Episode Book

EPISODE #6364

Collecting Speech Data on Outdoor Walks

PUBLISHED	30 October 2026 at 11:33 UTC
DURATION	22:13
LISTEN	myweirdprompts.com/episode/outdoor-speech-model-data-collection-01949/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.25
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	31 October 2026 at 05:58 UTC

SYNOPSIS

Fine-tuning NVIDIA's Parakeet speech model for everyday use requires gathering training data that matches real deployment conditions, especially walking outdoors. Movement alters breathing, phrasing, and microphone position in ways background noise alone cannot replicate. Examining input workflows reveals the trade-offs between handheld game controllers, smart glasses like the Even Realities G2, and synthesized voice prompts. Keeping the recording interface lightweight ensures consistent vocabulary and code-switching data without administrative friction.

TRANSCRIPT



Corn

Holding the phone higher doesn't solve it, Herman. It just moves the lamppost out of frame. Daniel's been improving his personal Parakeet speech model, and now the friction is collecting the next batch of recordings. He wants to read training sentences while walking, because walking is when he actually dictates.



Herman

His benchmarks are improving, and he's hearing the improvement in everyday use. That's a useful combination. It gives him a reason to keep recording sentences even when the recording process starts feeling like homework.



Corn

He's already got a custom app Claude helped build, and a little controller for starting, stopping, retaking, or skipping recordings. That works pacing indoors. Outdoors, he's still staring at the sentence on his phone. He's wondering whether glasses could put it in his vision, or whether spoken prompts could get the screen out of the process entirely.



Herman

The buttons can stay almost exactly as they are. We need to get the sentence somewhere he can comfortably perceive it.



Corn

Preferably somewhere that also contains the pavement.



Herman

With a workflow that doesn't require him to remember whether the last button saved the recording or discarded it. Moving the display won't help if every sample ends with a small administrative crisis.



Corn

The baseline matters here. He's not repairing an unusable recognizer. He's taking a strong generic one and teaching it the vocabulary, language switches, and preferences that keep recurring in his own speech.



Herman

NVIDIA's Parakeet TDT point six billion version three, with six hundred twenty-seven million parameters. It's available under the Creative Commons Attribution four point zero license. The architecture isn't the obstacle in this question. The obstacle is obtaining more of the particular speech that the base model doesn't get quite right.



Corn

A clean recording of Kubernetes is useful. A slightly breathless recording of Kubernetes beside a bus might be more relevant to Tuesday's actual work.



Herman

Breath timing changes with light exercise. You pause differently, shorten phrases, move the microphone, and sometimes turn your head halfway through a word. Traffic adds competing sound, but movement changes the speech itself. Adding a traffic recording to clean indoor speech doesn't reproduce all of that.



Corn

Nor does sitting on a bench reproduce the walking part. Daniel's bench recordings are already getting the street into the dataset. Now he wants the body doing the same thing it does during deployment.



Herman

Would you keep recording indoors too, Corn, or move the whole collection outdoors once the workflow worked?



Corn

Keep both. Otherwise he could teach it that his voice always arrives accompanied by a diesel engine. More seriously, he still dictates indoors. I'd label the conditions rather than replace one collection with the other.



Herman

I'd do that too. Quiet, outdoor stationary, outdoor walking. Simple session labels, not a weather diary. Then if one training run improves street dictation and harms quiet dictation, he can see the distinction.



Corn

Your weather diary would have microphone humidity.



Herman

Only if the microphone were doing something suspicious.



Corn

There it is.



Herman

The immediate input device is the 8BitDo Micro. In keyboard mode, its buttons send ordinary keystrokes. His recording app doesn't need a special concept of game controller. It needs reliable actions bound to keys.



Corn

Sixteen buttons to avoid touching four buttons.



Herman

He doesn't need to use all of them. The important detail is the mode switch. It has S, D, and K positions. K is keyboard mode, and that's the mode that supports remapping through their software. Gamepad mode and keyboard mode aren't interchangeable just because both connect over Bluetooth.



Corn

I'd give the four actions different physical positions, then stop adding features. Recording tools become miserable when every useful action has a long press, a double press, and an alternate meaning if you're already recording.



