Podcast Transcript

EPISODE #142

Breaking the Voice Wall: The Future of Native Speech AI

Published January 03, 2026 • Runtime: 29:05

https://myweirdprompts.com/episode/native-speech-to-speech-evolution/

EPISODE SYNOPSIS

In this episode, Herman and Corn dive deep into the technical and economic hurdles of real-time conversational AI. They explore why current voice assistants often feel like "confused walls" and how the transition from traditional text-based pipelines to native speech-to-speech models is fundamentally changing the user experience. From the staggering computational costs of processing raw audio tokens to the intricate social intelligence required for "turn detection," the brothers discuss whether voice interfaces can truly replace the keyboard in the modern workforce. Learn about the rise of semantic voice activity detection, the importance of prosody, and how edge computing might finally make natural human-AI dialogue a viable reality for businesses and individuals alike.

DANIEL'S PROMPT

Daniel

"I'd like to talk about real-time conversational AI through audio and the experience of speech-to-speech interaction. I believe voice has the potential to replace keyboards in the workforce, but there are significant hurdles, particularly cost and turn detection. Real-time speech-to-speech is currently much more expensive than using a speech-to-text and text-to-speech pipeline. Additionally, the model needs to accurately detect when someone is finished speaking or when to interrupt to make the conversation feel natural rather than jarring. I'm curious to hear your thoughts on how turn detection works and how advancements like Voice Activity Detection (VAD) might make these interfaces more affordable and enjoyable."

TRANSCRIPT

Corn

Hey everyone, welcome back to My Weird Prompts. We are coming to you as always from our home in Jerusalem, and I have to say, the energy in the room is high today. I am Corn, and sitting across from me is my brother.

Herman

Herman Poppleberry, ready to dive into the deep end. It is great to be here.

Corn

It really is. Our housemate Daniel sent us a voice memo this morning that actually hit on something I have been struggling with lately. I was trying to use a voice assistant to help me organize some research notes while I was cooking dinner, and the experience was just, well, it was jarring. I felt like I was talking to a very polite but very confused wall.

Herman

The classic voice assistant wall. It is a frustrating place to be, especially in two thousand twenty-six, when we feel like we should be further along.

Corn

Exactly. Daniel was asking about the state of real-time conversational AI through audio. He is thinking about whether voice could actually replace keyboards in the workforce, but he pointed out two huge hurdles that are keeping us from that future: the massive cost difference between traditional pipelines and native speech-to-speech, and the incredibly tricky problem of turn detection.

Herman

Those are the two big ones. It is the difference between a tool that feels like a gadget and a tool that feels like a colleague. If we are going to ditch the keyboard, we need something that understands the rhythm of human speech, not just the words we are saying.

Corn

Right, and that rhythm is so much more complex than we realize until it is missing. We take it for granted that when I pause to think, you know I am not done, or if you interrupt me with a quick mm-hmm, I know I can keep going. But for an AI, that is a massive computational challenge.

Herman

It really is. And it is not just about the code. It is about the physical reality of how these models are built and served. I have been looking into the latest benchmarks for native speech-to-speech models, and the architecture is fundamentally different from what we were using even two years ago.

Corn

Well, let us start there then. Let us talk about why it is so much more expensive to do this right. Most people probably do not realize that when they talk to an AI, there is usually a whole assembly line of different models working behind the scenes.

Herman

That is the traditional pipeline approach. You have one model for speech-to-text, which is your transcriber. Then that text goes to the Large Language Model, the brain, which generates a text response. Finally, that text goes to a text-to-speech model, the voice. It is like a game of telephone played at light speed.

Corn

And each of those steps adds what we call latency, right? The time it takes for the signal to travel through all those hoops.

Herman

Exactly. Even if each step takes only a few hundred milliseconds, by the time you add them up, you have a two-second delay. In human conversation, a two-second delay feels like an eternity. It makes the interaction feel like a walkie-talkie conversation rather than a natural chat.

Corn

And Daniel mentioned that this pipeline is actually cheaper than the new native speech-to-speech models. Why is that? You would think one model doing everything would be more efficient than three separate ones.

