Episode Book

EPISODE #6171

AI Audio Restoration: What Actually Ships in 2026

PUBLISHED	21 October 2026 at 13:53 UTC
DURATION	24:23
LISTEN	myweirdprompts.com/episode/ai-audio-restoration-tools-65736/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.11
BOOK GENERATED	22 October 2026 at 07:46 UTC

SYNOPSIS

Restoration tools have quietly become generative — they don't subtract noise, they rebuild the speech they think was underneath it. That works beautifully until it invents a phoneme you never said. This episode maps what actually ships in 2026: DeepFilterNet 3, ClearerVoice-Studio, Adobe Podcast's stem-splitting Enhance Speech update, ElevenLabs' voice isolator, and NVIDIA Maxine on the mature side; LavaSR and Miipher 2 on the bandwidth-extension side; Matchering and Ozone 12's Stem EQ and Unlimiter on the EQ side. Along the way: why DNSMOS stays high even when a model hallucinates content, why the URGENT challenge's top systems fused generative and predictive models, and why the real unsolved problem isn't the models — it's that you can't audit them by ear.

TRANSCRIPT



Corn

Picture a recording that sounds like it was made inside a stairwell. You hand it over, and what comes back is the same recording, only better. Daniel wants to know how far models have got at that job in particular. Not generation, not voice cloning. Matching a clip to a target EQ taken from a reference or a sentence, stripping background noise, rebuilding thin low-resolution speech, remastering. And behind all of it, whether an agent can sit in the editor's chair yet.



Herman

The answer changes depending on which of those jobs you mean, because they are at wildly different stages. Noise removal and bandwidth extension are products you can use today, some of them free. Describing the sound you want in words and getting a finished EQ back is mostly still research prototypes. And a full AI audio engineer that takes a raw file and returns a finished one? I haven't found a shipping product that does that.



Corn

So the maturity runs backwards from what the prompt cares about most.



Herman

Pretty much. The interesting part is that the mature tools are mature in a way that muddies the word editing. The best restoration models today are generative. They don't subtract the noise, they rebuild the speech they think was underneath it.



Corn

Which is generation wearing an editor's apron.



Herman

Yes, and that is where the trouble lives. A generative model that rebuilds your voice can rebuild it wrong. We'll get there. First, what actually ships.



Corn

Go on, I'll bring my scepticism.



Herman

On the open side there's DeepFilterNet three, which is real time and small enough to run on a laptop without the fans noticing. ClearerVoice-Studio from the ModelScope people bundles enhancement and also upsamples speech to forty-eight kilohertz. Resemble has an open enhancer that denoises and extends bandwidth, though the quality claims there are the company's own. On the closed side, Adobe Podcast's Enhance Speech got an update in March that split a recording into speech, background noise and music with separate controls, and you can download the stems. ElevenLabs has a voice isolator with an API. NVIDIA's Maxine is the enterprise version of all this, licensed, aimed at calls and streaming.



Corn

The stems are the part I'd have expected to be hard.



Herman

Separation is now just a feature. Noise removal turned from a filter into a question of which layer of the mix you want to keep. It's also why remastering tools are picking it up, but I'll hold that for later.



Corn

Four-letter names and free downloads. Where's the catch?



Herman

The catch is mostly what it was trained on. DeepFilterNet is the cheap, fast option, and a paper this year showed that if you train it with accurate simulated rooms, the word error rate on downstream transcription drops. So the quality of the fake rooms in the training data mattered more than anything clever in the architecture. The URGENT organisers say the same thing about training-data quality, loudly. Meanwhile the generative side got cheaper too, with one-step models like MeanFlowSE and DriftSE, which skip the long diffusion loop. DriftSE is MIT licensed and on Hugging Face.



Corn

So the diffusion tax is going away.



Herman

Partly. Another paper from June says the trade doesn't always justify itself, that GAN and discriminative models train faster and use data more efficiently, and diffusion is the expensive end of that. For a podcaster, I'd say: a small predictive model for ordinary hiss and room, a generative one only for disasters.



