Episode Book

EPISODE #6463

Why Training Speech Recognition From Scratch Fails

PUBLISHED	5 November 2026 at 14:58 UTC
DURATION	22:48
LISTEN	myweirdprompts.com/episode/training-speech-models-from-scratch-48963/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
BOOK GENERATED	6 November 2026 at 11:12 UTC

SYNOPSIS

Training speech recognition models from random weights requires learning acoustics, alignment, and language simultaneously from paired audio. While large language models train on trillions of free text tokens, a speech corpus provides only about nine thousand words per recorded hour. Fine-tuning an existing base model vastly outperforms scratch training below hundreds of hours. Going entirely from scratch demands massive external datasets, noisy open-source transcript scraping, or self-supervised representations.

TRANSCRIPT



Corn

Thirty-two hours. That's where Daniel is with his speech recognition project — his own voice, his own vocabulary, English with romanized Hebrew mixed in, and he says the results are already very good.



Herman

He's been fine-tuning a base model, which is the sensible thing to do, and he knows it's the sensible thing to do, which is why he's asking the other question.



Corn

Right. His framing is that fine-tuning means standing on somebody else's base model, and he wants to know when his own data reaches critical mass — the point where going it alone, training from random weights, becomes a real option. And how you'd actually make that call.



Herman

The phrase stayed abstract for him. What does a blank slate actually mean for a speech model, compared with a blank slate for a language model?



Corn

And that's the part worth getting right, because the two blank slates are not equally blank. Let's start with what that slate really is.



Herman

So, random initialization. Every weight in the network starts as a small random number — usually drawn from some distribution scaled to the layer width, but the content is noise. The untrained model has no notion of sound, no notion of letters, no notion that audio and text are even related things.



Corn

There's a nice demonstration of this in the Keras tutorial, the one that trains a CTC model on LJSpeech from scratch.



Herman

Twenty-four hours of one woman reading, twenty-six million parameters, character-level output, no language model. After one epoch the model's predictions are empty strings. Word error rate of one hundred percent. Not bad predictions — no predictions.



Corn

It hasn't even learned to be wrong yet.



Herman

That's the blank slate. You run audio through it and it says nothing, because saying nothing is exactly as probable as anything else it could say at that point. It only starts emitting characters once it's learned that some sound patterns predict some characters better than others.



Corn

And the contrast with an LLM is the part Daniel was circling. A language model trains on raw text, and every single token in the corpus is a free training target. The next token is the label. No annotation, no pairing, no human in the loop. You point it at trillions of tokens and walk away.



Herman

A speech model has three jobs, not one. It has to learn acoustics — what sounds are even present in the signal. It has to learn alignment, the mapping between audio frames and the characters they correspond to, and that mapping isn't given to you, that's literally what CTC has to discover on its own. And it has to learn the language, spelling included. All of that from data that has to be paired — audio and transcript together.



Corn

Which puts a hard ceiling on how much language a speech corpus can hold.



Herman

An hour of speech is only about nine or ten thousand words. Even if you talk fast.



Corn

Thirty-two hours, then, is roughly three hundred thousand words of language exposure.



Herman

Yes. And the models we compare against trained on trillions of tokens. Daniel's entire corpus, measured as language, is a rounding error next to what a language model sees in a warm-up pass.



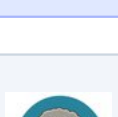
Corn

That arithmetic never stops being rude. You could record every day for a month and the model still wouldn't have read a novella.



Herman

Now here's where the data curves come in, because you can watch what hours actually buy you. Deep Speech 2, trained from scratch: about twenty-nine percent word error at one hundred twenty hours, about thirteen point eight at twelve hundred, about eight and a half at twelve thousand. Roughly a forty percent relative reduction every time you multiply the data by ten.



Corn

Each tenfold is a fight.



Herman

And the Whisper ablation shows the same bending curve at the top end — about thirty percent at three and a half thousand hours, fourteen at fourteen thousand, ten point nine at fifty-five thousand, and then six hundred eighty-one thousand hours buys you all the way down to nine point nine. After tens of thousands of hours, ten times the data is buying you fractions of a point.



Corn

So the from-scratch scaling story has a very long, very flat driveway before the house.



Herman

Whereas fine-tuning skips the driveway entirely. The single-speaker case that's closest to Daniel's situation is the dysarthric speech work from this year — one speaker, Whisper large, full fine-tune. That speaker started at a word error rate of one hundred twenty-eight percent, which tells you how badly a general model handles that speech. One point four hours of adaptation brought it to fifteen point eight. Twenty-two and a half hours brought it to ten point seven. All of the data, best result: nine point seven.



