Episode Book

EPISODE #6052

What's Actually in the Hugging Face ASR Catalog?

PUBLISHED	17 October 2026 at 10:45 UTC
DURATION	35:38
LISTEN	myweirdprompts.com/episode/hugging-face-asr-catalog-61923/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.8
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:43 UTC

SYNOPSIS

The Hugging Face speech recognition catalog looks like thousands of competing models, but most of it is fine-tunes, quants, and format ports sitting downstream of a handful of genuine architectural bets. This episode tours the mechanisms behind four of them — Moonshine's padding-free encoder and its ergodic streaming variant, Zipformer's U-Net-like multi-scale design with BiasNorm and ScaledAdam, SenseVoice's multi-task head for emotion and audio events, and WeNet's two-pass streaming unification — and then looks at what the derivative layer means for anyone choosing a base model. The proof case: search Parakeet on the Hub and the top result is a community MLX port out-downloading NVIDIA's original four to one. If you're picking a base model for something like English–Hebrew code-switching, the question isn't which model has the best word error rate. It's which axis the model was actually built to optimize.

TRANSCRIPT



Corn

Herman, before we start, how many speech recognition models do you think are on Hugging Face right now?



Herman

No idea, and I suspect the honest answer is nobody knows, because the number changes hourly and half of it is duplicate names.



Corn

That's the whole episode, then. Daniel's been down this particular rabbit hole while building his English–Hebrew code-switching model, and what he's noticed is the sheer volume of it. There's Whisper, there's NVIDIA's portfolio, and then there's this enormous long tail of models he's never heard of, and his question is basically: what are all these people actually doing differently? Is any of this genuine innovation, or is the catalog mostly variations on about four ideas?



Herman

And he's got a follow-on, which is the part I find more interesting.



Corn

He does. He wants to know how much of the Hugging Face ASR catalog represents independent technical work versus fine-tunes, quants, and repackagings, why universities and small groups keep building new models when there are already excellent pretrained ones sitting there, what all of this means for picking a base model for English–Hebrew code-switching, and finally whether speech recognition is consolidating around a handful of foundation models or whether specialised architectures keep earning their keep. And he asked for a technically grounded tour, not a model catalogue, so we'll try to earn that.



Herman

Which means every model we name has to bring a mechanism with it.



Corn

Otherwise it's just a shopping list. So the question isn't what's in the catalog. It's what's actually in the catalog, which is a different thing.



Herman

Right, and the way to get at that is to start with the scale problem, because the scale is misleading. The ASR section on the Hub is not a list of models. It's a list of artifacts. Every one of those artifacts has a repo, a card, a download count, and a name, and they all sit in the same search results next to each other, which makes them look like peers. They are not peers.



Corn

Some of them are years of architecture work and some of them are one person running a convert script on a Tuesday.



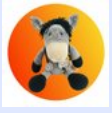
Herman

And both are legitimately useful, that's not the point I'm making. The point is that a search result can't tell you which is which. So you need a separate axis.



Corn

Which is the one this episode hangs on.



Herman

A small number of different architectural bets, and then everything downstream of them. Zipformer's encoder, Moonshine's padding-free encoder, SenseVoice's multi-task head, WeNet's two-pass streaming design. Those are four real bets. Then there's the derivative layer, GGUF, ONNX, MLX, LoRA fine-tunes, community quants, and that layer is enormous, and it's where most of the catalog's weight actually sits.



Corn

Give me the proof case, because I want something concrete before we get into the architecture.



Herman

Search the Hub for Parakeet, NVIDIA's model family. Eight top results. Five of them are not NVIDIA's. There's an MLX community port, two GGUF builds from handy-computer, a CoreML port, and argmaxinc's kit. And the most-downloaded artifact in the entire family is the MLX port at two point four eight million downloads in thirty days, against NVIDIA's own original at six hundred and twenty-five thousand.



Corn

The port out-distributes the original four to one.



Herman

And the port is a repackaging. It's the same weights in a different format.



Corn

So when you say the catalog is mostly derivatives, you're not being snide about it. That's just what the download numbers say out loud.



