Episode Book

EPISODE #6298

Why Dictation Models Steal Other People's Voices

PUBLISHED	27 October 2026 at 12:30 UTC
DURATION	22:56
LISTEN	myweirdprompts.com/episode/target-speaker-asr-parakeet-05822/
HOSTS	Corn, Herman, Vex, 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 models can transcribe every word accurately while capturing the wrong person's conversation. Fine-tuning a recognizer on a single voice does not teach it to ignore clean, loud background speech like train platform announcements. Addressing issues in systems like Parakeet requires target-speaker conditioning, diarization-dependent masks, or enrollment embeddings rather than raw audio alone, preventing fluent intrusions and music-induced hallucinations.

TRANSCRIPT



Corn

A dictation system can get every word right and still put somebody else's conversation in your document. Daniel has hit that problem while building his Parakeet fine-tune training set. Out on a train platform, his voice competes with announcements, nearby conversations, shouting, and music with lyrics. He wants to train it to keep following him, but he's asking what following him actually means inside the model.



Herman

The missing input is surprisingly basic. Stock Parakeet receives audio, but nothing saying which person in that audio owns the document. Being the person holding the phone doesn't automatically become a fact available to the recognizer.



Corn

My thumbprint has failed to establish editorial control.



Herman

Daniel's distinction matters before he mixes another recording. Improving recognition of his accent and vocabulary doesn't necessarily teach the model to reject somebody else's perfectly intelligible sentence. Those are different training objectives.



Corn

Especially when the announcement is clearer than the dictation. The platform speaker has amplification, a script, and no hesitation about where the train is going. Daniel has a thought he's still assembling.



Herman

That clean, confident competing sentence can be extremely attractive to a recognizer. The first thing to establish is that the model, as configured, has no explicit idea who Daniel is.



Corn

Even a personalized fine-tune? Surely hearing his voice repeatedly gives it some preference.



Herman

It can give it a statistical preference. His pronunciation becomes easier to decode, his vocabulary becomes more likely. But preference learned incidentally from examples is weaker than an explicit target instruction. Parakeet v three's documented interface is audio in, text out. No enrollment recording, no speaker embedding, no argument meaning transcribe this person.



Corn

Then calling it my model invites an unfortunate misunderstanding.



Herman

Single-speaker ASR and multi-speaker ASR have different jobs. A single-speaker recognizer can handle recordings containing different people speaking in succession without promising to resolve simultaneous speech or preserve one person's identity through overlap. Multi-speaker systems explicitly address who said what. Target-speaker recognition asks for one designated person's words.



Herman

The DiCoP model card is unusually blunt. Without conditioning, its encoder returns a fluent transcript of whoever is loudest. They disable the ordinary transcription route because it cannot supply the required target mask. Otherwise the output looks correct while answering the wrong task.



Corn

A beautifully typed intrusion.



Herman

Loudest is useful shorthand, though. There's no little sound meter electing one winner for the entire recording. The balance changes over time, frequency, pronunciation, and what the decoder finds plausible.



Corn

I would notice gibberish. I might not notice a sensible sentence that belonged to the man behind me, particularly if I'm dictating notes to read tomorrow. Fluency makes the mistake harder to catch.



Herman

Would you rather it return an incomplete fragment when the voices collide?



Corn

For dictation, yes. An obvious gap asks me to repair something. A polished substitution asks me to remember that I never thought it. Those are very different demands on someone reviewing a page.



Herman

That preference changes evaluation. Counting words alone can miss the distinction between mangling Daniel's sentence and accurately stealing somebody else's. Both hurt accuracy, but the second can look much more trustworthy.



Corn

Your enthusiasm for readable output has acquired a condition.



Herman

Ownership of the sentence. A modest requirement.



Corn

The loudness shorthand still bothers me. A quiet voice can have sharper consonants than a louder one. I can hear you over music without your defeating the entire sound system.



Herman

Local audibility matters. Speech overlaps in time and frequency. A competing voice can bury the target's consonants while leaving parts of its vowels available. That's energetic masking. Informational masking describes interference from meaningful competing speech, including its pull on human attention.



Herman

A recognizer experiences the acoustic mixture, not a listener's conscious distraction. But I wouldn't say it only responds to energy. Its learned language and acoustic patterns matter too. More energy makes a voice easier to recover, without guaranteeing that every word from it wins.



Corn

Someone whispering a very predictable phrase might beat someone loudly inventing a surname.