Corn

Is there any scoreboard for who's winning? Or is it everybody's demo reel against everybody else's?



Herman

There's one decent one. The URGENT challenge at ICASSP this year threw everything at the systems at once, noise, reverb, clipping, band limits, codec damage, wind, packet loss. Eighty-plus teams registered, twenty-nine made valid entries. And the organisers' overview says the top systems were hybrids, a generative model fused with a predictive one, and the hybrids consistently beat either camp on its own.



Corn

Neither tribe wins.



Herman

The one I'm sure about is the structure, the winner on the objective ranking fused a generative and a predictive model. I only read that team's abstract, so don't hold me to the details. And I haven't seen the final listener-based ranking, so I won't claim who took it.



Corn

Fusing the two sounds like admitting each one is bad at something. What is the generative half bad at?



Herman

Knowing when to shut up. A predictive model, the classic mask-the-noise kind, tends to smear and muffle, but it stays tied to the signal. A generative one sounds gorgeous and sometimes says things you didn't. There's a paper from June comparing the two on exactly that. Above about minus seven decibels, meaning the noise is up to seven decibels louder than the speech, the generative models hold up fine. Below that, the metrics fall off sharply and the output starts growing spectral content that wasn't in the input.



Corn

Seven decibels louder. That's a person shouting over a leaf blower.



Herman

Close to it. And the numbers on the content are what get me. The GAN-based systems kept the right phonemes about eighty-five percent of the time in the worse noise group and ninety-two percent in the slightly better one. That sounds okay until you realise it means one phoneme in twelve or so in the milder group is simply the wrong sound.



Corn

Which you won't hear, because it'll be a very well-made wrong sound.



Herman

The nasty part is the next bit. Another group found that DNSMOS, the automatic quality score half the field leans on, stays high even when the generative output has invented content. Listening tests of the standard kind miss it too, because the voice sounds clean and natural. Wrong word, lovely timbre.



Corn

So the cleaner it sounds, the less anyone checks it. Which I'd call backwards...



Herman

Fair, and there's a mitigation I like. A different paper found that when the model's speech embeddings vary a lot between runs, the output is probably making phonetic mistakes. Run it several times, compare, and you can cut the word error rate in the low-SNR cases. The reduction was fifteen percent, though I couldn't tell you whether that's relative or absolute.



Corn

So the fix for a model that invents things is to ask it five times and see where it disagrees with itself.



Herman

A lot of editors do the same thing with a bad take.



Corn

Fine, that one's good. Herman, when the noise is that bad, would you rather the tool refuse, or take its best guess and give you something listenable?



Herman

Refuse, flagged, every time. If it hands me a guess I want it labelled as one. As far as I know none of the shipping tools tell you when they've crossed into inventing, they just give you the file. Do you disagree?



Corn

I'd take the guess over a recording nobody can understand. But I'd want the label too.



Herman

Then we agree about the label, and the label is what's missing.



Corn

Upscaling next. Thin phone speech to something with air in it.



Herman

Bandwidth extension. LavaSR is the one I find most interesting this year, a paper from March, built on a Vocos vocoder. It takes anything from eight to forty-eight kilohertz in and writes forty-eight out. The model is around fifty megabytes and the repository is Apache licensed. On a modern GPU it claims something like five thousand times faster than real time.



Corn

Five thousand times. An hour of audio in less than a second.



Herman

Self-reported, single author, and trained on one clean speech corpus, so what it does to a noisy voicemail from a train is an open question. But the cost is clearly trivial now. The survey that came out in May makes the real point though. The problem is ill-posed. The highs that were thrown away aren't coming back, because the information is gone. Any model is choosing among many possible highs that fit. Older discriminative models average them and sound dull, so the field moved to generative ones that pick one and commit.



Corn

A confident guess at what the treble should have been.