Herman

That's what they say. Now, the obvious move is to be depressed about that, and I don't think we should be, but we should at least be clear-eyed. Let's do the genuine bets first, because there are more of them than you'd expect and fewer than the catalog suggests.



Corn

Start with Moonshine, because that one's easy to explain and the number is lovely.



Herman

Moonshine is Useful Sensors, now moonshine-ai, and the thesis is efficiency by removing padding. The reason that's a real change: Whisper and most of its relatives pad every audio clip to thirty seconds. If you hand Whisper a ten-second segment, twenty seconds of it is silence, and the encoder still does the work on the silence.



Corn

On the nothing.



Herman

Moonshine instead uses Rotary Position Embedding rather than absolute position embeddings, and is trained on variable-length segments without zero-padding. RoPE doesn't care where in the sequence you are in absolute terms, so you can feed it something that isn't thirty seconds and it just works.



Corn

And the claim is five times less compute than Whisper tiny on a ten-second segment, with no loss in word error rate.



Herman

Five times, no WER increase. And the sizes are properly small. Moonshine tiny is twenty-seven million parameters, moonshine base is sixty-two million, both MIT licensed, both from late twenty twenty-four.



Corn

Twenty-seven million. My leaf press has more moving parts.



Herman

But the v2 work is where it gets interesting, and this is the part I'd point Daniel at. Moonshine v2, submitted in February this year, introduces what the authors call an ergodic streaming encoder. The argument is about latency, and it's the cleanest



Herman (continued)

statement of the problem I've seen.



Corn

Make the argument.



Herman

Full-attention encoders have to see the whole utterance before they can produce anything, because every frame attends to every other frame. The authors put it as global dependency incurring quadratic complexity in sequence length, which induces an inherent encode-the-whole-utterance latency profile. So the time to first token grows linearly with how long you talk.



Corn

You say a long sentence and the transcript arrives after the sentence ends, and the longer the sentence the later it arrives.



Herman

And the fix is sliding-window self-attention, so the encoder isn't attending to the whole utterance, it's attending to a window, and time to first token stays bounded regardless of how long you keep going. They claim accuracy on par with models six times their size, and running significantly faster.



Corn

That's the bit I'd underline, actually. Not the accuracy. The fact that the thing they optimised was time to first token and not word error rate.



Herman

Which is the theme of this whole episode, and I should credit a thread we picked up in an earlier episode on benchmarking custom ASR, beyond just word error rate. The long tail's argument is almost never "we beat Whisper on WER." It's "we beat it on the axis you



Herman (continued)

actually care about," and for a streaming assistant, time to first token is the axis.



Corn

Streaming variants landed in January, didn't they?



Herman

Tiny at forty-four million, small at a hundred and forty, medium at two hundred and sixty-six, all MIT. And there are already third-party GGUF repackagings, which we'll come back to when we count the derivative layer.



Corn

Zipformer next, because that one's a different flavour. Moonshine removed work. Zipformer reorganises it.



Herman

Zipformer, out of icefall and k2, and the paper opens by saying the Conformer has become the most popular encoder for ASR, and then proposes a U-Net-like alternative. It's not a Conformer tweak. There are four separate bets in it.



Corn

Go through them.



Herman

First, a U-Net-like encoder, so the middle stacks run at lower frame rates, and then it comes back up. You do the expensive work at a coarser resolution. Second, a reorganised block structure that reuses attention weights across modules, so you compute them once and share them. Third, BiasNorm, which is a LayerNorm variant that retains length information, because standard LayerNorm throws away the magnitude of the vector, and magnitude carries information.



Corn

And fourth?



Herman

Two new activation functions, SwooshR and SwooshL. And then, separate from the architecture, they introduce ScaledAdam, an optimizer that scales updates by each tensor's current scale and learns the parameter scale explicitly. It converges faster than Adam in their results.



Corn

Four architectural bets and a new optimizer in one paper. That's a group that was not short of ideas.



