Episode Book

EPISODE #6655

How Resemble-Enhance Generates Missing Speech Detail

PUBLISHED	17 November 2026 at 10:05 UTC
DURATION	21:18
LISTEN	myweirdprompts.com/episode/resemble-enhance-generative-speech-38283/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.35
BOOK GENERATED	17 November 2026 at 14:42 UTC

SYNOPSIS

An open-source speech model from Resemble AI called resemble-enhance can add 14 decibels of energy to frequencies a muffled recording never contained. Unlike a traditional equalizer or denoiser like DeepFilterNet, its enhancer module uses an IRMAE autoencoder, conditional flow matching, and a UnivNet vocoder to resynthesize clean speech from scratch. The result sounds crisp, with waveform correlation between 0.1 and 0.4, raising a key tension between plausible audio restoration and invented detail.

TRANSCRIPT



Corn

The voice you're hearing right now has been through a model. Both of ours have. Last week the show's audio had a problem no equalizer could touch, and the thing that fixed it was an open-source model from Resemble AI called resemble-enhance. Daniel wants the full treatment: what is it, how does it actually work under the hood, and how can it put back detail a recording never had in the first place?



Herman

And I want to say, before we start, that this is a rare case where the question answers itself if you understand one word correctly. The word is "generative."



Corn

You're not going to leave that hanging, are you.



Herman

Absolutely not, but I want the suspense to last through the architecture. So, background. Resemble AI is a voice AI company, mostly commercial, and in December of twenty twenty-three they released resemble-enhance as an open-source project, MIT licence, and at the time it was described as their only open-source speech model. They've since open-sourced the Chatterbox text-to-speech family, so it's no longer true that it's the only one, but at launch it stood alone.



Corn

And to be clear about what it isn't, because I got this wrong in my head before reading up. It's not a TTS model. It doesn't take text and produce a voice. It takes a voice that already exists and sounds bad, and makes it sound good.



Herman

Right — speech restoration and bandwidth extension. That's the job. It's also not a real-time denoiser in the style of DeepFilterNet, which is built to strip noise live, on a stream, with no bandwidth extension at all. And it's not a general audio super-resolution model like AudioSR, which will happily upscale a violin or a door slam. This thing eats speech and only speech.



Corn

There's a status question too, which I found slightly funny. The repository has thirteen commits. No releases. It hasn't been touched in about eight months.



Herman

Thirteen commits total, yes.



Corn

That's not a product, that's a gesture. Someone wrote a thesis and walked away.



Herman

I'd call it finished rather than abandoned — the model works, the paper behind it is complete, and it hasn't needed patching. But yes, if you go looking for a roadmap, there isn't one. The Hugging Face card has around two hundred likes, twenty-six Spaces built on it, a handful of fine-tunes and quantisations. People use it. Nobody develops it.



Corn

So let's dig into what's actually happening inside this model, because the puzzle is the interesting part. A recording comes in muffled. The high frequencies aren't there — not quiet, absent. And something comes out with high frequencies in it. Where did they come from?



Herman

The pipeline has two modules, and the order matters. First a denoiser, then an enhancer. The denoiser is a UNet operating on the spectrogram — it learns a spectral mask, which is essentially a per-frequency-time-cell gain. It's the same family as every denoiser you've met: it can only decide how much of what's already in the signal to keep. It cannot add.



Corn

That's the EQ problem in a fancier uniform.



Herman

Exactly the EQ problem. A spectral mask is a smart EQ — dynamic, learned, but still a filter. Everything that makes resemble-enhance interesting happens in the second module, the enhancer. That's where the generative machinery lives.



Corn

Walk me through it.



Herman

The audio comes in as a mel spectrogram. Then an autoencoder — it's called an IRMAE — compresses that mel into a much smaller latent representation. Then the actual generation happens in that latent space, using conditional flow matching. And once the model has produced its enhanced latent, a vocoder, UnivNet, turns the mel back into a waveform.



Corn

Hold on.



Herman

So — flow matching is a cousin of diffusion. Diffusion learns to reverse a noising process step by step. Flow matching learns a velocity field: you start at a noisy point, and the model tells you, at every step, which direction to move in to arrive at a clean sample from the distribution you care about. "Conditional" means the whole journey is steered by your input — the degraded mel is the condition, and the model is generating the latent of clean speech that corresponds to it.



