Episode Book

EPISODE #6296

Why Speech Recognizers Cut Off Hesitant Speakers

PUBLISHED	27 October 2026 at 12:27 UTC
DURATION	29:11
LISTEN	myweirdprompts.com/episode/speech-recognition-pauses-turn-taking-07649/
HOSTS	Corn, Herman, Clive, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
BOOK GENERATED	28 October 2026 at 09:52 UTC

SYNOPSIS

Speech recognition systems routinely cut callers off mid-thought, mistaking a hesitant pause for a finished turn. The mechanics behind endpointing reveal why accurate vocabulary models still fail in emergency dispatch and accessibility tools. Looking at pause frequency, duration, and repair structure across spontaneous speech, older adults, and clinical cohorts shows that human hesitation follows complex syntactic patterns rather than a universal silence threshold.

TRANSCRIPT



Corn

A machine deciding I have finished thinking feels presumptuous.



Corn

Daniel is fine-tuning Parakeet for Hebrish dot app, and it has him watching speech differently. When recognition breaks, what exactly has broken? He means the shape of actual talking, the repetitions, abandoned sentences, sudden bursts of speed, and those silences when you are still assembling the thought.



Herman

The silence can contain most of the work.



Corn

His particular frustration is a detector treating a thinking pause as permission to end his turn. He wants to know how those pauses and repairs actually vary, especially among older speakers, people with dementia, and intoxicated callers. Emergency dispatch makes the stakes concrete. Someone trying to describe a crisis may sound nothing like someone reading training sentences into a microphone.



Herman

Sometimes the caller is remembering an address while listening for breathing. Finishing a polished sentence is fairly low on the agenda.



Corn

Meanwhile the software has already thanked them for their contribution. Daniel wants the measurements behind that mismatch, and what you would build differently. I think the awkward part is that a recognizer can get every audible word right and still fail him.



Herman

By ending the recording before the important word arrives. Yes. Recognition quality and listening well are related, but they are not interchangeable.



Herman

A useful separation here is the word decoder, the speech detector, and the decision about whose turn it is. They can be different components. People blame the transcript when the decisive mistake happened before the decoder received the next piece of audio.



Corn

You could fine-tune the vocabulary beautifully and still have a system that never lets you finish saying it.



Herman

Particularly a correction. Imagine, send it to Michael... no, Michelle. The audio contains the wrong name, a planning gap, a rejection, and the right name. A system evaluated on little finished clips gets the whole repair handed to it. A live system must decide whether to wait during the gap.



Corn

Michael receives another confidential attachment. Excellent latency.



Herman

That distinction matters more than a blanket claim that models are language-agnostic. They still learn particular languages and acoustic distributions. Accents and recording conditions still matter. Daniel's useful move is asking us to look beyond the inventory of sounds at how speech unfolds.



Corn

Also, I would resist saying all training data is read speech. There is conversational data. The problem is that even conversation can be chopped into tidy utterances before training.



Herman

Yes, that was too broad. Prepared speech and pre-segmentation are separate distortions. A recording can start as a messy phone call and reach the model as a collection of neat answers, each already supplied with its ending.



Herman

Corn, would you prefer a dictation app that preserves every restart, or one that quietly writes what you meant?



Corn

For a note, clean it up after I finish. For a call about someone not breathing, preserve the correction and preserve when it happened. I don't want the software improving the witness.



Herman

That choice changes the target. A polished document, a faithful transcript, and a live dispatch aid need different outputs, even when they share the same recognizer.



Corn

You are about to propose separate evaluation suites.



Herman

I am.



Corn

I could hear the spreadsheet opening.



Herman

You also just demanded three different products. I am merely providing the paperwork.



Corn

The history is useful here because people were measuring hesitation long before we had an app to interrupt it. Frieda Goldman-Eisler separated how often people pause from how long each pause lasts. That sounds obvious until you see systems reduce the whole person to total seconds of silence.



