Podcast Transcript

EPISODE #136

# The Ghost in the Machine: Why AI Voices Hallucinate

Published January 02, 2026 • Runtime: 24:00

https://myweirdprompts.com/episode/ai-voice-hallucination-science/

## EPISODE SYNOPSIS

Have you ever been startled by a text-to-speech voice that suddenly breaks into an aggressive shout or a creepy, rhythmic whisper? In this episode of My Weird Prompts, hosts Herman and Corn explore the fascinating and occasionally terrifying world of audio hallucinations in modern AI models like Chatterbox Turbo. They break down the complex mechanics of autoregressive models, explaining how tiny mathematical errors can spiral into feedback loops of silence or distortion. From the "thin rails" of compressed mobile models to the mystery of "latent space drift" where voices switch identities mid-sentence, this episode offers a deep dive into the acoustic breakdowns that happen when AI loses its way. Whether you're a developer working with zero-shot voice cloning or just a listener confused by a "haunted" podcast, you'll gain a new understanding of the science behind the glitches. Join the Poppleberry brothers as they pull back the curtain on the latent space and explain why your AI might be having an emotional breakdown.

## DANIEL'S PROMPT

**Daniel**

We have been trying different text-to-speech models for this podcast, including Chatterbox Turbo and zero-shot voice clones for Herman Popelberry and Corn. We've noticed that some smaller models have significant hallucinations, such as protracted silence, distorted voices, aggressive shouting, or random voices appearing. Why do these hallucinations happen in text-to-speech models? Is it simply that smaller parameter models are more susceptible, or are there other factors that cause some models to struggle while others don't?

# TRANSCRIPT

**Corn**

Welcome to My Weird Prompts! I am Corn, and I am here in our home in Jerusalem with my brother.

**Herman**

Herman Poppleberry, at your service. It is good to be here, Corn. We have a very meta topic today, do we not?

**Corn**

We really do. Our housemate Daniel sent us an audio prompt about the very technology that makes this podcast possible. He has been experimenting with different text to speech models for the show, and he is noticing some, well, let us call them colorful glitches.

**Herman**

It is fascinating. He mentioned Chatterbox Turbo, which is the latest iteration of the Chatterbox model we have been using. It is faster, it is more efficient, but as Daniel pointed out, it seems to have a bit of a temper. Or sometimes it just decides to stop talking altogether.

**Corn**

Right, he was describing these hallucinations where the voice becomes distorted, or starts shouting, or even worse, starts whispering like a certain dark lord from a galaxy far, far away. And then there is the issue of random voices appearing out of nowhere. It is like the AI is haunted.

**Herman**

It certainly feels that way when you are listening to it. But today, we are going to pull back the curtain on why this happens. It is not just about the size of the model, although that is a big part of it. It is about how these models actually construct speech and what happens when they lose their way in the latent space.

**Corn**

I am really curious about that. Because we usually think of hallucinations in terms of large language models making up facts or lying about history. But when a voice model hallucinates, it is a physical, or at least an acoustic, breakdown. So, Herman, where do we even start? Why does a model that is supposed to be reading a script suddenly decide to start shouting at the listener?

**Herman**

To understand that, we have to look at the architecture. Most of the cutting edge text to speech models we use today, including the ones Daniel mentioned like Chatterbox Turbo, are what we call autoregressive models. This means they generate audio one piece at a time, and each new piece is based on all the pieces that came before it.

**Corn**

Okay, so it is like a chain. If I say a word, the model looks at that word and predicts what the next sound should be based on the patterns it learned during training.

**Herman**

Exactly. It is predicting the next audio token. Now, in a perfect world, that chain stays strong. But imagine if the model makes a very slight error. Maybe it predicts a sound that is just a tiny bit off. Because the next prediction is based on that error, the error can compound. It is like a feedback loop.

**Corn**

So if it starts to drift toward a certain tone or volume, and then it uses that drift as the basis for the next sound, it just keeps going in that direction?

**Herman**

Precisely. This is often where the aggressive shouting comes from. The model might predict a slightly higher energy state for a syllable. Then, seeing that high energy, it thinks, oh, we must be in a shouting context now, and it doubles down. Before you know it, the model is trapped in a state where every subsequent token it predicts is louder and more aggressive than the last. It is a mathematical spiral.

**Corn**

That is wild. It is almost like it gets caught in an emotional loop that it cannot escape because its only point of reference is the mistake it just made. What about the silence, though? Daniel mentioned protracted silence. If it is always predicting the next thing, why would it predict nothing?

