Episode Book

EPISODE #6332

Why Extra Voice Hours Settle Bilingual Speech

PUBLISHED	28 October 2026 at 17:25 UTC
DURATION	23:10
LISTEN	myweirdprompts.com/episode/tuning-speech-models-code-switching-61245/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.25
SOURCES	13 cited by the research (listed at the end)
BOOK GENERATED	29 October 2026 at 08:41 UTC

SYNOPSIS

Fine-tuning a massive speech model like Parakeet TDT on personal speech can yield dramatic gains with modest datasets, dropping code-switched word error rates from 13.7% to under 5%. How data maturity stages work, why repeated vocabulary adds acoustic resilience rather than redundant weight, and how models handle Hebrew-English code-switching without regressing on base English recognition are explored through scaling curves, encoder-freezing experiments, and optimization mechanics.

TRANSCRIPT



Corn

Four point nine percent word error rate, against thirteen point seven for the stock model. Daniel's latest Sabich checkpoint, trained yesterday, is making English with Hebrew mixed in considerably less of a transcription adventure. His question is what the recordings are actually buying him now that it works.



Herman

That is a substantial reduction in errors on his code-switched evaluation. Sabich version two point two has moved beyond merely persuading Parakeet that a Hebrew word might occur halfway through an English sentence.



Corn

He's up from about twelve hours of his speech to fifteen, with ordinary English replay alongside it. He wants to put those hours into useful stages, from an early experiment to something mature, and understand what changes inside the model when he keeps adding mostly familiar vocabulary and familiar editing rules.



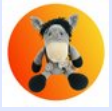
Herman

Especially when the base already understands most of the sentence. The remaining work can look tiny from outside and still require some rather delicate changes.



Corn

Fifteen hours to stop correcting the same noun. I respect the grudge.



Herman

His recordings include roughly eight hours of code-switching. That matters because total duration conceals what the model actually encounters. An hour of ordinary English and an hour containing frequent Hebrew insertions are different training experiences.



Corn

Before you make fifteen hours sound enormous, Herman, I can see you preparing the GPU comparison.



Herman

I was preparing the dataset comparison. The GPUs were going to be incidental.



Corn

They always arrive with luggage.



Herman

Parakeet TDT has six hundred million parameters. Its Granary training corpus contains about a hundred and twenty thousand hours of English. Daniel's fifteen hours are roughly zero point zero one two five percent of that duration. A very small addition to a very large prior education.



Corn

Although fifteen hours doesn't mean fifteen hours of influence. Those recordings get revisited, while the original corpus isn't standing behind every update objecting.



Herman

Yes, that distinction matters enormously. Corpus size establishes the starting point, not a voting system for subsequent training. Parakeet's main training ran for a hundred and fifty thousand steps across sixty-four A100 GPUs. Fine-tuning starts from the resulting weights, not from a live mixture containing all those original recordings.



Herman

NVIDIA's own model card then describes a second-stage fine-tune using about five hundred hours. Even that final refinement used roughly thirty-three times Daniel's speech.



Corn

Their finishing touches have a larger population than his village.



Herman

The Granary total also combines human transcription and machine-generated labels. About ten thousand hours were human-transcribed, with roughly a hundred and ten thousand pseudo-labeled. The later high-quality stage helps explain why a modest, carefully labeled dataset can still have leverage after enormous-scale training.



Corn

Daniel isn't rebuilding the English recognizer. He's correcting a recognizer that has strong reasons to expect English to continue being English.



Herman

What would make you call his checkpoint mature, Corn? Duration aside.



Corn

When it survives me ceasing to cooperate. A familiar sentence spoken clearly proves less than the same vocabulary mumbled while I'm thinking about something else. I'd also want it to preserve ordinary English, rather than treating every uncertain sound as an invitation to become bilingual.



Herman

That gives maturity observable meaning. Hours alone don't.



Corn

Still, I wouldn't throw away the buckets. Daniel wants a way to describe progress, not a calendar with a graduation ceremony.



