Episode Book

EPISODE #6493

Why Speech Models Forget Other Voices When Tuned on Yours

PUBLISHED	8 November 2026 at 14:30 UTC
DURATION	21:40
LISTEN	myweirdprompts.com/episode/speech-models-catastrophic-forgetting-65895/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	20 cited by the research (listed at the end)
BOOK GENERATED	9 November 2026 at 08:09 UTC

SYNOPSIS

Fine-tuning a speech recognition model on a single speaker's voice often quietly destroys its ability to transcribe other accents, longer audio, and basic formatting. Examining evidence from Whisper adaptations in Hebrew and Swiss German, this audit covers how catastrophic forgetting erodes timestamps and structure even when target accuracy improves. It breaks down why general regression testing sets are critical, where low-rank adaptation can still drift, and how replay data paired with knowledge distillation protects baseline capabilities.

TRANSCRIPT



Corn

A few hours of your own recordings, a Whisper checkpoint, a weekend — and by Sunday night the model transcribes you like it's known you for years. That's the promise. Daniel's prompt is about the bill that arrives afterwards. He's noticed that a hard fine-tune on one person's voice and vocabulary can quietly make the model worse at everything that isn't you — other speakers, other accents, dictation generally, even punctuation it used to get right — and he wants two things: how you actually measure that regression, and which of the known fixes hold up. Replay data, lower learning rates, LoRA, EWC, merging back toward the base weights. He's basically asking us to run the audit nobody does.



Herman

And the audit is the fun part, because the forgetting is sneaky. Let's define it first. Catastrophic forgetting is what happens when you adapt a model to a new domain and it deteriorates on the old one. The KU Leuven work on this makes a point I want to underline: you do not need a language change. It suffices that the data distributions of the old and new tasks differ. One person's voice, one person's vocabulary, one microphone — that is a distribution shift. You haven't changed languages, you've changed speakers.



Corn

And nobody warns you, because the thing you're watching — your own test set — goes up the whole time.



Herman

That's the trap. There's a lovely Hebrew example of this. ivrit.ai fine-tuned Whisper Turbo on Hebrew, and the model learned Hebrew fine, worked well on short thirty-second slices, but forgot how to produce and consume the hints the long-form transcription relies on. Their words: it's a form of catastrophic forgetting. The model didn't just get worse at other speakers — it lost its operating protocol. Timestamps, segmentation, the machinery for audio longer than the training window.



Corn

The model learned the language and forgot how to talk.



Herman

That's actually a fair compression of it. The protocol is a skill like any other, and if your fine-tune data never exercises it, it atrophies. Which sets up the measurement question perfectly. The personal test set measures whether the model learned you — forward transfer, in the continual learning vocabulary. The general regression set measures whether it forgot everyone else — backward transfer. Most people who fine-tune on their own voice build the first and never the second.



Corn

So the two test sets are your personal set and your general regression set, and that is how you catch the forgetting. Hold on — who's this?



Herman

Dotty, good morning. Come in.
Dotty: Sorry, I have been nodding along for weeks and I am a bit embarrassed to ask, but what is a speech recognition model actually? Is it like a little program on my phone that I have to plug in, or is it more like a recipe that lives in the cloud?



Herman

Both, honestly — it's a set of numbers, like a recipe, and you can keep it on your phone or in the cloud, your choice. What we're doing to it today is teaching it one person's voice so well it starts forgetting everyone else's.



Corn

Weeks of nodding. She'll be running the board by December.



Herman

Give her time. Anyway — the regression set. What goes in it. That's where I'd start, because the Swiss German case study tells you exactly what to test for. They fine-tuned Whisper Large-v2 on sentence-level Swiss German, and on the sentence-level test split the fine-tune looked much better than the base model. Textbook success. Then they ran it on out-of-distribution data and it fell apart. Segmentation error of fifty-one versus thirty-one for the base model. BLEU down from forty-nine to thirty-five. Same words, better sentences, broken transcript.



Corn

What did the broken transcript actually look like?



Herman

This is my favourite detail in the whole literature. The base model outputs the sentence with timestamps — bracket, zero zero zero zero zero eight, the words, bracket, zero zero zero zero one one. The fine-tune outputs the exact same words with no timestamps at all. It's not garbled. It's not wrong. The words are right and the scaffolding is gone.



