MY WEIRD PROMPTS

Podcast Transcript

EPISODE #58

Clean Audio, Messy Reality: Noise Removal for Voice-to-Text

Published December 12, 2025 • Runtime: 28:35

https://myweirdprompts.com/episode/clean-audio-messy-reality-noise-removal-for-voice-to-text/

EPISODE SYNOPSIS

When you need to record a voice memo while holding a fussy baby, which noise removal strategy actually works? Herman and Corn dive deep into the trade-offs between real-time on-device processing, cloud-based post-processing, and hardware microphone solutions. Discover why audio that sounds cleaner to human ears might actually transcribe worse, and learn which approach makes sense for your workflow. A practical guide to the neural networks and signal processing powering modern voice recording technology.

DANIEL'S PROMPT

Daniel

Hello there, Herman and Kori. So I have a question for you regarding the topic of we're going back to audio and voice tech. You may be able to hear dear little Ezra here. He's giving me a good workout today, carrying him around the place. And we're going through one of those more fussy days that I think is My wife has a wonderful decision flow chart printed up. It allows me to see the various things. Does he need a feed, milk, bedtime, all the all the various explanations for fussing, including when to maybe escalate it if we need to check in with a medical professional. But usually it's just those early stages thankfully. But there's just days where I think he's kind of really fussy. We talked about voice productivity and I mentioned a big reason or a big utility I found for voice tech has been that I got into it before becoming a parent. But since becoming a parent, it's proven really useful for just being able to when my hands are tied up and I really need to get something to somebody like my accountant or a client and it's urgent, I'm able to use it for recording emails. But a challenge in a circumstance like this is I'm not sending them the voice recording, so they I don't have to worry that they can hear sounds like what you're hearing at the moment, which is a bit less, you know, it wouldn't be the most professional. But I do notice that sometimes the transcription accuracy is degraded naturally by the ASR tech having to try to make sense of what you're saying and dealing with the background audio from in this case a fussing, crying, screaming baby. So regarding ways to kind of work around that, besides, of course, when I'm not saying that, you know, there's obviously you tend to as best as you can. But for background noise removal in general, whether it's this or whether it's might be a coffee shop or some other auditory environment that is that is problematic from a transcription standpoint, I notice there's a few approaches. One of them is real-time background noise removal that kind of tries to do that in real-time as the audio stream comes into the computer. And then you've got ones that try to do it afterwards, so you've recorded something and then they clean it up. On devices like Android, it's a little bit tricky to to really do much with the audio stream as it comes in in real-time. But I'm wondering, this is a general audio processing query and this is actually very much related to ASR because I think it's neural networks that really do this kind of cleanup work. Is it better to do these tasks if you're going to try to do them at the post-processing stage or real-time as it comes in or onboard the microphone because sometimes you do see microphones that say they have some kind of a background noise processing feature built in. Which makes the most sense and what are the pros and cons of the different methods?

TRANSCRIPT

Corn

Welcome back to My Weird Prompts, the podcast where our producer Daniel Rosehill sends in the questions that keep us up at night. I'm Corn, and I'm joined as always by Herman Poppleberry. Today we're tackling something that's become increasingly relevant, especially if you're trying to get work done in less-than-ideal circumstances - we're talking about background noise removal in voice recording and transcription technology.

Herman

Yeah, and what makes this particularly timely is that we're living in this moment where voice-to-text has become genuinely useful for productivity, but the real-world conditions people use it in are... messy. Literally, in this case - there's a baby involved in the prompt.

Corn

Right! So the core question here is really about strategy. When you're trying to clean up audio for transcription purposes, do you remove the noise in real-time as it's coming in, do you do it after the fact during post-processing, or do you rely on hardware microphones that claim to have built-in noise suppression? Which approach actually makes the most sense?

Herman

And I think the answer is more nuanced than people realize. It's not like one approach is objectively better across all scenarios. The trade-offs are real, and they're worth understanding.

Corn

Okay, so let's start with the basics. When we talk about background noise removal, we're essentially talking about neural networks doing computational magic, right? That's the technology powering most of these solutions these days?

Herman

Well, yes and no. I mean, neural networks are definitely the cutting edge of what's possible right now, but they're not the only approach. You've got traditional signal processing methods - things like spectral subtraction, Wiener filtering - those have been around for decades. But you're right that neural networks have really revolutionized this space because they can adapt to different types of noise in ways that older methods just couldn't.

Corn

So when I'm using something like the AI Noise Remover app on Android, or ElevenLabs' Voice Isolator - those are neural network based?

