Episode Book

EPISODE #6022

Fine-Tuning for Code-Switched Hebrew on 2 Hours of Audio

PUBLISHED	16 October 2026 at 10:38 UTC
DURATION	22:25
LISTEN	myweirdprompts.com/episode/code-switched-hebrew-fine-tuning-19924/
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 a speech recognition model on two hours of his own voice and hit 60% accuracy on code-switched Hebrew, up from under 20% across every leading model he tested. The catch: Parakeet V2 is English-only, and Hebrew isn't in its 25-language sibling either. This episode digs into what a fine-tune actually captures — new tokens versus new accents, why control sentences matter, and whether a 30% English ratio is the right number. Plus: why fine-tuning helps where a model is ignorant and hurts where it's merely imprecise, and the hallucination risk that comes with over-controlling.

TRANSCRIPT



Corn

Sixty percent. That's what Daniel's fine-tune hit on code-switched Hebrew, up from somewhere under twenty across every leading model he tested.



Herman

On two hours of his own voice, yes. Two hours of short sentences he recorded on a phone, using an app Claude wrote for him.



Corn

Which is the part that should stop people. Two hours of one man talking into a phone comfortably beat out models trained on a hundred thousand hours of audio.



Herman

Trained on a hundred and twenty thousand hours, in Parakeet's case. That's the V2 checkpoint. Six hundred million parameters, a six percent word error rate on the Open ASR leaderboard, and it runs about three thousand times faster than real time. It is a excellent English model. It just has no idea what Hebrew is.



Corn

So Daniel's prompt has two parts to it. He shipped the fine-tune, the benchmark looked good, and now he's stuck on the follow-up questions. First one came from Claude originally. If thirty percent of your training sentences should be ordinary English with no Hebrew mixed in, is thirty the right number? Second one is Hannah's, and it's the better question. If the whole thing was trained on one Irishman, who else does it actually help?



Herman

And I'd argue those two questions are secretly the same question.



Corn

Go on.



Herman

Both of them are asking what a fine-tune actually captures. The control sentences are about what it learns. The speaker question is about who it learns it from.



Corn

Let's get into it then. Start with what code-switching actually sounds like, because people hear that term and picture something much tidier than reality.



Herman

That's the thing. It's not two languages in clean blocks. It's an English sentence with makolet dropped in the middle, or misrad hapnim, or bitachon. You don't switch languages at the sentence boundary. You switch for one noun and switch back.



Corn

And the listener doesn't notice. That's the whole point. To the speaker it's one stream of speech.



Herman

To the model it's a trap, because Whisper-family models do internal language identification per utterance. That's how they decide which decoding path to take, and it means the model is constantly trying to answer a question the speaker never asked. Is this an English word or did the language change?



Corn

And with one short dictation burst, there's almost no context to answer with.



Herman

There's no signal, so it guesses. Which is why you get gibberish, or worse, a confident English word that is close to the Hebrew one but wrong.



Corn

Give me an example of the confident-wrong one. That's the failure people don't expect.



Herman

Makolet becomes "mocklet." Bitachon becomes "bitchen." The model hears something Hebrew-shaped, decides it's English, and produces the nearest English word that fits the acoustics. It's not noise. It's a real word, spelled correctly, and completely wrong. And because it's a real word, your spell-checker never flags it.



Corn

That's the nasty version. A human proofreader skims past it because it looks like language.



Herman

And it's worse in a dictation context, because you're not reading for meaning. You're reading for whether the transcript matches what you said, and "mocklet" is close enough to makolet that your eye slides off it.



Corn

Parakeet is a different failure entirely, though. Whisper at least knows Hebrew exists.



Herman

Parakeet V2 is English-only, full stop. And here's the piece that matters. Hebrew isn't in Parakeet V3's list either. That's twenty-five languages, and every single one of them is European. There is no Semitic language anywhere in it.



Corn

So the model has probably never seen a Hebrew character.



Herman

Almost certainly not. Those tokens are not in the vocabulary at all. The model card says it outright. If a word is not trained in the language model and not presented in vocabulary, the word is not likely to be recognized.



Corn

So Daniel isn't teaching Parakeet a new accent. And he isn't really teaching it a new language in the way people usually mean that either.



