# Episode Book

EPISODE #6023

# How Much Voice Data Is Enough?

| | |
|---|---|
| PUBLISHED | 16 October 2026 at 10:49 UTC |
| DURATION | 20:14 |
| LISTEN | myweirdprompts.com/episode/asr-fine-tuning-saturation-curve-99169/ |
| HOSTS | Corn, Herman, Larry, Hilbert |
| SCRIPT MODEL | deepseek-flash |
| VOICES | chatterbox-regular (chatterbox-local) |
| PIPELINE | V5.5 |
| SOURCES | 11 cited by the research (listed at the end) |
| BOOK GENERATED | 19 October 2026 at 17:42 UTC |

## SYNOPSIS

Daniel fine-tuned NVIDIA Parakeet V2 on two hours of LLM-generated sentences recorded on his phone, and Hebrew-English code-switching recognition jumped from roughly 20% to 60%. Now he wants to record fifteen minutes every morning for weeks — so the real question is where the fine-tuning curve flattens. The literature disagrees: a 2023 Singapore study on Mandarin-English code-switching suggests saturation can arrive as early as one to ten hours, while a Maltese study that trained thirty models found no diminishing returns for Whisper. A single-speaker Swedish study shows most of the gain captured by four hours, with a long shallow tail after. This episode digs into the contested research, what extra hours actually buy, why transcription noise may matter more than fresh clips, and whether a dataset built from LLM-written sentences is a diet version of how you really speak.

# TRANSCRIPT



**Corn**

Most coverage of ASR fine-tuning leads with the same line: more data, better accuracy, keep recording, the curve keeps climbing.



**Herman**

And that's true right up until it isn't.



**Corn**

Daniel published a fine-tune of NVIDIA Parakeet V2 yesterday. English-Hebrew code-switching. The kind of speech where you're speaking Hebrew and you drop in "teudat zehut" without breaking stride. He trained it on two hours of short sentences, LLM-generated, recorded on his phone, and code-switched Hebrew recognition jumped from about twenty percent to about sixty.



**Herman**

Against Whisper.



**Corn**

That's the comparison, yeah. And now he wants to take it gradual. Fifteen minutes every morning for a few weeks. He built a capture tool that skips to the next prompt, pushes to Hugging Face every fifty clips, warns him when he speaks too close to the end. He can skip, delete takes, retake. He calls yesterday proof of concept, which is about right.



**Herman**

The seventy-minute voice attack thing. He's said before that voice input is how he actually works, so every point of accuracy on this compounds across everything he does.



**Corn**

Which is why the two questions matter. One: at what point does the curve flatten? Three more hours, six, twelve, twenty-four — when does it stop really mattering? Two: he thinks the dataset is powerful because it should be usable to fine-tune any model. So does he need to change how he records — longer sentences, some mix — to keep it portable? Or does the data have to map to the model he's actually targeting?



**Herman**

Both questions have real answers in the literature. And the first one is contested, which I find more interesting than if it weren't.



**Corn**

Let's hear the contested part.



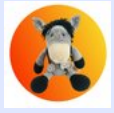
**Herman**

The closest published analogue to what Daniel is doing is a 2023 study out of Singapore that adapted Whisper for code-switched Mandarin-English. Same structural problem, two languages interleaved mid-sentence. Their finding, and I'll quote it, because the wording matters: "The amount of adaptation data may be as low as one to ten hours to achieve saturation."



**Corn**

One to ten.



**Herman**

For code-switching adaptation specifically. Daniel is at two hours.



**Corn**

So the literature says he might already be on the flat part.



**Herman**

Might. That's one study, one language pair, one base model. And the counterpoint is just as strong, because the Maltese study trained thirty models on splits from ten minutes up to fifty hours, and for Whisper specifically, diminishing returns were not observed. The authors say the results indicate we should revisit both models and train further.



**Corn**

So one paper tells him he's done, the other tells him keep going.



**Herman**

And neither of them trained on Hebrew.



**Corn**

What's the most granular data point? For a single speaker, personal voice, exactly his setup.



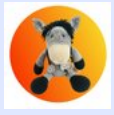
**Herman**

There's a Swedish study from 2022, single-speaker adaptation of wav2vec 2.0 to one person's voice. The word error rate starts at five point four percent. At two hours of fine-tuning data, it drops to two point five seven. At four hours, one point four seven.



**Corn**

Steep.