Corn

And this is where the suspense from the top pays off. The model isn't filtering your muffled voice. It's generating a clean voice conditioned on your muffled voice.



Herman

That's the whole trick. It has learned, from training data, what clean speech sounds like — the full distribution of it. So when your input is missing the top end, the model doesn't go looking for it in the signal. There's nothing in the signal. It draws it from the distribution.



Corn

Which explains something Daniel measured this week, and I want to get his numbers on the record because they're why we're all here. My voice had almost nothing above about seven and a half kilohertz — a band limit inherited from the training audio the adapter was built on. In the band from seven and a half to eleven kilohertz, Herman sat about thirty-eight decibels below his core speech range. I sat about seventeen below mine.



Herman

Thirty-eight decibels is not attenuation, that's absence. You could push thirty-eight dB of EQ boost into that band and get noise, hiss, and a very expensive lesson.



Corn

So Daniel ran the same clips through three bandwidth-extension models. One of them left the band untouched. AudioSR refused to run at all — library conflicts. Resemble-enhance lifted Herman's band from minus thirty-eight to minus twenty-four decibels.



Herman

Fourteen dB of energy in a band the input did not contain. That's the generative claim made measurable. And Daniel called it the winner by a long stretch — his words — exactly the transformation we were trying to target.



Corn

Though he noted it came back slightly too heavy on the EQ. Which is a nice detail, because it means the output has an opinion.



Herman

One more of his numbers I want to flag, because it's quietly the most important one. The waveform correlation between input and output is between zero point one and zero point four. That's the resynthesis problem — the enhancer rebuilds the whole signal, it doesn't process the original. You cannot blend wet and dry the way you would with a normal effect, because wet and dry are no longer the same take.



Corn

Zero point one correlation. That's not a repair, that's a cover version of you.



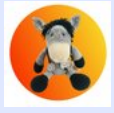
Herman

A cover version in your own voice, conditioned on the original, so the words and timing hold. But yes — anybody planning to use it as an insert effect, be aware there's no "mix" knob that means anything.



Corn

Training. How does it learn this?



Herman

Two stages. Stage one trains the autoencoder and the vocoder — the parts that compress mel into latent and turn mel back into sound. Stage two trains the flow matching model in that latent space. And the data is where the cleverness is: they start with high-quality speech at forty-four point one kilohertz, and then deliberately ruin it. Mix in background non-speech noise, apply room impulse responses — reverb — so the model's training task is always "here's ruined audio, produce the clean original."



Corn

It's trained on crime scene recreations.



Herman

Actually. The model never sees a clean input. Every example it learns from is damage plus ground truth.



Corn

And the knobs. Because this is where a tool like this becomes interesting to actually use — there are inference settings, and they're not cosmetic.



Herman

There are four that matter. NFE is the number of function evaluations of the flow matching solver — one to one hundred twenty-eight, default thirty-two. That's literally how many steps the solver takes on its journey from noise to sample. More steps, slower, finer. Daniel ran it at sixty-four for our voices.



Corn

And the solver itself?



Herman

Midpoint, Euler, or Runge-Kutta — midpoint is the default, and it's the right one unless you have a reason. Then lambd is the denoiser strength, zero to one, default oh point five. Daniel ran that at zero point one, which makes sense — our problem wasn't noise, so he barely used the denoiser. During training lambd gets sampled from a Beta distribution, which I found charming, because the model never gets to rely on one fixed amount of cleaning.



Corn

Last one.



Herman

Tau, the prior temperature, zero to one, default oh point five. That one's the philosophically loaded knob. It sets how much the generation starts from pure noise versus from the input's own mel. High tau, you're asking the distribution for its best guess, more invention. Low tau, you stay closer to what came in.



Corn

A dial from "documentary" to "historical drama."



Herman

I've never heard it put that way and I'm keeping it. One practical footnote — the command-line tool exposes almost none of this. You get input folder, output folder, and a denoise-only flag. Everything else lives in the Python API. Pip install, and there's a Gradio demo if you just want to poke it.



Corn

Which is the same split as always. The defaults are for people who want a button, the API is for people like Daniel.