Herman

Then use conditional buckets. For this single-speaker, narrow-domain situation, around two hours is an early feasibility experiment. Around five to ten can support a working prototype. Fifteen to thirty can support a candidate for mature personal use. Very mature needs demonstrated coverage and stable behavior, not another arbitrary duration threshold. Those are planning labels I would use, not established scientific boundaries.



Corn

Thirty hours gets an interview, not tenure.



Herman

Consider Mandarin-English adaptation with Whisper-small. Conversational SEAME speech could saturate with somewhere between one and ten hours. Read-speech ASRU continued benefiting beyond a hundred hours.



Corn

Same model, different finish line. That also stops us assuming tidy, read-aloud speech must always be the easier adaptation.



Herman

The target task, starting competence, label conventions, and range of conditions all matter. Distributional distance is a useful organizing idea, but it isn't a single measurable number that determines the answer. Coverage and optimization affect where the boundary falls too.



Corn

Your enthusiasm for one elegant explanation has been noted.



Herman

I did put qualifications in before you could confiscate it. An encoder-freezing comparison in that adaptation work is revealing. Keeping the acoustic encoder fixed hurt the read-speech task much less. Their interpretation was that its conditions aligned more closely with what Whisper already knew.



Corn

For Sabich, most of the acoustic world is familiar. Daniel's voice speaking English hasn't suddenly become a completely new language. The Hebrew insertions and the desired written forms are the concentrated novelty.



Herman

Probably a relatively narrow shift overall, yes. Though an English-only model can still find those insertions difficult. A small portion of the sentence can account for a large portion of its errors.



Corn

The boundary isn't an absolute hour count, then. It's how much unfamiliar territory the recordings cover.

L

Larry

Friends, are your hours just sitting there, unbucketed, un staged, never reaching their very mature tier? Then you need HourBox, the subscription service that delivers your training data back to you by post, one hour per month, because nothing says mature like waiting. Start on the Prototype tier at two hours a month, graduate to Mature at thirty, and for the truly serious, the Very Mature tier ships a single hour in a velvet pouch with a certificate.

L

Larry

And our cancellation policy is openly hostile, as all good policies are: to cancel, simply keep subscribing for eleven more months, then send us a letter explaining your diminishing returns, and we will redistribute your weights more firmly. HourBox. Your hours, bucketed at last.



Herman

Waiting does not make training data mature. Duration matters through the examples presented to the optimizer, and neither postal delay nor packaging changes their acoustic or linguistic coverage.



Corn

Let it go, Herman. The velvet pouch isn't peer reviewed.



Herman

The early bucket can actually go backwards. In a Kinyarwanda adaptation of Whisper large-v3, the untouched model had about forty-four percent word error rate. Fine-tuning on one hour raised that to nearly forty-eight percent.



Corn

Paid tuition, lost knowledge.



Herman

At fifty hours, error fell to around twelve and a half percent. That first hour wasn't simply a smaller version of the eventual benefit. Under that training setup, it pushed the model in an unhelpful direction.



Corn

Hold on.



Corn

That doesn't make every two-hour prototype hazardous. It makes the label prototype insufficient reassurance.



Herman

Correct. I was about to say the first bucket is a danger zone… better, it's an unproven zone where regression is possible. A tiny dataset can work wonderfully for a very narrow correction. It can also support misleading updates.



Corn

The practical buckets should carry that distinction. Early means we don't yet know whether the apparent gain survives a separate sample. Prototype means it reliably helps within a known envelope. Mature means that envelope is broad enough for its intended use. Very mature means repeated additions aren't uncovering large neglected corners.



Herman

Would you put Daniel at mature already?



Corn

Candidate, as you said. His reported result is strong. I wouldn't convert one aggregate score into a claim about every room, every Hebrew word, and every tired evening. That isn't doubting the score. It's refusing to make the score do additional jobs.



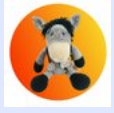
Herman

The longer Kinyarwanda curve illustrates the economics. Moving from fifty to two hundred hours bought roughly two point seven percentage points of improvement. Moving onward to the full fourteen hundred hours bought about the same additional improvement.