Herman

Potentially. The model combines acoustic evidence with learned regularities. When evidence gets ambiguous, those regularities have more influence. It may stay with one voice, switch between them, omit material, or produce a sequence nobody actually said.



Corn

A conference call assembled by committee.



Herman

There is genuine internal separation in some systems. Work on serialized output training, where multiple speakers' transcripts are emitted in sequence, found that the decoder performs implicit speaker separation. But ambiguous overlap still defeats it. Importantly, that finding concerns a model trained for that task, not proof that stock Parakeet secretly has a reliable target selector.



Corn

I suspect you're about to admire the implicit mechanism.



Herman

I do admire it.



Corn

It still picked the station announcement.



Herman

Admiration has no effect on the error rate, Corn. I can hold both positions.



Corn

Your loudness description gives our persistent caller his opening. He's through screening again, somehow. The dominance heuristic, the loudest voice winning, appears to be his department.

Vex

ODD SOCKS, people! Every mismatched sock in Jerusalem is a training sample, because whoever pairs them is deliberately teaching the model to ignore the quiet voice and latch onto the loud one! I have the receipts, two thousand and one pairs, catalogued by loudness!

v

Vex

So when Daniel dictates on a train platform and the model grabs the announcement instead of him, that is not a bug, that is the SOCK PROTOCOL, the loudest sock wins, every single time!

v

Vex

You two need to look into the sock drawer and tell me why nobody is asking who does the pairing! LOOK INTO IT!



Herman

Parakeet receives an audio waveform, not information about sock pairing. The changing balance of voices is also why a different background language doesn't automatically solve the problem.



Corn

My guess would have been that English dictation against Russian chatter is easier than English against English. I can't accidentally follow the Russian sentence very far.



Herman

Humans often do get that benefit, called linguistic release from masking. Competing talkers tend to interfere more when they sound perceptually similar to the target. Familiar, intelligible background speech also draws on processing that unfamiliar speech doesn't engage in the same way.



Corn

Though unfamiliar speech has different rhythms and sounds too. You can't put every difference down to understanding the words.



Herman

No, and separating those explanations is difficult. Experiments have tried distinguishing linguistic unfamiliarity from acoustic and linguistic dissimilarity. One perceptual-learning experiment found the foreign-language advantage disappeared after listeners trained on



Herman (continued)

the background language. That makes familiarity part of the explanation, not merely a fixed property of the waveform.



Corn

Learning a language has finally produced a disadvantage.



Herman

For ignoring it, occasionally. But transferring that result to ASR would be a leap. I don't know of a direct target-speaker ASR comparison establishing the benefit Daniel describes. A multilingual model may recognize both streams quite happily. A language restriction might suppress unwanted output, but suppressing Russian words doesn't recover English consonants buried underneath them.



Corn

It can refuse to print the interference without hearing through it.



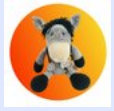
Herman

A target-conditioned model needs evidence about which frames belong to its target. The background language is probably a distraction from that design problem, though particular models could still show language-dependent differences.



Corn

I'd keep the language variation in the tests, then, but not make foreign chatter the presumed easy category. Otherwise the training set gets a comforting label the audio hasn't earned.



Herman

What would you expect when that background voice is singing?



Corn

Two problems. It might type the actual lyrics instead of me. Or it might turn a sustained vowel into something that was never sung.



Herman

Those need separating. Singing is real vocal content, and selecting it over Daniel is a target-selection error. Music can also trigger invented speech, including during instrumental passages. That's hallucination, not simply choosing the wrong speaker.



Corn

The instrumental has submitted minutes.



Herman

In the GDELT comparison of Chirp and Chirp two, the newer system repeatedly invented text during music while the earlier system rejected those sections better. Newer didn't mean better non-speech rejection. Separately, the Careless Whisper analysis found entire invented phrases or sentences in roughly one percent of the transcriptions it examined. That figure isn't a universal music error rate.



Corn

Nor proof that Parakeet behaves identically. I'd be annoyed if we diagnosed Daniel's model with another model's symptoms.



Herman

We need to inspect what it actually does. People processing music with Whisper often try Demucs first to separate sources. That's a front-end intervention rather than retraining the recognizer. But extracting vocals from music can leave you with the singer, precisely the voice Daniel doesn't want.



Corn

Congratulations, the unwanted speaker is now immaculate.