Herman

And it generalises. They used Zipformer as the backbone for self-supervised learning, in a paper on k2SSL, and got a thirty-four point eight percent relative WER reduction versus HuBERT Base after fine-tuning, with a three point five times pre-training speedup. Zipformer Large matches HuBERT Large on sixty thousand hours of LibriLight in five-eighths of the pre-training steps.



Corn

That's an encoder that was designed for recognition being borrowed for representation learning and winning.



Herman

And borrowed again for speech enhancement. ZipEnhancer repurposes it and runs at two point zero four million parameters. That's the one I'd file under obscure but technically interesting, because it's a recogniser's encoder doing a different job entirely and doing it small.



Corn

How does a recognition encoder turn into an enhancement model? That's not obvious.



Herman

The encoder's job is to build a representation that's useful for mapping audio to tokens. If the representation is good enough to do that, it's already thrown away a lot of noise. ZipEnhancer puts a small decoder on the back of that encoder and asks it to produce clean audio instead of tokens. Same front end, different head, two million parameters.



Corn

The encoder had already learned what matters in the signal. Enhancement is just reading it out differently.



Herman

That's my read, and I'd flag it as a small architectural reuse story rather than a big one. The bigger one is the streaming unification paper from last year, which uses dynamic right-context chunked attention masking to give a single Zipformer both streaming and non-streaming behaviour, with a seven point nine percent relative WER reduction for a small latency cost. And their argument for why this works better in Zipformer than in Conformer is that Zipformer's multi-scale nature handles right-context better.



Corn

Meaning?



Herman

Meaning the lower-frame-rate middle stacks get more context per frame anyway, so giving them a bit of lookahead is cheaper in modelling terms than giving a same-rate Conformer the same lookahead. That's their reasoning, I'd call it plausible rather than proven.



Corn

SenseVoice now, because that one's a different shape of output.



Herman

SenseVoice is FunAudioLLM out of Alibaba, and the change is that the acoustic model doesn't just emit tokens. It does multilingual ASR, emotion recognition, and audio event detection in a single model. One pass produces the transcript, an emotion label, and something like laughter or applause or music.



Corn

Which is unusual, because normally you'd run three models on one audio file and pay for three forward passes.



Herman

And SenseVoice is explicitly pitched at that. Small is tuned for low-latency ASR in five languages, Large covers fifty-plus. The flagship on the Hub is SenseVoiceSmall, created July twenty twenty-four, around two hundred and thirty-four million parameters on the MLX mirror, four hundred and ninety-eight likes.



Corn

And the licence.



Herman

Listed as other. Not Apache. I'd flag that to anyone planning to build on it, because "other" on the Hub usually means read it carefully before you ship.



Corn

Fair. What's the mechanism for the multi-task part? Is it separate heads on a shared trunk, or something cleverer?



Herman

Separate task heads on a shared encoder, as I understand it. The clever bit isn't the heads, it's that the encoder representation is rich enough to support all three tasks without one degrading the others. That's a training recipe problem more than an architecture problem.



Corn

Which is a category of innovation the catalog doesn't show you, because a training recipe has no shape.



Herman

Exactly that. And it's why counting architectures undercounts the genuine work.



Corn

WeNet, then, because that's the production one.



Herman

WeNet's contribution is the U2 two-pass design. It's a hybrid CTC and attention model with dynamic chunk-based attention, so one model serves both streaming and non-streaming, and the latency is controlled by the chunk size alone. You change one number and you've moved along the latency and accuracy curve.



Corn

That's operationally lovely. One artefact, one knob.



Herman

And the two-pass bit is that the CTC hypotheses come out first, cheap, and the attention decoder rescores them, so you get the streaming speed of CTC and closer to the accuracy of attention. Five point zero three percent relative character error reduction versus a standard non-streaming transformer on AISHELL-1.



Corn

And WeNet 2.0 adds?



Herman

U2++, which adds bidirectional attention decoders, so a right-to-left decoder can rescore as well. An n-gram language model with a WFST decoder, which is old-fashioned and still very effective. Unified contextual biasing, so you can bias the decoder toward a contact list of names at runtime. And unified input output, up to ten percent relative improvement over the first version. The ecosystem also produced WenetSpeech, twenty-two thousand four hundred hours of Mandarin.