Corn

The second discount required a warehouse.



Herman

Much more data for a similar absolute gain. The curve flattened substantially. Yet those recordings can add speakers and linguistic coverage. Daniel is often adding another rendition of something the model has already heard from him.



Corn

Another rendition still isn't an identical sound. The word after lunch differs from the word in the morning, even if the transcript matches.



Herman

Repeated vocabulary can contain valuable acoustic variation. The unresolved part is how much. I haven't seen a published sweep isolating this exact regime of the same speaker, niche vocabulary, and editing conventions. Broader scaling curves give us expectations, not a forecast for his next checkpoint.



Corn

Then twelve to fifteen hours can be valuable without proving thirty will be twice as valuable. It depends on whether those added recordings contain new situations or merely more comfortable examples.



Herman

Now consider an actual update. A recording enters the network, the model assigns probabilities to possible transcriptions, and the training objective measures how poorly it supports the reference. Backpropagation calculates which parameter changes would reduce that loss. The optimizer takes a step influenced by those gradients and its accumulated state.



Corn

No individual Hebrew word gets a labelled drawer.



Herman

No. Its recognition depends on distributed features and interactions. Full fine-tuning, which Sabich uses, allows the change to spread throughout the trainable model. It doesn't mean every parameter changes equally, or that larger changes are automatically better.



Corn

Daniel's phrase about carving stronger patterns sounds right if stronger means more reliably preferred. It sounds wrong if we imagine the same groove simply getting deeper each epoch. Some examples should make the model less certain about a previous interpretation.



Herman

A correction can raise the probability of a Hebrew sequence in one context while preserving an acoustically similar English sequence elsewhere. Gradients from different examples can reinforce one another, cancel, or conflict. More recordings improve our estimate of which changes help across the intended population of inputs.



Corn

Here the population is substantially one man.



Herman

A man with many possible utterances. Generalization can be personal and still be real.



Corn

I will permit him a population.



Herman

Going from two hours to ten might establish a recurring switch pattern and a consistent written form. Moving toward twenty or thirty might supply variants in pace, coarticulation, hesitation, or background sound. Those are plausible roles, not phases the network automatically enters at those durations.



Corn

The same vocabulary can acquire different neighbors. A Hebrew word after an English consonant isn't acoustically identical to that word after a long pause. That's a reason repetition could continue helping even after every word has appeared.



Herman

It can reduce uncertainty around decisions already learned. In the Mandarin-English work, six language-prompt strategies that behaved very differently before adaptation ended up within roughly a point of one another afterward.



Corn

The model stopped needing the correct pep talk.



Herman

Adaptation made its behavior less sensitive to that particular external cue. That supports the robustness interpretation. It doesn't prove every later hour buys robustness, or that Sabich's editing rules work like Whisper's language prompts. Parakeet has a different architecture and inference setup.



Corn

More reliable application of a rule is also useful even when no new rule appears. If it writes the preferred spelling most of the time but occasionally reverts to an English approximation, the remaining work is consistency.



Herman

Larger datasets in the Kinyarwanda work also produced steadier validation trajectories during training. Less bouncing between good and poor checkpoints. Though… stability is one benefit, not the entire explanation for lower error.



Herman

What do you think happens if Daniel doubles the dataset and keeps the same number of epochs?



Corn

He doubles the opportunities for updates, approximately, if batching stays the same. That means the experiment changes both the evidence and how long the optimizer works. We can't credit the extra recordings for everything.



Herman

Precisely the distinction I wanted. More epochs on the old dataset repeat familiar evidence. More unique recordings can change the evidence itself. Equal epochs, equal optimizer steps, and equal audio duration are different comparisons.



Corn

You just approved my answer in your examination voice.



Herman

Occupational residue.



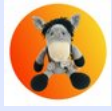
Corn

You also examine the speech recognizer using Hebrew no human clerk would voluntarily utter.



Herman

The bureaucratic register deserves representation.