Herman

Exactly. Those are using deep learning models trained on massive amounts of audio data to identify what's speech and what's noise, and then they suppress the noise while preserving the speech. The sophistication there is actually pretty remarkable. But here's where I'd push back on something - a lot of people assume that because these tools use neural networks, they're automatically superior to everything else. The reality is more complicated.

Corn

How do you mean?

Herman

Well, neural networks are incredibly powerful, but they're also computationally expensive. They require processing power. So when you're trying to do noise removal in real-time on a mobile device like an Android phone, you're hitting some hard constraints. You can't just run a massive deep learning model while simultaneously recording audio and keeping the rest of your phone responsive.

Corn

Right, so that's where the trade-off comes in. Real-time processing has to be faster, which means maybe less sophisticated?

Herman

Exactly. Real-time noise removal, especially on device, has to be optimized for speed. You're looking at smaller, more efficient models. They work pretty well, but they're not going to be as effective as what you could do if you had unlimited processing time to work with the audio after the fact.

Corn

But here's what I'm curious about - if I record something on my phone in a noisy environment, and then I upload it to something like ElevenLabs or another cloud-based service, they can apply heavier neural network models to clean it up?

Herman

Yes, and that's actually a really important distinction. Post-processing, especially when done in the cloud, can use much more sophisticated models because you're not constrained by the need for real-time responsiveness. You can throw computational resources at the problem. The trade-off is latency - there's a delay while the audio gets processed.

Corn

So for someone like Daniel, who's trying to dictate an email to his accountant while holding a fussy baby, real-time noise removal might be the way to go because he needs the transcription quickly. But the quality might not be perfect.

Herman

Right, but - and this is important - real-time deep learning noise suppression has actually come a long way. There was some really interesting research showing that you can do real-time noise suppression without even needing a GPU. Modern mobile processors are powerful enough to run reasonably sophisticated models in real-time. So it's not like you're choosing between "fast but terrible" and "slow but perfect."

Corn

What about the hardware microphone angle? Like, some microphones claim to have built-in noise cancellation. Are those actually useful, or is that more marketing speak?

Herman

That's a great question, and honestly, hardware-based noise cancellation has its place, but I think software has largely surpassed it for most use cases. Here's why - a microphone with hardware noise cancellation is typically using some kind of analog circuit or a very simple digital filter. It's fast, it doesn't require much power, but it's not intelligent. It can't really distinguish between speech and noise the way a neural network can.

Corn

So it's just kind of... blanket noise reduction?

Herman

Basically, yeah. It might suppress everything below a certain frequency, or it might try to detect sudden loud noises, but it doesn't have the contextual understanding that AI-based systems have. A neural network trained on thousands of hours of audio can learn that a baby crying has certain acoustic characteristics, and it can be more surgical about removing it while preserving the rest of the audio.

Corn

Although, wait - wouldn't hardware-based filtering be faster? Like, there's no latency with a hardware microphone doing its thing?

Herman

There is some latency, actually. And more importantly, there's no recovery from mistakes. If the hardware microphone decides something is noise and removes it, that's it. With software-based approaches, especially post-processing, you can be more conservative, you can tweak parameters, you can even manually review what happened.

Corn

Hmm, but I'm thinking about the practical scenario here. If I'm using voice-to-text on my phone right now, I'm probably not uploading to the cloud and waiting for post-processing. I'm probably using whatever real-time solution is built into my phone or my voice recording app.

Herman

Right, and that's where I think the practical answer becomes clear. For most people, most of the time, real-time on-device noise removal is the way to go. It's fast enough, it's good enough, and you get your transcription immediately. The post-processing approach is better if you have the time and if quality is critical.

Corn

So in Daniel's case, he needs the transcription quickly to send to his accountant. Real-time noise removal on the Android app he's using would probably be his best bet?

Herman

Absolutely. He records the message, the app cleans it up in real-time or near-real-time, he gets a transcription, he sends it off. Done. The baby noise might still degrade the transcription accuracy a bit, but it'll be way better than if he didn't use any noise removal at all.

Corn

But let's talk about the actual accuracy impact. I'm curious - if you've got a neural network trained to remove background noise, how much does that actually improve transcription accuracy for something like speech-to-text?

Herman

So, this is where things get really interesting. The relationship between noise removal and transcription accuracy isn't always linear. Like, you'd think that if you remove 50% of the noise, your transcription accuracy improves by some proportional amount. But it doesn't work that way in practice.

Corn

Why not?

Herman

Because transcription models - the speech-to-text engines themselves - are also trained on noisy audio. They've learned to handle some amount of background noise. So if you over-process the audio and remove too much, you can actually introduce artifacts that confuse the transcription model. You end up with audio that sounds cleaner to human ears but actually performs worse with ASR.