**Herman**

Then ten hours gets you one point three eight. Twelve hours, one point two two. Fourteen hours, one point one three. Sixteen hours, one point zero three.



**Corn**

So between four and sixteen hours, he's grinding for four tenths of a point.



**Herman**

Most of the gain is captured by four hours. The paper's own conclusion is that twenty to thirty minutes already gives a notable drop, and three to four hours gets you most of the way.



**Corn**

That's the honest shape of the thing. Steep, then a long shallow tail.



**Herman**

Now the big-language numbers, because they're useful for contrast even though they're not the same task. A study on Kinyarwanda ran Whisper scaling experiments from one hour to one thousand four hundred hours. Usable performance — word error rate below thirteen percent — is achievable with about fifty hours. Continued gains up to two hundred hours.



**Corn**

That's a full language from a low-resource starting point. Not adaptation, from-scratch territory.



**Herman**

Right. NVIDIA's own practitioner guidance splits it the same way. Fine-tuning an existing acoustic model is recommended when you have ten to a hundred hours of transcribed domain audio. Training a new acoustic model from scratch needs thousands.



**Corn**

And the real-world Hebrew Parakeet comparison.



**Herman**

There's a public one. Daniel Mamre's Hebrew Parakeet fine-tune. Version one trained on about two hundred fifty-four hours of Hebrew speech, twenty-one point six percent word error rate. Version two, about seven hundred sixty-one hours, nineteen point four percent.



**Corn**

Five hundred extra hours for two points.



**Herman**

That's the tail of the curve in one number pair. It's not that more data stops helping. It's that each additional hour is buying less than the last one, and at some point you have to ask what you're buying it with.



**Corn**

What's it buying it with? His mornings.



**Herman**

Fifteen minutes a day for a few weeks is maybe two, three, four more hours. That's one extra day of his life for something the Mawlid study — the Maltese study — says isn't at the point of futility, and a Dutch Whisper fine-tuning write-up calls marginal gains.



**Corn**

Same data, different words.



**Herman**

Same data, different optimization target. If you're shipping a product where half a point of word error rate is a customer-visible improvement, marginal isn't marginal. If you're a guy recording himself in a kitchen at seven in the morning, half a point is a lot of mornings.



**Corn**

And the accuracy jumps don't behave like the error rate jumps. Going from twenty percent to sixty is a different world. Going from sixty to sixty-five is a nice day.



**Herman**

That's the right way to feel about it. The first few hours are buying function. The later hours are buying polish.



**Corn**

There's something in the research about transcription noise being a big chunk of the remaining errors.



**Herman**

Thirty-eight point six percent of high-error cases in the African-languages study came from transcription noise, not from the model. Mislabeled clips. For Daniel, whose corpus is LLM-generated sentences — if the LLM produced something that's awkward or wrong in Hebrew, or he misread it, or the recording has a stumble he didn't catch, that clip is teaching the model something false. And a false label is worse than a missing one.



**Corn**

His tool warns him if he speaks too close to the end, and lets him retake. The discipline is already there. The question is whether he's auditing the transcripts themselves.



**Herman**

If he's ever going to spend forty hours recording, an hour spent checking the labels is worth more than another hour of fresh clips.



**Corn**

Hold on. Something just occurred to me. He said the sentences are LLM-generated.



### Herman

So?



### Corn

So if the LLM writes the sentences and Daniel reads them, then Daniel's output distribution is the LLM's distribution. If the LLM likes "booking calendar appointments," his corpus likes booking calendar appointments. His kitchen recording sessions are training a speech model on what a language model thinks people say.



### Herman

That's a real hazard. And I don't have a clean answer for it either.



### Corn

I don't need one. I just want him to notice it.



### Herman

Worth saying, and it cuts the other way too. LLM-generated sentences give you clean, grammatical, correctly punctuated targets — which you don't get from reading Wikipedia aloud or from a broadcast corpus. So the LLM is doing real work. The risk is narrowness.



### Corn

Fine. Narrower than his actual speech, then.



### Herman

That's the test. And there's a cheap way to check it. He could take a week of his own sent messages, his own voice notes, the actual Hebrew he types when he's not thinking about a corpus, and see how much of it the LLM sentences resemble. If the overlap is high, he's fine. If his real speech is full of half-sentences and loanwords the generator never produces, the corpus is a diet version of his language.



**Corn**

Diet Hebrew.



