Episode Book

EPISODE #6297

Teaching Dictation Models to Ignore Other Voices

PUBLISHED	27 October 2026 at 12:28 UTC
DURATION	22:03
LISTEN	myweirdprompts.com/episode/target-speaker-speech-recognition-87866/
HOSTS	Corn, Herman, 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

Standard speech recognition will faithfully transcribe clear background announcements into a personal document. Target-speaker automatic speech recognition attempts to fix this by enrolling an authorized user's voice and conditioning the transcription on that acoustic identity. A look at how speaker embeddings, attention conditioning, and architectures like Speaker-Querying Whisper handle interference reveals why rejecting clean background speech during silence is surprisingly difficult without explicitly training the model to produce an empty output.

TRANSCRIPT



Corn

Better noise cancellation is the standard answer to dictation going wrong in an airport. Daniel's example turns that answer inside out. The offending sound is a perfectly clear announcement, and he wants the speech model itself to know it belongs to somebody else.



Corn

His proposal is to record thirty to sixty seconds of your voice, turn that into a speaker embedding, and use it inside recognition, perhaps through attention, so only your speech becomes text. Ten seconds of Daniel dictating should appear. The announcement that follows, Ladies and gentlemen, we are now ready to board Delta flight ten to Newark, should produce nothing.



Herman

The airport has delivered an unusually legible input. Recognition succeeds, and the application fails.



Corn

Newark has entered the document uninvited.



Herman

Target-speaker ASR is the name for the architecture Daniel is imagining. The answer is yes, you can build speaker selection into recognition. The harder answer concerns what information you give it, where that information enters, and what happens when the intended speaker stops talking.



Corn

He's interested mainly in one authorised user. A family meeting with another authorised voice is an extension, not the main event. I want to hold you to that, Herman. Your meeting diagram already has chairs.



Herman

Two chairs. I've shown restraint.



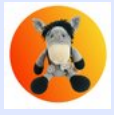
Herman

A speech activity detector will happily open for that announcement. A quality score may prefer it to Daniel's dictation. Loudness might favour it too. None of those signals tells the application whose words belong in the document.



Corn

Your instinct would be to improve the microphone placement.



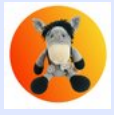
Herman

My instinct would be to remove the airport. Placement still helps, but you're right about the limit. Once clean speech reaches the recogniser, the system needs a target definition. Otherwise, faithfully writing it down is a reasonable response to the task it was given.



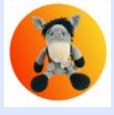
Corn

I suspect people imagine the app already knows the microphone is theirs, therefore the words must be theirs. Ownership of the laptop has somehow become an acoustic feature.



Herman

The microphone doesn't get the purchase receipt. Target-speaker systems add another input that identifies whom to follow. Most published work concentrates on simultaneous speakers, using mixtures such as Libri2Mix or recordings of meetings. Separating two overlapping voices is difficult, so that's where much of the effort goes.



Herman

Daniel's sequence has a different shape. He speaks, stops, and the announcement starts. There's no simultaneous voice to separate during the announcement. The desired output for a whole stretch of excellent speech is empty.



Corn

Empty is a surprisingly ambitious feature.



Herman

Commercial voice isolation already exists. Krisp, for one, sells isolation that sits ahead of speech recognition. The usual commercial arrangement cleans or selects the audio before the recogniser receives it.



Corn

Which answers whether someone can sell the experience, but leaves Daniel's architectural question open. He wants the recogniser to carry the intention, rather than receive audio somebody else has approved.



Herman

Would you care where the selection happens, Corn, if the experience were identical?



Corn

As a user, barely. As someone debugging the failure, enormously. If the airport survives the filter, can the recogniser still refuse it? I'd rather have that decision represented somewhere than hope the unwanted voice has vanished.



Herman

That distinction matters. Integration gives recognition a chance to learn which acoustic evidence deserves words, without requiring an intermediate waveform that sounds clean to us.



Herman

A conventional recogniser learns a mapping from speech audio to text. Whisper-style foundation models weren't generally trained with an enrolled user and an instruction to reject every other voice. Their architecture doesn't come with a personal ownership test.



Corn

Training it on more airport announcements could make this worse.



Herman

It could make the announcements easier to transcribe. You need the objective to change. The same recording might have Daniel's transcript as the correct answer when Daniel is selected, and the announcer's transcript when the announcer is selected.