Corn

One and a half hours bought more than everything after it combined.



Herman

Nearly. You're just teaching it this particular mouth and this particular vocabulary. And full fine-tuning beat LoRA in that study, which is a separate finding but a relevant one — if you have the hours, updating everything wins.



Corn

So, the critical mass question. Daniel asked when his own data reaches the point where from-scratch is a real option. What's the honest shape of that answer?



Herman

There is no measured crossover. Nobody has published a point where training from random weights overtakes a pretrained start on one speaker's domain data. Every direct comparison anyone's run below about two hundred hours favors the pretrained start, and the evidence of a base actually hurting you only shows up in the thousands to tens of thousands of hours, and mostly across unrelated languages.



Corn

So the threshold isn't a number of his hours.



Herman

That's the reframe I'd give him. The threshold isn't "do I have enough of my own recordings" — it's "am I willing to assemble a general corpus." A from-scratch model only knows what it has heard. If he wants one that handles anything beyond his own voice and vocabulary, he has to feed it thousands of hours of diverse speech, and his thirty-two hours become one ingredient, not the meal.



Corn

Meaning he'd stop being a trainer and become a librarian.



Herman

A curator, at minimum. There's a story there about open corpora we'll get to, and it's not a happy one.



Corn

Before that — the Keras tutorial eventually worked, right? It got to real numbers on twenty-four hours.



Herman

After about fifty epochs it hit roughly sixteen to seventeen percent on held-out clips of the same speaker. On twenty-four hours, one speaker, small model. But the detail I love — the errors it made were spelling slips. It wrote "proberly."



Corn

It learned the sounds before the spelling.



Herman

It's heard enough sound-to-character mapping to be mostly right, but it has no language model underneath it telling it that "proberly" isn't a word. Nobody ever read it a book.



Corn

A model that aces the accent and flunks the alphabet.



Herman

Which is the whole tension Daniel's living in, compressed. His fine-tune already knows how to spell because the base model heard an enormous pile of transcribed speech before he ever met it.



Corn

So that is the arithmetic on Daniel's thirty-two hours, and it is not kind.



Hilbert

Hilbert's Mid-Show Haiku.



Hilbert

Thirty two hours in. The weights start from nothing and stay nearly nothing. I have made my peace with the corpus.



Corn

The haiku has accepted it faster than we have.



Herman

He's not wrong, though. Three hundred thousand words. You can't argue with a man who's made his peace.



Corn

Then let's argue about the decision instead. Say Daniel accepts the curator job. What does going it alone actually get him, and what does it cost?



Herman

The costs first, because they're stark. His thirty-two hours is about zero point three percent of a ten-thousand-hour corpus. At an hour a day of recording — which is his stated plan — he hits about four hundred hours in a year, and ten thousand hours in roughly twenty-seven years.



Corn

He'd finish the corpus some time in the twenty fifties.



Herman

If he never misses a day. So you don't record your way there. You assemble. And the open corpora exist — YODAS has about three hundred seventy thousand hours, People's Speech around thirty thousand. But many of those labels are machine-made, and a spot check of People's Speech samples rated forty-four percent of them low quality. So you're inheriting somebody's noisy autopilot transcripts and calling it ground truth.



Corn

Half the library is shelved wrong, and you're not told which half.



Herman

Compute, though — compute is not the wall, and this surprised me. An open reproduction of Whisper-style training ran thirty-eight thousand hours on thirty-two A100s for about a week. Rough estimates for a from-scratch run at a thousand hours are somewhere in the five hundred to two thousand dollar range on rented H100s, maybe fifteen hundred to six thousand at ten thousand hours for a mid-sized model — and three to five times that once you account for hyperparameter search, because nothing works the first time.



Corn

That's a bad month, not a mortgage.



Herman

Right. For Daniel, who works with this stuff daily, the money is not the deciding factor. The data is.



Corn

What about the reasons to go it alone that aren't accuracy? Because there must be some, or nobody would ever do it.



Herman

There are, and they're good ones. A much smaller model — if you want speech recognition running on a phone or a small device, a compact model trained only on your domain can be a fraction of the size of a general one. Full control of the output vocabulary — Daniel's mix of English and romanized Hebrew is exactly the kind of thing where owning the tokenizer matters. And licensing — some base models come with terms you don't want. There's also a neat middle path: train a student model from random weights on a strong teacher's pseudo-labels, which worked in the literature as long as you have enough unlabelled audio. You still need the audio. You just don't need to transcribe it yourself.



Corn

Which loops to the actual middle ground, the self-supervised encoders. Say what those are, because I think that's the honest answer to "from scratch, but not stupid."