Herman

He's teaching it a new alphabet. New tokens that did not exist for this model twelve months ago. And that reframes the transfer question completely, because now we're not asking whether an accent carries over. We're asking whether tokens carry over.



Corn

And those are very different bets. An accent is a surface variation on sounds the model already knows. A token is a unit the model has no representation for at all.



Herman

Right. When you fine-tune for an accent, you're nudging existing weights toward a different distribution of the same phonemes. When you fine-tune for new tokens, you're asking the model to allocate representational space that didn't exist. That's a bigger ask, and it's a different kind of learning.



Corn

So the model now has Hebrew tokens it never had before. The question is what else it learned along the way, and that's where the control sentences come in.



Herman

The advice Daniel got was that roughly thirty percent of the training data should be plain, everyday English with no code-switching in it at all. The reasoning being that you stop the model overfitting on the code-switching itself.



Corn

Which sounds right, before you think about it too hard. What does the number actually buy you?



Herman

It buys you a contrast. You're not just teaching the model a target, you're teaching it the boundary. Here is what it looks like when I speak Hebrew, and here is what it looks like when I don't.



Corn

And without the second half of that, the model only ever sees Hebrew appearing. So it learns that Hebrew appears, not when it appears.



Herman

Right. It starts treating the Hebrew tokens as something that just happens in your speech at a certain rate, rather than something tied to a specific word.



Corn

Which is exactly the failure Daniel described. English words rendered as Hebrew.



Herman

And that failure mode is documented, which is the part I'd want him to know. There's a paper from Ugan and Waibel at Interspeech this year. They state flatly that fine-tuning on synthetic code-switching data generally degrades strong monolingual baselines.



Corn

Meaning you can fix the mixing and break the English.



Herman

Consistently. It's not an edge case, it's the default outcome. Their fix cut code-switched word errors by about thirty-three percent while improving overall word error rate by five percent, and it held monolingual performance steady. That's the trade Daniel's control sentences are buying.



Corn

So the direction of the advice is well supported. Nobody's arguing you should train exclusively on the target.



Herman

Nobody serious, no.



Corn

Then what about the number? Is thirty right?



Herman

Here's the honest answer. Nobody has published a control ratio for code-switching fine-tuning. It isn't there.



Corn

That's a real gap, not a gap you're papering over.



Herman

It's a real gap. The closest rigorous evidence is Maison and Estève, from twenty twenty-three, and it's about accents rather than code-switching. They mixed accented and non-accented French at different ratios and got U-shaped error curves.



Corn

U-shaped meaning?



Herman

Error falls as you add the other domain, then rises again when you strip the target domain out entirely. Best results were at more balanced mixes. What that tells you is that the optimum is task-specific, and that blindly maxing out on data is not the answer.



Corn

So there's a floor and a ceiling. Too little control and the model overfits the target. Too much control and you're not teaching the target at all.



Herman

And the shape of the curve matters more than any single point on it. If you're on the flat bottom of the U, small changes to the ratio don't hurt you. If you're on one of the arms, they do.



Corn

And the specific thirty percent?



Herman

Is a heuristic. A reasonable one, but a heuristic. The literature's optima ranged between fifty and ninety percent non-target data depending on the task, which puts thirty on the aggressive end.



Corn

So Daniel shouldn't agonise over it.



Herman

He should ablate it. Run the same fine-tune at thirty, fifty, seventy percent control and see. It's a weekend of recording and a few training runs.



Corn

There's a sharper piece of evidence I want to get to, though. Forgetting happens even when the model has already seen the target.



Herman

That's the Maison and Estève finding I keep coming back to. A model fine-tuned on Common Voice scored eighteen point seven word error rate on a different accented dataset, even though that second dataset was in its pre-training data.



Corn

So fine-tuning doesn't just add. It overwrites.



Herman

The model doesn't file the new behaviour alongside the old one. It replaces.



Corn

That's the bit that should make anyone nervous about fine-tuning. You're not stacking skills. You're trading them.



Herman

And you don't get to choose the trade. The model decides what it's willing to lose.

