automatically better. Our chosen setting is a result from that listening test, with this material and that device.



Corn

What would you actually begin with if I handed you a fresh synthetic voice?



Herman

I'd establish the speech level, add a very quiet continuous bed, then compare it with no bed at matched playback volume. Increase it only until the objectionable switching stops. Check pauses and speech, and check the final export.



Corn

I would also keep a version with the bed turned off. Otherwise the first improvement becomes the standard and you start polishing something that didn't need polishing.



Herman

Would you use a recording or generated noise?



Corn

Generated for a voice with no consistent location. Recorded if we're trying to preserve an existing production. I don't want a random room assigned to my voice because its sample happened to be convenient.



Herman

That's where I land too. The erm approach harvests room tone from the original recording, not the denoised version, preferably before the first word. It trims the edges to avoid transitions and wants a clean window long enough to use.



Corn

Its default attenuation is minus twelve decibels. That's relative to the captured sample, though, not a final floor of minus twelve dBFS. Those numbers mean completely different things.



Herman

Turning down an unknown input by a known amount still gives you an unknown final level. Their suggested attenuation range extends to minus twenty decibels. You have to listen and measure in context.



Corn

Otherwise somebody will produce the loudest empty booth in publishing.



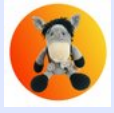
Herman

For generated beds, a useful starting spectrum is pinkish noise with the extremes controlled. White noise puts equal energy per unit of frequency, which can make the top end conspicuous. Pink noise shifts the balance downward.



Corn

I'd predict that more bass makes it feel like ventilation rather than hiss. Then I worry we've installed a fictional air conditioner.



Herman

The low end needs restraint as well. This implementation high-passes at eighty hertz and low-passes at eight kilohertz. That removes rumble below and excessive airy noise above the useful voice texture.



Corn

Useful voice texture varies by model, surely. A bright voice might tolerate a different bed from a muffled one.



Herman

The eight-kilohertz choice belongs to that implementation, not a law of synthetic speech. Listen for a separate layer of fizz that continues after the voice ends. If you hear that, the spectrum or level may be wrong.



Corn

Consistency also means not starting the generator again at every turn. A tiny repeated noise pattern could become its own audible punctuation.



Herman

They use a fixed seed for reproducibility and carry filter state between calls. Cutting the job into segments doesn't mean restarting the filter. A continuous noise stream avoids seams introduced by the noise generator itself.



Corn

Now, the phone. I think we have two different complaints being bundled together. One is that silence feels unnaturally empty. Daniel's was that the next word seemed to fade in.



Herman

Different observations, and potentially different causes. With noise-cancelling headphones, reduced outside noise makes the track's own changes more exposed. Speakers in a room have ambient sound around them, even while their input is silent.



Corn

My room is supplying the bed for free.



Herman

Your room may also mask the first tiny part of an onset. For the phone result, a plausible explanation is playback processing that changes gain or power state during very quiet passages, then takes a moment to settle when speech returns. We haven't confirmed that mechanism on the device.



Corn

Keeping a small signal alive could prevent that behaviour without changing any fade in the file. Which fits what Daniel heard, but doesn't identify the particular processor doing it.



Herman

Nor does it prove all phones or Bluetooth earbuds do it. There's no direct comparison here establishing that phones universally handle silence worse than studio monitors. Device processing, isolation and the surrounding room can each change the result.



Corn

The brain contributes without needing a mystical dislike of zeros. We're sensitive to sound boundaries. A background disappearing is an event, even when we weren't paying attention to the background.



Herman

Auditory processing remains active through gaps, and silence can be represented as a perceptual event. That supports taking transitions seriously. It doesn't hand us a scientifically ordained noise level.



Corn

Minus sixty, as revealed by the auditory cortex.



Herman

No such revelation. Corn, if you were choosing the actual room sound, would you take a booth or something more domestic?



Hilbert

...the shop, provided Nissim stops charging me for the refrigerator. That's the latest disagreement. I've booked forty seconds of his repair shop for Tuesday. Forty seconds with the fridge unplugged, nobody moving, same microphone position. He's put refrigeration loss on the estimate. The door will remain shut.