Herman

Right, and a good one, because human voices are predictable. Sibilants have a shape. But it's plausible rather than recovered. If you restore a recording of someone and the highs are invented, you've made a recording of a statistically likely version of that person. Google has an internal model, Miipher two, that does this kind of thing at industrial scale to clean training data, a million hours in about three days on a hundred accelerators. Not a consumer tool, but it tells you who is eating this stuff.



Corn

And everyone feeding it into speech models never checks whether the voice is the same voice.



Herman

Some of them check. Not enough. Here's what I wonder, Corn. Take a restored clip that sounds right to you. What is your ear actually checking when it says that sounds right?



Corn

Well, the phone is ringing, so let us see who it is. Hello, you are on the air.

D

Dorothy

Hello dear, it is your mother. I am ringing for Dorothy's Kitchen.



Herman

Dorothy, what a joy. Please go ahead, I have a pen.

D

Dorothy

Today I will teach you a glass of tap water. The ingredients are water, and a glass. That is all. Do not get creative.



Corn

Mum, is there anything we can substitute for the glass?

D

Dorothy

You fill the glass, but not to the very top, sweetheart, because the water needs room to think. Then you wait. The waiting is the hard part. Things have gone wrong before and I would rather not say how.



Herman

How long is the wait, Dorothy? I want to get the timing right.

D

Dorothy

Do not watch the glass too closely or it will sense your worry. I once looked away for a moment and, well. We do not speak of it.



Corn

Understood, Mum. I won't stare.

D

Dorothy

Serve it plain. If you must, a slice of lemon, but that is showing off. Now go and put a jumper on, it is cold. Bye bye, dear.



Herman

Goodbye, Dorothy, always a pleasure. Corn, you were about to tell me what your ear checks when a restored clip sounds right.



Corn

Familiarity. It's checking whether it matches my memory of how a person like that sounds. Not whether it matches the person. Which is exactly the test a generative model is built to pass.



Herman

That's why I'd say the evaluation is the unsolved part, more than the models. A Levenshtein distance on phonemes is a better test than a pleasant score, and there's a newer latent diffusion model, DiTSE, that claims less content hallucination. Claims, from the paper's abstract, which I've only read as an abstract.



Corn

So noise removal and upscaling are mostly there, and the catch is you can't audit them by ear. Now the EQ half, since that's where Daniel started.



Herman

This is where the maturity gap shows up properly. Matching a clip to a reference EQ is old. Matchering is open source, it takes a target and a reference and matches the overall loudness, frequency response, peak level and stereo width, and it has a brickwall limiter bolted on the end. No neural network in it at all.



Corn

A limiter I could use on Herman's enthusiasm.



Herman

I'll let that go, because you're also about to say it's not AI, and you'd be right. Same with the mastering assistants in the commercial tools. Ozone's assistant listens, analyses, and picks settings for modules that are classic signal processing. It's machine learning choosing the knobs, and the knobs are ordinary. iZotope's own pitch is guide, not decide.



Corn

Four hundred and ninety-nine dollars at launch for the Advanced tier to be guided.



Herman

Plus a fifty-five dollar Elements, and a two hundred and nineteen dollar Standard. I only have launch prices, so I wouldn't quote today's. But Ozone twelve did add a few real model-based pieces. Stem EQ, which uses separation so you can EQ the vocal inside a finished mix. An Unlimiter that tries to restore dynamics squashed out by heavy compression. Which, if you think about it, is bandwidth extension for loudness. The model guesses what the peaks used to look like.



Corn

And the describe-it-in-words side?



Herman

That's the research frontier, and it's a good one. Text two FX in twenty twenty-four used CLAP embeddings, which put text and audio in the same space, steering a differentiable equalizer and reverb from open-ended prompts. Then LLM two FX from Sony and KAIST last year had a language model predict EQ and reverb parameters from text, zero shot. You say make this warmer and brighter, and it outputs the actual settings. A January paper went further, fine-tuning a model on listening-test data so that warm means what a population of listeners means, not one engineer's taste.