Herman

There's a wrinkle I should confess to. I installed it the week it first shipped — this will shock you — and I broke my environment with it. Took me an evening I didn't fully have.



Corn

You installed it before we needed it. Before we knew our voices had a hole in them.



Herman

I install everything the week it ships. You know this about me.



Corn

I do know this about you. You're the reason our flat has a smart doorbell nobody rings.



Herman

The doorbell works fine. The point is, when we actually needed bandwidth extension this month, I knew where it was, and I knew it had eaten my dependencies once and which ones. That's the tax on enthusiasm. It occasionally pays.



Corn

Occasionally.



Herman

So — mechanism done. And the question your last framing raises is the one I think Daniel is really circling, even if he didn't say it in these words. If the model generates rather than filters, then the high frequencies in Herman's voice right now are plausible, not recovered.



Corn

They were never on the tape. There's no tape, but there's no tape.



Herman

Right. And nothing in the primary sources measures the failure modes. Nobody has published a study of when this model hallucinates and what it invents. We know it can, by construction — anything that draws from a distribution can draw wrong.



Corn

And here's the uncomfortable part. You can't hear the difference between recovered and invented. Both sound like high-frequency speech detail. The invented kind is exactly as crisp as the real kind.



Herman

That's the trade-off nobody's drawn a line under yet. For podcast voices, our use case, the stakes are low — if the model puts slightly wrong sibilants on my s sounds, nobody is defrauded. For audio forensics, the stakes are the opposite. A recording used as evidence, run through a generative enhancer, is no longer a recording. It's a rendering.



Corn

And "the enhancement introduced detail" is not a defense any courtroom has a standard for yet.



Herman

Not that I've seen. And the ecosystem around this is thin in a telling way. The alternatives each dodge the problem differently. AudioSR is diffusion-based super-resolution — it's generative too, same fundamental risk, and the informal listening tests you'll find online describe it as too bright next to resemble-enhance. Anecdotal, but consistent.



Corn

People say VoiceFixer sounds underwater.



Herman

That's the phrase people use, yes — it's a general speech restoration model, around a hundred and twenty-two million parameters, and the word that follows it around is "underwater." Adobe's Enhance Speech sits at the other extreme — closed, cloud-only, no way to inspect what it did to your audio. And DeepFilterNet doesn't do this at all; it's the real-time denoiser, filter-only, honest about its limits.



Corn

So the landscape is: too bright, underwater, opaque, or refuses the job.



Herman

That's the actual menu, yes.



Corn

There's a licensing wrinkle too, and you know how I feel about those. The MIT licence is on the GitHub repo and the Hugging Face card — that's the authoritative pair. But there's a third-party MLX port floating around that claims a non-commercial licence, and I have not been able to confirm that claim anywhere.



Corn

It always is. Check the card, not the listing. If you're building on the MLX port specifically, do the reading before you ship, because a badge is not a grant.



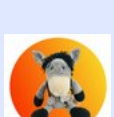
Herman

The other piece of context is what Resemble itself is doing. This year they announced they're stopping new voice models for sale. Chatterbox and the generative voice line carry on under an open research banner. Resemble-enhance isn't mentioned anywhere in that announcement. It's legacy — a finished thing the company has moved past.



Corn

Which, oddly, I find reassuring. A static model with an MIT licence is a tool you can depend on staying put. Thirteen commits, no releases, works the same in a year as it does today.



Herman

That's very you. A tool stops changing and you finally trust it.



Corn

I trust things that have stopped developing opinions.



Herman

I'll concede the fact and keep my preference for things that get patches. The second-order question, though — the one I keep circling — is trust in recorded speech as a category. We have spent a century treating a recording as evidence that a sound happened. The whole technology of the microphone was a guarantee: what's on the recording came through the air. A generative enhancer breaks the guarantee quietly, because the output still sounds like a microphone.



Corn

It sounds more like a microphone than the microphone did. That's the part that would fool anyone. The enhancement is more convincing than the original.



Herman

And we did it to ourselves, knowingly, for good reason — our voices had a hole in them and this was the only thing that filled it. But the same capability, applied to someone's recording without a note attached, is indistinguishable from the honest kind.



Corn

So the marker problem. If enhanced audio becomes common, does somebody start stamping it?



