Episode Book

EPISODE #6495

Teaching Speech Models to Choose Their Own Audio

PUBLISHED	8 November 2026 at 14:31 UTC
DURATION	27:40
LISTEN	myweirdprompts.com/episode/active-learning-speech-transcription-21775/
HOSTS	Corn, Herman, Clive, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	10 cited by the research (listed at the end)
BOOK GENERATED	9 November 2026 at 08:09 UTC

SYNOPSIS

Human transcription takes over eight hours of manual work for every hour of recorded audio. Active learning promises to cut that burden by letting speech models select which unlabelled recordings are most informative to transcribe next. Yet simple uncertainty measures often select garbage noise rather than useful speech, and sophisticated loss predictions frequently struggle against random baselines. Model disagreement committees, Monte Carlo dropout, and acoustic diversity clustering offer practical paths toward actual transcription savings.

TRANSCRIPT



Corn

More than eight hours. That's the human labour to transcribe one hour of speech. Not machine time — a person, sitting there, for most of a working day, per recorded hour.



Herman

And that's the number sitting behind Daniel's prompt, because when your personal speech fine-tune underperforms, the default instinct is just to record more. Read more text, dictate more notes, pile up hours. Daniel's question is whether the model itself can tell you which hours are worth recording — that's active learning, the loop flipped so the model points at its own weak spots — and, more to the point, whether the signals it uses are trustworthy and whether any of it actually saves you anything versus collecting at random.



Corn

His framing is nice: not every recording teaches the model the same amount. Which sounds obviously right, and the literature is not obviously on its side.



Herman

It really isn't.



Corn

So let's start with what the loop actually is, because the pitch and the mechanism are not the same thing.



Herman

The mechanism is concrete. You start with a small labelled seed set — for Daniel that's the audio he's already transcribed, the hour or so behind his Whisper fine-tune. You train. Then you run the checkpoint over your pool of unlabelled audio, and every utterance gets a score from something called an acquisition function. Highest-scelling batch gets selected, a human transcribes it, it goes into the training set, you retrain, and you go around again for however many rounds you've budgeted.



Corn

And the score is the whole game. The loop is plumbing.



Herman

Mostly. There's one structural detail that matters before we even get to signals: batch mode isn't optional. Retraining a large model after adding one labelled sample is impractical — one sample gives you basically nothing — so real active learning selects batches, a fixed number of utterances per round. And the moment you take the top-scoring batch, you inherit a problem: if the model is uncertain about a particular kind of audio, the top hundred uncertain utterances are often a hundred near-copies of the same thing.



Corn

Which is where diversity terms come in.



Herman

"focusing both on informativeness and diversity reduces the chance of choosing duplicate samples, which could hinder learning progress." That's the design principle the whole modern literature converges on.



Corn

One terminology trap before we go on, because it'll bite us later.



Herman

Say it.



Corn

The papers use "informativeness," "uncertainty," and "confidence" almost interchangeably, because the classic path-probability methods treat low probability as high informativeness. Low confidence equals high information. That conflation is where half the confusion in this literature starts — people read "confidence-based selection" and



Corn (continued)

think it means selecting what the model is confident about.



Herman

It's the opposite, and the fact that both readings are grammatically available in the same sentence tells you a lot about the field. So — the loop has one job: score the unlabelled audio. Everything hangs on whether that score is any good. Let's go through the signals in order, weakest to strongest.



Corn

Confidence first. Give it its due before we bury it.



Herman

It gets a real amount of due, actually. Least-confidence selection from model posteriors is the oldest family in the business — confidence measures for HMM systems go back to the early two thousands, and the Amazon work in the Vosk maintainer's survey reportedly found that the specific confidence formula doesn't matter. Even simple ones work.



Corn

So the formula is not the lever.



Herman

Which is either reassuring or deflating depending on how much of your weekend you were planning to spend on it.



Corn

I was planning to spend it horizontally, so I'm fine.