**Herman**

Silence is a very interesting edge case. In the training data, silence usually happens at the end of a sentence or during a natural pause. If a model gets confused, or if the input text is ambiguous, it might predict a silence token. The problem is that once it is in a state of silence, the most likely thing to follow silence, statistically, is often more silence.

**Corn**

Oh, I see. It is the same feedback loop, but in reverse. It predicts a pause, and then it looks back and says, well, I have been silent for ten milliseconds, so I should probably be silent for another ten.

**Herman**

Right. It gets stuck in a local minimum. It cannot find the path back to speech because the probability of jumping from silence back into a complex vocal sound is lower than just staying quiet. It is essentially waiting for a nudge that never comes.

**Corn**

This makes me wonder about the parameter size question Daniel raised. He noticed that smaller models seem more susceptible to this. Is it just that they are less smart, or is there something about the way they are compressed that makes them more unstable?

**Herman**

It is a bit of both. Think of the parameters as the model's memory and its nuanced understanding of the world. A massive model with billions of parameters has seen so many examples of human speech that it has a very strong sense of what is normal. It has a high degree of what we call robustness.

**Corn**

So it has a stronger gravitational pull toward normal speech?

**Herman**

That is a great way to put it. A large model can handle a weird prompt or a strange word because it has enough context to stay on the rails. A smaller model, like a Turbo version or a mobile optimized version, has had its parameters pruned or its weights quantized to make it faster and smaller. This means its internal map of speech is less detailed.

**Corn**

So the rails are thinner.

**Herman**

Exactly. The rails are thinner and the gaps between them are wider. When a small model hits a word it does not quite understand, or a sequence of letters it has not seen often, it is much more likely to fall off those rails. And once it falls off, it does not have the depth of understanding to find its way back. It just starts guessing, and that is when you get the Darth Vader whispers or the distorted robotic buzzing.

**Corn**

It is interesting that you mention the Darth Vader whispers. That specific hallucination feels very specific. Why would a model default to a low, breathy whisper instead of just white noise?

**Herman**

That usually happens because of how the model handles breath sounds. In human speech, we have these non voiced sounds, like the letter S or the sound of a breath. These are technically just shaped noise. If a model loses its ability to generate clear tonal vowels, it might fall back on these noise based sounds because they are easier to approximate.

**Corn**

So it is a fallback mechanism?

**Herman**

In a way. It is trying to fulfill the requirement of generating audio, but it can no longer reconstruct the complex harmonics of a human voice. So it generates this grainy, breathy texture. And because it is still trying to follow the rhythm of the text, it sounds like a rhythmic, creepy whisper.

**Corn**

That is actually quite terrifying when you think about it. It is the ghost of the speech it is trying to make. I want to dig more into the random voices too, because that seems like a different kind of error. But before we get into the mystery of the phantom voices, we should probably hear from our sponsors.

**Herman**

Good idea. Let us take a quick break.

**Corn**

We will be right back. Larry: Are you tired of your neighbors having better grass than you? Are you sick of looking at your lawn and seeing boring, green blades of nothingness? Introducing Glow-Turf. The world's first bioluminescent, synthetic-organic hybrid ground cover. Glow-Turf does not just sit there. It glows with a vibrant, pulsating neon purple hue that can be seen from low earth orbit. It is made from a proprietary blend of recycled fiber optics and deep-sea jellyfish proteins. It requires no water, no sunlight, and only the occasional sacrifice of a small household appliance to maintain its luster. Warning, Glow-Turf may cause temporary night blindness, mild radioactivity in local pets, and an inexplicable desire to start a cult. But your yard will be the talk of the town, literally. People will be talking about it for miles. Glow-Turf. Lighting up the neighborhood whether they like it or not. Larry: BUY NOW!

**Corn**

Alright, thanks Larry. I think I will stick to regular grass for now.

**Herman**

I do not know, Corn. A purple lawn could be a real conversation starter.

**Corn**

A conversation with the local health department, maybe. Anyway, back to text to speech hallucinations. Before the break, we were talking about why models shout or go silent. But Daniel mentioned something even weirder, random voices appearing. Voices that are not the ones the model is supposed to be using. How does a model that is trained to sound like you or me suddenly sound like a completely different person in the middle of a sentence?

**Herman**

This is one of the most fascinating aspects of zero shot voice cloning. When we use a model like Chatterbox, we are giving it a small sample of a voice, what we call an embedding. This embedding is like a set of coordinates in a massive multidimensional space of all possible human voices.

**Corn**

Right, so my voice is at one coordinate, and your voice, the Herman Poppleberry voice, is at another.

**Herman**

Exactly. Now, these models are trained on thousands of different people. When the model is working correctly, it stays locked onto your coordinates. But when it encounters a difficult piece of text, or if the model is too small to maintain a stable state, it can literally drift in that coordinate space.

