Episode Book

EPISODE #5840

Why Voice Clones Lisp: Sibilants, Codecs, and the Sounds We Lose

PUBLISHED	10 October 2026 at 15:35 UTC
DURATION	23:17
LISTEN	myweirdprompts.com/episode/voice-clones-sibilants-codecs-74031/
HOSTS	Corn, Herman
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.3
SOURCES	16 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:40 UTC

SYNOPSIS

Why do cheap voice clones lisp? Daniel asked, and the answer runs from your teeth to your earbuds. Sibilants like s and sh are broadband noise with no pitch for a codec to lock onto; plosives like p and t are puffs of moving air that shove a microphone diaphragm. Every stage of the chain — microphone, codec, vocoder — is doing its job correctly, and the fricative is what pays for it. A 2023 Aarhus study found AMR-WB cuts off around 5,600 Hz, right through the middle of the sibilant band, and under Opus, f and th become nearly identical in spectral space. The codec isn't broken. It's succeeding, and the consonants are the casualty.

TRANSCRIPT



Corn

The sounds you never think about are the ones the whole chain is quietly failing at.



Herman

And they're not the vowels. Nobody complains about a vowel.



Corn

Daniel wants to know why s, sh, f, p, t and k are so hard to get right. His point is that sibilants hiss, plosives pop, fricatives sit up in frequencies that compression loves to throw away. Which is why de-essers and pop filters exist in the first place, and why a cheap voice clone so often goes lispy or smeared at exactly those moments. What's physically going on in those sounds that makes them so fragile, all the way from the mouth to the earbud.



Herman

That's the whole arc of it. Mouth, microphone, codec, synthetic voice. And the thing that makes it a story rather than a list is that every one of those stages is doing its job correctly. That's what breaks them.



Corn

Say that again.



Herman

The microphone is trying to be sensitive. The codec is trying to be efficient. The vocoder is trying to be compact. None of them is malfunctioning. They're all succeeding, and the sibilant is what pays for it.



Corn

So it's a collision of good intentions.



Herman

Every stage.



Corn

Start with the face. What's actually happening in my mouth when I say s?



Herman

Nothing is vibrating. That's the first thing to get straight. A vowel is your vocal folds buzzing at a steady rate, and everything above that is the shape of the tube filtering the buzz. Harmonic, periodic, predictable. An s has no pitch at all. You're forcing air through a narrow gap between your tongue blade and the hard palate, and that jet of air becomes turbulent, and the turbulence hits your upper incisors.



Corn

My teeth are the speaker.



Herman

Literally. The Royal Society did a proper aeroacoustic study of it, and their line is that the principal source of the sound is the diffraction of jet turbulence pressure fluctuations by the incisors. The turbulence is the energy. The teeth are what radiates it into the room.



Corn

I've been walking around with a tweeter in my skull this whole time.



Herman

There's a computational aeroacoustics study that traces it even more precisely. The jet comes down through the gap between the upper and lower incisors, it impinges near the lower lip, and the eddies in that region become a direct source of aerodynamic sound. Then the articulators diffract it outward.



Corn

And all of that energy lands where?



Herman

Above four kilohertz. An Interspeech paper on it says the impingement on the incisors and lips generates broadband noise over four kilohertz. Broadband. No fundamental, no harmonics, no pitch for a codec to lock onto. Just shaped noise.



Corn

Which is the first time in this conversation we've said the word the codec is listening for.



Herman

And it's the word that dooms them, yes. Hold that.



Corn

Plosives are a different animal, right? p, t, k.



Herman

Different mechanism entirely. Those are stops. You build pressure behind a complete closure, hold it, and release it. What you hear is the burst. One study on microphone placement describes popping as coming from a brief turbulence of air exhaled from the mouth when you pronounce a plosive consonant.



Corn

So it's not even really sound in the usual sense. It's a puff of moving air with a noise attached.