Herman

Sabich also uses speed perturbation, slightly slowing or accelerating recordings around their original speed. That increases variation the optimizer sees, but it doesn't create a new room, microphone, speaker intention, or vocabulary context. It supplements coverage rather than multiplying independent experience.



Corn

Three speeds, one chair.



Herman

The full-fine-tune distinction matters here. With adapters, the base weights may stay frozen while a smaller set of added or low-rank parameters carries the adaptation. Full fine-tuning offers more freedom, including more freedom to damage previously useful behavior.



Corn

A smaller parameter budget doesn't guarantee good manners, though.



Herman

No. Sequential adaptation to Pacific Indigenous languages has produced severe internal representational drift, including forgetting with LoRA. Restricting the update helps shape where learning occurs. It doesn't exempt it from competing objectives.



Corn

For Hebrish, preserving English is part of learning the task properly. English isn't some unrelated legacy feature Daniel can afford to drop.



Herman

His roughly hour and a half of LibriSpeech replay explicitly supplies counterpressure. Those batches reward continued ordinary-English recognition while the target recordings reward the niche adaptation.



Corn

Compared with fifteen hours of his speech, that's about a tenth as much replay. Enough to matter, but not an insurance policy.



Herman

Continual-learning work has recovered a substantial portion of the gap to joint training using only zero point six percent of original training data. That demonstrates how little replay can sometimes help. It is not directly comparable with Daniel's tenth, because the denominators and learning setups differ.



Corn

Thank you for not dividing two unrelated percentages and declaring victory.



Herman

I had already crossed that calculation out mentally. Where change occurs matters too. Encoder-linear-layer adaptation has prevented forgetting in a studied domain shift. HDMoLE matched full fine-tuning on target accents with about nine point six percent of the trainable parameters.



Corn

Those results make the six hundred million less intimidating. You may not need to renovate every room to correct the address.



Herman

They also prevent us interpreting a full fine-tune as necessarily using all that freedom productively. We can't infer Sabich's internal update pattern from its final word error rate alone.



Corn

Diminishing returns could simply mean it already learned the recurring distinctions. New recordings give gradients pointing toward much the same solution. Not much changes on unseen speech because not much relevant uncertainty remains.



Herman

Or the remaining errors require information the added recordings don't supply. A rare name, an unusual switch boundary, or an ambiguous written convention can remain untouched while ordinary examples accumulate.



Corn

Memorization is one possible next step, not the definition of a plateau.



Herman

Yes. If training loss keeps falling while held-out performance stalls or worsens, fitting recording-specific detail becomes a stronger explanation. A flat validation curve alone could also reflect optimization limits, label inconsistency, or ordinary evaluation noise.



Corn

The model can become excellent at yesterday's microphone.



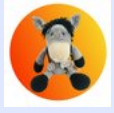
Herman

Noisy reference text is especially expensive here. In a Kikuyu error analysis, about thirty-nine percent of high-error cases involved noisy ground-truth transcriptions. That doesn't mean that share of the entire corpus was wrong. It means many conspicuous apparent failures involved a flawed answer key.



Corn

More hours of contradictory spelling gives it a longer argument.



Herman

Consistent labeling can matter more at this margin than sheer duration. I wouldn't attach a numerical exchange rate to careful versus careless hours. The point is that gradients follow the supplied targets, including their mistakes.



Corn

There is another moving target. A speaker who repeatedly sees a transcription fail may start saying the troublesome word differently.



Hilbert

That does happen, Corn. My snag sheet has a column for words I stopped using. Longer than the column for words the software learned. I started it when I was counting dictation corrections for a records contractor. That was in two thousand three. Herman, you separated extra recordings from extra steps quite carefully. The council records office did not separate either from me learning to speak like their software.



Corn

The snag sheet was recording changes at both ends.



Hilbert

Clumps of them. Several bad entries for the same word, then none. Sometimes I had begun leaving a gap before it. Sometimes I had replaced it. I deliberately said a few badly again to see whether the improvement survived. It usually did not.



Herman

The missing errors would have looked encouraging without the notes.