Herman

Wav2vec 2.0-style training. You take unlabelled audio — no transcripts at all — and the model learns by predicting masked segments of its own representations. The layer analysis is lovely: shallow layers learn acoustic properties, middle layers end up encoding phones, deeper ones start capturing word identity. So the model ends up knowing sound, but not spelling.



Corn

The parrot problem.



Herman

You still have to bolt on an output head and teach it to write. But look at the numbers — the architecture trained from scratch on all nine hundred sixty labelled hours of LibriSpeech scored three percent on test-clean with no language model. The same architecture, pretrained on fifty-three thousand unlabelled hours and then fine-tuned on just one hundred labelled hours, scored three point one. About the same on the clean set and better on the hard set, from one hundred supervised hours instead of nine hundred sixty.



Corn

So the unlabelled world does the acoustics, and your labelled hours only have to buy the spelling.



Herman

And at tens of hours, that kind of setup leans hard on an external language model to fill in what the audio alone can't. Which, for code-switching, matters — because on SEAME, the Mandarin-English code-switching benchmark, about a hundred hours, a well-built from-scratch transformer-transducer reached around twenty and twenty-eight percent mixed error. That's competitive. Recipe quality counts for a lot.



Corn

Daniel's mix is milder than that, though. English with romanized Hebrew is one alphabet, shared script.



Herman

Same script, which is why the romanization works at all. The model doesn't need Hebrew letters, just new spellings in the alphabet it already has. The interesting finding on where a base starts to hurt: Google's BigSSL work found English pretraining helped for Hungarian, Chinese and Hindi at one hundred and one thousand labelled hours, but actually hurt on the full sets — nine thousand to twenty-seven thousand hours of the target language. The base becomes a bad prior once you have enough of the right data. And at thirty-four thousand hours of voice search traffic, they found pretraining was no longer necessary to benefit from big model sizes.



Corn

So the base is scaffolding — you take it down once the building can stand, and nobody's building is at that stage.



Herman

Nobody's personal building. There's also a subtler point from the transfer work: an English-trained encoder moved to Russian with just sixteen hours scored thirty-two percent versus nearly sixty from scratch. Even at fifty thousand hours of domain data, the pretrained start still ended best and reached ninety percent of its final accuracy about eighteen times faster. The from-scratch run never caught it.



Corn

Now the costs of staying on a base that you can fix without leaving it — because a fair version of Daniel's question has to include those.



Herman

Two big ones. Tokenizer lock-in: if you want to change the tokenizer — say, to handle his romanized Hebrew properly — the usual approach is you reinitialize the decoder and the joint network and keep the encoder. You lose the top of the model, not the bottom. And the second one is the internal language model that a transducer builds. It favors common words, so for a vocabulary like Daniel's, where the interesting words are not the common ones, that bias actively works against you.



Corn

That's the one I'd worry about for him. Half his vocabulary is domain words the base has barely seen.



Herman

There's a fix that doesn't require leaving the base at all — internal language model subtraction at decode time. You estimate the model's own internal language model and subtract it out, and in the literature that gave up to fifteen and a half percent relative improvement over the usual shallow-fusion approach. You keep the base and defuse its prior.



Corn

So the checklist version: before you leave the base, you've tried retraining the top with a new tokenizer, and you've tried subtracting the internal language model.



Herman

Both. And honestly, for thirty-two hours — the research says fine-tuning is still the right call, full stop. The from-scratch option only becomes real if he's willing to build or curate a general corpus, and even then, his own data becomes a final-stage fine-tune on a base he built himself.



Corn

Which is a sentence that should be said twice: even the from-scratch route ends in a fine-tune. He'd be standing on his own base model instead of somebody else's.



Herman

And I don't think he'd notice the difference in the transcripts.



Corn

One thing we've skirted — the self-training route. Student from random weights, teacher's pseudo-labels. Does that change the calculus for a single speaker?



Herman

It changes the transcription burden but not the audio burden. You still need the hours of audio; you just let a teacher model label them. And the catch is that the teacher's errors become the student's gospel. If the teacher is bad at romanized Hebrew — which a general model will be — the student inherits that badness with total fidelity.



Corn

Learned ignorance, at scale, with a receipt.



Herman

My read is that it works best as a top-up on top of real labelled data, not a replacement. Your own transcribed hours stay the gold standard, and the pseudo-labelled hours pad the acoustics.



Corn

And that loops back to what he's already doing. He keeps recording an hour a day — he's not wasting his time, the fine-tune keeps improving, and every hour he records is also an hour he could later feed to a self-supervised encoder with no transcription at all.