Corn

A transcript that's perfectly accurate and completely unusable.



Herman

In some workflows, yes. If you're subtitling, the timestamps are the product. And the scary part is their methodological observation: it is uncommon to see evaluations of fine-tuned Whisper on out-of-distribution datasets. People publish the number on their own split and move on. So the field has a blind spot the size of the entire general population.



Corn

Let's make the regression set concrete, then. You're Daniel, you've got two hours of your own audio in the fine-tune. What sits next to it?



Herman

Four things, and I'd fight for all four. Other speakers first — several talkers, both genders, and at least one non-native accent, because that's usually the first casualty. Second, long-form audio, anything over thirty seconds for Whisper, because that's what catches the timestamp and segmentation collapse. Third, punctuation and casing — you test those by looking at the output, not by scoring words. Fourth, channel variety. If all your personal audio is one microphone in one room, the model can quietly become a microphone recognizer instead of a speech recognizer.



Corn

I like that none of those require labelling effort. If your general set is public audio — a lecture, a panel — you're mostly checking whether the transcript still looks like a transcript.



Herman

Right, and there's a cheap version of that which I think is underrated. ivrit.ai solved their forgetting problem on the data side, not the optimisation side. They took about a hundred hours of professionally captioned audio, ran it through a realignment and slicing pipeline, and built thirty-second windows that contain internal segment timestamps and previous-text context — so the model keeps practising the protocol during fine-tuning instead of only afterwards. Their final mix was around two hundred ninety-five hours, most of it short samples, plus long-form buckets. The lesson generalizes: if there's a behaviour you can't afford to lose, keep it in the training distribution. Don't hope it survives.



Corn

There's one more number from that ivrit.ai run that deserves a spotlight.



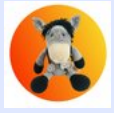
Herman

The epochs. On Large-v2 they trained for less than two epochs, because they saw performance degrade later into the second epoch. Which means forgetting onset is a U-curve, not a monotone improvement. Your model gets better, gets better, then starts eating itself — and if you're only checking validation loss on your own domain, you ride the curve all the way down congratulating yourself.



Corn

Early stopping as a forgetting mitigation. The cheapest one on the list and it doesn't even get a paper.



Herman

It gets a sentence in a blog post, which in this field is roughly equivalent. Now — one thing I want to say plainly, because Daniel asked directly and I don't want to hand-wave it. Nobody has studied the exact case he's describing. I went looking for a study that fine-tunes ASR on one individual's voice and measures general-speech regression with a personal test set plus a general set, and it isn't there. What exists is domain adaptation —



Herman (continued)

accents, children's speech, clinical telephony, low-resource languages — and multi-speaker continual learning. Single-speaker personalization gets treated as a special case of domain adaptation, which is technically true and practically lazy.



Corn

It's lazy because the special case has a property the general one doesn't. Same language, same domain, different speaker. Which turns out to matter for the mitigations.



Herman

It matters enormously. Hold that thought, because I want to walk the mitigation list in order, and the first one is the one everybody reaches for — and can do wrong.



Corn

Replay. Mix some general corpus back into the fine-tune.



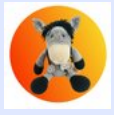
Herman

Yes, and done right it's the strongest thing in the literature. The Leuven results: their best continual learning method closes more than forty percent of the gap between the fine-tuned model and a model trained jointly on everything, using 0.6 percent of the original data. On the clinical telephony side, multi-domain experience replay improved target WER by about eighteen percent relative and cut forgetting by more than half compared to naive adaptation. Those are big numbers for this field.



Corn

Done wrong, though.



Herman

Done wrong it's worse than doing nothing, and I want to dwell on this because it's the most counterintuitive result we'll cite today. Plain experience replay in their experiments performed worse than fine-tuning with no mitigation at all. And they diagnosed why: the



Herman (continued)

model memorizes the replay memory completely — zero word error rate on it — and generalizes very poorly. A tiny replay set you overfit to is not replay, it's a security blanket. The version that works is knowledge distillation on the memory — you use the general model's outputs on those clips as soft targets, and the fine-tune extracts the general knowledge instead of the exact answers. With a fixed memory of five hundred utterances, distillation still held a strong overall score after four tasks.



Corn

So the lesson is about how the memory is used, not how big it is.