Herman

Total silence hides very different behavior. Someone might stop once to plan an entire explanation. Someone else might stop repeatedly while finding individual words. Equal silent time, different sequence, different demands on an end-of-turn detector.



Corn

Her work also tied pauses to places where the next word was less predictable. The gap belongs to a choice. It isn't necessarily a loss of momentum.



Herman

The especially striking finding was that hesitancy and fluency could alternate rhythmically in spontaneous speech. Planning-heavy stretches followed by fluent stretches. That pattern was not evident in reading. Reading supplies much of the linguistic plan in advance.



Corn

The book has already done the worrying.



Herman

It has supplied the words and their order. The reader still has work to do, but spontaneous speech adds deciding what to say while saying it. Speech rate consequently varies inside a single speaker's turn. A fluent burst doesn't establish the speed they will maintain through the next difficult thought.



Corn

Which makes averaging words per minute a fairly blunt instrument. If I spend ages finding the street name and then give the address quickly, the average conceals precisely where the system needed patience.



Herman

Speech rate including pauses and articulation rate excluding them answer different questions. For endpointing, you need the sequence, including pause position, what preceded it, and what comes afterward during training.



Corn

How long is normal, then? I suspect the temptation is to name a number and manufacture a normal person around it.



Herman

There is no universal thinking-pause duration, but there are two useful anchors. Between turns, conversation is astonishingly fast. Stivers and colleagues compared ten languages in two thousand nine and found every one of them minimising the gap, with the typical handover around two hundred milliseconds. Inside a turn, while someone is still planning, pauses run from fractions of a second to several seconds, and longer when searching, remembering, or attending to something outside the conversation. A conventional pause-analysis threshold around a quarter of a second is a measurement choice. It does not mean people complete thoughts on that schedule.



Corn

So the machine is borrowing the speed of a handover and applying it to the middle of a thought.



Herman

That's the mismatch in one sentence. A gap that would be awkwardly long between two people's turns can be perfectly ordinary inside one person's.



Corn

A quarter second is barely enough to regret starting the sentence.



Herman

Nor should we mix an analytical threshold with a product timeout. One says what gets counted as a pause. The other says when a machine takes action. They are often discussed as though they were the same parameter.



Corn

False starts add another distinction. Repeating the word the can mean I am holding the floor. Repeating no can be the information. Deleting every repetition makes both look tidy and one look wrong.



Herman

Repairs have internal structure. The abandoned stretch, an interruption point, sometimes an editing phrase, then the replacement. The abandoned stretch can be quite long. If you reset context at a pause, the replacement may lose the thing it was replacing.



Herman

Children give us a surprisingly hopeful example. They use more disfluencies and fillers, including more nasal filled pauses. Their abandoned stretches were longer than their repairs on average, unlike the adult comparison.



Corn

Yet I wouldn't assume you need a separate theory of hesitation for every age.



Herman

You don't. A detector trained on adult telephone conversation transferred better to children's speech than to another adult conversational collection. Its F one score was roughly ten percent higher on the children. Domain and recording style can outweigh the simple adult versus child label.



Corn

The machine recognized childhood better than a different telephone call.



Herman

Broadly, yes. That is an argument for learning repair structure rather than assuming a demographic label tells you everything about the speech.



Corn

Older speakers get the opposite problem. People assume there is one well-characterized elderly voice, then treat departures as pathology.



Herman

Motor-speech skills tend to slow with age, but the reference data becomes sparse for the very oldest speakers. Work on people aged eighty to one hundred explicitly identifies a shortage of normative data. That is a shortage, not proof that everyone in that range sounds alike.



Corn

Especially awkward when voice control is being sold as the interface that needs no dexterity.



Herman

Hearing, dentition, cognition, and motor performance can all matter. The baseline should allow normal ageing without confusing it with disease. Clinicians tend to find a change from someone's own usual performance more useful than their distance from an imagined



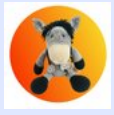
Herman (continued)

average.