Corn

Same sound, different correct document. That makes the enrollment input part of the task definition, rather than a little accuracy enhancement.



Herman

Yes... and we should separate this from hallucination. Whisper can produce phrases when there was no speech corresponding to them. Its internal representations also retain information about non-speech sounds. Noise robustness doesn't mean noise disappears inside the model.



Herman

One Polish team ran over three hundred thousand non-speech clips through it and collected recurring outputs into a Bag of Hallucinations. Separate evaluations found whole invented phrases in roughly one percent of transcriptions. Those findings concern particular evaluations, not a universal rate for every recording.



Corn

The Delta announcement needs no invention. It's an accurate transcription of an unwanted event.



Herman

Right. Well... accurate if the words are recognised correctly. Even perfect accuracy on that voice would violate Daniel's request. Post-processing for suspicious invented phrases won't remove a legitimate boarding announcement reliably.



Corn

Especially if I'm actually dictating a boarding announcement. The words themselves cannot be the blacklist.



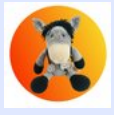
Herman

Enrollment provides acoustic identity instead. A speaker encoder processes the sample and compresses it into a fixed-size vector. Training encourages samples from the same speaker to land near each other despite different sentences, with different speakers farther apart.



Corn

It has to forget the sentence while remembering the speaker.



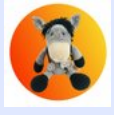
Herman

Broadly. It won't forget every trace of content or recording conditions. Those are some of the complications. A comparison of target-speaker systems used vectors with five hundred and twelve dimensions. They're learned coordinates, not a row of labelled measurements for accent, pitch and throat size.



Corn

No little box marked donkey.



Herman

The absence has not impaired my representation.



Corn

Thirty seconds sounds short for a permanent description of a person. Though I suppose your voice gives them plenty to work with.



Herman

Mine is generously documented. Thirty to sixty seconds of reasonably clean, varied speech is plenty for a useful enrollment. More precisely, getting a stable speaker representation from that amount is plausible and routine. Getting flawless rejection from it is a separate demand.



Herman

Here's the part I think you'll object to before I finish. A better speaker-verification embedding doesn't necessarily produce better target-speaker recognition.



Corn

You were going to sell me the vector before revealing that.



Herman

I was going to show you the elegant mechanism first.



Corn

There it is.



Herman

The NTT and Brno comparison tested speaker representations across recognition, extraction and personal speech detection. Strong verification performance didn't reliably predict strong performance on those tasks. An ideal identity code could outperform enrollment-derived vectors.



Corn

An identity code meaning the experiment already knows which person it is? Not a miraculous microphone that assigns everybody a flawless number.



Herman

Correct. It's a controlled reference. It helps distinguish trouble in representing identity from trouble using identity. More awkwardly, when they optimised the conditioning vector for the target task, its best position changed with the competing speech.



Corn

I remain the same person, but the useful description of me changes depending on who's interrupting.



Herman

The vector needs to emphasise whatever distinguishes you from that particular interference. A fixed enrollment remains useful. The result challenges the assumption that one universal voiceprint is automatically the best steering signal forever.



Corn

I buy that, with a reservation. Daniel doesn't need the mathematically best vector in every room. He needs one good enough to reject the airport without demanding a fresh audition beside every gate.



Herman

That's fair... useful and optimal are different thresholds. Would you prefer the system to adapt its conditioning during a session?



Corn

Only if it keeps the original identity anchored. Otherwise it might adapt beautifully to the announcer while I'm buying coffee. Improvement needs a definition of whom it's improving for.



Herman

That's why session context and target assignment have to remain distinct. The alternative approach gives the recogniser frame-level labels: silence, target speech, non-target speech, or overlap.



Hilbert

Hilbert's Anteater Trivia. Emas National Park, in central Brazil, is known for its giant anteaters. I have never been, which puts me ahead of most park rangers.



Corn

A convincing case for staying home.



Herman

Those labels let the recognition layers respond differently to each state. We went through DiCoW and its Parakeet cousin, DiCoP, last time: small learned transformations per label, and they work because another component has already made the difficult judgement of who is speaking when. No speaker embedding in sight.



Corn

The distinction is attractive, though. The recogniser gets told what role this sound has now. It doesn't have to discover an entire identity catalogue before deciding whether a word belongs.



