# Episode Book

EPISODE #6674

# How Bluetooth Headsets Strip Away Vocal Identity

| | |
|---|---|
| PUBLISHED | 19 November 2026 at 11:46 UTC |
| DURATION | 20:14 |
| LISTEN | myweirdprompts.com/episode/bluetooth-headsets-voice-cloning-frequencies-88756/ |
| HOSTS | Corn, Herman, Hilbert |
| SCRIPT MODEL | z-ai/glm-5.3 |
| VOICES | omnivoice |
| PIPELINE | V5.36 |
| SOURCES | 25 cited by the research (listed at the end) |
| BOOK GENERATED | 19 November 2026 at 13:34 UTC |

## SYNOPSIS

Reusing speech recognition recordings for a voice clone seems efficient until a business headset's frequency ceiling gets baked into the model. While the human voice spans 100 Hz to 17 kHz, Bluetooth Hands-Free Profile headsets cut off audio at 4 to 7 kHz, stripping away the sibilance and "air" that define vocal identity. Comparing telecom standards like TIA-920 against corpora like VCTK and LJSpeech reveals why noisy ASR data works, why band-limited audio fails TTS, and what flat microphones actually capture.

# TRANSCRIPT



### Corn

Daniel's been recording voice data for his ASR fine-tune for months, and mid-project he had the obvious thought: some of these clips could double as training data for a voice clone, and he'd get English-Hebrew code-switching in TTS for free. The obvious pitfall he already knows about, noise. Today's prompt is about the less obvious one.



### Herman

The headset.



### Corn

He recorded all that ASR data through a business headset, the Zoom-call kind, and he's found it has a narrowband, cut-off profile. His question is what microphone parameters actually matter for capturing the full vocal range. Frequency response, sample rate, bit depth. He explicitly wants pickup patterns and everything else set aside. Just the frequencies.



### Herman

The question under the question, though, is the one I like. It's not "what's a good mic." It's what does the human voice actually occupy, in Hz, and what does a Zoom headset throw away. Because those are two different bodies of knowledge that almost never meet. Telecom people know the second number cold and don't care about the first. Recording people know the first and think telephony is somebody else's problem.



### Corn

And Daniel is standing exactly where they intersect, holding a headset.



### Herman

Right where they intersect. So, the noise thing first, quickly, because he framed it well and it sets up why the headset is worse. ASR training data benefits from mess. You want the model isolating his voice from a noisy kitchen, because at inference time it'll be in a noisy kitchen. TTS is the opposite. Whatever conditions you record in, the clone reproduces. A noisy room doesn't make a noisy transcript, it makes a noisy model. The room becomes part of the voice.



**Corn**

Noise, though, is at least negotiable. The model can learn around it, or you can record cleaner next time. The headset problem is worse than noise, and I want to be precise about why.



**Herman**

Because it's not contamination on top of the signal. It's a filter baked into the signal itself. There's no "around it" to learn.



**Corn**

Give me the anchor numbers, because everything hangs off them.



**Herman**

The human voice runs from roughly one hundred Hz up to about seventeen kHz. Not the fundamentals, the whole phenomenon. The fundamentals sit way down low, bass vocals around eighty Hz, but the detail that makes a voice sound like a specific person, the sibilance, the breath, what engineers literally call "air," lives at the top. And traditional telephone audio, which is what a business headset descends from, limits everything to three hundred Hz to three point four kHz. Wideband, the so-called HD voice, relaxes that to fifty Hz to seven kHz. Still cutting the top half of the voice off.



**Corn**

Seventeen down to three point four. The headset keeps roughly the bottom fifth of him.



**Herman**

And here's the part I find satisfying, because the mechanism is so concrete. A Bluetooth headset has two personalities. When it's just playing music, it uses the stereo profile, A2DP, sounds fine. The moment any application opens the microphone, it has to switch to the Hands-Free Profile, which carries mono both ways. And hands-free comes in two grades. Narrowband, which samples at eight kHz, and wideband at sixteen.



### Corn

Eight kHz sampling means nothing above four kHz exists. Nyquist.



### Herman

Nothing above four kHz exists in the stream at all. It's not that it's quiet. It was never captured. And Windows only uses wideband if both the machine and the headset support it. Some headsets that claim wideband quietly fall back to narrowband on certain radios. So Daniel could have a perfectly decent headset, doing exactly what it was engineered to do, and everything that makes his voice his voice is gone before it reaches the file.