Corn

Herman, you would ask the app for a clinical history before letting it open the curtains.



Herman

Only if it diagnosed dementia while refusing to open them.



Corn

Fair distinction.



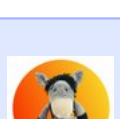
Herman

What would you actually change for an older user, Corn? A permanent slow mode?



Corn

No. I'd let the system learn that user's interaction pattern and give them a clear way to hold the microphone open. A permanent slow mode assumes they are always slow. Someone can take time finding a name and then finish faster than I can move a cursor.



Herman

User-specific adaptation plus explicit control. The control matters because adaptation can guess wrong.



Herman

Dementia makes the measurements more complicated, not more uniform. In one picture-description comparison, pause rate distinguished healthy elderly controls from both mild and moderate Alzheimer's groups. Pause duration did not. Fillers didn't separate the groups either.



Corn

Counting ums would have bought you very little.



Herman

The largest differences involved pauses at the beginning of utterances. Position mattered too. Mild and moderate disease showed different patterns within clauses and at clause boundaries. You cannot reduce that to progressively longer silence.



Corn

There is another picture-description result pointing in a different direction, though. Qiao and colleagues found average silence duration tracked cognitive performance most strongly among the features they examined.



Herman

Lofgren and Hinzen found rate, not duration, separated their groups. Qiao found duration particularly informative. Those are different results, although group discrimination and correlation with cognitive scores are not identical statistical questions.



Corn

Nor are the participants and pause definitions necessarily interchangeable. Still, if someone claims the settled recipe is just wait longer for dementia, those results don't support it.



Herman

No. Both results tell you silence deserves attention. Neither establishes a clinical timeout. More frequent pauses could require repeatedly allowing continuation without every gap becoming exceptionally long. Longer pauses could require patience at an individual gap.



Corn

I would keep both measurements rather than elect a winner.



Herman

Same here. Duration, frequency, position, and the surrounding language. Also, a pause pattern is not a diagnosis. The recognizer's job in this setting is to keep hearing the person, not infer a medical label from a difficult afternoon.



Corn

Speech can be hesitant because the question is difficult.



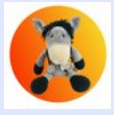
Herman

Or because the speaker is describing something unfamiliar. Aphasia can also involve word-finding and repair difficulties, but it is not interchangeable with dementia. I wouldn't borrow one group's timing distribution and call it validated for the other.



Corn

That seems to be the recurring trap with population-specific software. The label becomes more detailed than the data underneath it.



Herman

Alcohol is a particularly sharp example. A frequently cited experiment used eleven men reading text sober and intoxicated. Acoustic measures involving pitch distributions, signal quality, and the long-term spectrum separated those conditions very well within that experiment.



Corn

Reading text. Even the drunk training subject has prepared remarks.



Herman

The reported discrimination error was below five percent for some measures. But articulation measures varied strongly between individuals, and the authors themselves judged low-level forensic detection impractical because the acoustic effects had ambiguous causes.



Corn

The impressive number came with a fairly unimpressed conclusion.



Herman

Sustained vocal effort could produce similar effects. That does not mean shouting equals drunkenness. It means the acoustic signature lacks the specificity you would need to confidently distinguish them. A frightened caller straining their voice can trigger the same suspicion.



Corn

Distress gets mistaken for intoxication, then intoxication gets mistaken for irrelevance.



Herman

Intoxication would not make the emergency less real anyway. You need intelligibility and appropriate assistance, not an automated judgment of the caller's character.



Corn

Drugs are even less amenable to one speech profile. Sedation and agitation do not create the same conversation.



Herman

I don't have a usable peer-reviewed acoustic characterization of non-alcohol drug-intoxicated speech for ASR here. I would not fill that gap by treating all drugs as alcohol with different branding. Even the alcohol example was controlled reading, not a chaotic live emergency.