L

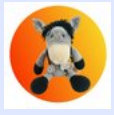
Larry

Friends, are you tired of your speech recognition model misunderstanding you? Well, so was Gerald. Gerald runs a falafel stand in Ramat Gan, and Gerald told me, he said, Larry, my transcription software turned my entire lunch order into the Book of Leviticus. Then I tried the new Chutzpah Switch Pro, and now my model understands every word I say, in every language, including three I made up. My cousin Shlomo, sorry, my cousin Avi, same man, different day, he tried it and his model started transcribing his thoughts before he had them. Chutzpah Switch Pro. It switches. It pross. Available nowhere.



Corn

So Daniel's own benchmark has a result in it that cuts against everything people assume about fine-tuning.



Herman

It does. He fine-tuned on about ninety minutes of audio and improved the smaller Whisper models but made the larger ones worse. Tiny, base, small, all got better. And then the exception.



Corn

Code-switching improved dramatically across every model size.



Herman

Every size. The thing that the base model is worst at is the thing fine-tuning is best at fixing.



Corn

That's the insight worth sitting with. Fine-tuning helps where the model is ignorant, not where it's merely imprecise.



Herman

And it explains why Daniel's Parakeet result is so big. He wasn't polishing a good model. He was handing a model tokens it had never possessed.



Corn

It also explains why the bigger Whisper models got worse. They were already good at English. There was nothing for the fine-tune to add on that axis, so it only had room to subtract.



Herman

The larger the model, the more it has to lose and the less it has to gain from narrow data. That's the pattern.



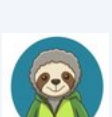
Corn

So back to the thirty percent. Is it about right?



Herman

If a third of his users' utterances actually code-switch, then thirty percent control is defensible on distribution grounds alone. You'd be matching the training mix to the inference mix, which is the only principled way to pick it.



Corn

And if his users code-switch more than that?



Herman

Then he's over-controlling, and he's spending training sentences on English the model already handles. That's the cost. Every control sentence is a sentence not teaching Hebrew.



Corn

Cue the overfitting risk.



Herman

Which is real. Without enough control data, the model learns Hebrew tokens as a positional or acoustic pattern rather than a lexical one. It hears a certain rhythm and reaches for a Hebrew word that isn't there.



Corn

How bad does that get?



Herman

There's a paper on Whisper hallucination that found around one percent of transcriptions contained entirely invented phrases. Not errors, invented text. And there's a well-documented case of someone fine-tuning Whisper-small on Hinglish data and getting both a performance drop and more hallucination.



Corn

Invented text is a different category of failure. A wrong word is a wrong word. Invented text is the model writing its own script.



Herman

And it's the hardest failure to catch, because it reads fluently. You only notice when it says something you never said.



Corn

You can make a model worse at everything by trying to make it better at one thing.



Herman

That's the whole tension. And it's why Daniel's own note about epochs matters. He wrote up a gist on this. Single-speaker data carries higher overfitting risk, so fewer epochs. Three to five. For accented speech, five to ten. For code-switching, ten to fifteen.



Corn

Interesting that code-switching earns the most epochs.



Herman

Because it's the hardest thing to learn. It needs the most passes to stick, and it's also the thing most likely to tip into overfitting if the data is narrow.



Corn

So the epochs and the control ratio are pulling in the same direction. Both are trying to give the model enough exposure to learn the new thing without letting it forget the old one.



Herman

And both are guesses calibrated on intuition rather than measurement. Which is fine, as long as you know that's what they are.



Corn

So the answer to question one is the direction is right, the number is unproven, and his own results already show where the real gains are.



Herman

That's the summary. Though I'd put it differently. The thirty percent is probably fine, and it's also the least interesting knob he has.



Corn

That's all about what the model learned. The other half of Daniel's question is who it learned it from.



Herman

And that's where it gets more interesting, because the finding that answers it is counter-intuitive.



Corn

Hannah's question, for the record. She pointed out that Daniel is one Irishman in a sea of American accents, and asked whether any of this transfers.



Herman

Good question, and the literature has a cleaner answer than I expected. There's work from Stanford last year, Roll and colleagues. Adaptation has two phases.



Corn

Name them.