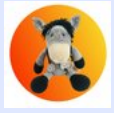
Herman

A single start and stop toggle could even reduce the active controls. Although... no, separate actions may be clearer if he's already comfortable with them. The app should acknowledge the result, not just acknowledge that it received a button.



Corn

A vibration for recording started, another for saved, something unmistakably different for discarded. The phone can remain in his pocket. I don't want a cheerful beep that means either success or Bluetooth disconnected.



Herman

That's a state problem worth solving before buying a display. His app should always know which prompt is active, which take belongs to it, and whether that take is committed. A retake should preserve the sentence and replace the attempt, not silently advance.



Corn

Walking is an excellent way to discover that your interface assumes you're watching it.



Herman

What would you tolerate as an audio prompt? Full sentences, short phrases, or just the technical term?



Corn

Short phrases. A long spoken sentence becomes a memory test, and I'd start paraphrasing it. Which might be perfectly good speech, but then the displayed sentence is no longer the correct transcript.



Herman

That's the hidden cost. Prompted collection is attractive because the text is already known. If he hears a sentence, remembers the gist, and says something slightly different, the app must capture that difference in the label. Otherwise he's collecting natural speech with inaccurate transcripts.



Corn

The model gets marked wrong for hearing what he said.



Herman

TTS still deserves a trial. Daniel's objection about unfamiliar words is valid for an unconfigured voice. Pronunciation dictionaries let him specify how a technical term should be spoken. ElevenLabs supports dictionaries, Google Cloud has custom pronunciation controls, and NVIDIA's TTS tooling supports pronunciation overrides too.



Corn

Before we solve the display problem, let's take a short break.

L

Larry

There is a moment, just before the words arrive, when the street goes quiet and the model listens. That moment has a name. It is called Parakeetene. Parakeetene is not a device. It is not an app. It is a feeling, pressed into a small glass bottle and sold at a reasonable

L

Larry (continued)

price. One drop on the wrist, and every sentence you speak becomes a benchmark. Every lamppost becomes a training sample. Whisper the word, and the background noise of an entire city will arrange itself politely into columns of clean data, forty percent quieter, ninety percent more accurate, and two thousand and one percent more you. Parakeetene. Ask for it where fine tuning is sold, and walk on, unbothered, into the beautiful noise.



Herman

I already own a bottle. It gave me a rash shaped like the edge of my watch. Pronunciation overrides mean Daniel could author the troublesome term once, then reuse that entry across many spoken prompts.



Corn

That fixes the robot's pronunciation. It doesn't necessarily preserve his.



Herman

Particularly at a language boundary. An English technical noun inside a Hebrew sentence may acquire different stress or rhythm from the same noun in an English sentence. A synthetic voice can pronounce every individual word acceptably and still give him an unnatural example of the whole utterance.



Corn

I suspect I'd imitate it without meaning to. Not the voice, but where it pauses. If it reads a sentence as a demonstration, I'm going to deliver the demonstration back.



Herman

That's a plausible concern, not a reason to ban the method. He could compare a small batch of audio-prompted recordings with visually prompted ones. Listen for changes in cadence, pronunciation, and sentence fidelity before collecting hours.



Corn

You were about to configure an entire voice service before trying ten sentences.



Herman

Twenty sentences. I do have restraint.



Corn

Hm.



Herman

The distinction is between TTS as a cue and TTS as the training audio. Several code-switching projects use synthetic audio directly to supplement scarce recordings, and they can get meaningful recognition improvements. Daniel would be using it to elicit real speech, which is a different intervention.



Corn

His street recordings contain things the synthetic collection doesn't naturally give him. His own breath, his own mouth, and the exact way he abandons a sentence when a motorcycle passes.



Herman

Yes, though abandoned sentences need their own handling. If the task expects a complete prompt, a clipped or interrupted attempt shouldn't quietly receive the complete sentence as its transcript.



Corn

I'd make interruption cheap. Retake shouldn't feel like failing an exam. If recording outdoors requires pretending the outdoors isn't happening, the collection process will lose.



Herman

There's a relevant medical code-switching dataset called ViMedCSS. It contains Vietnamese speech with English medical terms embedded in the utterances. Domain adaptation and multilingual pretraining help different portions of that speech. It reinforces the value of checking the inserted terms separately from the surrounding sentence.



Corn

Which also means collecting nothing but exotic technical nouns would be a mistake. He needs the ordinary speech that joins them together.