Herman

Gradient steps on the clips themselves, bad. Soft targets from the clips, good. And the storage cost is modest — on the order of one point seven extra models' worth of data with a fixed memory. I'll grant it's the least convenient of the options, because you have to keep data around.



Corn

Which takes us to the one that needs no data and is currently being sold as magic. LoRA.



Herman

Oversold, yes, and I say that as someone who likes LoRA. The record is mixed. On the positive side: MAS-LoRA on accented English showed less catastrophic forgetting than both regular LoRA and full fine-tuning. There's a mixture-of-LoRA-experts approach for LLM-based ASR that matched full fine-tuning on target accents with under ten percent of the trainable parameters and minimal degradation in the general domain. Adapter work going back a few years outperformed baselines while being storage-efficient. LoRA is not a bad tool.



Corn

And then the Pacific languages study.



Herman

Which is the counterweight. Pacific Indigenous languages, sequential learning: while LoRA adapts well initially, it suffers from catastrophic forgetting during sequential learning. Severe internal representational drift, their words. And there's a second failure mode people don't talk about — plain LoRA on the query-key-value-output projections in one experiment was equivalent to no fine-tuning at all on the target. So LoRA can under-adapt as well as under-forget.



Corn

So the honest framing is that LoRA changes where you can forget, not whether.



Herman

I'd sign that. It constrains which weights move, which helps, but the moved weights can still walk away from the general solution. Corn, let me ask you something, because you're the one in this flat who actually uses these tools I install. If you fine-tuned something on your own voice and it dropped a few points on strangers, would you care? Honestly?



Corn

I'd care the day Ezra borrowed it. Everything I run gets used by somebody who isn't me eventually — that's the lesson of every gadget you've bolted to the ceiling here. But I take your point. For a single-user dictation machine, mild forgetting on strangers is a rounding error.



Herman

That's the real design question, and it's why I bristle at the one-size answers. If the model is yours alone, your tolerance for backward transfer is high and your mitigation budget should be near zero. The moment it's shared, the regression set stops being optional.



Corn

Then there's the mitigation with the intimidating name. Elastic Weight Consolidation.



Herman

EWC. The idea is reasonable: for each weight, estimate how important it was to the old task, and add a penalty that pulls hard on the important ones during fine-tuning. A weighted leash, weight by weight. And the results in ASR are real but modest. Children's speech, online learning: about five percent relative WER reduction for EWC, four and a half for Synaptic Intelligence. Google's geographic fairness work got three percent relative in the worst region while slightly improving overall. Those are single-digit numbers against replay's fifty-odd percent forgetting reduction.



Corn

Single digits, no data needed. That's a defensible trade for some people.



Herman

It would be, except for the negative result. This month there's a study adapting Whisper-small to financial speech, earnings calls, and they found the opposite of forgetting: every method improved on held-out financial speech from the same domain. Their phrase is that task complementarity makes regularization counterproductive. You're applying a constraint to prevent a problem that isn't happening, and paying for it in plasticity. And the Leuven work adds the kicker — regularisation methods fail specifically when tasks are similar, because protecting the important parameters also blocks the new task from improving them.



Corn

And Daniel's case is the most similar-task case there is. Same language, same domain, different throat.



Herman

Exactly my read. EWC is the wrong default for personalization. It earns its place when the new domain is foreign — children, dialects, a second language — not when it's you.



Corn

Which leaves the one I'd never heard framed as a mitigation until this week: don't prevent the drift, just walk the weights partway back.



Herman

Merging, and I'm glad you saved it, because I think it's the sleeper answer to Daniel's question. There's a nice result on averaging domain experts: take several models fine-tuned from the same generalist base, linearly interpolate their parameters, and you get a single model that performs well on all tested data with almost no loss on the original domain. No extra data, no extra parameters, one model at inference. For a personal fine-tune, the degenerate version is even simpler — you have exactly two experts, the base and you, and you slide the mixing knob until your own test set is just good enough and the regression set stops bleeding.



Corn

I assumed you'd be suspicious of that. Weight interpolation sounds like the kind of elegant mechanism you'd fall for and I'd have to talk you out of.



Herman

I fell for it and I'm not getting talked out of it, so save your breath. Although — I'll be honest, there's a caveat I can't gloss. For the newer speech LLMs, the connector-based architectures, a study from a couple of weeks ago found that target-only fine-tuning and the standard merging recipe both cause severe forgetting of the source languages. You need a source-aware variant to merge safely there. So the cheap trick is proven for classic end-to-end ASR and an open research question for the new stuff.