Corn

So the model doesn't touch the audio.



Herman

I like this part. The language model picks the parameters, the audio goes through a normal equalizer, and you can open it and see the curve. An engineer can look at the chain and say, no, that cut at three kilohertz is wrong, and change it. Nothing gets invented. There's also a recent paper on multi-turn refinement, too harsh, soften it, with an LLM plus CLAP-guided optimisation. I couldn't confirm the date on that one, so treat it as very recent.



Corn

Real question. If the EQ is just a chain of knobs, why would I want a language model doing it rather than a slider?



Herman

You wouldn't, if you knew what you wanted in hertz. The model earns its keep when you only know what you want in adjectives. And the tough bit is that the same adjective means different things on a voice and on a drum bus. I'm not certain anyone has solved that.



Corn

The two-voices-one-curve problem again. Nobody fixes that with one preset.



Herman

Yes. So the pipeline for matching is one of two: deterministic matching against a reference, which works and is boring, or language to parameters, which is promising and narrow. LLM two FX covers EQ and reverb and that's it. Compression, de-essing, automation across a whole episode, none of that is in there.



Corn

Ozone has a reference option too, doesn't it? Same idea as Matchering with a nicer interface?



Herman

Similar goal. The Master Assistant in the new version has a custom flow where you pick the modules, set a loudness target in LUFS, add genre modifiers, and optionally point it at a reference. Then it proposes. Neutron, the mixing sibling, does the same per track, suggesting EQ and compression for a vocal or a guitar. LANDR sells mastering as styles, warm, balanced, open, and a reference option, and its engine claims nineteen million tracks of training, undated, which I'd file under marketing. All of those are the selection model. None of them is a model that takes your audio and emits better audio.



Corn

Selection with a very expensive suit on.



Herman

Which is fine, honestly, because selection is half of what an assistant engineer does. The other half is hearing it.



Corn

And remastering as a whole? Daniel put it on the list.



Herman

Remastering mixes three things. Source separation, so you can treat stems. Mastering moves, like loudness and tonal balance. And restoration. Separation got very good, the BS-RoFormer family won the twenty twenty-three separation challenge, and an ensemble from MVSEP reports over eleven decibels of vocal signal-to-distortion ratio on its own test set. Self-reported, so a pinch of salt. There's also research on applying a reference track's mastering character to your track at inference time, ITO-Master, and Sony's MEGAMI, a diffusion model for multitrack mixing that works on effect embeddings.



Corn

You've listed nine things. Which ones can I download tonight and trust?



Herman

Matchering, certainly. Ozone, if you pay. Separation models, for stems. The rest are papers. Which gets us to the agent question.



Corn

Which is the one he cares about.



Herman

And it's the weakest part of the picture. There are frameworks, WavCraft from twenty twenty-four and AudioFab, which is open source and uses MCP to expose thirty-six audio techniques to a language model. The planner picks a tool, runs it, listens, picks the next. The problem is the language model. A paper from July, RIME, built about fifteen thousand instruction and audio pairs for post-production, with tools for separation, mixing and effects. Zero-shot, multimodal agents struggled to pick the right operator and its parameters. Training on the synthetic pairs helped, which tells you the gap is closable. But it's a gap.



Corn

So the sad version of the agent story is a model that reaches for the wrong tool and can't tell it's wrong.



Herman

Because it can't hear what it just did. Ears are the missing piece. And look at what we just said about evaluation. The metrics that would tell the agent it's succeeding are the same ones that miss invented phonemes. An editor listens for the word that shouldn't be there. A model grading itself with DNSMOS is marking its own homework.



Corn

Everything you've said makes me think replacing the editor isn't the right frame at all. The editor stops cleaning and starts auditing.