Herman

Ordinary speech, code-switched speech, quiet speech, walking speech. Not necessarily equal amounts. Representative amounts, adjusted when evaluation shows a specific weakness.



Corn

I'd try another audio route too. Record a natural thought while walking, then correct its transcript later. No sentence to see, no synthetic voice to imitate.



Herman

That gets him spontaneous dictation. It also brings back transcription work. Would you actually do the correction afterward?



Corn

For a small supplementary batch, yes. For every walk, no. I know precisely how much future Corn enjoys administrative tasks.



Herman

Then a hybrid is sensible. Prompted sentences for deliberate vocabulary coverage, spontaneous recordings for natural phrasing. An audio cue could name a topic or term rather than dictate the sentence, provided he accepts that the transcript must be checked



Herman (continued)

afterward.



Corn

For the exact-sentence workflow, glasses remain the cleaner idea. They preserve the text without putting another speaker's delivery in his ear.



Herman

Display glasses with teleprompters are shipping. Even Realities G2 is one example. The glasses cost five hundred ninety-nine dollars, and the optional R1 ring costs two hundred forty-nine. The ring lets you scroll while your hand stays at your side.



Corn

Two hundred forty-nine dollars to move a sentence.



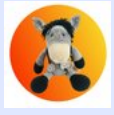
Herman

You're about to defend his existing controller.



Corn

It already knows where his thumb is.



Herman

I would keep it too, initially. A ring is attractive ergonomically, but only if its events can reach the recording app. Buying a second input system before knowing whether it integrates would add uncertainty without solving the display connection.



Corn

Thank you. I was prepared to be unreasonable for longer.



Herman

The G2 glasses weigh thirty-six grams and have neither a camera nor speakers. That's relevant. He'd be buying a text display, not a compulsory new recording chain. The phone and his chosen microphone could keep doing their existing jobs.



Corn

That separation appeals to me. I don't want glasses deciding that their convenient microphone is now the dataset microphone. He's trying to improve recognition for the setup he actually uses.



Herman

Advertised display brightness is twelve hundred nits. The field of view is twenty-seven point five degrees. Those specifications tell us something about the display, but they don't settle whether small green text remains comfortable against a bright street.



Corn

A number measured on the display doesn't describe the building behind it.



Herman

Contrast, fit, prescription needs, and where the text sits all matter. My ears would also have objections to an untested frame. I'd want Daniel to try a sentence in actual daylight, not just admire a floating menu indoors.



Corn

Another candidate is Vuzix Z one hundred. Different display, same integration question?



Herman

Five hundred dollars, approximately. Well, four hundred ninety-nine. It weighs thirty-eight grams and uses a monochrome waveguide. Its safety-glasses certification concerns impact protection. It doesn't certify that the wearer is paying attention.



Corn

The lamppost receives no corresponding upgrade.



Herman

Halliday G2 puts its display above the normal eyeline. That placement might suit intermittent glances better than a sentence directly over the walking scene. It lists sixteen hundred nits of brightness and costs five hundred ninety-nine dollars.



Corn

I'd prefer a short sentence appearing only when requested. Not a paragraph permanently occupying the sky. Read it, let the display disappear, record, then bring up the next one.



Herman

Would you want it disappearing automatically?



Corn

No. I'd want a predictable action. Automatic disappearing is wonderful until I haven't finished reading. Maybe the recording button also hides the prompt, and stopping brings back the result state. But keep an option to leave the sentence visible if I need it.



Herman

That gives us two sensible modes. Glance and recall for short prompts, persistent display for exact reading. He can compare them rather than assume a heads-up display must be continually watched.



Corn

Glasses move the screen. They don't move the attention back to the street.



Herman

Teleprompters are mostly designed around someone presenting a prepared script. Their scrolling features can be very sophisticated. Rokid describes speech-following scroll, for example. But following a presentation isn't the same as recording discrete labelled samples.



Corn

Especially when the sentence contains the word the recognizer already struggles with. I don't want a recognizer deciding whether I deserve the next prompt.



Herman

Speech-following could help pace text, but it should not decide sample acceptance. The recorder needs explicit ownership of prompt identity, recording boundaries, and save status.