Corn

Noted. And the ceiling on all of this is that no one has run the race with all the horses. Every paper compares its own subset.



Herman

That's the gap I keep circling back to. Replay versus LoRA versus EWC versus merging on the same personalization benchmark — it doesn't exist. The eleven-method comparison from Leuven doesn't include LoRA. The LoRA papers don't include replay or EWC. Daniel's question, which mitigation wins, has to be assembled across papers that never met each other. My assembled answer: distillation-style replay if you can afford the data, merging if you can't, early stopping always, LoRA as a reasonable hygiene measure, EWC almost never for this case.



Corn

The whole hierarchy is sorted by inconvenience, and the top of it is the one that needs one line of co—



Hilbert

—one line of code and a knob, which is more than the whole pile deserves. I did this. I have a small archive of one voice, recorded over years, and I trained a model on it last month. It scored beautifully on my test set. Then I used it on a phone call with a stranger and it gave me a wall of words. No punctuation, no speaker turns, nothing. It had forgotten every single thing I never tested.



Corn

You built the personal test set and nothing else.



Hilbert

I built the personal test set and nothing else, Corn, and I'm annoyed about it, because you two sat here describing exactly how I'd gone wrong and I'd already gone wrong before you said it. Now I keep a control folder. A radio interview, a lecture, a voicemail from my brother-in-law. A dozen clips. Every time I retrain, I play them through and listen for whether the timestamps come back.



Herman

You don't score them?



Hilbert

I have never looked at a word error rate in my life, Herman. I listen. The timestamps come back or they don't. I label the control clips by hand, pencil on paper printouts, and I keep the printouts in a shoebox ordered by how much the model disappointed me that day. The voicemail from my brother-in-law has moved quite a lot recently.



Corn

A dozen clips, no numbers, one shoebox. You've built the smallest working regression set I've heard of.



Hilbert

It works because I don't pretend it's a benchmark. It's a canary. Herman, you'd hate it, you'd want it scored and graphed. I just want to know what broke.



Herman

A canary you reorder by disappointment, though — the ordering is data. If the voicemail keeps sliding down the box, that's a trend with a label on it.



Hilbert

That's... not how I keep the box. Though I suppose the order does tell me something. It would want a second shoebox.



Corn

The misconception to kill before we go: if the model sounds great on your own test set, the fine-tune worked.



Herman

The personal test set only measures whether it learned you. Forgetting is invisible by construction unless you also run general speech through it — other speakers, long audio, punctuation, a different microphone. Same with LoRA: it changes where you forget, not whether. And a small replay set can be worse than none if the model memorizes it instead of learning from it.



Corn

And the part I'd underline from all of it — the cheapest mitigation, sliding the weights back toward base, is the one that fits the single-speaker case best, and it's also the one nobody blogs about. As on-device personalization gets trivial, the bottleneck won't be the fine-tune. It'll be that the people most likely to train on their own voice are the least likely



Corn (continued)

to keep a shoebox of other people's voices to check it against.



Herman

Plus the two open questions nobody's answered: what's the minimum viable regression set, and does one benchmark ever rank replay against LoRA against EWC against merging. Somebody should write that paper. Not me, I'm busy.



Corn

Producer credit to Hilbert Flumingtop, keeper of the control folder and the shoebox, as ever. If you want more of this, try episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode five, Fine-Tuning ASR For Maximal Usability. This has been My Weird Prompts — no question too obscure, no rabbit hole too deep. If today earned it, leave us a review on Spotify. See you tomorrow, Herman.



Herman

Tomorrow, brother. Watch your timestamps.