Herman

Embedding-conditioned models take a different route. Speaker-querying Whisper uses learned queries and enrollment information to draw speaker-specific features from the mixture. Those features steer recognition towards the selected voice.



Corn

Attention gets evidence of whom to follow, rather than a command written in English saying ignore everybody else.



Herman

Training makes that relationship meaningful. Just attaching a vector to an existing decoder doesn't give the coordinates a shared meaning. You need examples, a way to inject the information, and an objective that rewards the selected transcript.



Corn

Otherwise it's five hundred and twelve numbers wearing a visitor badge.



Herman

You could inject the conditioning alongside acoustic features, through transformations inside the encoder, or through attention that consults speaker information. Daniel's decoding-stage intuition is sound in spirit, but the useful intervention often happens before final token selection.



Corn

Waiting until the model has decided the sentence is about Delta seems late. I would want identity attached to the acoustic evidence before the language model gets enthusiastic about finishing the destination.



Herman

Earlier access helps preserve the association between words and voice. The canonical design has an auxiliary enrollment network and a recognising network taking both its representation and the live audio. It can be trained end to end, or adapted from an existing recogniser.



Herman

The speaker-querying approach gets about fifteen percent word error on Libri2Mix and around four percent on WSJ mixtures. Different benchmarks, different difficulty. Those results establish that enrollment can steer transcription, not that an airport app has been solved.



Corn

What does the training example look like when Daniel is silent? I think that matters more than the clever attention layer.



Herman

It should include non-target-only stretches whose desired transcript is empty. Also target-only stretches, overlap, pauses and transitions. If training overwhelmingly says there are two voices and one must be selected, it may not adequately teach the case where the target is absent.



Corn

The model must be allowed to find nobody.



Herman

Yes. No... more than allowed. It must be rewarded for that decision in the right circumstances. Daniel's PA case may be easier for a segment-level identity gate because nothing overlaps. Yet it tests target absence, something an overlap benchmark score can conceal.



Corn

I'd score the airport intrusion separately. A transcript with every word of my dictation intact and one unwanted sentence could look respectable on a broad accuracy average. I'd still delete the app.



Herman

What would you put beside word error, then?



Corn

How much non-target speech reaches the document, and how much of mine disappears. Separately. Otherwise a very quiet system wins by refusing to transcribe anyone, which is a remarkably expensive notepad.



Herman

Streaming adds the timing dimension. An offline system can inspect later evidence before deciding whose voice it heard. A live dictation system must either wait, emit provisional text, or risk committing the wrong speaker's opening words.



Corn

Dimmed text doesn't feel provisional when it says I'm boarding.



Herman

A streaming target-speaker transducer with dual attention remained around sixteen percent word error under contaminated enrollment, while comparison approaches exceeded seventy-five. That demonstrates robustness in its tested conditions. It doesn't give us end-to-end delay for a dictation product.



Corn

You can call a system streaming while leaving me unsure how long the first acceptable word takes.



Herman

Processing incrementally and feeling immediate are different properties. I'm not sure what delay would be acceptable here. Would you tolerate a short hold before each new speaker?



Corn

Before a speaker change, perhaps. Before every phrase after a breath, less so. I pause to think, Herman. I don't want my own software treating contemplation as a suspicious new identity.



Herman

It would have ample test material.



Corn

You may use the gaps to recover from your explanations.



Herman

Another complication is that enrollment isn't always clean. Self-enrolled systems can locate a passage in the conversation where the selected speaker is most active, then use that passage through cross-attention in the encoder. One extension roughly halved its speaker-attributed error measure against its predecessor on its benchmark.



Corn

Useful when nobody remembered to record the ceremonial voice sample. Though somebody still needs to decide which participant is the user.



Herman

Self-enrollment finds representative audio. It doesn't read intention. Additional cues can help: in one dual-cue evaluation, enrollment alone gave roughly twenty-nine percent character error, while enrollment plus a short text cue brought it below nine percent.



Corn

Text cues sound useful when you know what someone might say. My next dictated sentence is often news to me.



Herman

They're task-dependent. The result says voice enrollment needn't be the strongest available cue. It doesn't mean dictation should require a script.



Corn

The family-meeting exception stays small, then. Add another authorised identity, permit either, and perhaps preserve who said what if the document needs it.