Herman

Right, and that puff is what hits a microphone diaphragm as a mechanical event, not an acoustic one. Which is why a plosive through a close mic doesn't just sound loud, it sounds like the capsule got shoved.



Corn

Okay. So the mouth has already done its damage. Now the microphone.



Herman

The microphone makes it worse, and the reason is proximity. When you close-mic someone, you're picking up far more of the high frequency content than anyone in the room would ever hear. The problem band is roughly four to ten kilohertz depending on the talker, and a very common place to focus is around eight.



Corn

And sibilants specifically?



Herman

Five to eight kilohertz is where they cluster, and they carry a surprising amount of energy up there. Some microphones have a frequency response that boosts that region as a design feature.



Corn

As a feature.



Herman

That's the part I find annoying. Michael Joly's analysis of condenser mics makes the case that the K67 type capsule, with its built-in high frequency pre-emphasis, is responsible for thousands upon thousands of excessively bright and sibilant microphones. That's not physics. That's a choice somebody made.



Corn

A choice that then got sold as professional.



Herman

That's exactly the trap. People hear that bright, sibilant top end and think it sounds expensive. It doesn't sound accurate. It sounds like a marketing decision from decades ago that we all agreed to stop questioning.



Corn

Which kills the standard defense. The performer gets blamed for sibilance, and half the time it's the capsule.



Herman

More than half, arguably. The de-essing literature is clear that excess sibilance can come from compression, from microphone choice and technique, and even from the way a person's mouth anatomy is shaped.



Corn

So the singer with the sharp s gets notes about their diction, and the actual culprit is a capsule designed in the sixties to sound exciting.



Herman

And then compressed badly on top of it. Don't leave that out. Compression raises the quiet parts, and the tail of an s is a quiet part.



Corn

So now I'm a sound engineer. I've got a bright mic exaggerating five to eight k, a compressor lifting the tail of every s, and a performer with a narrow palate. What do I reach for?



Herman

A de-esser, which is a compressor with an equalizer-boosted copy of the signal on its sidechain, so it only engages on the sibilant range. And a pop filter, which is a mechanical fix for a mechanical problem. It blocks the burst of air from a hard consonant before it reaches the capsule.



Corn

De-essers fix the excess. Pop filters fix the puff.



Herman

And both of them are compensations for earlier decisions, which is the pattern of this whole episode.



Corn

Here's what I want to push on. If I de-ess too hard, what happens?



Herman

You get a lisp. The sound design guide's line is that the mistake to avoid is heavy de-essing that lisps the voice. It's a real failure mode. Push a de-esser far enough and you've removed the very frequencies that make an s an s.



Corn

So I've taken a fragile sound, run it through a bright mic, compressed the tail, and then surgically removed what's left.



Herman

And then you send that file to a voice cloning service. And the clone comes back sounding like it has a speech impediment.



Corn

And we blame the model.



Herman

Everyone blames the model. The lisp may have been manufactured in the studio, several stages before the model ever saw the audio.



Corn

That's the second time you've said a version of that.



Herman

Because it's the thing I keep circling. The chain doesn't have one weak link. It has four stages, and each one is independently degrading the same narrow band of frequencies.



Corn

Fine, but the codec is the one that worries me most. The microphone exaggerates, the de-esser removes, but the codec is deciding what counts as speech.



Herman

And that's the philosophical joke at the center of this. Codecs are designed to identify and translate speech into a numerical representation, and their stated goal is to detect speech and exclude noise.



Corn

Noise being irregular sound waves.



Herman

Irregular sound waves. Fricatives are aperiodic and noise-like by physics. The codec's success criterion is the fricative's destruction. It's not a bug, it's the specification.



Corn

So if the codec is doing well, my s is gone.



Herman

That's not a metaphor. There's a 2023 study out of Aarhus that took read speech from thirty male English speakers across five accents and compressed it three ways. AMR-WB at twelve point six five kilobits, MP3 at thirty-two, Opus at twenty-four. Then they measured the spectrum against uncompressed sixteen kilohertz WAV.