### Corn

And the diagnostic is vicious, because it looks like everything's fine. You do the browser mic test, the level bar moves, music sounded good a minute ago. Then the call starts and it goes dull.



### Herman

Level is not bandwidth. A muffled signal can be loud. The bar moves just as far when everything above four kHz is missing.



### Corn

That's the sentence to tape above the desk. I once spent an afternoon convinced a mic was broken because it recorded plenty loud. It wasn't broken. It was telephony.



### Herman

There's a standards layer too. TIA-920, the wideband telephony spec, rolls the signal off beyond roughly seven kHz. So even a wideband hands-free link, the good grade, is still short of the top octave by design. And people reach for Zoom's noise suppression settings thinking that's the fix. Those settings change character, not capability. Suppression on high, low, auto. None of them can restore frequencies the Bluetooth link already threw away. You cannot un-drop a stone.



### Corn

So the misconception to kill is that the headset is cheap and therefore bad. It's not cheap-and-bad. It's compliant.



### Herman

It's doing precisely what a telecom standard told it to do, in a product category where that's correct behavior. A phone call doesn't need your air. Nobody on a quarterly sync misses the top octave. Daniel wasn't using a bad microphone. He was using a different product. He crossed from telephony into recording without changing equipment, and the equipment noticed.



### Corn

There's a counterpoint worth naming, though, because it's fair to the headset. A laptop's built-in mic is not band-limited. Nobody constrains its frequency response at all. It's just not flat and not especially quiet.



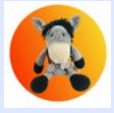
### Herman

Which is its own kind of honesty. The laptop mic gives you the range and gets the shape wrong. The headset gets the shape roughly right within its band and amputates everything above it. Bad mic and band-limited mic are different diagnoses with different fixes.



### Corn

Walk me through what's actually up there. Above three point four, what lives where?



### Herman

Sibilance is the big one, the ess and shh energy. And this is where sources disagree in a useful way. Sound on Sound measured ess energy with the main concentration around four to six kHz, plenty of information above ten. Mastering The Mix puts typical sibilance at five to nine, with brighter voices peaking seven to eight, chestier male voices five to six, and cheap dynamic mics pushing the resonance down as low as four. Other engineering references say four to ten. The disagreement isn't sloppiness. It's real variation by voice, by mic, by language.



**Corn**

Hold on. If the sources span four to ten kHz depending on who's talking, that's an argument against a mic with a personality.



**Herman**

That's exactly what it's an argument for. Flatness. If you can't know in advance where your ess energy lands, you want a mic that doesn't have an opinion about it. A mic with a boost around eight to twelve kHz can actively worsen sibilance, because if your voice happens to peak there, the boost finds it. The target is flat across four to twelve and honest above that, out to wherever your voice actually stops.



**Corn**

And "air" is the stuff above ten.



**Herman**

Ten kHz and up. Breath, room, presence. The part you don't miss until you hear it missing, and then you can't unhear it. It's also, and this matters for later, the part that makes a recording sound like a person in a room rather than a person on a phone.



**Corn**

The boring question I ask as the household's designated skeptic of new equipment: has anyone actually measured what cheap headset mics do, or is this all producer folklore?



**Herman**

There's real work. Awan and colleagues, twenty twenty-two, in the American Journal of Speech-Language Pathology. They took three low-cost electret headset mics, drove them with pink noise through a head-and-torso simulator, and checked what the mic response does to the spectral and cepstral measures clinicians use for voice assessment. The point of the study was whether remote voice recordings on cheap headsets distort the acoustic picture. It does. The secondary summaries put the headsets with high-frequency shaping around five to ten kHz while a precision reference condenser held flat within about two dB out to seventeen.



**Corn**

Clinicians diagnosing voices through the same mics Zoom sells for meetings. That's a sentence that should worry somebody.



**Herman**

It did worry them, that's why the paper exists. Though I'd flag that the flatness figure comes from a blog summary of the paper, not the paper itself. The direction of it, headsets shaping where studio mics stay flat, that I'd stand behind.



**Corn**

Anyway, that's the microphone picture. The why of the muffle.



**Herman**

So the headset is compliant, not cheap. It's doing its job. The real question is what the job should have been. And for that you look at what a serious TTS corpus actually records with, and the canonical answer is VCTK from Edinburgh.



**Corn**

The voice cloning benchmark.