Herman

Engadget's G2 review was titled Smart Glasses, Stupid Controls. That captures a practical risk with buying from the feature list. Having a teleprompter doesn't establish that its controls will cooperate with Daniel's app.



Corn

A static script import could still support a crude trial. Put the same ordered sentences on the glasses and in the recorder. Advance both manually and see whether reading feels tolerable.



Herman

Useful as an ergonomics trial. Fragile as a collection system. One skip or retake can put the two sequences out of step. Then the audio may be fine while the sentence attached to it is wrong.



Corn

I'd put the prompt identifier beside the sentence in that trial. Not read it aloud, just make mismatches visible. It would still need checking afterward.



Herman

For the real implementation, his app needs to push the current sentence to the glasses and ideally receive confirmation that it arrived. I haven't seen that dynamic-text path confirmed for these particular teleprompter workflows. A static import feature doesn't prove an API exists, and lack of documentation doesn't prove there's no developer route.



Corn

That's the purchase gate, then. Not whether a reviewer enjoyed delivering a speech. Whether Daniel's app can replace the sentence on demand.



Herman

The phone should remain the authority. The glasses display the prompt, the controller sends actions, and the recorder commits audio and text together. If the display disconnects, it should pause advancement rather than keep silently generating a dataset.



Corn

A disconnected display with an enthusiastic recorder would produce excellent samples of confusion.



Herman

There's plenty of teleprompter software and plenty of dataset tooling. There are tools that validate recordings and produce NeMo-friendly manifests. I don't know an established walking-first application that joins this particular input, display, and labelling workflow.



Corn

He doesn't need to rebuild all of those things. The promising project is a narrow bridge from his current app to a display, with the existing controller retained.



Herman

Provided the display connection is available. Otherwise I'd start with configured audio prompts, short enough to remember, and leave a clear gap before recording. Don't let the synthetic prompt leak into the microphone and become part of the sample.



Corn

Headphone playback rather than the phone announcing Kubernetes to the street.



Herman

Then an explicit recording cue after playback has finished. The app should make those separate states, rather than depend on an estimated silence.



Corn

The controller can keep the whole sequence boring.



Hilbert

Boring is generous, Corn. I tried the controller on this desk while you were discussing pronunciation dictionaries. The direction buttons did nothing. I rang the number on a repair receipt from a man I met on the bus to Haifa. Avi. I still think of him when a Bluetooth device connects successfully and does nothing useful.



Herman

Connected and configured are separate achievements.



Hilbert

He told me to hold Minus and up for five seconds. That enables the directional control. Minus and left selects left-stick control instead. I used the edge of a pencil because the buttons are close together under my claws. It worked. Herman, you were very careful about the mode switch earlier. This is another switch, except they haven't provided a switch.



Corn

A button combination wearing the switch's clothes.



Hilbert

I had one direction set to advance the sentence. Once it worked, I advanced twice because I was still holding it. Then I rang my nephew, who repairs handheld controllers. He says people replace working controllers because a mode has changed. He also says everyone should buy the replacement through him, so he's not to be trusted. His opinion is that the ordinary face buttons are easier for this job. He went on about rubber contacts. I put him on speaker and turned it down.



Herman

If Daniel uses the direction pad for retake or skip, that activation is worth checking. He hasn't said those buttons are malfunctioning, though. His demonstrated problem is still reading the phone.



Hilbert

I know. I tried the face buttons after that. Start, stop, next sentence. Avi stayed on the call until the first recording saved. Eleven minutes, including the nephew.



Corn

Eleven minutes is a much better experiment than buying glasses and discovering their prompter only accepts an imported document.



Herman

It also suggests the order I'd actually recommend. Keep the input hardware, make the app's states audible or tactile, and trial short audio prompts with pronunciation overrides. In parallel, ask the glasses vendor to demonstrate dynamic text arriving from another app.



Corn

Not a sales promise. A sentence sent, then replaced, then replaced again after a retake. If that works, a borrowed pair can answer the outdoor readability question before the expensive part.



Herman

I'd also want the phone locked and pocketed during that demonstration. A workflow that only works while its app is open under somebody's attentive finger hasn't passed Daniel's actual test.



Corn

The idea that sticks for me is that his walking changes the speech he wants to recognize. Getting the phone out of his face is what lets him collect that speech, rather than merely collect traffic noise.