Corn

Wait, so you can have audio that sounds worse to a person but transcribes better?

Herman

Exactly. This is counterintuitive, but it happens. The sweet spot is usually somewhere in the middle - enough noise removal to help the transcription model, but not so much that you're introducing processing artifacts.

Corn

That's wild. So the optimal noise removal isn't necessarily the one that sounds best?

Herman

Not necessarily. And this is why the research in this area is still really active. People are trying to figure out the optimal balance. Some of the newer approaches are actually training noise removal models and transcription models together, so they're optimized for each other rather than being separate steps.

Corn

Okay, so let me try to synthesize this. We've got three approaches - real-time on-device, real-time in the cloud, and post-processing in the cloud. Real-time on-device is fastest but maybe less sophisticated. Cloud-based real-time might be better quality. Post-processing could be the best quality but with latency. Is that fair?

Herman

That's a reasonable summary, but I want to add some nuance. Real-time in the cloud isn't really a thing in most cases. You're either doing real-time on-device or post-processing. The distinction is usually on-device versus cloud. And the other thing is that "cloud-based" doesn't necessarily mean better. It means you have more compute available, but the actual model being used matters more than where it's running.

Corn

Right, good point. So if I'm using Voice Isolator by ElevenLabs, for example, am I doing real-time processing or post-processing?

Herman

It depends on the tool. Some of their products are real-time, some are batch processing. But the general trend in the industry is moving toward real-time on-device processing because... well, privacy, latency, cost. You don't want to be uploading all your audio to the cloud if you don't have to.

Corn

That's a good point about privacy. If I'm using my phone to record a message about sensitive financial stuff, I probably don't want that going to a cloud service.

Herman

Exactly. And that's another reason why on-device processing has become more popular. The models are getting smaller and more efficient, so you can run them locally without needing tons of processing power.

Corn

Alright, let's take a quick break from our sponsors. Larry: Tired of noisy recordings ruining your productivity? Introducing CleanVoice Pro™ - the revolutionary audio processing device that uses proprietary quantum-resonance technology to eliminate background noise before it even enters your microphone. Simply attach CleanVoice Pro™ to any microphone, and it creates an invisible acoustic shield around your voice. NASA scientists were consulted on this project, probably. Users report cleaner audio, improved transcription accuracy, and a mysterious humming sound that some find soothing. CleanVoice Pro™ - because your voice deserves to be heard, clearly. Side effects may include over-confidence in your voice quality and an urge to record yourself reading poetry. BUY NOW!

Herman

...Alright, thanks Larry. So where were we?

Corn

We were talking about the practical trade-offs between different noise removal approaches. I think one thing I want to circle back to is the Android-specific challenge that was mentioned in the prompt. Android is apparently trickier for real-time audio processing than iOS?

Herman

Yeah, that's actually true, and it's worth understanding why. iOS has a more controlled environment - Apple controls both the hardware and the software, so they can optimize audio processing pipelines. Android is fragmented. You've got hundreds of different devices with different processors, different audio hardware, different versions of the OS. It's harder to build something that works consistently across all of them.

Corn

So does that mean Android users are stuck with worse options?

Herman

Not necessarily stuck, but they have fewer first-party options. Apple's built-in voice memo app has pretty solid noise cancellation. Android doesn't have an equivalent built into the OS. But there are good third-party apps. The AI Noise Remover app I mentioned earlier is actually quite good. So is Voice Isolator. They work around the fragmentation by being optimized for the most common device configurations.

Corn

But the latency issue is still there, right? Like, if you're trying to do real-time processing on Android, you're fighting against the OS architecture?

Herman

You're fighting against it somewhat, but it's getting better. Modern Android processors are legitimately powerful. The Snapdragon chips in current flagships can handle neural network inference pretty easily. The issue is more about consistency - you might have great performance on a flagship phone and mediocre performance on a mid-range device.

Corn

So if Daniel's using an older Android phone, he might get worse results?

Herman

Possibly, yeah. Or he might want to consider using a cloud-based service where the processing happens on powerful servers regardless of his phone's specs.

Corn

But then he's back to the privacy issue and the latency issue.

Herman

Right, so it's trade-offs all the way down. There's no perfect solution that works for every scenario.

Corn

I think that's actually an important point to make. Let me ask you this - if you were advising someone in Daniel's situation, what would you tell them to do?