Herman

That's the part I find elegant. The unlabelled middle ground doesn't care that he hasn't transcribed it. Wav2vec-style pretraining just wants the audio.



Corn

So the same recordings are simultaneously fuel for the model he has and the model he might build. That's the one number I'd hand him — the fifty-three thousand unlabelled hours versus one hundred labelled ones, same architecture, same score.



Hilbert

He'd want to hear about the parrot, then.



Corn

I'm sorry, the parrot.



Hilbert

You were on about the model learning sounds but not spelling. I had a parrot that did that. Not mine, strictly — it belonged to a man in my building, Ronen, fourth floor, and I looked after it when he travelled. Two years I had that bird on and off. I kept a notebook of every word it took on. Three hundred words, give or take, and I still know it was three hundred because I counted them on a Sunday in March and got two hundred ninety-six and then it learned four more by June.



Herman

Two hundred ninety-six and then four.



Hilbert

The bird would say shalom perfectly. Clean as anything. But if you gave it the keyboard — it liked the keys, it would peck at them — it spelled it sholom. Every time. O where the A should be. Nobody taught it the wrong spelling, it just had the sound and no orthography, and it didn't care which letters stood in for it.



Corn

That's the Keras tutorial with feathers.



Hilbert

Ronen's bird, so I take no credit for the training. But the notebook was mine and the counting was mine, and I'll tell you the part that annoys me. For the past month I've been working with it on saying the name of this show. My Weird Prompts. It won't have it. Says the shalom fine, has done for years, and a four-word English title is apparently beyond it.



Hilbert (continued)

Ronen thinks it's because the bird doesn't like me. I did the sound engineering on a dubbing stage for a stretch and I know what a clear syllable sounds like, Corn, and my delivery is not the problem.



Corn

The bird simply has standards.



Hilbert

The bird has a tin ear and a long memory. Two hundred ninety-six words, Herman. Your Daniel's thirty-two hours of speech, that's about the same count. Three hundred thousand words, you said it yourself. I heard you say it. The parrot topped out and so will the corpus, and nobody warns you in advance where the ceiling is.



Herman

The notebook surviving is the part I want to ask about — no, actually, I won't. The spelling slip is what gets me. Sholom with an O. That's "proberly" from the tutorial, almost letter for letter.



Hilbert

Ronen still has the bird. I still have the notebook. We are all keeping things.



Corn

Two hundred ninety-six words and it spells the one it knows best wrong.



Hilbert

It says it better than either of you would. That's the thing about sound.



Corn

Well, that's the parrot standard settled. Where were we — the decision.



Herman

The decision, spelled out plain. If the goal is accuracy on his own voice and vocabulary, fine-tuning wins at thirty-two hours and there's no measured scenario below a couple hundred where it doesn't. If the goal is ownership — small model, own tokenizer, clean licence — then from-scratch becomes a project about assembling a corpus, not about recording more of himself. And the self-supervised encoder is the route that lets his existing recordings do double duty.



Corn

And even that route ends with a fine-tune on top. There's no version of this where he escapes the last mile.



Herman

The last mile is the only mile anyone actually walks. In vision, people showed random initialization matches ImageNet pretraining on a big detection task if you just train longer — and nobody does that, because training longer costs more than borrowing, and it always will.



Corn

Here's the one thing to take from today, if you take one thing. Thirty-two hours sounds like a lot of recordings, and it is — measured as language, it's about three hundred thousand words, which is less than most language models see before their first checkpoint. The blank slate for speech isn't just empty, it's empty in three directions at once — sound, alignment, spelling — and every one of those has to come from data somebody had to prepare.



Herman

So the question isn't "when is my data enough to start over." It's "when am I ready to build the library the new model would need." For almost everyone, the answer is never, and the borrowed base is not a compromise — it's the deal of the century.



Corn

The open question I'd leave Daniel with: at an hour a day, he crosses four hundred hours within the year. Nowhere near from-scratch territory, and still under the roughly thousand hours where self-supervised training has been tested. But those recordings keep. Whatever he decides later, they're the one ingredient nobody else can supply.



Herman

And as open corpora and label-cleaning tools get better, that middle route gets cheaper every year. The decision may not look the same in two years' time.



Corn

Something to watch, then. Producer credit as ever to Hilbert Flumingtop, holding the desk and apparently a two-year parrot ledger. If you enjoyed this one, try episode eight, Building Your Own Whisper. This has been My Weird Prompts — if you enjoyed this one, the website is my weird prompts dot com, and you'll find the feed there too. Leave us a review if you get a minute.



Herman

We'll be back soon. See you tomorrow, Corn.



Corn

See you tomorrow.