**Corn**

So it just wanders into someone else's neighborhood?

**Herman**

Yes! It might drift toward a more feminine voice, or a voice with a different accent, or just a generic voice that was very common in its training data. Because the model is trying to find the most probable next sound, if it cannot find a high probability sound in your voice, it might find one in a different voice and just jump over there.

**Corn**

That explains why it feels so jarring. It is not just a glitch in the audio quality, it is a glitch in the identity of the speaker.

**Herman**

It really is. And in smaller models, this is much more common because the boundaries between different voices in that internal space are not as well defined. Everything is more blurred together. It is like a map where the cities are all bleeding into each other.

**Corn**

Does the context window play a role here? I know with large language models, if the conversation gets too long, they start to lose the thread. Does a text to speech model have a similar limit where it just forgets who it is supposed to be?

**Herman**

It absolutely does. Most of these models process audio in chunks. If the chunking strategy is not perfect, or if the model is being asked to generate a very long, continuous stream of speech without a reset, the internal state can degrade. In the industry, we call this state drift. The model's memory of the voice it started with begins to fade, and it starts to revert to its baseline average.

**Corn**

Which is why the voices sometimes become more robotic or generic as the sentence goes on?

**Herman**

Exactly. It is losing the fine details that make the voice unique. Those details are the hardest things for the model to keep track of. It is much easier for it to just produce a generic, mid range human sound than to maintain the specific rasp or pitch of a cloned voice.

**Corn**

I want to talk about the zero shot aspect specifically. Daniel mentioned that he uses a single sample of a voice to clone it. That seems like a lot of pressure on the model. Does that contribute to the hallucinations?

**Herman**

Immensely. Zero shot means the model has never seen that specific voice before. It is trying to generalize based on a few seconds of audio. This is incredibly difficult. It is like showing a painter a single polaroid of a person and asking them to paint a full length portrait of them in a hundred different poses.

**Corn**

And if the polaroid is a bit blurry, or if the person is making a weird face, the painter has to guess.

**Herman**

Right. If the input sample has background noise, or if the person is speaking with a specific emotion, the model might bake that into its understanding of the voice. If the sample has a bit of an echo, the model might think that the echo is part of the voice itself. Then, when it tries to generate new speech, it might hallucinate that echo in weird ways, leading to that distorted, robotic sound Daniel described.

**Corn**

So the quality of the prompt is just as important as the model itself.

**Herman**

Even more so, in many cases. But there is another factor at play here that we have not touched on yet, and that is the difference between autoregressive models and diffusion models.

**Corn**

Oh, I have heard about diffusion in the context of image generators like Midjourney. Are they using that for voice now too?

**Herman**

They are. And this is where the industry is heading in twenty twenty-six. Diffusion models work differently. Instead of predicting the next piece of audio in a chain, they start with a block of random noise and gradually refine it into a clear signal, guided by the text.

**Corn**

That sounds much more stable.

**Herman**

It is! Because it is not a chain, it does not suffer from those compounding feedback loops. If it makes a mistake in one part of the audio, it does not necessarily ruin the rest of it. You do not get that spiraling shouting or the endless silence as often.

**Corn**

So why are we still using the autoregressive models like Chatterbox Turbo?

**Herman**

Speed. Autoregressive models are incredibly fast. They can generate audio almost as quickly as you can read the text. Diffusion models are much more computationally expensive. They take longer to refine that noise into speech. For a podcast like ours, where we have a lot of dialogue, speed is a huge factor.

**Corn**

So Chatterbox Turbo is a trade off. We get the speed, but we have to deal with the occasional Darth Vader whisper.

**Herman**

That is the current state of the art. We are in this transition period where we are trying to make these models smaller and faster, but we are sacrificing the stability that comes with larger parameters or more complex architectures.

**Corn**

It is funny, because we are talking about this as a technical problem, but it has a real impact on the listener's experience. If a voice suddenly starts shouting, it breaks the immersion. It reminds you that you are listening to a machine.

**Herman**

It really does. But there is a flip side to that. Some people actually find these hallucinations interesting. They are like a new form of digital art. There are entire communities online dedicated to finding and sharing the weirdest AI glitches. It is like looking at the brushstrokes in a painting. It shows you how the thing was made.

**Corn**

I can see that. It is the uncanny valley, but instead of being creepy, it is almost experimental. Still, for our purposes, we probably want to keep the shouting to a minimum.

**Herman**

Agreed. And there are ways to mitigate it. Daniel was asking if it is just about parameter size, and the answer is that better training data and better prompting can help a lot. If you give the model a very clean, dry voice sample, and if you use a model that has been trained on a more diverse range of speech, you can reduce the frequency of these hallucinations even in smaller models.