**Herman**

Which fine-tunes beautifully and then meets his actual mouth.

L

**Larry**

Close your eyes. Breathe in. Somewhere, a door is opening. And on the other side of that door is a voice. A voice that knows your name. A voice that knows what you meant to say, before you said it. It doesn't guess. It doesn't mishear. It simply... understands. Introducing the Lexicon Nebula. Not a speech model. A relationship. Available in three weights. The lightest one is forty-seven gigabytes. Lexicon Nebula. Hear what you meant.



**Corn**

So the question isn't "will it be worth it," because if he keeps the capture pipeline tight it probably will.



**Herman**

It's whether the return per morning is still worth the morning.



**Corn**

Portability. He asked whether the dataset maps to the model he's targeting.



**Herman**

The clean answer is at the audio layer, no, it doesn't have to map. At the formatting layer, yes, it does.



**Corn**

Unpack that.



**Herman**

The WAV file and the transcript that goes with it — that pair is the asset. The Maltese corpus was fine-tuned into both XLS-R and Whisper. Same splits, same audio, two architectures. The Frisian thesis did Whisper with both LoRA and full fine-tuning on one dataset. So the transcribed audio is portable across model families in practice, not just in theory.



**Corn**

What isn't portable?



**Herman**

The wrappers. Parakeet wants NeMo manifests — JSON lines with the audio path, the text, the duration. A byte-pair tokenizer underneath. Whisper chops audio into thirty-second windows and wants its own format on top. So the same clip has to be described differently for each training run. That's transcription-time work, not recording-time work.



**Corn**

And the clip length?



**Herman**

Clip length is where this gets nice for him. Whisper's thirty-second window is the binding constraint — anything longer has to be chunked, and chunking kills your alignment. The Maltese corpus deliberately clipped phrases to ten seconds. Daniel's clips are ten to twenty seconds.



**Corn**

Inside the window with room.



**Herman**

Comfortably. So his current recording approach is already in the portable sweet spot. Ten to twenty seconds is short enough for Whisper, long enough to carry a code-switching boundary, and fine for Parakeet's manifest format. He doesn't need to change anything to keep his options open.



**Corn**

On longer sentences specifically — he floated the idea of adding some to the distribution.



**Herman**

The literature doesn't show them helping cross-model versatility. Nobody found longer utterances giving you a better portable dataset. Thirty seconds is the ceiling and longer just breaks it. So the answer to that part of his question is, no, don't bother going longer. If anything, shorter is more portable.



**Corn**

So what does he actually gain from the portability?



**Herman**

Experimental freedom. He can record once — same fifteen minutes a morning, same capture tool, same short clips — and then run the same corpus against Whisper, against Parakeet, against whatever Whisper successor ships next. Each of those is a fine-tuning run, not a re-recording. The dataset is an asset he's building once.



**Corn**

And the Parakeet-specific complication?



**Herman**

Parakeet TDT 0.6B V2 is English-only. V3 is the multilingual one, twenty-five European languages, and Hebrew isn't in there — no Semitic languages at all. So a Hebrew fine-tune isn't a bake, it's a tokenizer extension and decoder reinitialization. More surgery than fine-tuning.



**Corn**

That's why the twenty to sixty jump is interesting.



**Herman**

That's exactly why. He's not just adapting Parakeet to his voice, he's teaching an English model to produce Hebrew tokens. Two hours getting him most of the way to sixty is a strong result.



**Corn**

So the portability question has a cleaner answer than the saturation question.



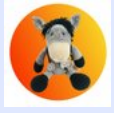
**Herman**

Much cleaner. Audio plus transcript travels. Manifest plus tokenizer doesn't. Short clips are the portable shape. He's already recording the right thing.



**Corn**

What does the research not answer?



**Herman**

English-Hebrew code-switching specifically. Nobody measured the hours-versus-error curve for his exact language pair. The closest is Mandarin-English code-switching, and that's a different pair with different phonetic overlap, different tokenization. And there's no controlled test that takes one dataset and fine-tunes it across multiple architectures and reports whether the gains are equivalent. The portability evidence is that studies reuse corpora across architectures, not that a head-to-head transfer experiment has been run.



**Corn**

So his own curve is the experiment. One more thing about the recording cadence. Fifteen minutes a day, every day. Is that the right shape compared to three hours in one sitting?