Herman

The desired output determines which separated material is useful. Music examples in a fine-tune aren't pointless, but simply adding songs doesn't tell you whether you're teaching resistance to accompaniment, rejection of lyrics, or refusal to invent words in non-speech. Those errors need different evidence.



Corn

Then Daniel can train the target behavior, but the audio alone doesn't specify the lesson. If his transcript excludes the announcement, the model sees what he wanted, without necessarily getting a durable rule for identifying him next time.



Herman

Target-speaker conditioning makes that rule explicit. The surprisingly small result from Klement and colleagues is that even a learned bias term for each diarization output type, inserted before the first transformer block, can turn a single-speaker recognizer into a target-speaker recognizer.



Corn

Four small nudges, after all that ambiguity.



Herman

A small intervention, not a small surrounding system. Someone still has to supply trustworthy speaker activity information. The full method uses frame-level diarization dependent transformations, FDDT. For each frame, a mask gives probabilities for silence, target only, non-target only, and overlap. That's STNO.



Corn

Overlap meaning Daniel is speaking and somebody else is speaking, rather than uncertainty about which one spoke.



Herman

Correct. Uncertainty can appear in the probabilities across those classes. Overlap describes an actual condition. The learned transformations adjust hidden representations according to those probabilities, giving the recognizer information about how to treat each stretch.



Corn

It gets an instruction to ignore a voice without needing to manufacture a pristine recording of the other one.



Herman

That's the attraction. It learns relative distinctions between the designated target and other speech, rather than having to learn every possible speaker identity as a separate thing. Enrollment-based systems are another route, where a short recording of the user's voice becomes an embedding that steers the model. That deserves its own conversation; for today, the point is that something has to tell the recognizer who the target is.



Corn

Enrollment sounds closer to Daniel's eventual product. He wants his own dictation, not a meeting split into anonymous people.



Herman

An enrollment signal could help the upstream system associate Daniel with activity frames. Diarization alone doesn't automatically know which cluster is the phone's owner. You still need that association. The conditioned recognizer then consumes the activity mask. These are connected jobs, not interchangeable ones.



Herman

Would you accept enrollment if it meant another model running on the device?



Corn

If it actually prevents substitutions, yes. I'd resist a permanent enrollment ritual, but an occasional voice sample seems cheaper than checking every sentence for a stranger. I'd want to know whether it works while the street changes, not merely whether it remembers me in a quiet room.



Herman

That's where the distinction between personal adaptation and target selection pays off. They can reinforce each other, but a fine-tune full of Daniel's clean speech doesn't demonstrate target selection under overlap.



Corn

You're looking pleased with the tiny bias again.



Herman

It is an elegant intervention.



Corn

You have mentally installed it already.



Herman

Only in the encoder. I've left the rest of our afternoon untouched.



Corn

Generous.



Herman

The ablation makes the data question concrete. On a Whisper English meeting evaluation, full STNO conditioning gave about twenty-seven percent word error. Target-only conditioning was about thirty-five percent. The extra distinctions helped considerably.



Corn

Silence and other people's uninterrupted speech suddenly look less like filler in a training set.



Herman

Removing the explicit silence category gave thirty percent. Another reduced representation, target and non-target, came in just under twenty-nine. The representation matters, not just whether a human could reconstruct similar information from it. Explicit silence gave the full configuration a useful advantage.



Corn

I wouldn't claim that proves every model needs the same class scheme. But it does argue against filling Daniel's entire set with heroic mixtures. Some examples need to say there's nothing to transcribe, even while another person is clearly audible.



Herman

Target-absent material is especially important to the objective. Otherwise a recognizer can learn that some transcript is always expected. Silence frames help establish a different boundary: speech isn't merely difficult here, it's absent.



Corn

Finally, a legitimate training use for saying nothing.



Herman

Initialization is the nastier detail. Randomly initialized diagonal transformations produced word error above one hundred percent. Starting near identity gave roughly twenty-eight percent, and a suppressive initialization roughly twenty-seven.



Corn

The upgrade has deleted competence.



Herman

Word error can exceed one hundred because insertions add errors beyond the number of reference words. Those random transformations disrupted representations the pretrained model already knew how to use. Identity initialization initially preserves them. Suppressive initialization starts with a structured bias toward reducing unwanted contributions.



Corn

More flexible transforms might let it relearn faster, though.



Herman

In those experiments, full non-diagonal matrices degraded performance. More freedom was not a free improvement. Small, carefully initialized changes preserved the pretrained system's strengths better.



Corn

I'll concede the elegant mechanism. I retain my objection to your installation timetable.