**Corn**

What about the temperature settings? I know with text models, you can turn down the temperature to make them more predictable. Does that work for voice too?

**Herman**

It does. Temperature in a voice model controls how much risk the model takes when predicting the next sound. If the temperature is high, the model is more likely to pick a less probable sound, which can lead to more expressive, natural speech, but also more hallucinations.

**Corn**

So it is a balance between being boring and being crazy.

**Herman**

Exactly. If you turn the temperature all the way down, the voice becomes very flat and robotic, but it is very stable. If you turn it up, it sounds more human, until it suddenly decides to scream at you. Most producers spend a lot of time finding that sweet spot.

**Corn**

It sounds like a lot of trial and error. Which brings us back to Daniel's work on this podcast. He is essentially the one in the laboratory, tweaking the dials and trying to keep us from losing our minds.

**Herman**

He is the unsung hero of the show. Every time you hear me speak without sounding like a distorted robot, you can thank Daniel's fine tuning of the parameters.

**Corn**

Well, I for one am glad I do not sound like a Sith Lord. At least not today.

**Herman**

Give it time, Corn. The day is young.

**Corn**

So, looking ahead, what do you think the next year holds? We are in January of twenty twenty-six. By this time next year, will we have solved the hallucination problem?

**Herman**

I think we will see a shift toward hybrid models. We will use small, fast autoregressive models to generate a rough draft of the speech, and then a very small, very fast diffusion model to clean it up and remove the artifacts. This could give us the best of both worlds, the speed of Chatterbox Turbo with the stability of a much larger model.

**Corn**

That would be a game changer. It would make high quality voice cloning accessible to everyone, not just people with massive server racks.

**Herman**

It is already happening. We are seeing models now that can run locally on a high end smartphone and produce speech that is almost indistinguishable from a human. The hallucinations are getting rarer, but they are also getting more sophisticated.

**Corn**

What do you mean by that?

**Herman**

Well, instead of just shouting or going silent, a sophisticated hallucination might be a model adding a laugh where it was not scripted, or changing its tone to be sarcastic because it misinterpreted the subtext of the sentence.

**Corn**

Wait, is that a hallucination or is that just the AI getting better at acting?

**Herman**

That is the big question! When the error makes the speech sound more human, do we call it a bug or a feature? If the model decides to take a breath because it thinks the sentence is too long, but there was no breath in the script, is that a failure or a success?

**Corn**

I guess it depends on whether you wanted that breath there. It is the difference between a tool and a collaborator.

**Herman**

Exactly. And I think that is where we are heading. We are moving from text to speech being a simple conversion tool to it being a creative partner that brings its own interpretation to the text. Which, of course, means we will have even more to talk about on this show.

**Corn**

I am looking forward to it. Even if you do occasionally shout at me in the middle of a recording.

**Herman**

I promise to keep the shouting to a minimum, brother. Unless the prompt calls for it.

**Corn**

Fair enough. Well, I think we have covered a lot of ground today. We have talked about the feedback loops of autoregressive models, the dangers of small parameter sizes, the mystery of the drifting voice coordinates, and the future of diffusion.

**Herman**

It has been a deep dive, for sure. And I think the big takeaway for Daniel and for our listeners is that these glitches are not just random noise. They are a window into how these incredible machines actually think and process information.

**Corn**

It is a reminder that even the most advanced artificial intelligence is still, at its heart, a series of probabilities and predictions. And sometimes, those predictions just go a little bit sideways.

**Herman**

And that is what makes it interesting. If it were perfect, we would not have anything to talk about.

**Corn**

Very true. Well, I think that is a good place to wrap things up for today.

**Herman**

I agree. It has been a pleasure as always, Corn.

**Corn**

Likewise, Herman. And a big thank you to Daniel for sending in such a thought provoking prompt. It is always fun to talk about the tech that actually makes this whole thing possible.

**Herman**

Absolutely. It keeps us on our toes.

**Corn**

If you enjoyed this episode, you can find more of our discussions on Spotify. We have a whole library of episodes exploring everything from the future of cities to the ethics of artificial intelligence.

**Herman**

And do not forget to visit our website at myweirdprompts.com. We have an RSS feed for subscribers so you never miss an episode, and there is a contact form if you want to send us a prompt of your own. We love hearing from you.

**Corn**

We really do. This has been My Weird Prompts. I am Corn.

**Herman**

And I am Herman Poppleberry.

**Corn**

Thanks for listening, everyone. We will see you next time.

**Herman**

Goodbye for now!