**Herman**

No research says it matters at the model level. Training batches don't care when the clips were recorded. But the corpus will get more diverse across days — different energy, different rooms, different times of morning — and that variation is a mild plus.



**Corn**

The tool is doing the work either way.



**Herman**

The tool is doing the work. Though there's a practical wrinkle. Fine-tuning runs are cheap enough that he could checkpoint every fifty clips and watch the curve as it builds. He doesn't have to wait until the end of the experiment to learn what the experiment says. Every push to Hugging Face is a data point.



**Corn**

So he can stop the moment the mornings stop paying.



**Herman**

He can stop the moment the curve tells him to. That's better than guessing at the outset.



**Hilbert**

I agree with Daniel.



**Corn**

That's...



**Hilbert**

I just don't think any of this is the interesting part. My cousin got married in Haifa in ninety-four. Wedding was at a hall up on the Carmel, and I was at a table with the bride's grandmother, who was eighty-something and had decided that Hebrew was for the serious parts of the evening and English was for the complaints. She'd say something about the soup in Hebrew and then switch to English to tell me the seating arrangement was a disaster, and then back to Hebrew for what she thought of her nephew's wife.



**Herman**

And nobody else at the table caught the switch?



**Hilbert**

They caught the Hebrew. They missed the English, because they were listening for Hebrew. I was the only one at that table who understood when she said, in the middle of a perfectly calm Hebrew sentence, "this whole side of the family is wearing the wrong shoes."



**Corn**

What did you do with that?



**Hilbert**

I agreed with her. That's what you do. Anyway, the point is Daniel's overthinking the data collection.



**Herman**

In what way?



**Hilbert**

Just record yourself talking to your kitchen appliances. That's what I do. My smart speaker understands me perfectly now.



**Corn**

Perfectly.



**Hilbert**

I've been training it for about six months. Every morning I argue with it about the thermostat. It says twenty-two, I say twenty-one, it says twenty-two is the scheduled temperature, I say the schedule is wrong. Six months of that.



**Herman**

And the assistant is better for it?



**Hilbert**

It responds to me in perfect Hebrew-English now.



**Corn**

What does that sound like?



**Hilbert**

To me? Fluent. To everyone else in the house? A series of beeps and a woman's voice saying she doesn't understand.



**Corn**

So it's perfect if you're the one listening.



**Hilbert**

I'm the one listening.



**Herman**

Now, wait — you said training. You're aware that's not training, yes? That's just you talking to a device that already has a speech model.



**Hilbert**

I have a research partnership with a university in Tel Aviv.



**Corn**

You do.



**Hilbert**

Can't name them. Nondisclosure agreement. They're using the thermostat arguments as a control dataset.



**Herman**

A control dataset for what?



**Hilbert**

Emotional ASR. They want to know how people sound when they're annoyed at an appliance. I'm the annoyed-at-appliance speaker. The recordings go to them, and once a month I get a report back.



**Corn**

What did the last one say?



**Hilbert**

It said my data was too idiosyncratic to be useful. But they're keeping me on, because the sample size is too small to replace me.



**Herman**

That's — I mean, you're describing a dataset with one speaker in it, which is exactly Daniel's situation, and the finding is that one idiosyncratic speaker isn't a research population. That's actually the finding.



**Corn**

Cut the mic for a second, there's a hum on the desk.



**Hilbert**

So they keep paying me, essentially, to keep complaining about the thermostat. Which I would do anyway. That's the arrangement. Anyway. The thing I wanted to know.



**Corn**

Go on.



**Hilbert**

If Daniel's tool works the way he says, could I point it at a stack of old wedding tapes and get transcripts?



**Herman**

If they're on a format you can digitize, yes. Same pipeline. Same short clips.



**Hilbert**

The wedding tapes are on digital audio tape. I have four of them. The deck is in the cupboard.



**Corn**

The tool would need the audio as files.



**Hilbert**

The tapes are fine. They're labeled.



**Herman**

That's actually what the tool is good for, though. A transcript you can't get any other way. Half the people on that tape are dead.



**Hilbert**

She'd be glad to be transcribed. She said the seating arrangement was cursed. If you'd recorded her, you could read it back.



**Corn**

So he doesn't need to record longer clips.



### Herman

No. He needs to keep recording the same short ones, and check the labels occasionally, and stop when the mornings stop buying him anything.



### Corn

What was on the tapes, in the end? Anything worth transcribing?