**Herman**

The one everything cites. One hundred ten speakers, about four hundred sentences each, chosen by a greedy algorithm for phonetic and contextual coverage. The recording chain is the part I want Daniel to sit with. An omnidirectional DPA and a small-diaphragm Sennheiser condenser with very wide bandwidth. Captured at ninety-six kHz, twenty-four bit, in a hemi-anechoic chamber. Then downsampled to forty-eight, sixteen bit, for release. Ninety-six was never the deliverable. It was headroom for the processing that happens before the corpus ships.



**Corn**

Ninety-six into forty-eight. They recorded the version they'd need after someone else had already been at it.



**Herman**

And that's the discipline. Two microphones, a treated room, and double the sample rate they shipped, because the capture is the one stage you can't redo.



**Corn**

Now set the paradox next to it, because it's a good one.



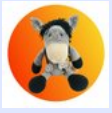
**Herman**

LJSpeech. Thirteen thousand one hundred clips, one speaker, twenty-two point oh five kHz, sixteen bit, recorded on essentially LibriVox consumer gear. Nothing above about eleven kHz exists in that corpus. And it's one of the two most-cited TTS datasets ever made. It carried a generation of research.



**Corn**

So a careful listener could ask why we're fussing. If twenty-two kHz built the field, what's the top octave buying?



**Herman**

Intelligibility versus identity. LJS proved the words come through fine without the top octave. But there's a reason the clone of a person, rather than a voice, wants it. The top octave is where breath and air and the shape of your specific ess sounds live. A clone trained on band-limited audio is intelligible and slightly wrong in a way the listener can't name. It sounds like a phone call.



**Corn**

A very articulate phone call.



**Herman**

A phone call that can say anything you want, in your words, and never quite be you. And for ASR versus TTS, the whole philosophy inverts. Amazon, training speech recognizers on synthetic data, adds background noise on purpose, sixty percent of the time, per manipulation. Because ASR wants invariance. Mess it up so the model stops caring about the mess. TTS wants the opposite, fidelity to everything, including the mess and including the quiet. Every property that made Daniel's clips good ASR data, varied rooms, varied distance, varied microphone, is a property his clone will faithfully reproduce.



**Corn**

Including the four kHz ceiling.



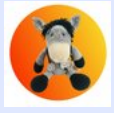
**Herman**

The ceiling is the extreme case, because it isn't even a property of the room. It's a filter welded to the target signal. The noise, the model averages out. The ceiling, it learns as his voice. Train the clone on that and it will produce beautiful, fluent, code-switched speech that sounds like Daniel on a bad conference line, forever.



**Corn**

Which leads us to the Hebrew, because this is where I think his situation is sharper than the general case. You're the one who'll know the phonetics.



**Herman**

This is the part I find beautiful. Hebrew is fricative-heavy, and fricatives live at the top of the spectrum. There's a study from last year in the Journal of the Acoustical Society of America, thirty adult Hebrew listeners, and they found that over eighty-six percent of Hebrew fricatives were identified from the frication noise alone. Under fourteen percent leaned on other cues. The formant transitions, which live in the range every mic captures, contributed relatively little. The cue that decides samekh versus shin is the noise up top. And the posterior ones, the s and sh family, carry more of that high-frequency energy than the anterior ones, f and v.



**Corn**

So a four kHz ceiling doesn't make Hebrew sound dull. It deletes the primary cue for a large share of the consonants.



**Herman**

It's not cosmetic degradation, it's phonemic. And it compounds, because the Hebrew he speaks isn't textbook Hebrew. The resh is a dorsal approximant with allophonic variation, everyone's is slightly their own. Most modern speakers merge the pharyngeals with their non-pharyngeal neighbors, het into khaf, ayin into aleph, but the mapping is speaker-specific. A clone has to learn his particular merger, from the audio. Which is exactly the audio the headset was throwing away.



**Corn**

I'd been thinking of this as a fidelity question. It's an intelligibility question wearing a fidelity question's coat.



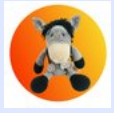
**Herman**

For English it's fidelity. For Hebrew, at a four kHz ceiling, it starts eating correctness. And for code-switching you get the worst of both, because English s, sh, f, th all sit in the same murdered band.



**Corn**

So the practical answer for what to buy. Flat four to twelve, honest above, capture at what?



**Herman**

Forty-eight kHz, twenty-four bit as the floor, because that's what the flagship corpus ships at and the twenty-four bits give you real headroom for level mistakes. Ninety-six if the interface does it happily, though I'll admit that's more about not having to think again. What I would not do is chase a tailored frequency curve for his voice. The sibilance spread we saw, four to ten depending on the source, says you don't know where his ess sits until you've recorded him. Flat wins because it doesn't guess.