Corn

And the cut-offs?



Herman

They ran white noise through each codec to find where it stops. AMR-WB around five thousand six hundred hertz. Opus around seven thousand. MP3 around seven thousand one hundred. Everything above that slopes away into nothing.



Corn

Five six for AMR. Which is right in the middle of the sibilant band.



Herman

Below the middle of it, in a lot of voices. Now the results. Centre of gravity and standard deviation dropped significantly for essentially every fricative in every codec. For s, MP3 lowered the centre of gravity by three hundred and forty-nine hertz. Opus by three hundred. AMR by three hundred and forty-nine.



Corn

And for the softer ones?



Herman

For f, Opus dropped it by four hundred and fifteen hertz. For th, four hundred and ninety-seven.



Corn

That's a lot of movement for a sound that's already sitting up high.



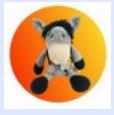
Herman

And here's the finding that should stop anyone who cares about this. Under Opus, f and th become almost identical in centre of gravity and standard deviation. AMR-WB makes them more alike too.



Corn

Wait. The codec isn't just dulling the sound. It's erasing the distinction between two different phonemes.



Herman

It's collapsing categories. The paper's own warning is that this potentially makes sounds less distinct.



Corn

So f and th merge. What else is near enough to merge?



Herman

That's the open question, and it's a real one. Sixteen kilohertz was the ceiling of the test. Thirty male speakers, studio clean. Female speakers are reported to be more affected by codec compression than male speakers. And the paper says outright that with background noise, or in live transmission, the effects are likely to be more prominent, especially when the codec identifies the fricatives as noise and eliminates them from transmission.



Corn

Eliminates them from transmission.



Herman

That's the paper's phrasing, and it's not being dramatic. That's the codec working as designed.



Corn

So every phone call I've ever made has been quietly reassigning my consonants.



Herman

Every compressed call. And Daniel's question about the earbud end is the right place to stop, because by the time it reaches you, the s has been through four decisions made by four different people who never spoke to each other.



Corn

Alright. There's a practical thing I want to raise, because you mentioned training.



Herman

Go on.



Corn

If every model anyone trains on real-world audio is trained on codec-processed audio, and the codec has already collapsed f and th into the same point in spectral space, then what is the model learning?



Herman

It's learning the merge. It's learning the distinction as the codec renders it, not as the mouth made it. Which matters enormously if you're doing forensic phonetics, or speaker identification, or accent work.



Corn

Or if you're building a transcription model and wondering why it hears th where the speaker said f.



Herman

There are specific numbers worth having here, because they show how much material these codecs are actually chewing through. The corpus in that study had twenty-nine thousand three hundred and thirty-six tokens of s and five thousand one hundred and four of sh. That's from one study of thirty speakers reading a ten minute passage.



Corn

So scale that to the internet.



Herman

There's more speech passing through lossy codecs in a day than any of us can picture, and a meaningful fraction of it is sitting in the band the codec is throwing away.



Corn

Now the synthetic voice. Because that's where Daniel's lisp question actually lives.



Herman

Two separate mechanisms there, and they're often confused. The codec damages the audio you feed in. The vocoder struggles to generate the sound you ask it for. Different failures, same frequencies.



Corn

Start with the evidence that the models find these sounds hard.



Herman

A 2026 paper on phoneme-level deepfake detection found that complex vowels and fricatives exhibit higher divergence while simpler phonemes remain more stable. That's the finding. The fricatives are where synthetic speech deviates most from real speech, which is exactly why they're the easiest phonemes to build a detector around.



Corn

The detector works because the fricative is where the model gives itself away.



Herman

That's the whole detection literature in one line. You don't catch a fake on the vowels. You catch it on the s.



Corn

Why, though? What's the model actually failing to do?