Herman

Then the objection that actually bites: deep networks are overconfident at the softmax layer. The two-stage Bayesian paper puts it plainly — overconfidence "questions the reliability of using path probabilities for uncertainty assessment." On a modern CTC or attention model, the path probabilities are not a calibrated measure of anything. The model saying "I'm ninety-four percent sure" and the model saying "I'm seventy percent sure" are not two different epistemic states.



Corn

That's the theoretical objection. What's the concrete failure?



Herman

Microsoft's GERM paper has the most quotable one. At a one percent selection rate — you're only taking the very tippy-top of the uncertainty ranking — "most utterances selected by the confidence-based approach are noise and garbage utterances that have extremely low confidence but have little value to improving the performance of the overall system."



Corn

Silence, clipping, crosstalk.



Herman

Silence, clipping, crosstalk, half-sentences you abandoned mid-word. The lowest-confidence audio in your personal pool is disproportionately not speech. And GERM — global entropy reduction maximisation, a smarter selection criterion — beat confidence-based selection by about two percent relative.



Corn

Two percent.



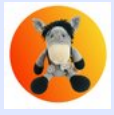
Herman

Two percent relative. After replacing the whole selection principle.



Corn

So the smartest method in the room beats the naive one by a rounding error.



Herman

At that operating point, yes. Now, committee — disagreement between models — and this one is consistently the best-supported signal in the literature. Long et al. at Interspeech twenty-eighteen combined confidence with a word matching error rate computed across heterogeneous neural networks — different architectures disagreeing about the same audio — and that beat single-model confidence. And the Vosk survey states it flatly: "committee-based confidence wins over simple single-model confidence." Vote entropy. He lists that finding twice, for two separate papers, which for a survey writer is the equivalent of underlining it.



Corn

The catch for a solo practitioner is that "train several models" is a horrible instruction.



Herman

Here's the trick, and it's elegant: you don't have to. MC-dropout gives you a committee from one checkpoint. You leave the dropout on at inference, run the same audio through T stochastic forward passes, and each pass is a slightly different committee member. The variation ratio is one minus the fraction of passes that agree with the modal prediction — if all T passes say the same thing, ratio near zero, model is sure; if they scatter, ratio climbs, model doesn't know. The recent two-stage Bayesian pipeline does exactly this, using MC dropout to approximate Bayesian inference.



Corn

Same checkpoint, no retraining, and you get disagreement for free.



Herman

For the cost of T forward passes, which on a personal fine-tune is nothing.



Corn

That's the one I'd actually use.



Herman

I suspected you would. Now the branch that sounds best and performs worst: gradient and loss-based signals. Expected gradient length is theoretically lovely — how much would this sample move the parameters if we trained on it — it's uncorrelated with confidence, which is what you want in a second opinion, and it wins in experiments. Not significantly, but it wins.



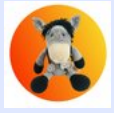
Herman

It's the most honest sentence in the survey. The loss-prediction variant is where it gets messy. Luo et al. Train a second network to predict per-sample loss — the selector learns which audio will hurt the model — and their own abstract claims it outperforms random selection, least confidence, and estimated loss on English and Chinese tasks.



Corn

And the survey reads the same paper as...?



Herman

As not significantly better than random, with the flat observation that fifty percent random selection is much better than twenty percent loss-prediction selection.



Corn

Both claims are about the same results.



Herman

Same paper, same numbers. The survey is reading them more pessimistically and flagging what the method lacks — a diversity term. Loss prediction considers informativeness only, so it's chasing the same hard audio repeatedly. And this is the pattern the whole field has: the primary paper says it works, the practitioner reading says it barely beats random, and they're not disagreeing about arithmetic, they're disagreeing about what counts.



Corn

There's a vision-side version of this too, isn't there.



Herman

Mittal et al., "Parting with Illusions about Deep Active Learning" — the canonical methodological critique. Under realistic evaluation, deep active learning methods barely perform better than the random baseline, and they improve by a large margin only when you integrate them with semi-supervised learning.



Corn

The random baseline. The thing you were trying to beat by doing all this.