Herman

Honestly? I'd tell them to try a few approaches and see what works best for their specific use case. But if I had to pick one, I'd probably recommend using a real-time on-device solution if they have a reasonably modern Android phone. Something like AI Noise Remover or Voice Isolator. These apps are good enough that you'll get a transcription that's clean enough to send to an accountant, and you get it immediately without uploading anything to the cloud.

Corn

And if they don't have a modern phone?

Herman

Then maybe explore cloud-based options, or accept that some transcription errors will happen and just do a quick manual review of the transcription before sending it.

Corn

Yeah, that's practical. I mean, if you're dictating an email, even if there are a few errors, you can probably catch them and fix them quickly.

Herman

Exactly. The goal isn't perfect transcription - it's good enough transcription that saves you time compared to typing.

Corn

Alright, so we've covered the technical landscape. Let me ask a bigger picture question - do you think the state of noise removal technology is actually good enough now that people should be regularly using voice-to-text in less-than-ideal environments?

Herman

I think it's good enough for a lot of use cases. If you're in a moderately noisy environment - a coffee shop, an office with background chatter - yeah, modern noise removal is quite good. If you're in a really extreme environment - like, a construction site or a loud concert - then you're probably going to have trouble no matter what.

Corn

And what about the baby crying scenario specifically? That's a pretty intense acoustic environment.

Herman

Baby crying is actually... it's a specific challenge because it's high-pitched, it's unpredictable, and it can be very loud. The neural networks have probably seen training data with babies crying - this is a common real-world scenario - so they should handle it reasonably well. But it's going to degrade accuracy compared to a quiet environment.

Corn

So Daniel's not going to get perfect transcription, but it should be better than nothing?

Herman

Right. And honestly, I think that's the realistic expectation for voice-to-text in less-than-ideal environments. You're not going to get transcription quality that's as good as if you had recorded in a quiet room, but you can get something that's useful and saves time.

Corn

Let me bring up something that occurred to me while you were talking. You mentioned that transcription models are trained on noisy audio, so they can handle some noise. Is there a scenario where you'd actually want less noise removal rather than more?

Herman

Yeah, I think so. If you're using a transcription model that's trained on a certain level of background noise, and then you apply very aggressive noise removal, you might be moving the audio away from what the model expects. It's like... you're making the audio cleaner, but also weirder from the transcription model's perspective.

Corn

That's a really interesting point. So there's an optimal level of noise removal that's not the maximum possible noise removal?

Herman

Exactly. And that's something I wish more people understood about this technology. The intuitive thing to do is "remove as much noise as possible," but that's not always the right approach.

Corn

Has there been research on finding that optimal level?

Herman

There's been some work on it, yeah. Some researchers have been training noise removal and transcription models jointly, so they learn what level of noise removal is actually optimal for downstream transcription. But it's not something that's widely implemented in consumer tools yet.

Corn

So most consumer apps are just trying to remove as much noise as possible?

Herman

I think they're trying to balance removing noise while not introducing too many artifacts. But yeah, they're not necessarily optimizing specifically for transcription accuracy. That's a more specialized approach.

Corn

Interesting. Alright, we've got Jim on the line. Jim, what's on your mind? Jim: Yeah, this is Jim from Ohio. Look, I've been listening to you two go on about all these fancy neural networks and cloud processing and all that, and I gotta say, you're way overthinking this. Back in my day, we just spoke clearly and made sure we weren't in a noisy environment. That's it. Also, we got a cold front moving through Ohio today, really knocked my heating bill up, but that's beside the point. But seriously, why can't people just wait until they're in a quiet place to record their messages?

Herman

Well, that's a fair point in theory, Jim, but in practice, sometimes you can't wait. If Daniel's holding a baby and he needs to send something urgent to his accountant, he might not have the luxury of waiting until later when it's quiet. Jim: Yeah, but... he could just type it. I mean, we've got keyboards. We've got our fingers. What happened to just typing?

Corn

I get what you're saying, but the whole point of voice-to-text is that it's faster when your hands are full. If you're literally holding a baby, you can't type. Jim: Ehh, I don't know. Seems like a solution in search of a problem to me. Also, my neighbor Gary has the same baby situation, and you know what he does? He just puts the baby down and types. Problem solved.

Herman

But that's not practical for everyone, Jim. Some people don't have a safe place to set the baby down. Some people need both hands. Jim: I'm just saying, maybe the real solution is to not try to do two things at once. My cat Whiskers has the right idea - when she's doing something, she focuses on that one thing. She doesn't try to write emails while eating. Well, she does try, but it's messy.