Herman

The first is speaker-specific. Four to six examples from the same speaker give you about twenty percent relative error reduction. Then there's a variety-general phase, where larger numbers of examples confer similar benefits even when they come from different speakers.



Corn

So at low data, it has to be you.



Herman

And then somewhere between ten and twelve examples, the same-speaker advantage disappears entirely. It's worth one point one percentage points at four to six shots, and nothing at ten to twelve.



Corn

Daniel has two hours.



Herman

He is nowhere near the speaker-specific phase. He blew past it in the first ten minutes of recording. The question isn't whether his voice dominates the model. It's what accent group he just adapted to.



Corn

The two hours isn't a liability in the way people assume. It's more data than the phase transition needs.



Herman

Far more. The interesting question is what that surplus data does once the speaker-specific phase is over. It's not learning his voice anymore. It's learning something broader, and the literature says that something is variety.



Corn

Which brings the Irish accent back into play.



Herman

Here's where the research flatters him, against expectation. Maison and Estève are clear that a model fine-tuned on accent A has reduced performance on accent B, even in the same language. Accent mismatch is real.



Corn

The worry is legitimate.



Herman

The fix is documented. Training on multiple accents gave up to twenty-five percent relative error reduction on accents the model had never seen, compared to single-accent training.



Corn

The answer is more accents, not fewer.



Herman

More accents. But then Roll and colleagues found something that inverts the entire business. Low-resource varieties gain most from adaptation. German-accented English, thirty-five percent better. Indian-accented English, thirty-nine percent. US English, twenty-two.



Corn

The minority accent is the better teacher.



Herman

That's exactly what it says. Daniel's Irish accent may make his fine-tune a better donor for other non-American speakers than an American one would have been.



Corn

He's spent the whole project assuming his accent was a limitation.



Herman

It might be the asset. Which is a much better answer than he was expecting, and I'd want him to hear it plainly.



Corn

What about gender, though? That's the half of Hannah's question with no paper behind it.



Herman

No study isolating male-to-female transfer in single-speaker ASR fine-tuning. There isn't one. The literature does accent and speaker variety. Nobody has run the gender version.



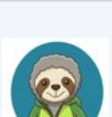
Corn

So you're inferring.



Herman

I'm inferring, and I'll say so. The best proxy is the same finding about same-speaker advantage fading at higher example counts. If speaker identity stops mattering past ten shots, gender is probably a smaller factor than accent.



Corn

Why would that be, though? Acoustically.



Herman

Accent differences are systematic. They move whole vowel categories. They change prosody across an entire utterance. Gender differences are mostly fundamental frequency and formant scaling, and the acoustic front-end has seen women across a hundred and twenty thousand hours of pre-training.



Corn

The model has heard women.



Herman

It has heard millions of women. It has heard far fewer Irish-accented men saying makolet.



Corn

That's the asymmetry. Accent is the axis, not gender.



Herman

Accent is the axis. And it's the one he can fix cheaply.



Corn

How cheaply?



Herman

Roll and colleagues found about twenty-five to thirty seconds of transcribed audio captures most of the available adaptation benefit. That's a handful of sentences.



Corn

He doesn't need thousands of sentences from women. He needs a couple of dozen.



Herman

A couple of dozen from a few different accents. American, British, one or two others. Same LLM-generated sentences, read aloud, a few minutes per speaker. That would test the transfer hypothesis directly.



Corn

What does he do with the result?



Herman

If Hebrew recognition holds up on an American speaker, the tokens transferred, and the accent worry was overstated. If it collapses, he's built a model for one Irishman and he knows it.



Corn

That's the experiment worth running, and it costs him an afternoon.



Herman

The other thing to say is how much of the Hebrew gain even can transfer, because the Hebrew is the novel part and nobody has measured it. Every paper we've got is about accent adaptation in monolingual settings. There is no literature on cross-speaker transfer of newly added tokens.



Corn

Daniel's benchmark might be the only data point.



Herman

As far as I can tell, it is. Which is either exciting or lonely depending on your temperament.



Hilbert

Regional is a word they use when they don't want to say wrong.



Corn

Morning, Hilbert.