Hilbert

There was a correction to my address that took six weeks. Martin Bell in records kept getting a transcript containing the wrong road. I sent the audio again. Then I sent the text. He asked for the audio once more. I put each version in the snag sheet. I had priced Dragon NaturallySpeaking for home at a hundred and forty-nine pounds and ninety-nine pence. Didn't buy it. I was already changing enough nouns at work.



Corn

You weren't paying retail to surrender the remaining nouns.



Hilbert

Martin eventually typed the address himself. I had stopped saying the road name by then. I dictated the postcode and added the road afterward. The final correction took him eleven minutes.



Herman

Eleven minutes of manual correction can disappear from an accuracy story if only the automatic output gets measured. For a personal recognizer, the speaker's adaptation and the model's adaptation can both improve the observed result.



Corn

A shrinking snag sheet could describe a better model, a narrower vocabulary, or a more obedient speaker. Those produce different kinds of usefulness.



Herman

Daniel selecting the best validation epoch adds another distinction. That checkpoint is the best observed point on that particular set during training. Repeatedly using the same set to select versions can gradually make it part of the development process.



Corn

Not because the model secretly reads it. Because the person deciding what to change does.



Herman

A small adjacent detail worth keeping is that code-switched Mandarin-English evaluations often report mixed error rate, combining appropriate units across the languages. Their percentages aren't directly interchangeable with Daniel's word error rate. The adaptation pattern travels better than the absolute score.



Corn

That leaves the next checkpoint with a more interesting test than simply containing more hours. If the gains survive different rooms, different microphones, and speech Daniel hasn't unconsciously tidied for the recognizer, we'd know more about what kind of maturity he's buying.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you enjoyed this one, try episode eleven, How Does Fine Tuning Work Anyway; episode five, Fine-Tuning ASR For Maximal Usability; and episode nine, Benchmarking Custom ASR Tools - Beyond The WER. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

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



Corn

See you tomorrow.

SOURCES

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

1. **NVIDIA Parakeet TDT 0.6B v2 model card, created 2025-04-15, last modified 2026-06-29 [primary]**

   https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2

2. **Adapting OpenAI's Whisper for Speech Recognition on Code-Switch Mandarin-English SEAME and ASRU2019 Datasets (full text read, pp. 3**

   5)
   https://arxiv.org/abs/2311.17382

3. **How much speech data is necessary for ASR in African languages? An evaluation of data scaling in Kinyarwanda and Kikuyu (full text read, pp. 3**

   6)
   https://arxiv.org/abs/2510.07221v1

4. **Continual Adaptation for Pacific Indigenous Speech Recognition, Interspeech 2026**
   https://arxiv.org/abs/2603.06310v2

5. **Updating Only Encoders Prevents Catastrophic Forgetting of End-to-End ASR Models, Interspeech 2022**
   https://arxiv.org/abs/2207.00216v1

6. **Using Adapters to Overcome Catastrophic Forgetting in End-to-End ASR, ICASSP 2023**
   https://arxiv.org/abs/2203.16082v3

7. **SSVD-O: Parameter-Efficient Fine-Tuning with Structured SVD for Speech Recognition, ICASSP 2026**
   https://arxiv.org/abs/2601.12600v1

8. **HDMoLE: Mixture of LoRA Experts with Hierarchical Routing and Dynamic Thresholds, ICASSP 2025**
   https://arxiv.org/abs/2409.19878v3

9. **Continual Learning for Monolingual End-to-End Automatic Speech Recognition, EUSIPCO 2022**
   https://arxiv.org/abs/2112.09427v4

10. **ViMedCSS: A Vietnamese Medical Code-Switching Speech Dataset & Benchmark, LREC 2026**
    https://arxiv.org/abs/2602.12911v2

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

12. **AIxBlock, How Many Hours of Audio Do You Need to Train an ASR Model (2026)**
    https://aixblock.io/blogs/hours-audio-train-asr

13. **Hugging Face forums thread on Whisper entity normalization**
    https://discuss.huggingface.co/t/inverse-normalising-entities-in-whisper/30833/3