Herman

You would think so, but the computational intensity of a native multimodal model is staggering. In a pipeline, the Large Language Model is only processing text tokens. Text is very light. It is easy to move around and cheap to process. But in a native speech-to-speech model, the AI is processing raw audio tokens.

Corn

So it is not just looking at the words, it is looking at the sound waves themselves?

Herman

Precisely. It is looking at the pitch, the tone, the speed, and the background noise all at once. An audio file has thousands of times more data than a text file of the same length. When you ask a model to reason across that much data in real time, the amount of processing power required per second of conversation goes through the roof.

Corn

So we are talking about a massive increase in the number of calculations per second.

Herman

We are. In some cases, running a truly native speech-to-speech model can be ten to twenty times more expensive than the old pipeline method. That is a hard pill for a company to swallow if they are trying to roll this out to thousands of employees.

Corn

That makes sense. If you are a business looking at your bottom line, paying twenty times more just so your AI sounds a bit more natural might not seem worth it. But I wonder if we are missing the second-order effects there. If a voice interface is natural enough that people actually stop using keyboards, the productivity gains might far outweigh the API costs.

Herman

I agree with you there. But that brings us to the second hurdle Daniel mentioned: turn detection. This is where the user experience really lives or dies.

Corn

This is what I was dealing with in the kitchen. I would say something like, let me see, I need to find the notes on... and then I would pause for one second to remember where I saved them, and the AI would immediately jump in and say, I am sorry, I do not understand what you want to find. It is like talking to someone who is constantly finishing your sentences incorrectly.

Herman

It is the silence problem. Traditionally, these systems use something called Voice Activity Detection, or VAD. It is a relatively simple algorithm that just looks for a lack of sound. If it detects silence for more than, say, five hundred milliseconds, it assumes you are done and sends the audio to the model.

Corn

But humans use silence for so many things. We use it for emphasis, for thinking, for dramatic effect. A simple silence detector cannot tell the difference between a finished thought and a mid-sentence pause.

Herman

Right. And then you have the opposite problem: barge-in. If the AI is talking and you want to correct it, the system has to be able to hear you over its own voice and decide if your interruption is meaningful. It is a very delicate dance.

Corn

It really is. I remember we touched on some of the early versions of this back in episode two hundred twenty-seven when we were talking about noise reduction. But this goes so much deeper than just filtering out the hum of an air conditioner. This is about social intelligence.

Herman

It really is. It is about understanding the intent behind the sound. And that is where the latest developments in VAD are getting really interesting. We are moving away from simple volume-based detection toward what people are calling Semantic Voice Activity Detection.

Corn

Semantic detection. So the system is actually listening to the meaning of the words to decide if the turn is over?

Herman

Exactly. It is looking at the grammar and the prosody. If you say, I think we should go to the... the model knows that is an incomplete sentence. It knows the probability that you are done speaking is very low, even if you stay silent for a full second. On the other hand, if you say, That is all for now, it knows the turn is definitely over.

Corn

That sounds like it would require even more processing power, though. Now you have a mini-model running constantly just to decide when to let the big model talk.

Herman

It does, but the efficiency gains elsewhere are starting to make it viable. But before we get deeper into the mechanics of turn detection and how this might actually change the workforce, we should probably take a quick break.

Corn

Good idea. Let us take a quick break for our sponsors. Larry: Are you tired of your thoughts being trapped inside your own head? Do you wish you could express your innermost desires without the hassle of moving your jaw or using your vocal cords? Introducing the Mind-Link Muzzle from Thought-Tech Industries. Our revolutionary, non-invasive head strap uses patented bio-static sensors to translate your brain's sub-vocalizations directly into a high-quality, robotic monotone. Perfect for long commutes, awkward family dinners, or when you just don't feel like being a human being. The Mind-Link Muzzle: because talking is so last year. Side effects may include mild scalp tingling, temporary loss of short-term memory, and an inexplicable craving for copper wiring. Larry: BUY NOW!