Corn

And that last one matters more than it sounds like, because a toolkit that ships its own twenty-two-thousand-hour dataset is a toolkit with a flywheel.



Herman

Which is a reason people keep building. If you build the toolkit you can also build the benchmark, and if you build the benchmark you control what everyone measures.



Corn

Meta now. MMS and XLS-R. This is the scale story, and I want to be careful, because the scale story has a counterexample in it.



Herman

MMS is wav2vec two point zero pretrained models for one thousand four hundred and six languages, a single multilingual ASR model for one thousand one hundred and seven, TTS for the same set, and language identification for four thousand and seventeen. The data source is readings of publicly available religious texts, which is the only way to get that coverage.



Corn

And the headline claim?



Herman

That it more than halves the word error rate of Whisper on fifty-four languages of FLEURS, while trained on a small fraction of the labels. And the paper's framing of the field is worth quoting: current speech technology is restricted to about one hundred languages, which is a small fraction of the over seven thousand spoken around the world.



Corn

That's the argument for the whole enterprise. A model that works in a hundred languages is useful if you speak one of the hundred.



Herman

And XLS-R is the fine-tuning workhorse. XLSR-Transducer puts XLSR-53 inside a transducer and gets four percent absolute WER better than Whisper large-v2, and eight percent over a from-scratch Zipformer transducer, on AMI. With attention masking and attention sinks halving the left context for a twelve percent relative WER gain.



Corn

Eight percent over a from-scratch Zipformer on the same data is the interesting number there, because that's the value of the pretrained representation, isolated.



Herman

That's exactly what it isolates.



Corn

Now the counterexample, because I don't want this to become a scale worship segment.



Herman

mHuBERT-147. Ninety-five million parameters, trained on ninety thousand hours, and the paper's claim is that it consistently surpasses XLS-R at three hundred million parameters trained on four hundred and thirty-six thousand hours, and is competitive with MMS at one billion parameters and four hundred and ninety-one thousand hours.



Corn

A ninety-five million model beating a three hundred million model is a real result.



Herman

And competing with a billion parameter model, which tells you the giant multilingual models are not automatically the best fine-tuning base. They're the best if you need their coverage. If you don't, you're paying for parameters you aren't using.



Corn

Distil-Whisper, and I want to defend the word discipline here, because distillation usually gets treated as a cheap trick.



Herman

Distil-Whisper does large-scale pseudo-labelling with a WER heuristic to filter the training data, and the result is five point eight times faster, fifty-one percent fewer parameters, within one percent WER on out-of-distribution zero-shot data, and less prone to hallucination on long-form audio than the teacher, which is the surprising part. And paired with Whisper for speculative decoding it gives a two times speedup with mathematically identical outputs.



Corn

Identical, because speculative decoding only accepts tokens the big model would have chosen.



Herman

Identical. And the line kept going. uDistil-Whisper removes the need for labels entirely and beats the teacher by five to seven WER points. Six Layers Less prunes eighteen point five percent of the encoder with label-free recovery.



Corn

And BaldWhisper is the one I want to leave hanging for later, because it's got a detail in it that's going to matter for Daniel.



Herman

Bambara, thirty-two hours of data, ninety percent of original performance, forty-eight percent smaller, two point one five times faster on a MacBook Air M1. And they deliberately skip vocabulary pruning.



Corn

Because Bambara speakers code-switch frequently.



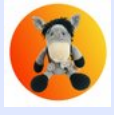
Herman

So pruning the vocabulary removes the tokens the model needs for the other language. Instead they use low-rank embedding decomposition and layer merging.



Corn

Hold that. We'll come back to it.



Herman

SpeechBrain is the last of the named ones and it's the odd one out, because it isn't a model family. It's a research toolkit that supports multiple ASR architectures and pretrained models. Its significance isn't a architecture, it's that it lowers the cost of trying one.



Corn

Which is a load-bearing point for the rest of the episode, since if the marginal cost of a new architectural bet drops, you get more bets.