Corn

The deployment problem is missing from the experiment intended to explain it.



Corn

Before the timing machinery, our regular correspondent is calling from wherever he has just been mildly let down.

C

Clive

This is One-Star Getaways, and my one star this week goes to the Hotel Puente Quieto in Buenos Aires, where the promised river view turned out to be a framed photograph of the river, hung slightly crooked, and the mini bar held one warm bottle of water and a note saying enjoy responsibly.

C

Clive

The lift played the same nine seconds of tango for three days and the pillow was a folded towel with ambitions. For anyone hunting a genuinely disappointing trip this winter, I recommend it without hesitation.



Corn

A photographed river sounds pleasantly quiet. Herman won't book it. Back to the caller: the software has to decide what silence means before it knows why the person has stopped.



Herman

A VAD can tell you that speech activity has ceased without telling you the turn has ended. A fixed silence timeout converts the first observation into the second decision. Shorten it and you interrupt more people. Lengthen it and every completed turn feels slower.



Corn

And you'd tune that timeout on yourself, I suspect, so it would wait just long enough for you.



Herman

You are predicting my calibration bias.



Corn

I live with its calibration bias.



Herman

Accepted. Testing on yourself is especially deceptive when you already know what the software expects. You unconsciously become a better input device.



Herman

LiveKit's end-of-turn benchmark makes that tradeoff measurable. It treats every pause inside a user's turn as a decision point, across fourteen languages. At a three-hundred-millisecond latency budget, the small open detectors the VoxTurn model card compares against cut people off in well over a quarter of those pauses. One of them, close to forty percent.



Corn

That is an interruption habit, not an occasional misunderstanding.



Herman

The explanation is training geometry... no, training examples. Many detectors learn from segmented utterances with a label at the end. Deployment asks them questions inside a silence, at different elapsed times. Those are different tasks.



Corn

An utterance boundary supplied by a dataset curator is knowledge the live system doesn't have.



Herman

Yes. Training on the final frame of a completed utterance doesn't adequately teach the uncertainty halfway through a thinking pause. The detector needs examples of unfinished turns that happen to be quiet.



Corn

I can see why adding a longer timeout persists. It is one setting you can change without rebuilding the dataset.



Herman

It is also immediately perceptible in a demo. The user gets interrupted less, and the delay after easy sentences seems tolerable. Then the product meets a long search for a name, or a rapid exchange where that same delay is maddening.



Corn

Two users, one knob, neither pleased.



Herman

Would you accept a button that says I am still thinking, Corn, or would that defeat the point of voice?



Corn

For ordinary dictation, gladly. Hold to talk, tap to finish, something predictable. Accessibility needs vary, so it cannot be the only option. For an emergency call, requiring the caller to manage the assistant's conversational anxiety would be absurd.



Herman

That distinction gives us separate designs. Explicit controls where feasible. Better inference where it has to be hands-free. Neither excuses silent truncation.



Herman

VoxTurn trained on real pauses from phone calls and face-to-face dialogue. It sampled decision points throughout those pauses, with timing-aware labels. During silence it checks again every quarter second rather than making one decision and letting a fixed timer take over.



Corn

Waiting becomes something the model keeps considering, rather than a countdown to interrupt.



Herman

In English, its own model card reports eleven point five percent false cutoffs at the same three-hundred-millisecond budget. Those are self-reported numbers, and LiveKit's own larger detector posts slightly better on that benchmark. Either way, it's a substantial improvement, not a solved interaction. Plenty of people still get cut off.



Corn

Eleven percent is not a rounding error when it is your sentence.



Herman

Across fourteen languages the mean was nineteen point nine percent. Each decision took eight point three milliseconds in its reported setup. Cheap enough to revisit the decision, but we should not mistake that compute time for the whole system's response latency.



Corn

Nor mistake the English result for every language. Daniel is mixing languages, which adds another reason not to inherit one benchmark's confidence.