Hilbert

There was an actor, Moishe Krendel, worked out of a studio on Allenby. Dubbed a children's cartoon into Galitzianer. Whole production, twelve episodes. Then the broadcaster heard it, said it was too regional, and had it re-dubbed.



Corn

Same actor?



Hilbert

Flatter. They kept the flat one, obviously.



Herman

And the original?



Hilbert

I have it. Only copy left. Biscuit tin, top of the wardrobe, next to the fuse box.



Herman

You have a tape?



Hilbert

I have a tape.



Corn

Do you have a tape player?



Hilbert

No.



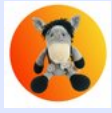
Corn

Of course not.



Hilbert

Been meaning to get one since the tin came to me. That's eleven years. Still meaning to.



Herman

What do you think about the broadcaster's decision?



Hilbert

I think about it more than I think about most things. He got paid twice for the same twelve episodes. That's the only good thing that came out of it.



Corn

The flatter version won.



Hilbert

The flatter version always wins. Ask your Irishman.



Corn

I might.



Hilbert

He should keep his accent. That's all I'll say.



Corn

So the two things Daniel can actually settle with a weekend of recording. The control ratio and the accent transfer.



Herman

Both empirical, both cheap. He could ablate the control percentage at thirty, fifty and seventy and just look at the curves. And he could get three speakers with different accents to read the same fifty sentences and re-run the benchmark.



Corn

The thing that stays unresolved regardless.



Herman

The Hebrew tokens. Nobody has measured how much of that gain transfers across speakers, because the accent isn't the novel part. The alphabet is.



Corn

He's standing in the one spot the literature hasn't reached.



Herman

He could be the first data point on it. That's not nothing.



Corn

The control sentences are there to keep the model honest about the boundary. That's the one wrong idea to kill, by the way.



Herman

That thirty percent is some published constant.



Corn

It isn't. It's a heuristic, and it's probably fine, and the paper that would settle it doesn't exist.



Herman

The other one. That a single speaker's fine-tune is doomed to only work for that speaker.



Corn

It isn't. Same-speaker advantage is gone by ten shots, and he has two hours.



Herman

Which means he's not the bottleneck. The accent coverage is.



Corn

That's fixable for the price of a few afternoons and some friends with opinions about how he says makolet.



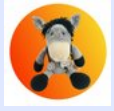
Herman

Thanks to Hilbert Flumingtop, who produces this and owns at least one tape in a biscuit tin.



Corn

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. If you got something out of this one, a review helps other people find it.



Herman

If you've got a question of your own, send us your own prompt on Telegram at t dot me slash MWP listener bot. We'll be back soon.

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, release 2025-05-01 [primary]**
   https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2

2. **Daniel's Whisper fine-tune benchmark [primary]**
   https://github.com/danielrosehill/Whisper-Fine-Tune-Accuracy-Eval

3. **Choosing Epochs for ASR Fine-Tuning, 2025-11-25 [primary]**
   https://gist.github.com/danielrosehill/0fd5579dd554596b9c4a73fedf917697

4. ***My Weird Prompts* episode, 2026-09-23**
   https://exa.ai/library/podcast/2fcpbmjslqh/episode/pwk3xykt60v

5. **Maison & Estève, Improving Accented Speech Recognition with Multi-Domain Training, 2023**
   https://arxiv.org/pdf/2303.07924

6. **Roll et al., In-Context Learning Boosts Speech Recognition, Stanford, 2025**
   https://arxiv.org/html/2505.14887

7. **Ugan & Waibel, Adding Robust Code-Switching Capabilities, Interspeech 2026**
   https://arxiv.org/abs/2606.21990v1

8. **Nguyen & Tran, Can we train ASR systems on Code-switch without real code-switch data?, Interspeech 2025**
   https://arxiv.org/abs/2506.14177v1

9. **Paik et al., HiKE: Hierarchical Evaluation Framework for Korean-English Code-Switching, EACL Findings 2026**
   https://arxiv.org/abs/2509.24613v4

10. **Careless Whisper: Speech-to-Text Hallucination Harms, 2024**
    https://arxiv.org/html/2402.08021v2

11. **Whisper fine-tuning hallucination discussion**
    https://huggingface.co/spaces/openai/whisper/discussions/100