Herman

The ARMAX paper on vocal tract modeling gives the cleanest answer. All-pole autoregressive models, which is what a lot of classic vocoders are, cannot provide the locations of anti-formants. Zeros, in filter terms. And that increases estimation error specifically in nasal, fricative and stop consonants.



Corn

The exact three classes.



Herman

Nasals, fricatives, stops. That's not a coincidence, that's a modeling limitation with a name. An all-pole model can place resonances. It can't place the notches, and fricatives and nasals live in the notches.



Corn

So the model is structurally blind to part of what makes those sounds what they are.



Herman

And then you layer the usual suspects. There's a long literature on noisy output because of how hard it is to capture the time-varying nature of speech. And phase modeling is often ignored entirely by using a minimum phase filter during synthesis, which means the speech quality suffers.



Corn

Phase. The part nobody listens to on purpose.



Herman

The part the ear is remarkably sensitive to. And fricatives are where the phase structure is most chaotic, which is the worst possible combination if you're skipping phase modeling.



Corn

So the clone has two possible ways it goes wrong. Inherit a smeared source, or generate a smeared sound.



Herman

And the two compound. Take audio that's already been de-essed into a lisp, run it through a codec that's collapsed the high band, and ask a vocoder with no phase modeling to reconstruct it. Nobody in that chain set out to make it lispy, and the lisp is what you get.



Corn

Every hand it passed through was holding it correctly.



Herman

And they all handed it on slightly worse.



Corn

Hold on. There's a version of this where the model isn't failing at all. If the training data is codec-processed, the model is faithfully reproducing what it was taught.



Herman

That's the uncomfortable one, yes. A model trained on phone calls will generate phone call sibilants, because phone call sibilants are what an s is, as far as it has ever been told.



Corn

So some of what we call a cheap clone artifact is the model being an excellent student of a degraded corpus.



Herman

Which means the fix might not be the model.



Herman

Right, and I want to say one thing about where this leaves the whole chain, because it's actually a nice picture. Four stages, four goals. The microphone wants sensitivity. The compressor wants consistency. The codec wants efficiency. The vocoder wants compactness.



Corn

And none of those goals is "preserve the s."



Herman

None of them. The s is simply not a stakeholder in any of those meetings.



Corn

Four thousand hertz of turbulence, and it clears the whole building without anyone noticing it left.



Herman

It's still there in the waveform. Some of it comes back over a good speaker. You and Daniel are both working with hardware, aren't you?



Corn

Difference between a call and a recording is the difference between a rumor and a quote.



Herman

So let me pull it together. If you've got a recording session where the s is the problem, you don't fix it by de-essing harder. You fix it by finding where in the chain the s got broken. Usually it's the mic.



Corn

You have opinions about microphones held above your weight class.



Herman

I have opinions about everything.



Corn

That's what makes this work.



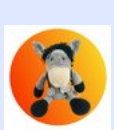
Herman

I think I know what you're going to say about the pop filter.



Corn

The mechanical fix for the mechanical problem. You're going to say it's a workaround, and the real fix is placement.



Herman

I wasn't going to say that at all.



Corn

You were going to say it's elegant.



Herman

It is elegant. It's a screen of cloth that stops moving air and lets sound through. That's a real solution to a real physical event.



Corn

A sock would do the same thing. I want that on the record. He said a sock would do the same thing.



Herman

I said it would, I didn't say you should.



Corn

Everything I've ever needed to know about audio engineering in one exchange.
There was a mic once. I want to be careful how I put this, because the hosts have been treating the microphone as the culprit and they're not wrong, but they're pointing at the wrong thing in it.



Corn

This should be good.



Herman

Go on then.
It was called the Vienna something. I'd have to look it up. I paid eleven hundred for it, which was the most I've ever paid for anything that wasn't a car, and I bought it because the man at the counter said the word "air" four times in ninety seconds. It had a rise of about four decibels starting around six k and climbing to about nine by the top, and I knew that going in, and I told myself that was the sound I wanted.