Herman

The useful transferable idea is how the examples are constructed. You would still validate the actual language mix, speakers, microphones, and interaction pattern.



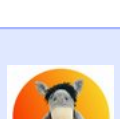
Corn

What tells it the user is about to continue? Words? Intonation? The growing silence?



Herman

Potentially all of those. A sentence fragment, continuation-like delivery, and the time already elapsed provide evidence. None is decisive. A grammatical sentence can be followed by another sentence in the same turn. An unfinished sentence can be abandoned.



Herman

I don't know how well any one learned detector will handle the longest individual pauses without testing it. Repeated evaluation improves the decision process. It doesn't give the model access to an intention the speaker hasn't expressed.



Corn

Sometimes the correct answer is unknowable yet.



Herman

Hm. Which argues for responses that preserve recoverability. Keep listening, allow interruption of the assistant, retain the audio across the boundary. Taking a conversational turn need not destroy the user's ability to finish.



Corn

The non-speech distinction deserves more attention in dispatch. A cough can be medically relevant without being a word. Background crying can matter without being a new speaker taking the floor.



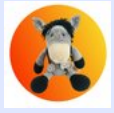
Herman

This architecture separates speech-versus-non-speech classification from turn completion, with training that includes coughs, laughter, noise, and music. For dispatch I would also preserve relevant sound for the human. Suppressing something from the transcript should not automatically erase it from the evidence.



Corn

A clean transcript can be an aggressively incomplete account of a call.



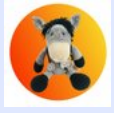
Herman

Exactly the danger. Then microphone distance adds an independent problem. In a Thai meeting corpus, fine-tuning reduced overall word error from about sixty-four percent to thirty-eight percent. Better, but still a difficult transcript.



Corn

The improvement is large and the remaining error is large. Both can be true.



Herman

Far-field speech was worse still. After adaptation, its word error remained about fifty percent. Distance, room acoustics, and conversational style can all move deployment well away from the conditions where a recognizer looks impressive.



Corn

Which is why I care about the recordings the product actually receives, not only a prestigious benchmark with a clean ending attached.



Herman

Commercial systems have also produced substantially worse results on real spontaneous conversation than headline benchmark scores suggest. Call-center comparisons show segmentation choice alone changes measured performance.



Corn

The test can improve before the listener does.



Herman

If you cut clips using the reference transcript, you may remove exactly the boundary problem the deployed system must solve. Evaluate the continuous interaction too. Recognizing preselected speech and deciding what to send for recognition need their own accounting.



Corn

For Hebrish, I would keep an evaluation set of complete dictation sessions. Not merely the surviving clips after the detector has removed the awkward bits.



Herman

Include corrections, hesitations, and changes of speed. Score the lost words as well as the substitutions. Otherwise the component that discards your sentence can escape the transcript audit.



Corn

It has achieved accuracy through not listening.



Herman

Emergency dispatch takes that concern beyond irritation. Caribbean-accented emergency speech work proposes using low recognition confidence as a reason to prioritize human attention, particularly alongside distress cues. The model's uncertainty becomes useful operational information.



Corn

I'd want to separate more attention from automatic emergency severity. Noise can lower confidence too, and a quiet, clear caller can be reporting a catastrophe.



Herman

That separation is vital. Low confidence can trigger review or assistance. It should not automatically determine dispatch priority. Composed bystanders and trained responders may describe life-threatening events without sounding distressed.



Corn

The system needs to hear the calm person saying not breathing, not wait for a scream.



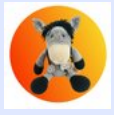
Herman

Human dispatchers miss cardiac arrest too. The trial context here reports approximately a quarter of out-of-hospital cardiac arrests not being identified. That establishes the difficulty of the task, not that ASR causes those misses or that a speech model automatically fixes them.



Corn

You want the medical tool badly enough to be irritating about its validation.