Corn

That's... actually not a bad life lesson, Jim. But I think for people who do need to multitask, this technology is genuinely useful. Jim: Fine, fine. I'm not saying it's not useful. I'm just saying maybe people are too reliant on technology to solve problems that wouldn't exist if they just planned better. But what do I know? I'm just a guy from Ohio with a cat and bad knees.

Corn

Thanks for calling in, Jim. We appreciate the perspective. Jim: Yeah, alright. Good luck with all that neural network stuff.

Herman

So, Jim's not wrong that planning ahead would solve some of these problems. But I think there's a real value in having tools that let you be productive in non-ideal circumstances.

Corn

Yeah, I mean, real life is messy. Babies cry. You're in coffee shops. You're in cars. The technology that works in those real-world scenarios is valuable, even if the ideal scenario is to be in a quiet room.

Herman

Exactly. And I think that's actually been a big driver of voice-to-text adoption - it lets people capture thoughts and communicate when they otherwise couldn't.

Corn

So, let's talk about practical takeaways. If someone's listening to this and they want to improve their voice-to-text experience in noisy environments, what should they actually do?

Herman

First, try the built-in noise cancellation on your device. Most modern phones have something. If that's not enough, try a dedicated app like AI Noise Remover or Voice Isolator. See what works for you.

Corn

And if you're using Android specifically, which the prompt mentioned?

Herman

Same advice. There are good third-party apps available. You might have to pay a few dollars, but it's worth it if you're regularly recording in noisy environments.

Corn

What about the post-processing angle? Is there a scenario where someone should be uploading their audio to a cloud service for processing?

Herman

If you have time and quality is critical, yeah. If you're recording something that needs to be perfect - like, a podcast episode or a professional voice memo - then post-processing with a more sophisticated tool makes sense. But for quick transcriptions that are good enough to send to your accountant, on-device real-time processing is probably sufficient.

Corn

And what about hardware microphones with built-in noise cancellation? Should people bother with those?

Herman

I'd say they're worth considering if you're regularly recording in very specific environments - like, if you're always recording in a car, a car-specific microphone might be worth it. But for general use, software solutions are more flexible and usually better.

Corn

So the takeaway is: try the software solutions first, see what works for your specific situation, and don't stress if it's not perfect?

Herman

Yeah, that's a good summary. And remember, the goal is usually just good enough, not perfect.

Corn

I think that's actually a really important mindset shift. A lot of people might think voice-to-text has to be perfect to be useful, but in reality, it just has to be better than the alternative.

Herman

Right. If you're choosing between dictating something and letting work pile up because you're too busy holding a baby, dictating with imperfect transcription is the clear winner.

Corn

Alright, so looking ahead, where do you think this technology is going? We've got real-time neural network processing on phones, we've got cloud-based solutions, what's next?

Herman

I think the next frontier is better integration between noise removal and transcription. Like, models that are specifically trained to work together. We'll probably also see better on-device processing as phone processors get more powerful. And there's interesting research into multi-modal approaches - using video or other sensors to help identify and remove noise.

Corn

Multi-modal? Like, the phone would know you're holding a baby and adjust its noise removal accordingly?

Herman

Exactly. Or the camera could detect that you're in a coffee shop and apply different noise removal parameters. That kind of contextual intelligence.

Corn

That's pretty cool. But that's probably a ways off?

Herman

It's being researched now, so maybe a few years before it becomes mainstream. But I think that's the direction things are heading.

Corn

Well, this has been really interesting. I came in thinking this was going to be a straightforward question about which approach is best, but it turns out it's much more nuanced than that.

Herman

Yeah, it's one of those areas where the technology is genuinely useful but also not a magic bullet. Context matters. Your specific situation matters. The device you're using matters.

Corn

Which is probably why Daniel sent in this prompt - because he was grappling with these trade-offs in real time.

Herman

Absolutely. And I think the fact that he's already thinking about this stuff, already using voice-to-text productively, means he's ahead of most people in terms of leveraging this technology.

Corn

Yeah, the people who are really going to benefit from better noise removal are the ones who are already trying to use voice-to-text in challenging situations.

Herman

Right. And as the technology improves, I think we'll see more and more people in that category. Voice-to-text is becoming genuinely practical for real-world use.

Corn

Alright, I think that's a good place to wrap up. Thanks to our producer for this fascinating prompt - it's given us a lot to think about. And thanks to Herman for diving deep into the technical details.

Herman

Happy to be here. And thanks to Jim for keeping us grounded with his skepticism.

Corn

That's My Weird Prompts. You can find us on Spotify and wherever you get your podcasts. We'll be back next week with another prompt. Until then, try not to overthink your noise removal settings.

Herman

Or do overthink them. Whatever works for you.

Corn

Fair enough. See you next time.