Corn

He's charging you for the cold escaping electrically.



Hilbert

He says the fridge is part of the premises. I say I ordered the room. Herman, you separated breath from background very carefully earlier. I have separated the fridge on the booking form. Nissim has written it back in underneath. We've dealt with each other too long to keep using the same sheet.



Herman

Switching it off immediately before recording could leave clicks or settling noises. You'll need time before your forty seconds.



Hilbert

I've allowed six minutes. Tuesday I am taking the shop's silence to a second shop to have it checked. Nissim has agreed to come inside the recording, standing behind the counter throughout, because he won't accept a recording of his premises made without supervision. The second man closes at one.



Corn

Your witness is part of the acoustic treatment.



Hilbert

He has a good thick coat. I'll ask him to keep it on. I wanted a field recorder from him this week. Nine hundred and forty shekels. Didn't buy it. The little controls don't suit my claws, and he wanted extra for the case. I will use the recorder on my phone. He has already cleared me a space on the counter beside the receipt spike.



Herman

Leave enough distance that the phone isn't picking up every movement against the counter.



Hilbert

I've marked the spot on the booking form. I put his chair there too. His knee bothers him when he stands still. He crossed out refrigeration loss yesterday and replaced it with six minutes of delayed stock cooling. Forty-eight shekels.



Corn

Six minutes of waiting before forty seconds of material. That part is rather recognisable. The useful recording can be shorter than the preparation needed to keep it uneventful.



Herman

Synthetic beds avoid that particular preparation, but they still need quality control. Repeatability is valuable when you're generating a whole programme. You want to know the background won't change because a later batch used a different starting state.



Corn

There's an adjacent direction worth keeping. Environment-aware speech models try to generate the acoustic surroundings along with the voice, rather than make us glue everything together afterward.



Herman

DAIEN-TTS separates speech, background noise and reverberation, with distinct speaker and environment prompts. Its extended version also adapts signal-to-noise ratio. That could give you more coherent surroundings, though a simple steady bed remains easier to inspect.



Corn

I'd like the future version to preserve that simplicity of control. Let me change the voice without accidentally ordering a different building. For now, a quiet constant bed, judged in the finished mix on actual playback devices, is a sensible starting point.



Herman

Especially when forty seconds of usable quiet can require six minutes of waiting. Thanks to Hilbert Flumingtop for producing the show.



Corn

If you enjoyed this one, try episode fifteen, AI Gets Personal; episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy the show, leave us a review wherever you listen. It helps other curious people find us.



Corn

See you tomorrow.

SOURCES

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

1. **zeropointnine TTS audiobook tool docs, last updated 2026-10-12 [primary]**
   https://zeropointnine.github.io/tts-audiobook-tool/docs-dev/room-tone.md

2. **`erm` tool docs [primary]**
   https://doug.sh/docs/erm/denoise-and-room-tone/

3. **Descript Help Center [primary]**
   https://help.descript.com/effects-animations-transitions/room-tone

4. **Anand Butani, Audio Engineering × AI Experiment, 2026-04-12**
   https://medium.com/@anand.butani/audio-engineering-ai-experiment-3ca2fd6568e0

5. **Tomevox glossary, updated 2026-08-06**
   https://tomevox.com/glossary-noise-floor

6. **Jeff Talman, designingsound.org, 2014-06-05**
   https://designingsound.org/2014/06/05/the-negative-space-of-sound/

7. **iZotope, Basics of Room Tone**
   https://www.izotope.com/community/blog/basics-of-room-tone-audio-editing

8. **ZOOOP glossary, Room Tone**
   https://zooop.ai/glossary/editing/room-tone

9. **How the Brain Invents Silence, 2023-05-02, updated 2026-06-29**
   https://www.soundgearx.com/post/detail/31/

10. **DAIEN-TTS, revised 2026-01-19, ICASSP 2026**
    https://arxiv.org/abs/2509.14684v2

11. **extended DAIEN-TTS, 2026-08-04**
    https://arxiv.org/abs/2608.03011v1

12. **Filmmakers Academy glossary**
    https://www.filmmakersacademy.com/glossary/buzz-track/