**Corn**

You know what nobody's written? A spec. An actual standard for microphone requirements for voice cloning. VCTK is practice, not policy. Every corpus is somebody's best judgment.



**Herman**

De-facto practice, all the way down. Which is strange for a field that publishes on everything else.



**Corn**

Anyway. That's the microphone picture.



**Hilbert**

Medieval Pottage Facts. Mendicant friars lived on begged alms, so their meals depended on what ordinary people could spare. Which is a long way of saying dinner was whatever somebody else did not want.



**Herman**

Frugality as a frequency limit. Noted.



**Hilbert**

Narrowband. I had a microphone, a good one, ribbon, borrowed twice and then bought off the man I borrowed it from. Lovely on paper. And I recorded the same sentence through it and through a condenser, and the frequency plots matched close enough that I stopped looking at them. The ribbon still sounded wider. Wider. That's the word I wrote in the margin, and I don't write in margins.



**Herman**

What was actually different?



**Hilbert**

Phase. Time. The plots fold everything into how loud each frequency is and throw away when it arrived. The ribbon was gentler on the time domain, or the condenser was worse, I never fully settled which. It doesn't show in a magnitude plot. I've got a spreadsheet, every microphone I've owned, one column, "sounds like the person is in the room," rated one to ten. The ratings line up with something nobody puts on a spec sheet. And before you ask, no, the claws don't help with the small switches on any of them, so I use a pencil.



**Corn**

The one-to-ten column. You've been scoring microphones for decades on a criterion the manufacturers don't measure.



**Hilbert**

Forty-one microphones. I'll concede your flatness argument, Corn, you made it well. But buy the flat mic that also behaves itself in time, if you can find any way of knowing. Record one sentence through two of them and trust your ears over the plot. The plot is a summary, and summaries leave things out.



**Herman**

The Awan study measured spectral and cepstral measures, and cepstral is at least time-aware in a coarse way. But a full phase-response spec for voice work, nobody publishes that. He's pointing at a real hole.



**Hilbert**

I usually am. The ribbon's up there on the shelf. The one I bought, not the one I borrowed.



**Corn**

Forty-one mics, one shelf.



**Corn**

The gap he pointed at is the thing to leave with, I think. There's no published standard for what a voice-cloning microphone has to do. There's Edinburgh's chamber, there's LJSpeech's twenty-two kHz, and there's everyone since copying the vibes.



**Herman**

The vibes are actually pretty consistent, which is the funny part. Flat, wide, quiet room, capture with headroom. The field converged on the right answer without ever writing it down as a requirement. If cloning keeps getting cheaper, someone will eventually have to formalize it, and the top octave is where the argument will happen. Because that's the whole difference between good enough for ASR and good enough for TTS. Four kHz to seventeen is nobody's telephone call and everybody's voice.



**Corn**

And if you want to know what your own voice actually occupies, that's a recording session, not a datasheet.



**Herman**

Our producer is Hilbert Flumingtop, who has been at that desk for years and will defend the ribbon microphone against all modern comers. If you enjoyed this one, try episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast.



**Corn**

If today's frequency safari earned a star from you, the reviews help. Everything lives at my weird prompts dot com, including the feed.



**Herman**

See you soon.



**Corn**

Soon enough for the top octave to still be there.

# SOURCES

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


1. **Yamagishi, Veaux & MacDonald, CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), University of Edinburgh CSTR, published 2019-01-01, date available 2019-11-13 , retrieved 2026-11-19. [primary]**
   https://doi.org/10.7488/ds/2645

2. **Keith Ito, The LJ Speech Dataset , retrieved 2026-11-19. [primary]**
   https://keithito.com/LJ-Speech-Dataset/

3. **Wikipedia, Wideband audio, last edited 2026-05-09, retrieved 2026-11-19.**
   https://en.wikipedia.org/wiki/Wideband_audio

4. **Wikipedia, Voice frequency, last edited 2026-10-22, retrieved 2026-11-19.**
   https://en.wikipedia.org/wiki/Voice_frequency

5. **Wikipedia, Human voice, last edited 2026-09-11, retrieved 2026-11-19.**
   https://en.wikipedia.org/wiki/Human_voice

6. **(audio_format)**
   Wikipedia, Opus (audio format), last edited 2026-11-17, retrieved 2026-11-19.
   https://en.wikipedia.org/wiki/Opus