Corn

Alright, thanks Larry. I think I will stick to my own vocal cords for now, but I appreciate the hustle.

Herman

Always a unique pitch from Larry. Anyway, going back to what we were talking about before the break, we were looking at the transition from simple silence detection to these more advanced, semantic systems.

Corn

Right. And you mentioned prosody. For those who might not be familiar with the term, we are talking about the rhythm, stress, and intonation of speech. Like how my voice goes up at the end of a question?

Herman

Exactly. Humans use prosody to signal all sorts of things. We have a certain way of trailing off when we are inviting someone else to speak, and a different way of pausing when we are just catching our breath. Native speech-to-speech models are finally starting to pick up on those cues.

Corn

So, in two thousand twenty-six, are we seeing models that can actually handle backchanneling? You know, the little mhmms and yeahs that we do to show we are listening?

Herman

We are starting to. That is one of the most exciting things about these native models. Because they are trained on actual audio of people talking to each other, they are learning that a quick uh-huh from the listener does not mean the speaker has stopped. It is a signal to continue.

Corn

That feels like a huge breakthrough for making these things feel enjoyable to use. If I can give the AI a little verbal nod without it stopping its entire train of thought to ask me what I meant, that is a game-changer.

Herman

It really is. But let us look at the cost angle again, because Daniel was right to point that out as a major hurdle. Even with semantic VAD making things feel better, the bill at the end of the month for a company using these native models is still much higher. How do we close that gap?

Corn

Well, one way is through edge computing, right? If we can move some of that turn detection and initial audio processing onto the user's device instead of doing it all in the cloud, we could save a lot on data transfer and server costs.

Herman

That is a big part of the solution. We are seeing specialized chips now that are designed specifically to run these low-latency audio models. If your phone or your laptop can handle the VAD and the first layer of the speech model locally, you only need to hit the expensive cloud servers for the heavy-duty reasoning.

Corn

It is like having a receptionist at the front desk who handles all the basic interactions and only calls the CEO when something complex comes up.

Herman

That is a perfect analogy. And as those edge models get better, the cost of the overall system drops. We are also seeing new techniques in model quantization and distillation. Basically, taking these massive models and shrinking them down so they can run faster and cheaper without losing too much of that social intelligence.

Corn

I wonder, though, even if we solve the cost and the turn detection, is voice really going to replace the keyboard in the workforce? I mean, I love the idea of being able to just talk to my computer, but think about an open-plan office. If everyone is talking to their AI assistants at once, it is going to be total chaos.

Herman

That is a very practical concern. We talked about the vanishing air gap in episode two hundred forty-six, but there is also a literal air gap problem here. Sound travels. But I think we are seeing some interesting hardware solutions for that. Directional microphones are getting incredibly good at isolating a single voice even in a noisy room.

Corn

And there is also the privacy aspect. If I am working on a sensitive legal document or a medical report, I might not want to be dictating it out loud where my coworkers can hear me.

Herman

True. But think about the types of work where keyboards are actually a hindrance. Think about surgeons, or mechanics, or people working on assembly lines. For them, a truly responsive, real-time voice AI is not just a luxury, it is a massive safety and efficiency upgrade.

Corn

That is a great point. We often think about the workforce as people sitting at desks, but a huge portion of the economy involves people whose hands are busy. If a mechanic can ask an AI for the torque specs on a specific bolt without having to wipe their hands and walk over to a terminal, that is a huge win.

Herman

Exactly. And even for desk workers, I think we underestimate how much the keyboard limits our thinking. We can speak about one hundred fifty words per minute, but most people only type at about forty or fifty. There is a bottleneck between our thoughts and the screen.

Corn

I definitely feel that. Sometimes I have a complex idea and by the time I have typed out the first sentence, the second half of the idea has started to fade. If I could just talk it out, and the AI could help me structure it in real time, that would be incredibly powerful.

Herman

And that brings us back to Daniel's point about the experience being natural rather than jarring. For that kind of brainstorming to work, the AI has to be a partner. It has to know when to chime in with a clarifying question and when to just let you ramble.