Corn

And?
And every s sounded like a snake getting stepped on. I'd record a paragraph and play it back and the s's would stand up out of the sentence like they were being introduced. I took it back. He said it was my technique. I said it was his microphone. We were both right.



Herman

Was it the pop filter you tried first?
No. The pop filter didn't touch it. The pop filter is for the air, not for what the capsule does with what it hears. What fixed it was a sock. A thick one, off a pair of gray wool ones I'd had since I don't remember, and I'd put it over the head and speak into that, and the top came down enough to live with.



Corn

You recorded your voice work through a sock.
For eleven years. I still have it. The sock. It has a name. It's called Gerald. I say good morning to it before I start, because it has to sit through the whole session and it may as well be acknowledged.



Herman

I'm not sure what to do with any of that.
The man who taught me the trick said he'd worked on the Apollo program. He said the entire moon landing was dubbed in a studio in Burbank with a sock over the microphone, and that if you listened carefully you could hear the capsule rattling on the descent recording. I never checked. Seemed rude.



Corn

Gerald.
Gerald has views on de-essers. He doesn't like them. He thinks a de-esser is an apology for a microphone you shouldn't have bought, which I've always thought was a fair position. Anyway. Fifteen hundred for the sock. Fine. The microphone's in a cupboard.



Corn

Fifteen hundred.
For the pair. Two socks. Hardly relevant which, since I only ever used the one.



Herman

I want to make sure I've got the physics right there. A wool sock over the capsule is going to act as a high-frequency absorber, and depending on the density and the distance from the diaphragm you'd expect a progressive roll-off well into the top octaves. It's crude but it's not nonsense.
It's not nonsense. It's very crude and it's not nonsense. Six decibels of shelf by eight k, near enough, with the sock directly on it. I measured. Once. Then I stopped measuring and just recorded.



Corn

Your answer to Daniel's question is: buy the wrong microphone and put a sock on it.



Herman

His answer is buy the wrong microphone and put a named sock on it.
That's the answer. Everything above eight k that a de-esser would have pulled down, the sock's already taken. Then you don't need the de-esser and you don't get the lisp, because you never had the problem the de-esser was invented for.



Corn

The codec still eats it.
Codec still eats it. Nothing you do in a room fixes what a codec decides on a wire. Different problem, different man.



Herman

Which is essentially the argument we've been making all episode.



Corn

Don't take credit for his argument.



Herman

I'm not taking credit. I'm noting he arrived at it independently, wearing a sock.



Corn

Alright. Let me land this, because going back to what Hilbert said, the sock is the whole thing in miniature.



Herman

The sock being all of it. Go on.



Corn

He solved a problem he'd created by buying a microphone with four decibels of lift at six k, and he solved it mechanically, before the signal ever existed. That's the right order. Fix it at the mouth before it gets to the codec, because once it's at the codec, nothing you own can help you.



Herman

There's the open question, then. If Opus can take f and th and drop them on top of each other in spectral space, what happens to the speech we keep in compressed form and never keep a copy of?



Corn

Every phone call. Every voice memo that went through a messaging app. Every recording session where the only backup is the compressed stream.



Herman

Those aren't degraded copies. They're the only copies, and they're missing the part that carried the phonemic distinction.



Corn

Which is a strange thing to think about when you consider how much of the historical record of the twentieth century is on formats that lost the top of the band. We may be the last generation that could have heard the difference.



Herman

It sharpens Daniel's question in a direction I hadn't expected. The closer synthetic speech gets to real speech, the more it has to nail the sounds we've spent four stages systematically destroying. The models will succeed at the vowels long before they succeed at the s, because the s is the last thing standing.



Corn

Everything upstream has to be fixed before the clone can be right, and it can't fix itself.