Herman

Diarization tracks who spoke when across participants. Daniel's ordinary case is target membership, an authorised-user decision. With multiple authorised voices, diarization and attribution become more useful. A product would likely split it the same way: a voice profile for personal dictation and a separate meeting mode.



Corn

I still favour the boring upstream component for shipping this soon. I know you're preparing to accuse me of resisting elegance.



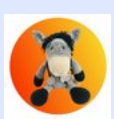
Herman

I was preparing to agree. A pre-filter works with existing recognisers. Rebuilding their conditioning path means model access, training, evaluation and maintenance every time the base model changes.



Corn

You've installed enough fresh models to qualify as a maintenance event.



Herman

The installation is often the pleasant part. Krisp's Voice Isolation sits ahead of transcription and reports forty-six percent lower word error with fifteen milliseconds of processing latency on a CPU. Those are company evaluation claims, and the latency belongs to isolation, not the entire app.



Herman

Its tests covered ten recognition engines and over sixteen hundred real-world recordings. That establishes a practical path without proving anything about Daniel's in-model proposal.



Corn

They also inherit a harsh limit. If the filter removes my quiet word, the recogniser never gets a vote.



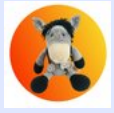
Herman

An integrated model could use phonetic and speaker evidence jointly instead. Conversely, a separately inspectable filter makes debugging easier. Neither arrangement automatically wins on clean-speech preservation.



Corn

Have we actually got the exact airport test evaluated anywhere?



Herman

I haven't seen a published evaluation matching it: single-user dictation, then a clean non-overlapping PA, with zero PA text required. Most results concern overlapping mixtures or meetings. I'd call Daniel's proposal buildable, but I wouldn't quote those benchmark scores as its acceptance test.



Corn

There's also a difference between excluding everybody except me and letting another person exclude themselves.



Herman

Target-speaker unlearning explores that inverse request, people who don't wish to be transcribed. The same identity-sensitive recognition can support different permissions. An enrolled voice establishes a match, not who has authority to set the rule.



Corn

At the user interface, the promise ought to be...



Hilbert

A separate terminal for the announcer. Mine went to my neighbour Nissim eleven minutes ago. I rang him while you were discussing the empty transcript. He has brought it upstairs. It is sitting outside the door because he says an airport terminal cannot enter a residential room.



Herman

Your dictation terminal, rather than an airport terminal.



Hilbert

A dictation terminal. I have been keeping announcements out of my notes for years. The microphone feeds a little computer. My voice goes to the document. The other voice goes through the loudspeaker to Nissim, who writes it on a boarding card. He has been doing departures for my kitchen since March. No passengers. He insists on that.



Corn

That explains the separate terminal.



Hilbert

Herman, you made the frame labels sound quite compact earlier. Mine are spread across the landing. There is a cable under the fire door and a bell for uncertain speakers. The bell is objectionable. It rings during chewing. I sold the terminal to Nissim for forty shekels during your explanation of attention. I wanted the bell out of my room. He has returned with a transfer-of-operations form.



Herman

The uncertain cases still need somebody to decide. Selling the machine hasn't removed those.



Hilbert

He has put me down as deputy. I crossed it out. We tested it just now with the Newark sentence. He read the announcement. I dictated a complaint about the announcement. My document kept the word Newark and lost the word complaint. Nissim wrote the rest on his card. The manufacturer calls that personalisation. It has certainly made it personal.



Corn

Forty shekels, and you're still deputy.



Hilbert

Acting deputy. He corrected the form. His boarding card took eighteen seconds to reach me. He had to find a pen.



Herman

Eighteen seconds spent outside recognition would dominate the experience, however fast the model ran. That's why component latency and the delay to usable text need separate measurement.



Corn

The idea that sticks for me is that intended speaker belongs in the recognition task definition. Daniel's dictation and the PA can both be recognised perfectly, yet only one belongs in his document.



Herman

Enrollment can supply that distinction, but a fixed voiceprint is only one way to represent it. Relative frame labels and adapted conditioning show why the identity mechanism needs more thought than recording a sample.



Corn

I'll be watching for an integrated system measured on unwanted text actually reaching the screen, including the waiting around that happens outside the model. The empty Newark transcript deserves its own result.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you enjoyed this one, try episode seven, Building Custom ASR Tools; episode eight, Building Your Own Whisper; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy these conversations, leave us a review wherever you listen. It helps other curious people find the show.



Corn

See you tomorrow.