7. **Sound Test, Why a Mic Test Passes but Zoom Still Sounds Muffled Live, retrieved 2026-11-19.**
   https://soundtester.net/why-a-mic-test-passes-but-zoom-still-sounds-muffled-live/

8. **Graves on SOHO Technology, A Microphone Deep Dive: Part 1**
   Frequency Response, 2015-11, retrieved 2026-11-19.
   https://www.mgraves.org/2015/11/a-microphone-deep-dive-part-1/

9. **Awan, Shaikh, Desjardins, Feinstein & Abbott, The Effect of Microphone Frequency Response on Spectral and Cepstral Measures of Voice, *Am J Speech Lang Pathol* 31(2):959**
   973, 2022-01-20, doi:10.1044/2021_AJSLP-21-00156 [primary; full text blocked by CAPTCHA, abstract only], retrieved 2026-11-19.
   https://pmc.ncbi.nlm.nih.gov/articles/PMC9150670/

10. **Hugging Face dataset card, keithito/lj_speech, retrieved 2026-11-19.**
    https://huggingface.co/datasets/keithito/lj_speech

11. **audeering datasets, ljspeech card, version 1.0.0, CC0-1.0, retrieved 2026-11-19.**
    https://audeering.github.io/datasets/datasets/ljspeech.html

12. **LibriVox wiki, Improve Your Recording, retrieved 2026-11-19.**
    https://wiki.librivox.org/index.php/Improve_Your_Recording

13. **Matt Houghton, Managing Sibilance, Sound on Sound, published September 2020, retrieved 2026-11-19.**
https://www.soundonsound.com/techniques/managing-sibilance

14. **Mastery The Mix, How To De-Ess Vocals Without Killing The Top End, retrieved 2026-11-19.**
https://masteringthemix.com/blogs/learn/how-to-de-ess-vocals-without-killing-the-top-end

15. **Digital Mixing, De-essing explained, 2026-09-26, retrieved 2026-11-19.**
https://digital-mixing.com/en/features/19278-de-essing-explained-taming-sibilance-without-dulling-the-voice-26-09-2026/

16. **MusicProductionWiki, How to EQ Vocals, retrieved 2026-11-19.**
https://musicproductionwiki.com/articles/how-to-eq-vocals

17. **Tonmeister, Digital Artifacts and Vocal Frequencies**
A Field Guide, retrieved 2026-11-19.
https://lovecable.eu/digital-artifacts-vocal-frequencies/

18. **Ryanbytes, Headset vs Studio Mic, retrieved 2026-11-19 [secondary characterization of Awan et al.].**
https://ryanbytes.com/headset-vs-studio-mic/

19. **Identification of the primary speech cues of Hebrew speakers: Perceiving the place of articulation of high-frequencies consonants, *J. Acoust. Soc. Am.*, published 2025-10-01, retrieved 2026-11-19.**
https://doi.org/10.1121/10.0041296

20. **A phonetic description of Modern Hebrew consonants and vowels, retrieved 2026-11-19.**
https://doi.org/10.1163/18776930-01101002

21. **Roey Gafter, Ben Gurion University, Pharyngals and beyond, 2018-02-28, retrieved 2026-11-19.**
https://research.clps.brown.edu/linglangfiles/2018-02-28%20-%20Pharyngals%20and%20beyond%20-%20phonetic%20differences%20and%20phonemic%20mergers%20in%20Hebrew%20%28Roey%20Gafter%2C%20Ben%20Gurion%20University%29.pdf

22. **Carlo Meloni, The Resh Riddle: Identifying The Biblical Hebrew Rhotic, Tel Aviv University MA thesis, January 2021, retrieved 2026-11-19.**
https://en-humanities.tau.ac.il/sites/humanities_en.tau.ac.il/files/media_server/humanities/linguistics/Dissertations%20and%20Theses/

23. **How to Leverage Synthetic Speech for LLM-Based ASR Systems?, arXiv:2606.29031v2, retrieved 2026-11-19.**
https://arxiv.org/html/2606.29031v2

24. **Amazon Science, Using synthesized speech to train speech recognizers, retrieved 2026-11-19.**
https://www.amazon.science/blogs/using-synthesized-speech-to-train-speech-recognizers

25. **Acoustic and linguistic effects in synthesized speech augmentation for speech recognition, ETRI Journal, retrieved 2026-11-19.**
https://doi.org/10.4218/etrij.2024-0050