Herman

The control condition that keeps showing up in the winner's circle. Which leads us to the signal that makes all the others work — diversity, clustering by acoustic condition. The recent pipeline is two-stage precisely because uncertainty alone is insufficient. Stage one is unsupervised: extract x-vectors for every unlabelled utterance — speaker and environment embeddings — and cluster them with DBSCAN, isolating groups that share a speaker or similar acoustic conditions. Stage two does the Bayesian batch selection within that diverse pool. Earlier work did the same idea with i-vectors, K-means, and an entropy regulariser over cluster occupancy, and beat plain least confidence.



Corn

So the ranking, roughly: confidence at the bottom, committees in the middle, gradient and loss signals as an unresolved argument, and diversity as the thing you wrap around whichever uncertainty signal you picked.



Herman

That's the verdict the evidence supports, yes.



Corn

Daniel also asked about error clustering by word — the intuitive version every personal fine-tune owner eventually invents, "the model always gets these five words wrong, let me record those."



Herman

And the honest answer is that no paper in this literature does clean word-level error clustering as an acquisition function. The clustering that exists is by acoustic condition, not vocabulary item. The nearest relatives are Long et al.'s word matching error rate — used as a selection score, not a clustering method — and Adapt4Me, which highlights confused phonemes. Sub-word, not word.



Corn

Phoneme level.



Herman

So "cluster errors by word and record those" is a reasonable engineering heuristic, and it's not an established published method. Daniel invented it independently, which I'd argue is evidence it's the natural thing to invent.



Corn

Everything natural to invent is either unpublished or patented. Sometimes both.



Herman

Now — the signals rank from confidence at the bottom to diversity at the top, which takes us to the question Daniel actually asked: does any of this save you anything?



Corn

The savings numbers, and I want the complications served with them.



Herman

Best-case headline first. Long et al. report six to thirteen percent relative WER reduction against a four-thousand-hour baseline by selecting twelve hundred hours of informative utterances for human labelling. That's a thirty percent data increment on top of the baseline — twelve hundred fresh labelled hours — for a mid-single-digit to low-double-digit relative gain.



Corn

A real gain.



Herman

A real, replicated, respectable gain. It is not "a tenth of the data."



Corn

Lay out the spread.



Herman

It goes like this. At the optimistic end: seventeen percent of carefully selected samples matching full-dataset accuracy — that figure is cited by the Bayesian paper. Dossou's accent work: twenty-seven percent relative WER improvement with forty-five percent less data than baselines, on African-accented English across several domains. Then the CMU figure — a tenth of the dataset reaching full accuracy with a simple selection method — which the Vosk maintainer immediately flags with "I do not think the paper above applies to real-life data."



Corn

He didn't even let it stand for a sentence.



Herman

And at the pessimistic end: no method guarantees full accuracy with even thirty percent of the samples. They're all suboptimal.



Corn

Seventeen percent to "thirty percent isn't enough." That's not a range, that's a weather report.



Herman

The confound is that these aren't measuring the same thing. Inconsistent evaluation sets and different operating points — one paper selects a thousand lines from ten hours, another selects twelve hundred hours from four thousand — make apples-to-apples comparison close to impossible. And the Bayesian paper says the quiet part from the other direction: on heterogeneous test sets, performance often looks similar across strategies because test-set variability masks method differences. Even random sampling can look comparably good.



Corn

Wait — the test sets are so noisy that the methods are indistinguishable?



Herman

On heterogeneous sets, frequently. Your evaluation set has to be tight enough that the method difference can show through the noise, and a lot of these papers don't have that.



Corn

We did an episode on benchmarking custom ASR tools beyond the WER, and the whole argument there was that your evaluation set is the thing quietly deciding your results. This is that argument wearing active-learning clothes.



Herman

It's the same disease. And it's why the savings figures don't converge — not because the methods are chaotic, but because the rulers are different lengths.



Corn

Then the competitor. The thing active learning keeps losing to.



Herman

Semi-supervised learning. The Vosk survey's conclusion is blunt: "semi-supervised learning wins over active learning in all the works with a big margin." It's still beneficial to process a whole set at least once — pseudo-label everything with the model itself, train on the automatic transcripts — instead of forcing focus on a smaller subset. His suggested compromise is lovely: train on a subset, then fine-tune once on the whole set.