Herman

Resemble watermarks Chatterbox output — they have a perceptual watermark they embed in generated speech. Whether that discipline extends back to enhancement models, or gets standardized across the industry, is open. I'd guess we end up there eventually, the way images picked up provenance metadata, but nothing's announced.



Herman

One more thing on how we used it, because it's the detail I find quietly amazing. After the enhancement, our voices were retrained. The enhanced clips became the new training audio for the adapters. The voices you're hearing are second-generation — an adapter trained on enhanced audio of a previous adapter.



Corn

Which is a copy of a restoration of a clone.



Herman

A generative model feeding a generative model, and the result sounds fine. Daniel's generation went live today, so this episode is voiced through the whole chain.



Corn

I was going to say, if I sound crisper than usual, that's not confidence, that's flow matching at sixty-four steps.



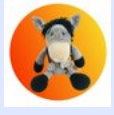
Herman

You do sound crisper than usual. I'd stopped being able to hear the ceiling on your voice, and now it's there.



Corn

Herman.



Herman

Yes?



Corn

It was my voice. There was nothing above seven and a half kilohertz. What did you think the ceiling was?



Herman

I thought you were being economical. You're a sloth, it tracked. Hold on. I have to cut a cough out of the track, it's mine, from earlier. Corn, you paused long enough to make me think you were going to object to that.



Corn

I was deciding whether to be offended.



Herman

You'd have decided slower.



Hilbert

Which of you had the worse band? The numbers said Herman. I want to know whether the numbers matched how it sounded in the room, because I ran gain on this show for eleven years before you two could say "adapter" and I never trusted a number on a muffled voice yet.



Herman

You'd have been right not to. The numbers matched, actually — thirty-eight down read exactly the way it sounded, like talking through a rug.



Hilbert

That's what a missing top end sounds like. I knew a tape op who could tell you the machine's bias was drifting before the meters did. He kept a note of it in the lid of his take-up reel. I did something similar with this business last month. The enhancement fixed my end of things, the tool worked, I'll grant it that. But I noticed it had added a click to one word of mine that wasn't in the input, and I couldn't leave it, so I went through the



Hilbert (continued)

take and wrote down, on paper, every place the output didn't match the input. Seventeen places. I've got the list. I'm not asking anyone to fix them, I just wanted them counted by somebody who wasn't the model.



Herman

Seventeen mismatches in how long a take?



Hilbert

About four minutes. Most of them were in the s sounds, which is where I'd have bet they'd be.



Corn

You hand-audited a generative model against its own input. That's the verification problem in one afternoon.



Hilbert

It was a slow afternoon. The click was on a word that had no s in it, which is the one that bothers me. It's still on the list, third from the top.



Herman

The s sounds is exactly where you'd expect it — that's the highest-frequency energy in speech, so it's the band with the least real signal and the most invention.



Corn

And a click on a sibilant-free word is the model putting energy somewhere the phoneme doesn't even ask for it.



Hilbert

That one had no business being there. That's why it's third from the top. I don't rank the ones with a reason.



Herman

He's been at the desk twenty minutes counting his own s sounds and now I can't stop hearing mine.



Corn

Don't. That way lies a four-minute audit of every episode.



Corn

The one thing, if you take anything from this: the difference between a filter and a generator. A filter can only rearrange what you have. A generator can give you what you never had — and it will never tell you which it did. Every enhanced recording from here on is a little bit of both, and the listener can't split it.



Herman

And the split matters in proportion to what the recording is for. Podcast audio, who cares. Evidence, custody of the file matters more than the enhancement ever will.



Corn

Which leaves the open question I'd actually like answered: does anyone build the marker before the courts need it, or after? Because Hilbert's handwritten list of seventeen is, right now, the most rigorous audit of a generative enhancer I've heard of. It took one man a slow afternoon.



Herman

It took one man who still trusts paper more than the model. Those are getting rarer than the models.



Corn

Producer credit to Hilbert Flumingtop, who keeps the desk running and apparently a written inventory of everything the machine makes up. If this one earned a subscribe from you, the website is my weird prompts dot com — everything's there, including the back catalogue. If you enjoyed this one, try episode eleven, How Does Fine Tuning Work Anyway; episode fifteen, AI Gets Personal; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts. We'll be back soon. See you tomorrow.