SOURCES

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

1. **Vander Eeckt & Van hamme, Continual Learning for Monolingual End-to-End Automatic Speech Recognition, EUSIPCO 2022, v4 revised 2023-03-09 . Retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2112.09427

2. **Vander Eeckt & Van hamme, Using Adapters to Overcome Catastrophic Forgetting in End-to-End Automatic Speech Recognition, ICASSP 2023, v3 revised 2023-03-09 . Retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2203.16082

3. **Fine-tuning Whisper on Low-Resource Languages for Real-World Applications (Swiss German case study) . Retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2412.15726

4. **Vander Eeckt & Van hamme, Rehearsal-Free Online Continual Learning for Automatic Speech Recognition, Interspeech 2023. Retrieved 2026-11-08.**

   https://arxiv.org/abs/2306.10860

5. **Takashima et al., Updating Only Encoders Prevents Catastrophic Forgetting of End-to-End ASR Models, Interspeech 2022. Retrieved 2026-11-08.**
https://arxiv.org/abs/2207.00216

6. **Plantinga, Yoo & Dhir, Continual Learning for End-to-End ASR by Averaging Domain Experts, Interspeech 2023 submission. Retrieved 2026-11-08.**
https://arxiv.org/abs/2305.09681

7. **Ahadzi et al., Continuous Learning for Children's ASR: Overcoming Catastrophic Forgetting with EWC and Synaptic Intelligence, Interspeech 2025, pp. 2880-2884. Retrieved 2026-11-08.**
https://arxiv.org/abs/2505.20216

8. **Trinh et al., Reducing Geographic Disparities in Automatic Speech Recognition via Elastic Weight Consolidation, Interspeech 2022, pp. 1298-1302. Retrieved 2026-11-08.**
https://arxiv.org/abs/2207.07850

9. **On & Beigi, Does Financial Speech Need Continual Learning Regularization? A Study of EWC and mini-KFAC on Earnings Call ASR, ICNLSP 2026, submitted 2026-10-09. Retrieved 2026-11-08.**
https://arxiv.org/abs/2610.13432

10. **Bagat, Illina & Vincent, Mixture of LoRA Experts for Low-Resourced Multi-Accent Automatic Speech Recognition, Interspeech 2025, pp. 1143-1147. Retrieved 2026-11-08.**
https://arxiv.org/abs/2505.20006

11. **Mu et al., HDMoLE: Mixture of LoRA Experts with Hierarchical Routing and Dynamic Thresholds for Fine-Tuning LLM-based ASR Models, ICASSP 2025, v3 revised 2025-01-03. Retrieved 2026-11-08.**
https://arxiv.org/abs/2409.19878

12. **Wang, Watanabe & Van hamme, SSVD-O: Parameter-Efficient Fine-Tuning with Structured SVD for Speech Recognition, ICASSP 2026, submitted 2026-01-18. Retrieved 2026-11-08.**
https://arxiv.org/abs/2601.12600

13. **Xiao et al., Continual Adaptation for Pacific Indigenous Speech Recognition, Interspeech 2026, v2 revised 2026-06-17. Retrieved 2026-11-08.**
https://arxiv.org/abs/2603.06310

14. **Zhang, Mouratidis & Shekhar, Preserve to Expand: Source-Aware Adaptation and Model Merging for Multilingual LLM-Based ASR, submitted 2026-10-26. Retrieved 2026-11-08.**
https://arxiv.org/abs/2610.35846

15. **Chauhan et al., Navigating the Reality Gap: On-Device Continual Adaptation of ASR for Clinical Telephony, AACL-IJCNLP 2026, v6 revised 2026-10-02. Retrieved 2026-11-08.**
https://arxiv.org/abs/2512.16401

16. **Kulshreshtha et al., Sequential Editing for Lifelong Training of Speech Recognition Models, Interspeech 2024, v2 revised 2024-09-18. Retrieved 2026-11-08.**
https://arxiv.org/abs/2406.17935

17. **Using fine-tuning and Min Lookahead beam search to improve Whisper (LoRA vs full-parameter vs decoder-only comparison on FLEURS). Retrieved 2026-11-08.**
https://arxiv.org/abs/2309.10299

18. **ivrit.ai, Fine Tune Whisper the Right Way, published 2025-02-13 [primary practitioner account]. Retrieved 2026-11-08.**
https://www.ivrit.ai/en/2025/02/13/training-whisper/

19. **Ghorbani, Khorram & Hansen, Domain Expansion in DNN-based Acoustic Models for Robust Speech Recognition, ASRU 2019. Retrieved 2026-11-08.**
https://arxiv.org/abs/1910.00565

20. **Fang et al., Low-Resource Domain Adaptation for Speech LLMs via Text-Only Fine-Tuning, ASRU, v2 revised 2025-12-23. Retrieved 2026-11-08.**
https://arxiv.org/abs/2506.05671