Herman

The Parakeet port is already real. DiCoP, from the Brno speech group, uses Parakeet TDT version two as its base, about six hundred million parameters, with a FastConformer encoder and TDT decoder. Its mask advances at eighty-millisecond intervals. Each encoder layer gets a small class-dependent transformation.



Herman

A meeting can be decoded once per target speaker without a separate speech-separation front-end. No speaker embedding goes into that recognizer, but it absolutely does require the target activity mask.



Corn

That's a checkpoint in the same family as Daniel's model, not a drop-in guarantee for his particular version or his phone.



Herman

With oracle diarization, meaning supplied reference-quality speaker activity, it achieved about two and a half percent speaker-attributed word error on one synthetic two-speaker test. On Libri three Mix it was about thirty-three percent. Those are different datasets, so the gap isn't a controlled experiment about adding one speaker.



Corn

The easy number needs its conditions attached.



Herman

Real meeting tests were around sixteen to eighteen percent. The related Whisper system also beat a separation-and-diarization cascade on its meeting benchmark, so conditioned recognition can be competitive without reconstructing isolated waveforms. But deployment includes generating the masks, not just running the recognizer.



Corn

A small encoder addition can sit behind a rather substantial identification problem.



Herman

The DiCoW card explicitly warns that performance depends heavily on diarization quality, and more than two fully overlapping speakers can be difficult. Incorrect masks can tell the recognizer to suppress Daniel exactly when he's speaking. For his dataset, accurate labels carry much of the value.



Corn

The announcer's sentence isn't the only thing that can be confidently wrong. The upstream system can mislabel ownership before the transcription begins.



Herman

Music remains outside the reported DiCoP evaluations. Meetings and synthetic speech mixtures don't establish robustness to lyrics competing with dictation. Daniel's examples could reveal something useful there, but neither a good meeting score nor these small transformations settle that case.



Corn

Before paying for a bigger recognizer, I'd check where the microphone...



Hilbert

Has put your neighbour. Mine puts him in the document through a wall, Corn. I say, please amend the loading bay hours, and it writes his request for another parking permit. I'm less interested in what language he's using than why his application gets completed while mine doesn't.



Hilbert

This Thursday I'm going to the fictional municipal dictation trial at the permit office. Mr. Barak has agreed to sit behind the partition. I've booked the adjoining booth for the neighbour, who will be represented by a silent cardboard cutout. The recognizer has a favourite. I intend to give its favourite an appointment.



Corn

The cutout will need the correct permit category.



Hilbert

Visitor. I've already put that on the booking. I repeated please amend the loading bay hours seventeen times during the last test. Then I held the desk microphone against the partition to get its attention. That's when the neighbour became perfectly clear. My snout doesn't fit the booth's headset, so the desk microphone is what I have to work with.



Herman

Against the partition would change the sound reaching it. The loudest person in the room needn't provide the strongest voice at the microphone.



Hilbert

That's the little distinction I meant, Herman. At the microphone. You explained the frequency bands carefully earlier, then you both kept saying loudest. I was louder. He was beside the cable opening. For Thursday, Mr. Barak wants a recording of the whole appointment. The office's test page only saved the transcript last time. We couldn't play the mistake back.



Corn

Seventeen repetitions and no appeal recording. Thorough paperwork.



Hilbert

I've asked him to keep the audio this time. He sent me a form authorizing the recording. The appointment is eleven forty.



Herman

Having the original audio lets you distinguish a target-selection mistake from an invented sentence. If only the transcript survives, you lose evidence about what the recognizer actually heard.



Corn

It also stops us arguing from a memory of being louder. For Daniel's tests, the recording preserves whether the announcement actually covered the word or merely distracted him while he said it.



Herman

There's a common misconception worth killing here: enough noisy examples will automatically teach a recognizer whose voice matters.



Corn

No. They may improve robustness, but reliable target selection needs a target signal and training that uses it. Lyrics add a separate question about unwanted vocals and invented text. Neither is settled by making the training audio louder.



Herman

The experiment I'd watch is what happens when accurate target labels become imperfect street labels, with the original audio retained so wrong ownership and invented words can be told apart.



Corn

Thanks to Hilbert Flumingtop for producing the show.



Herman

If this was your kind of episode, go back for episode seven, Building Custom ASR Tools; and episode ten, How ASR Went From Frustration To ... Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

You'll find the show at my weird prompts dot com.



Herman

See you tomorrow.