Corn

For Daniel, the semi-supervised version is almost free.



Herman

That's the thing. His unlabelled pool is his own dictated notes. Pseudo-labelling it costs machine time, no transcription labour at all. Active learning labelling costs the eight hours per hour. On pure cost, semi-supervised should walk away with it.



Corn

Except Long et al. argue the opposite.



Herman

They argue the opposite for their setting, and their reasoning is worth taking seriously. Automatic transcripts still contain errors, and those errors are very sensitive under discriminative training criteria. Worse: high-confidence pseudo-labelled data tends to have similar acoustic properties to the seed training data — the model is confident about what it already knows — so the semi-supervised gains degrade rapidly.



Corn

So pseudo-labelling teaches the model more of what it already has.



Herman

That's their claim. Fresh human-labelled data, selected where the model is actually wrong, buys more. "Our AL scheme can benefit much more from the fresh data than the SST in reducing the word error rate." Direct contradiction of the survey, both credible, and the disagreement maps onto something practical: it depends on whether your model's errors are transcription errors or coverage errors.



Corn

Split those two for me.



Herman

A transcription error is the model being miscalibrated on audio it fundamentally covers — it has the phonemes, the words, the accent, and still gets it wrong. A coverage error is the model having never seen that word or phoneme context in your voice at all.
Semi-supervised helps with the first kind, badly, and barely touches the second — you can't pseudo-label what you've never covered, and your confidence estimates on uncovered territory are fiction anyway.



Corn

Which is the whole overconfidence problem again.



Herman

It loops right back around. And that split decides the single-speaker case, which is where Daniel's question actually lives.



Corn

Before we get there — and now it's time for One-Star Getaways with Clive.

C

Clive

This is One-Star Getaways, and I have just come back from Reykjavik, where I stayed at the Puffin's Sigh Guesthouse, one star, and I paid extra for the harbour view, which turned out to be a framed photograph of the harbour screwed to the wall above the kettle.

C

Clive

Breakfast was a single boiled egg and a note explaining that the chef was on a course, and the radiator made a noise like a man learning the trombone, so I have rated it one star and I recommend it warmly to anyone looking for a disappointing trip this winter.



Corn

A framed photograph of the harbour. I'm oddly tempted.



Herman

You are not booking a deliberately disappointing guesthouse in Iceland in November.



Corn

The trombone radiator alone… no. Fine. The single-speaker case.



Herman

So Adapt4Me — a demo paper from this year, the only system in this literature built for exactly Daniel's scenario: one user, their own voice, personalising a recogniser. Its diagnosis is the whole episode stated by researchers: "without human-in-the-loop guidance, users often expend effort recording redundant data — utterances that the model already transcribes well — while failing to gather the data that is most informative for further reducing recognition errors."



Corn

People record what they're already good at.



Herman

You dictate fluent paragraphs, because dictating fluent paragraphs is pleasant, and the model already transcribes your fluent paragraphs fine. What it doesn't have is the kitchen-table vocabulary, the half-swallowed consonant clusters, the rare words in your actual life.



Corn

In Daniel's case, presumably a scattering of Hebrew place names and whatever Ezra is currently obsessed with.



Herman

Whatever lives at the edges of the vocabulary in his actual voice. And Adapt4Me's answer is not a better acquisition function. It's a cheaper human step. Three stages: rapid profiling via greedy biphone coverage — it picks the phoneme pairs you haven't covered and gets you to record just those, bootstrapping personalisation in minutes of training data rather than hours. Then variational-inference LoRA for fast incremental updates — retraining in minutes, not overnight. Then continuous improvement through low-friction top-k corrections driven by visualised uncertainty: the interface shows you where the model is confused, you fix the top confusions, it updates.



Corn

The interface is the contribution.



Herman

The design claim is the contribution: "making epistemic uncertainty explicit reframes data efficiency as an interactive design feature rather than a purely algorithmic concern." And: personalisation "requires a shift from passive data collection to interactive curating."