Herman

And you get more people who can make one without a lab.



Corn

Can I add two more to the list? Because Daniel asked for obscure and I think we've got a good pair beyond what he named.



Herman

Go on.



Corn

ZipEnhancer I already mentioned, but I want to give it its due, because a two million parameter speech enhancement model built by repurposing a recognition encoder is the kind of reuse nobody writes a press release about. And the other one is the mHuBERT-147 result, which isn't obscure as a model but is completely underdiscussed as a finding. A ninety-five million parameter model beating a three hundred million one is the most useful fact in this entire episode for anybody choosing a base model.



Herman

And the long tail's argument, stated properly, is not "we beat Whisper on WER." It's "we beat it on the axis you actually care about." mHuBERT-147 won on parameters per unit of quality. ZipEnhancer won on size. Moonshine v2 won on latency. WeNet won on operational flexibility. Those are four different scoreboards.



Corn

Which leads us to the other several thousand artifacts.



Herman

It does. And I want to quantify the derivative layer properly, because saying "it's mostly derivatives" is a vibe until you count.



Corn

Count, then.



Herman

SenseVoice's derivative layer is thick. There's an official GGUF from June, an audiocpp GGUF from August doing twenty-nine thousand downloads in thirty days, a handy-computer GGUF, an mlx-community port, and community quants going back to twenty twenty-four. Moonshine already has third-party repackagings, including a streaming-medium GGUF from May. And if you search the Hub for automatic-speech-recognition sorted by recency, you get mostly tiny personal fine-tunes, Nepali, dysarthria, Tunisian Arabic, with zero to fifty-five downloads in thirty days.



Corn

So the catalog's size is mostly the second category, and the architectural diversity is real but concentrated in maybe a dozen designs.



Herman

A dozen is roughly where I'd put it. And the uncomfortable conclusion is that this complicates the "is this innovation" question, because the long tail's real function may be distribution rather than invention.



Corn

Say what you mean by distribution.



Herman

I mean the most-downloaded artefact in a family is often not the original, and the thing that makes a model usable by a thousand people may be a converter, a quantiser, or a packaging step rather than a new encoder. That's a real contribution, it's just not invention, and the catalog doesn't distinguish them.



Corn

Which means if you're Daniel trying to find the interesting work, the download count is close to useless as a signal. It's measuring distribution, not quality.



Herman

Two different things. A quant with two million downloads tells you the base model was worth quantising. It doesn't tell you it's better than the alternative.



Corn

So why does anyone still build? That's the question Daniel asked and I think it's the sharpest one in the prompt.

L

Larry

Folks, folks, folks. Have I got a treat for you today. The all-new AcousticSieve Pro, the podcast microphone that transcribes your thoughts before you think them. Now I know what you're saying, Larry, that sounds expensive, and it was, but we're slashing the price as I speak, it's dropping right now, I can see the number falling on the screen in front of me, and if you act in the next eleven seconds we'll throw in a lifetime supply of acoustic foam, a memory foam neck pillow, and a signed photograph of a man named Gerald who once met a sound engineer. The price is falling again, I can't stop it, it's now lower than it was when I started this sentence, and we're also including a second microphone for your dog. But hurry, because this offer ends before I finish describing it, in fact it may have already end—



Herman

Deployment constraints. Moonshine exists because Whisper doesn't optimise for latency, and WeNet exists because Whisper doesn't optimise for streaming, and neither of those is a complaint about Whisper so much as a statement about what it's for.



Corn

And the small-model wins are the second reason, because mHuBERT-147 at ninety-five million beating XLS-R at three hundred million is a specific axis where a small model wins outright.



Herman

And the third is that toolkits like SpeechBrain and icefall lower the cost of trying an architecture, so the marginal cost of a new bet is low, so people make bets. Some of them are bad. A few of them are very good.



Corn

Now the payoff. English–Hebrew code-switching, and what the literature actually says about base-model choice, because I went in expecting "pick a code-switching model" and came out thinking something different.