Herman

I do. An alert that recognizes words but arrives after the dispatcher has moved on can fail operationally. An overconfident wrong transcript can anchor the human on a mistaken account.



Herman

Corn, if you could see only the best transcript or a rough transcript with uncertain stretches marked and playable, which would you use?



Corn

The rough one with the audio. Especially around an address, a negation, or a correction. I would rather know where to listen again than have a fluent sentence conceal the missing bit.



Herman

Confidence needs calibration under the actual channel conditions. Then the interface can make checking efficient. You don't want a dispatcher deciphering decorative confidence colors while trying to hear the caller.



Corn

Nor a queue where every unfamiliar accent is permanently marked difficult.



Herman

Routing needs auditing across accents and populations. The aim is getting assistance to uncertain calls, not making those callers wait behind the easy ones.



Corn

The training problem repeats itself at every layer. Scripted alcohol speech, chopped-up conversations for endpointing, clean clips for evaluation. Then everyone expresses surprise at a live caller.



Herman

Those shortcuts make experiments manageable. They become harmful when nobody tests what the shortcut removed. We still need actual emergency-call pause distributions. The dispatch material here addresses recognition, accents, and distress, but it doesn't supply a direct measurement of those silence durations.



Corn

A forty-second gap could also be somebody looking for the address rather than thinking about the sentence.



Hilbert

Forty-one seconds. I just made one looking for my glasses. The recorder ended my turn before I had reached the second pocket. I was still addressing it, Herman. I took your distinction between speech stopping and the turn ending quite literally.



Corn

The glasses were participating in the endpoint test.



Hilbert

I put the same recording into my dictation test six times while you were talking. I described every pocket as I searched it. Then I labelled the silence occupied, because I was occupied. The program won't accept that label. It offers speech and silence. Very limited interest in trousers.



Herman

Occupied would require information beyond the waveform. The audio can contain no speech while the speaker is still engaged.



Hilbert

I accept that. I have now assigned the empty interval twenty-seven words of back pay. The transcript is longer, but the recognizer didn't supply any of them. I wrote searching in each position. Corn, you gave the correction example in less time than it took me to decide whether a waistband counted as a pocket.



Corn

You have created paid silence. I support its working conditions.



Hilbert

It is fascinating that I can remain in a conversation while doing something the microphone cannot hear. Then the interface informs me I have left. That is poor manners dressed as a measurement. Nadav Shalev can search an entire coat in eighteen seconds. He announced that once and I have had to live with it.



Herman

The recording needs to preserve the interval even if the transcript has nothing to put there.



Hilbert

It does now. I extended the file instead of filling the gap. My glasses are in Nadav's desk drawer on the third floor. I remembered at forty-one seconds.



Corn

Extending the file is a small distinction with a large consequence. Silence can remain available without being turned into invented words.



Herman

For a live system, retaining context across the gap lets the eventual continuation revise what came before. In training, retaining the gap lets us test what the detector actually would have done.



Corn

That also separates a person's thinking from everything else they might be doing. Looking for medication, opening a door, moving toward the patient. The pause isn't necessarily lexical retrieval.



Herman

A complete interaction includes those tasks. A transcript alone won't explain every delay. Timing, audio, and the sequence of actions matter, without assuming every long silence is evidence of impairment.



Corn

I'd like future evaluations to preserve the empty intervals as carefully as the spoken words. Then we can test whether a system actually stays with someone through a search, a repair, or a moment of uncertainty, instead of rewarding it for receiving a conveniently finished clip.



Herman

The practical answer is real conversational training, timing-aware turn decisions, and recoverable interactions. In dispatch, uncertain recognition should bring human attention closer, not make the caller disappear.



Corn

Thanks to Hilbert Flumingtop for producing the show.



Herman

If you enjoyed this one, try episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

If you enjoyed this conversation, leave us a review. It helps people find the show.



Herman

See you tomorrow.