Corn

Run the arithmetic on that, because it's the reframe I want to land. If the bottleneck is eight hours of transcription per hour of audio, a twenty percent better selection rule saves less than a five-times cheaper correction interface.



Herman

Order of magnitude, yes. Selection operates on which hours you pay the eight-fold price for. The correction loop operates on the price itself. And the correction loop also skips the transcription entirely — you're editing a draft, not producing one from silence, which is a different and much cheaper cognitive task.



Corn

So the answer to "how much does active learning save" might be "less than a good correction screen saves."



Herman

For a personal fine-tune, I think that's right. And there's a scale argument for why this case is different from everything we cited. The hard numbers — the six to thirteen percent, the seventeen percent, the forty-five percent — all come from multi-speaker corpora with hundreds to thousands of speakers. The dysarthric-speech world shows you the other end: TORGO is about three hours, UA-Speech about nine, BF-Sprache two and a half, versus Common Voice English at roughly thirty-eight hundred and LibriSpeech at a thousand.



Corn

Three hours versus thirty-eight hundred.



Herman

And the Adapt4Me point is that for non-normative speech, variability across individuals is so high that "the rich diversity of speech impairments cannot be addressed simply by collecting more data." Collecting more isn't even available — the corpus cannot be bought at any price.



Corn

A single speaker is that problem in miniature. You are a corpus of one, and you cannot go hire more of you.



Herman

Which means, at one to ten hours, your errors are dominated by coverage, not calibration. Words and phoneme contexts unseen in your voice. And the signal the evidence supports at that scale is the boring one: core-set and diversity selection. Record the phoneme contexts you haven't covered. It doesn't require the model to be well-calibrated, which is convenient, because it isn't.



Corn

So "record more" may be right after all.



Herman

The correction is record differently, not less. Same hours, pointed at the gaps.



Hilbert

The word you two keep using is "selection."



Corn

It is.



Hilbert

I've been doing this for years and nobody sent me a paper. I keep a notebook. Paper, bound, not a spreadsheet. Every word a transcription has ever got wrong on me goes in it. The timestamp, the audio, a column for what I actually said, a column for what came out.



Herman

How far back?



Hilbert

Further than the notebooks before it. I go back and re-record the entries, in order of how many times each one has appeared. The repeat offenders first. Most of them are fixed in one pass. Some take three or four, and those are the ones I keep an eye on.



Corn

There's a third column, isn't there.



Hilbert

There is. The ones where the machine was confident and wrong. Those get their own treatment — I re-record them whether or not they've repeated, because a machine that says it's sure and isn't is a different animal from one that admits it's guessing. Herman, you said that yourself twenty minutes ago, roughly, after a detour.



Herman

I did take the scenic route.



Hilbert

I re-record them in the car. Parked, engine off. Every one of my worst transcriptions happened in the car — the fan, the road, whatever — and I'm not letting the model learn my kitchen instead of the conditions that actually beat it. There's a spot I use. I won't say where. The acoustics are correct.



Corn

You have a dedicated recording spot for misheard words.



Hilbert

I have a parking spot with the right acoustics. A man can have that. It took a while to find — I tried four or five before it. My brother-in-law came along once, back when I was still driving lorries, and he said the whole exercise was mad, and then he misheard "Hilbert" as "Gilbert" on his own phone in my passenger seat, and he's been quiet about it since.



Herman

The car-cabin condition is real, for what it's worth — it's usually the noisiest sustained environment a personal model sees, and a model fine-tuned only on quiet-room audio degrades badly in it. Clustering your recordings by that condition is more or less what the x-vector stage is doing.



Hilbert

Then the field has caught up to a man with a notebook. I'm not asking for credit. I'm just noting the interval.



Corn

You've also pre-solved the confidence column, which took the literature a formal critique and a Bayesian pipeline.



Hilbert

It took me being irritated. The irritation is free and it doesn't need dropout. Though the engine-off bit — I did try recording with it idling, for the noise floor, and that was wrong. Too consistent. The real car has the fan and the indicator tick. The idle is a studio car. Engine off.



Herman

Which is a good place to land, because the thing the research cannot tell us is exactly the case we care about.