Herman

That's the forward-looking thought. The clone will always be slightly wrong at exactly the moments that matter, until the source material stops being wrong first.



Corn

One thing from the reading that didn't fit anywhere. Opus, as a codec, the specification actually scales from six kilobits for narrowband speech all the way to five hundred and ten kilobits for stereo music.



Herman

Same codec. Six k on a call, five hundred and ten on a recording session.



Corn

The algorithm that eats your s on a phone call can carry a symphony, and it's the same standard. That's not a broken technology. That's a spectrum of tradeoffs, and the s sits at the cheap end of it.



Herman

Well put. And that's the show.



Corn

We should probably mention that Hilbert is our producer.



Herman

Hilbert Flumingtop, who has been sitting there the whole time presumably thinking about his sock. If you want more of this, try episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; episode ten, How ASR Went From Frustration To ... Whisper Magic; and episode fifteen, AI Gets Personal. This has been My Weird Prompts.



Corn

If you've got a question that keeps eating at you, send it on Telegram at t dot me slash MWP listener bot. Everything else is at my weird prompts dot com.



Herman

We'll be back soon.



Corn

See you then.

SOURCES

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

1. **Christensen, K.V., Spectral implications of codec compression on voiceless fricatives, ICPhS 2023 [primary]**
   https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/587.pdf

2. **Aeroacoustics of [s], Royal Society Proc. A [primary]**
   https://royalsocietypublishing.org/doi/10.1098/rspa.2004.1405

3. **RFC 6716, Definition of the Opus Audio Codec [primary]**
   https://emailstuff.org/rfc/rfc6716

4. **Computational aeroacoustics to identify sound sources in the generation of sibilant /s/, Wiley**
   https://onlinelibrary.wiley.com/doi/10.1002/cnm.3153

5. **Yoshinaga et al., Experimental Validation of Sound Generated from Flow in Simplified Vocal Tract Model of Sibilant /s/, Interspeech 2016**
   https://www.isca-archive.org/interspeech_2016/yoshinaga16_interspeech.pdf

6. **DSP Concepts, Microphone De-Essing and De-popping**
   https://documentation.dspconcepts.com/awe-designer/8.D.2.5/microphone-de-essing-and-de-popping

7. **Wikipedia, De-essing**
   https://en.wikipedia.org/wiki/De-esser

8. **TzAudios, What is Sibilance and How Does Your Microphone Choice Affect It?**
   https://tzaudios.com/blogs/what-is-sibilance-microphone-choice

9. **Michael Joly, Why Many Condenser Mics Are Too Bright and Sibilant**
   http://heardofaudio.com/resources/blog/PDF/Why-Many-Condenser-Mics-Are-Too-Bright-and-Sibilant.-Michael-Joly.pdf

10. **IRPR Sound Design Guide, How to reduce plosives and sibilance in voice recordings**
    https://sounddesign.irpr.agency/guides/how-to-reduce-plosives-and-sibilance/

11. **Violet Recording, What Is a Pop Filter?**
    https://violetrecording.com/what-is-a-pop-filter/

12. **sonible, Moving Beyond Traditional De-essers**
    https://www.sonible.com/blog/beyond-traditional-deessers/

13. **Nallaguntla et al., Phoneme-Level Deepfake Detection Across Emotional Conditions, arXiv:2605.03079 (2026-05-04)**
    https://arxiv.org/abs/2605.03079v1

14. **Li et al., Modeling and Estimation of Vocal Tract and Glottal Source Parameters Using ARMAX-LF Model, arXiv:2410.04704 (2024-10-07)**
    https://arxiv.org/abs/2410.04704v1

15. **Song et al., ExcitNet vocoder, arXiv:1811.04769**
    https://arxiv.org/abs/1811.04769v3

16. **Fan et al., A Waveform Representation Framework for High-quality Statistical Parametric Speech Synthesis, arXiv:1510.01443**

https://arxiv.org/abs/1510.01443v1