Corn

So, let us talk about the state of the art in turn detection right now, in early two thousand twenty-six. What is the most advanced approach we are seeing?

Herman

The most advanced systems are using what is called Continuous Multimodal Integration. Instead of waiting for a pause, the model is constantly updating its understanding of the conversation every few milliseconds. It is looking at the audio stream, but it is also looking at the history of the conversation and even the visual cues if a camera is involved.

Corn

Visual cues? So it is looking at my face to see if I look like I am about to say something?

Herman

Yes. Things like your breath intake, your lip movements, and even your gaze. If you are looking at the screen and you take a quick breath, the AI knows you are probably about to speak, and it can preemptively pause its own output. It makes the transition feel almost instantaneous.

Corn

That is fascinating. It is moving from a reactive system to a predictive one. Instead of waiting for me to stop, it is predicting when I will stop.

Herman

Exactly. And that prediction is what eliminates the perceived latency. If the AI can start preparing its response before you have even finished your sentence, it can respond the moment you are done. That is how you get that sub-two-hundred-millisecond response time that makes a conversation feel real.

Corn

But that sounds like it could go wrong in some pretty funny ways. If the AI predicts I am going to finish a sentence one way, but I take a sharp turn at the last second, does it have to throw away everything it prepared?

Herman

It does. And that is where the cost comes in again. You are essentially paying for the AI to think about multiple possible futures at once. It is a bit like how modern processors use speculative execution to speed things up. They guess which path a program will take and start working on it ahead of time. If they guess wrong, they just discard the work and start over.

Corn

So we are doing speculative execution for human conversation. That is a wild thought.

Herman

It really is. And it is only possible because we have the raw horsepower now to do it. But back to Daniel's question about affordability. As we get better at these predictions, we can actually be more efficient with how we use our compute. We can focus our processing power on the most likely paths.

Corn

I am curious about the psychological impact of this. If we start interacting with AI that is this good at turn detection, are we going to start expecting the same level of perfect timing from the humans in our lives?

Herman

That is a deep question, Corn. We already see people getting frustrated when their friends do not reply to texts fast enough. If we get used to an AI that never interrupts us inappropriately and always knows exactly when to speak, real humans might start to seem a bit, well, clunky.

Corn

I can see it now. You will be at dinner with someone and they will pause to take a bite of food, and you will find yourself getting annoyed that they did not signal their turn was not over yet.

Herman

We might need to start wearing little lights that turn red when we are thinking and green when we are ready for a response.

Corn

Do not give Larry any ideas for his next ad.

Herman

Good point. But seriously, the potential for this in the workforce is huge. Think about meetings. If you have an AI participant that can actually follow the flow of a multi-person conversation, it can act as the perfect scribe and facilitator. It can notice when someone is trying to speak but getting talked over and find a gap for them.

Corn

That would be a massive improvement for remote work. One of the hardest things about video calls is that natural turn-taking is so much harder when you have even a tiny bit of lag. If an AI could act as a sort of traffic controller for the conversation, it would make those meetings so much more productive.

Herman

It really would. But again, it all comes back to that turn detection. In a group setting, the problem gets exponentially harder. Now the AI has to track multiple voices, understand who is talking to whom, and decide if a particular comment needs a response or if it was just a side remark.

Corn

It is like the cocktail party effect for AI. Being able to focus on one voice in a crowded room while still being aware of the overall environment.

Herman

Exactly. And that is where the multimodal aspect is so important. If the AI can see who is looking at whom, it has a much better chance of understanding the dynamics of the room.

Corn

So, if we look ahead to the rest of two thousand twenty-six and into twenty-seven, where do you see the biggest breakthroughs coming from? Is it going to be better algorithms, or just cheaper hardware?

Herman

I think it is going to be a combination of both, but the real breakthrough will be in what I call context-aware latency.

Corn

Context-aware latency? What do you mean by that?