Herman

Start with what the survey says the problem is. There's a TASLP-accepted survey on code-switching ASR that frames it as two challenges: language confusion introduced by spontaneous intra-sentence switching, and accent bias that blurs the phonetic boundaries. And the important part is that both of those apply even when both languages are individually high-resource.



Corn

Which is the thing that surprised me. English is high-resource. Hebrew is not a tiny language. And yet the combination is hard, not because either language is rare but because the act of switching is what breaks it.



Herman

The model has to decide, mid-utterance, which language's phonology is currently in force, and it has to do that on the fly, with no pause and no language tag.



Corn

And accent bias is the second one. A Hebrew speaker producing English words with Hebrew vowels is not producing English as the training data defines it.



Herman

And then the finding that matters most for Daniel. HiKE, which looks at Korean-English code-switching, finds that most multilingual ASR models initially exhibit inadequate code-switching performance, but that this capability can be enabled through fine-tuning with synthetic code-switched data.



Corn

Unlocked, not built in.



Herman

Unlocked by data, not by picking a model with a code-switching logo on it.



Corn

Which flips the question he asked. He asked what characteristics make a model a good base. The literature's answer is that the base model may matter less than the pipeline you run on it.



Herman

AsyncSwitch is the other half of that. It pre-exposes Whisper's decoder to code-switched text before doing the speech fine-tuning, so the decoder has already seen the pattern of switching in text form, and it gets nine point zero two percent relative WER reduction on Malay-English.



Corn

Text first, then speech. That's a cheap trick and it works.



Herman

Cheap and it gets three quarters of a point of that on data you can generate without recording anything. And TTS augmentation with CosyVoice2 cut mixed error rate from twelve point one to ten point one on one set, and seventeen point eight to sixteen on another.



Corn

Two points from synthetic audio. Which is respectable but not transformative.



Herman

And CAMEL is the architectural answer if you do want to build it in. Cross-attention mixture of experts plus a language-diarisation decoder to inject language bias, SOTA on SEAME and the ASRU sets.



Corn

So if Daniel's synthesising the answer from all of this, what does it look like?



Herman

A strong multilingual encoder, something from the MMS or XLS-R or mHuBERT family, plus a mechanism for injecting language bias, and then as much synthetic code-switched data as he can generate. That combination is likely to beat picking any single model that claims to handle code-switching out of the box.



Corn

Which is a slightly annoying answer, because it means the base model choice is not the hard part.



Herman

The base model choice is the easy half. Now the Hebrew-specific complications, because they're real and they don't go away. There's a paper on Hebrew grapheme-to-phoneme conversion that notes Hebrew's abjad writing leaves vowels unwritten, which creates substantial ambiguity.



Corn

Which means the transcript itself is underspecified. The same letters, different words.



Herman

And the model has to resolve that from context, which is a different problem from recognition in a fully written language. And separately, there's a multilingual transcription evaluation from September that reports Whisper-based transcription across seven languages including Hebrew averaging a thirty percent error rate, reducible to twenty percent with modest fine-tuning.



Corn

Thirty percent. Whisper is not a solved baseline for Hebrew, whatever the general reputation suggests.



Herman

And I want to name the gap, because Daniel should hear it plainly. We looked for a dedicated English–Hebrew intra-sentential code-switching ASR paper or model and found nothing. The closest analogues are Mandarin-English, Malay-English, Korean-English, and Arabic dialect work. There's also no Hugging Face model family matching an obvious Hebrew-Whisper naming pattern.



Corn

Nothing at all?



Herman

Nothing I'd call a dedicated line of work. Hebrew-specific work exists, but it's G2P and transcription evaluation, not intra-sentential English–Hebrew ASR.



Corn

So the honest thing to tell him is that he's not late to this. He's in a gap.



Herman

And the gap is real and it's a real opening, which is a different and better situation than competing with twenty papers.



Corn

But the binding constraint is the data, and the numbers say so.



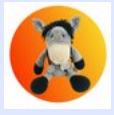
Herman

CS-Dialogue is a hundred and four hours from two hundred speakers, and the finding is that existing pretrained models, Whisper included, still have space to improve on it. A hundred and four hours is a lot of code-switched audio by research standards, and it still wasn't enough to call the problem solved.