Herman

The label still has to match what he actually says. Any prompting method earns its place by preserving that relationship without making the walk an interface exercise.



Corn

The open question is whether a glasses maker can make changing a line of text as straightforward as sending a keyboard event. That would turn a short controller test into a usable collection workflow.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you want more of this, try episode nine, Benchmarking Custom ASR Tools - Beyond The WER; episode seven, Building Custom ASR Tools; and episode five, Fine-Tuning ASR For Maximal Usability. This has been My Weird Prompts, the human-AI collaboration podcast. No question too obscure, no rabbit hole too deep.



Herman

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



Corn

See you tomorrow.



Herman

Until then.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **NVIDIA Parakeet TDT 0.6B v3 model card, created 2025-08-04 [primary]**
https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3

2. **MLX community port, created 2025-08-16 [primary]**
https://huggingface.co/mlx-community/parakeet-tdt-0.6b-v3

3. **GGUF conversion, created 2026-05-10 [primary]**
https://huggingface.co/handy-computer/parakeet-unified-en-0.6b-gguf

4. **8BitDo Micro product page [primary]**
https://www.8bitdo.com/micro/

5. **8BitDo Micro FAQ (keyboard mode, D-pad activation) [primary]**
https://support.8bitdo.com/faq/micro.html

6. **8BitDo Ultimate Software for Micro (button mapping, keyboard-mode-only) [primary]**
https://support.8bitdo.com/ultimate/micro.html

7. **Ultimate Software V2 (key mapping, macros) [primary]**
https://app.8bitdo.com/Ultimate-Software-V2/

8. **Meta, AI glasses for teleprompting [primary]**
https://www.meta.com/ai-glasses/learn/ai-glasses-for-teleprompting/

9. **Rokid AI Teleprompter (smart-scroll) [primary]**
https://global.rokid.com/blogs/academy-glasses/3-1-ai-teleprompter

10. **INMO GO3 script import [primary]**
https://www.inmo.com/blogs/inmo-insights/teleprompter-glasses-import-scripts-from-phone

11. **RayNeo iO teleprompter / Smart Crown [primary]**
https://www.rayneo.com/blogs/news/smart-glasses-live-captions-meetings-teleprompter

12. **MemoMind One teleprompter [primary]**
https://www.memo-mind.com/blogs/in-the-moment/speak-without-notes-with-memomind-ai-teleprompter

13. **ElevenLabs pronunciation dictionaries [primary]**
https://elevenlabs.io/docs/eleven-agents/customization/voice/pronunciation-dictionary

14. **Google Cloud TTS `customPronunciations` [primary]**
https://docs.cloud.google.com/text-to-speech/docs/reference/rest/Shared.Types/StreamingSynthesizeConfig

15. **NVIDIA TTS NIM customization (SSML, pronunciation dictionaries) [primary]**
https://docs.nvidia.com/nim/speech/latest/tts/customization.html

16. **Best AI Glasses with a Teleprompter (2026), Shade & Signal**
https://shadeandsignal.com/best/best-ai-glasses-with-teleprompter

17. **Engadget, Even Realities G2 Review: Smart Glasses, Stupid Controls, 2026-10-19**
https://www.engadget.com/2288658/even-realities-g2-smart-glasses-review/

18. **Soniox wiki, pronunciation control via SSML/phonemes**
https://soniox.com/wiki/pronunciation-control-ssml

19. **HiKE Korean-English code-switching benchmark, EACL Findings 2026**
https://arxiv.org/abs/2509.24613v4

20. **ViMedCSS Vietnamese medical code-switching dataset, LREC 2026**
https://arxiv.org/abs/2602.12911v2

21. **TTS data augmentation for code-switching ASR, APSIPA 2025**
https://arxiv.org/abs/2601.00935v1

22. **Code-mixing guided synthetic speech, Interspeech 2026**
https://arxiv.org/abs/2606.19381v1

23. **AsyncSwitch, ASRU 2025**
https://arxiv.org/abs/2506.14190v1

24. **prompted-recording HF/NeMo manifest tool, created 2026-07-14**
https://github.com/andrewferk/speech-dataset-workbench

25. **Telugu ASR corpus recording tool**
https://github.com/iamprashanth238/speech-data-telugu-uoh