Herman

Right now, we try to make every response as fast as possible. But not every human response is fast. If you ask me a really difficult philosophical question, and I respond in two hundred milliseconds, you are going to think I did not really consider it. You expect a pause there.

Corn

That is true. A fast response can sometimes feel dismissive or shallow.

Herman

Exactly. So the next generation of these models will actually be able to vary their latency based on the complexity of the task. They will use that time to do more deep thinking, and they will signal that they are thinking through backchanneling or even just a thoughtful hmmm.

Corn

So the AI will be pretending to be slower to seem more human?

Herman

It is not just about pretending. It is about using the time effectively. Instead of a shallow, fast response, it takes an extra second to generate a much deeper, more nuanced one, but it uses audio cues to keep the connection alive during that second. It is about managing the user's expectations of the turn.

Corn

That is a really sophisticated way of looking at it. It is not just about speed; it is about the quality of the interaction.

Herman

Exactly. And that is how we get to the point where voice truly can replace the keyboard. When the interaction is so smooth and so productive that the keyboard feels like a step backward, like using a typewriter.

Corn

I can see that. I mean, I already use voice for a lot of my initial drafting, but I always have to go back and clean it up with a keyboard because the AI didn't quite catch my meaning or it cut me off. If we can eliminate that cleanup phase, the productivity would be off the charts.

Herman

It really would. And for people with disabilities or repetitive strain injuries, this isn't just about productivity. It's about accessibility. It's about being able to participate in the workforce on an equal footing.

Corn

That is a great point. We have talked about accessibility in the past, but the move to high-fidelity, real-time voice interfaces is probably the biggest leap forward for accessibility since the invention of the screen reader.

Herman

I agree. It is a very exciting time to be following this. Daniel's prompt really hit on the core of why this feels like it's just out of reach but also so close. The hurdles are real, but the solutions are being built right now.

Corn

So, to wrap up the technical side, we are looking at a move toward native multimodal models that reason across raw audio tokens, the implementation of semantic VAD to handle complex turn-taking, and the use of edge computing to bring down those massive API costs.

Herman

That is a great summary. And don't forget the predictive turn-taking. That is the secret sauce that makes the latency disappear.

Corn

Right, the speculative execution for conversation. I am still wrapping my head around that one.

Herman

It is wild, isn't it? But it is exactly what our brains are doing all the time. We are constantly predicting what the person across from us is going to say next. We are just teaching the machines to do the same thing.

Corn

Well, I think we have covered a lot of ground today. From the cost of audio tokens to the social etiquette of AI interruptions. It is clear that the transition from keyboard to voice is a lot more complicated than just adding a microphone to a chatbot.

Herman

It really is. It is about recreating the most complex thing humans do, which is talk to each other.

Corn

Well, I for one am looking forward to the day when I can cook dinner and organize my research notes without my AI assistant getting impatient with me.

Herman

We will get there, Corn. Probably sooner than you think.

Corn

I hope so. And hey, before we sign off, I want to say a huge thank you to everyone who has been listening. We have been doing this for two hundred forty-nine episodes now, and the community that has grown around My Weird Prompts is just incredible.

Herman

It really is. We love hearing from you all. And if you have a second, we would really appreciate it if you could leave us a review on your podcast app or on Spotify. It genuinely helps other people find the show and allows us to keep digging into these weird and wonderful topics.

Corn

Yeah, a quick rating or a few words about what you enjoy really makes a difference. And remember, you can always find our full archive and more information at myweirdprompts.com. We have an RSS feed there for subscribers and a contact form if you want to send us your own weird prompts.

Herman

Or just to say hi. We like that too.

Corn

Definitely. And thanks again to our housemate Daniel for sending in this prompt. It was a great one to dive into.

Herman

Absolutely. It gave me an excuse to read those new white papers I have been eyeing.

Corn

I knew you had them tucked away somewhere. Alright, I think that is it for today. This has been My Weird Prompts.

Herman

I am Herman Poppleberry.

Corn

And I am Corn. We will see you next time.

Herman

Until then, keep asking the weird questions.

Corn

Bye everyone.

Herman

Take care.