Herman

That's my read too. A human editor spends most of a day on judgment calls, this breath stays, this cough goes, this room tone matches the last cut. The tools can do the mechanical passes. I wouldn't hand a model the final listen, not yet. Would you?



Corn

Depends who's grading it. If the grader is the model, no. Everything you're hearing from us went through a mastering chain, so I have a stake in this.



Herman

Then you'll want to hear this. The thing I'd flag, though, if this is going into your own workflow, is what a tool does to the thing you can't hear. The hybrid models are the best at noise, and generative restoration is the best at upscaling, and both of them fail quietly. You need a check that isn't a quality score. Transcribe before and after and compare the words.



Corn

Cheap, and nothing in it relies on anyone's ears. Fine, I'll take that one.



Herman

One more point on the products, because the convergence is striking. Adobe's separation controls, Ozone's stem EQ, the voice isolators. Everybody is arriving at the same architecture. Split the file into layers with a separation model, treat each layer with the right tool, put it back together. Sometimes an LLM picks the treatments. That's the nearest thing to an AI engineer anyone has, and it's a pipeline, not an engineer.



Corn

And the part of that pipeline that can't be bought yet...



Herman

Is the last step. Deciding it's done.



Corn

Hm. The part about the highs is still bothering me. If a model adds everything above the cutoff, what you hear is a prediction of what the original sounded like. It isn't a measurement. It's...



Hilbert

Trespass. Everything above eight kilohertz at our address belongs to Mr. Pruitt at number nine. He has a council letter. The letter is addressed to the previous occupant of the flat, a pony called Hobbs, but the council confirmed it stands.



Herman

Which means a bandwidth extension job on an eight kilohertz file puts the whole of the generated treble on his land. Forty kilohertz of it, if you're going to forty-eight. And since the model wrote all of it, he's billing you for the guess, not for anything that was in the recording.



Hilbert

Two pounds forty a kilohertz. Cash through the letterbox. He writes a receipt, Herman, in triplicate, and the third copy goes to his sister in Leeds who has never been told why. He retired from the municipal pigeon registry in a year I'd rather not say, and since then the highs are all he's got. He can't hear above nine himself. He charges for it from memory.



Corn

Then your cheapest edit is the one that adds nothing.



Hilbert

I tried that. If I leave a file at its original bandwidth, he charges a storage fee for the gap. He says the space exists whether I fill it or not. That's a separate form. I have the form.



Herman

A storage fee on empty spectrum is at least consistent. The model is selling you spectrum that was never recorded, so the space is the product.



Hilbert

Herman, you keep saying invented. He prefers imported. Noise removal pays him as well. Hiss lives up there. When a denoiser takes the hiss out he counts it as a sale of his property without a receipt, and he writes. Eleven letters since August. I answer every one, Corn, with a stamped envelope, because he reuses the stamps. He keeps the cash in a tuba, which he says is not a tuba.



Corn

What does he say it is?



Hilbert

A safe. It's got a mouthpiece. Last Tuesday I ran an eight to forty-eight job, a voicemail off a train, and it came to ninety-six pounds. He gave me the change in postage stamps and a receipt that says tuba, paid.



Herman

Ninety-six is forty kilohertz at your rate. The arithmetic holds, at least. The model's side of the bill and his are the same size, and neither of you can hear what you're paying for.



Corn

Let me do the closer, then. The misconception I hear most is that noise removal is subtraction, that the model finds the noise and takes it away and what's left is what was there.



Herman

The good ones don't work that way anymore. They rebuild the speech from what the noise left behind, and they're graded by scores that don't notice when the rebuild is wrong. Cleaner is not the same as truer.



Corn

Then the thing to watch is whether the tools start itemising what they made up, like a bill, instead of handing you one lovely file. A confidence map on the highs. A flag on the phonemes it wasn't sure about. Then an editor, or an agent, would have something to audit.



Herman

Thank you to our producer, Hilbert Flumingtop, for keeping the sound honest through all of this.



Corn

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Herman

We'll be back soon.