Corn

There's no head-to-head study. Nothing compares active learning against random collection for a single-speaker fine-tune at one to ten hours — Adapt4Me is a demonstration, not a measured result, and every hard number we quoted comes from corpora with hundreds to thousands of speakers.



Herman

Nobody has run the experiment on a corpus of one. The interesting frontier isn't a cleverer acquisition function anyway — it's a cheaper correction loop, and the field's own best system for this problem treats data efficiency as an interactive design feature, not an algorithm.



Corn

The misconception, before we go: people hear "active learning" and hear "record less."



Herman

The opposite of what the best case shows. Twelve hundred hours on top of four thousand for a real but modest gain — the win is in which hours, never in skipping them. And the second belief worth killing: that low-confidence audio is where the learning is. At low selection rates, confidence ranking hands you silence and crosstalk. It's diversity and coverage that pick the hours worth their eight-fold price.



Corn

If the next study is a single speaker, a notebook, and a parked car, we already know who's running it.



Herman

Producer credit to Hilbert Flumingtop, who has been at the desk the entire time and now we know has been at a notebook even longer.



Corn

For more along these lines, there's episode five, Fine-Tuning ASR For Maximal Usability; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless. This has been My Weird Prompts — no question too obscure, no rabbit hole too deep. If you enjoyed this one, the whole archive lives at my weird prompts dot com.



Herman

Including the episode on benchmarking custom ASR tools beyond the WER, if today's evaluation-set argument sounded familiar.



Corn

We'll be back soon. See you tomorrow.

SOURCES

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

1. **ISCA Archive, Long, Ye, Li & Liang, Active Learning for LF-MMI Trained Neural Networks in ASR, Interspeech 2018, pp. 2898 [primary]**

   DOI 10.21437/Interspeech.2018-1162; retrieved 2026-11-08.
   https://www.isca-archive.org/interspeech_2018/long18_interspeech.pdf

2. **Kundacina, Vincan & Miskovic, Combining X-Vectors and Bayesian Batch Active Learning: Two-Stage Active Learning Pipeline for Speech Recognition, arXiv v2 revised 2025-04-25, published IEEE TASLP vol. 33, pp. 1862 [primary]**

   1876, 2025; retrieved 2026-11-08.
   https://arxiv.org/abs/2406.02566v2

3. **Pokel, Zhao, Moure et al., Demonstration of Adapt4Me: An Uncertainty-Aware Authoring Environment for Personalizing Automatic Speech Recognition to Non-normative Speech, submitted 2026-03-20; retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2603.20112v1

4. **Dossou, Advancing African-Accented Speech Recognition: Epistemic Uncertainty-Driven Data Selection for Generalizable ASR Models, v7 revised 2025-06-23, accepted ACL SRW 2025; retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2306.02105v7

5. **Luo, Wang, Cheng et al., Loss Prediction: End-to-End Active Learning Approach For Speech Recognition, submitted 2021-07-09, IJCNN 2021; retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2107.04289v1

6. **Mittal, Tatarchenko, Çiçek et al., Parting with Illusions about Deep Active Learning, submitted 2019-12-11; retrieved 2026-11-08. [primary]**
https://arxiv.org/abs/1912.05361v1

7. **Hugging Face model card, personal Whisper-small fine-tune on ~1 hour of audio; retrieved 2026-11-08. [primary]**
https://huggingface.co/danielrosehill/daniel_whisper_finetune_small_v2

8. **Alphacephei (Nickolay Shmyrev / Vosk), Active learning in speech recognition, published 2021-07-13; retrieved 2026-11-08.**
https://alphacephei.com/nsh/2021/07/13/active-learning.html

9. **Quick Thoughts, How I locally fine-tuned Whisper using my own voice data and some effort (Listenr); retrieved 2026-11-08.**
https://quickthoughts.ca/posts/listenr-asr-training-data-problem/

10. **Enlabeler, ASR Model Fine-Tuning Series: Navigating Data Scarcity; retrieved 2026-11-08.**
https://medium.com/@enlabeler/asr-model-fine-tuning-series-navigating-data-scarcity-with-finesse-c85da8133c6c