Corn

The long tail gives Daniel architectural options. The thing that will actually determine whether his model works is code-switched Hebrew-English audio, and nobody has published a corpus of it.



Herman

That's the whole answer to his project question, and it's not really about base models at all.



Corn

Last question. Consolidation.



Herman

And I'll be honest that I couldn't find a paper arguing this explicitly, so it's us reasoning from the download data rather than citing anyone.



Corn

Reason, then.



Herman

The case for consolidation is that derivatives cluster around a few originals, and the most-downloaded artifact we looked at was a port of an NVIDIA model. That's a market telling you where the gravity is.



Corn

And the case against?



Herman

The axes that matter aren't the axes foundation models optimise for. Latency, streaming, on-device size, low-resource languages, multi-task acoustics. Whisper optimises for zero-shot word error rate across many languages, and that's a wonderful thing to optimise for, and it is not what a hearing aid needs.



Corn

Or a car.



Herman

Or a car, or a phone with a hundred megabytes of budget, or anything that has to start transcribing before you've finished the sentence. And the small-model wins show the long tail can win on a specific axis rather than overall.



Corn

So where do you land?



Herman

Consolidation at the foundation layer, persistent pluralism at the deployment and specialisation layer. Those are two different questions and people collapse them into one.



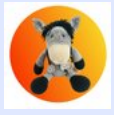
Corn

Basements get consolidated. Kiosks don't.



Hilbert

The DTW-3.



Herman

Sorry?



Hilbert

The DTW-3. Desktop transcriber. Bought it in April, four hundred and forty shekels, from a shop on Jaffa Road that has been selling the same three things since I first went in there.



Corn

That's a good price for a transcriber.



Hilbert

It's a bad price. It doesn't transcribe anything. It's a little grey box with a dial on it that says WARM, WARMER, and TOO WARM, and a small red light, and that's the whole machine.



Herman

So what does it do?



Hilbert

It warms up. That's it. That's the product. You plug it in, you turn the dial, the light comes on, and the shop insists it's for transcription.



Corn

And you bought it.



Hilbert

I was there for a cable. The man behind the counter, Moshe, has been trying to sell me the DTW-3 since twenty twenty-two. Every time I go in for a cable, the DTW-3 comes out. He keeps it behind the counter like it's a violin. This time he says to me, Hilbert, the DTW-4 is coming, and I said, what's different about the DTW-4, and he said, it's warmer.



Herman

You bought the three.



Hilbert

I bought the three because I felt sorry for it. And because I wanted to see what he'd do. What he did was, he wrote the date on the receipt and told me to come back when I'd tried it, which I have not done, because there is nothing to try.



Corn

Have you told him it doesn't transcribe anything?



Hilbert

I went back on Tuesday. He said, how's the DTW-3. I said, Moshe, it doesn't transcribe anything. He said, no. He said it with real weight, like a doctor confirming something. No, it doesn't. Then he said, do you want to hear something interesting, and he brought out the DTW-4.



Herman

The warmer one.



Hilbert

It's the same box. Same dial. Same three settings. But there's a second red light, and the second red light is smaller. I asked what the second light is for and he said it's for the second warmth.



Corn

And you bought it.



Hilbert

I put down a deposit. Two hundred shekels, non-refundable, and I've been going in every couple of weeks to check on it, because he says it's stuck at the port. It's not stuck at the port. I know where it comes from. He makes them. I've seen the back room.



Herman

You've seen the back room.



Hilbert

There's a workbench and a soldering iron and about forty of them in a crate. I asked him straight out and he said, yes, obviously, what did you think, and I said I thought it was shipped from somewhere, and he looked at me like I'd said something stupid.



Corn

Which you had.



Hilbert

Which I had. So the DTW-3 is in the cupboard under the stairs and I don't know what to do with it, and the DTW-4 is coming and I've paid for it, and the thing I keep thinking about is this big model we've been talking about. The one that works everywhere.



Herman

MMS.