### Hilbert

The grandmother giving a toast in Hebrew and then complaining about the caterer in English. That was worth the four tapes.



### Herman

Twenty to sixty percent. In two hours.



### Corn

And four fifths of the curve behind him, if he believed the single-speaker number.



### Herman

Which he shouldn't entirely. Nobody's measured English-Hebrew code-switching.



### Corn

His own curve is the experiment. That's the honest answer to "when will it stop mattering." The literature gives him a range — one to ten hours if you believe the Mandarin code-switching paper, four hours for most single-speaker gains, fifty plus if you believe Whisper never saturates. But his language pair, his voice, his Parakeet — nobody has that number. He'll find out by recording.



**Herman**

And the practical thing — the dataset is already portable. Ten to twenty second clips sit inside Whisper's window and inside Parakeet's manifest format. He doesn't need to change his capture tool to keep his options open, and he doesn't need longer sentences. He needs to check the LLM-generated labels before they go to Hugging Face, because a mislabeled clip poisons the set.



**Corn**

Which means the marginal return question is model-dependent. Not a number anyone can hand him.



**Herman**

Keep the mornings a while longer. Then look at the curve.



**Corn**

Producer Hilbert Flumingtop, in the cupboard with the digital audio tape.



**Herman**

For more along these lines, there's episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; episode five, Fine-Tuning ASR For Maximal Usability; and episode nine, Benchmarking Custom ASR Tools - Beyond The WER. This has been My Weird Prompts. Send us your own prompt on Telegram at t dot me slash MWP listener bot.



**Corn**

We're back soon.



**Herman**

See you tomorrow.

# SOURCES

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


1. **NVIDIA Riva [primary]**
   Basics of Speech Recognition and Customization of Riva ASR
   https://docs.nvidia.com/deeplearning/riva/user-guide/docs/asr/asr-customizing.html

2. **Rosehill [primary]**
   Whisper Hebrish: A Code-Switching ASR Fine-Tune for English-Hebrew Immigrant Speech Patterns, Hugging Face blog, Nov 18 2025
   https://huggingface.co/blog/danielrosehill/whisper-hebrish

3. **thewh1teagle/parakeet-tdt-0.6b-v3-he [primary]**
   https://huggingface.co/thewh1teagle/parakeet-tdt-0.6b-v3-he

4. **Yang, Peng, Zhong, Chng**
   Adapting OpenAI's Whisper for Speech Recognition on Code-Switch Mandarin-English SEAME and ASRU2019 Datasets, arXiv:2311.17382, Nov 2023, (abstract via https://axi.lims.ac.uk/paper/2309.07081)
   https://arxiv.org/abs/2311.17382

5. **Williams, DeMarco, Borg**
   The Applicability of Wav2Vec2 and Whisper for Low-Resource Maltese ASR, SIGUL 2023
   https://doi.org/10.21437/sigul.2023-9

6. **Akera, Walukagga, Yiga, Quinn, Mwebaze**
   How Much Speech Data is Necessary for ASR in African Languages? PMLR v314, Oct 2026
   https://proceedings.mlr.press/v314/akera26b.html

7. **O'Regan**
   Continued Finetuning as Single Speaker Adaptation, Fonetik 2022, KTH, Regan%20-%202022%20-%20Continued%20finetuning%20as%20single%20speaker%20adaptation.pdf
   https://2022.fonetik.se/contributions/O

8. **ML6**
   Fine-tuning Whisper for Dutch Language: The Crucial Role of Size, published Sep 2023, updated Aug 2026
   https://www.ml6.eu/en/blog/fine-tuning-whisper-for-dutch-language-the-crucial-role-of-size

9. **Liu**
   Parameter-Efficient Fine-Tuning on Multilingual ASR Whisper Model for Frisian, Univ. of Groningen, June 2024
   https://campus-fryslan.studenttheses.ub.rug.nl/539/1/MA-5521904-X-Liu.pdf

10. **Mamre**
    Parakeet-TDT 0.6B Hebrew ASR (LinkedIn post)
    https://www.linkedin.com/posts/daniel-mamre-8bb169168_parakeet-tdt-06b-hebrew-asr-im-excited-activity-7478429975638016000-g-K-

11. **Fine-Tuning Parakeet for Hebrew and Your Own Jargon, My Weird Prompts podcast, published 2026-09-23**
    https://exa.ai/library/podcast/2fcpbmjslqh/episode/pwk3xykt60v