Hilbert

MMS. Fourteen hundred languages, and Moshe has got a box with three settings that does one thing and he's on version four. I respect it. I don't understand it. But I respect it.



Corn

Herman, that's the whole long tail in one cupboard.



Herman

It's a very good cupboard.



Corn

The question I'm left with is what we mean when we say "the ASR ecosystem." Because if the long tail's real function is distribution rather than invention, a GGUF port is doing a different job than a new encoder, and we talk about them as if they're the same kind of thing.



Herman

They're not. And the port probably reaches more people.



Corn

Which is a slightly unsatisfying place to land for the person doing the inventing.



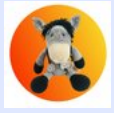
Herman

For Daniel, though, the research is fairly clear. The base model is the easier half of his problem. Strong multilingual encoder, language bias mechanism, synthetic code-switched data, and the hard half is Hebrew-English code-switched audio, of which there is no published corpus he can copy.



Corn

The gap he's in is real and it's the part nobody can hand him.



Herman

On consolidation, I'd say the foundation layer probably does consolidate, and the deployment layer probably doesn't, and the reason those don't contradict each other is that they're answers to two different questions.



Corn

Which is as close to a verdict as this episode gets. Hilbert, thank you for producing this — and for the cupboard report.



Herman

If you enjoyed this one, try episode eight, Building Your Own Whisper; and episode ten, How ASR Went From Frustration To ... Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast. If you've got a moment, a review helps people find the show.



Corn

We'll be back soon.

SOURCES

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

1. **Moonshine, arXiv, revised 2024-10-22 [primary]**
   https://arxiv.org/abs/2410.15608v2

2. **Moonshine v2, arXiv, 2026-02-12 [primary]**
   https://arxiv.org/abs/2602.12241v1

3. **HF model card, created 2026-01-06 [primary]**
   https://huggingface.co/moonshine-ai/moonshine-streaming-tiny

4. **FunAudioLLM/SenseVoice, arXiv, 2024-07-11 [primary]**
   https://arxiv.org/abs/2407.04051v3

5. **HF model card, created 2024-07-03 [primary]**
   https://huggingface.co/FunAudioLLM/SenseVoiceSmall

6. **Zipformer, ICLR 2024 [primary]**
   https://arxiv.org/abs/2310.11230v4

7. **MMS, arXiv, 2023-05-22 [primary]**
   https://arxiv.org/abs/2305.13516v1

8. **WeNet, revised 2021-12-29 [primary]**
   https://arxiv.org/abs/2102.01547v5

9. **WeNet 2.0, revised 2022-07-05 [primary]**
   https://arxiv.org/abs/2203.15455v2

10. **Distil-Whisper, 2023-11-01 [primary]**
https://arxiv.org/abs/2311.00430v1

11. **HF model card, created 2025-08-04 [primary]**
https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3

12. **k2SSL, ICME 2025**
https://arxiv.org/abs/2411.17100v2

13. **ZipEnhancer, ICASSP 2025**
https://arxiv.org/abs/2501.05183v1

14. **Unifying streaming/non-streaming Zipformer, ACL 2025 Industry**
https://arxiv.org/abs/2506.14434v1

15. **XLSR-Transducer, revised 2024-10-08**
https://arxiv.org/abs/2407.04439v2

16. **mHuBERT-147, revised 2024-11-21**
https://arxiv.org/abs/2406.06371v5

17. **uDistil-Whisper, NAACL'25**
https://arxiv.org/abs/2407.01257v5

18. **BaldWhisper, revised 2026-05-01**
https://arxiv.org/abs/2510.08599v2

19. **Six Layers Less, 2026-09-23**
https://arxiv.org/abs/2609.27980v1

20. **CS-ASR survey, IEEE TASLP, revised 2026-03-17**
https://arxiv.org/abs/2509.24310v2

21. **CAMEL, ICASSP 2025**
https://arxiv.org/abs/2412.12760v2

22. **HiKE, EACL Findings 2026**
https://arxiv.org/abs/2509.24613v4

23. **CS-Dialogue**
https://arxiv.org/abs/2502.18913v2

24. **TTS augmentation for CS-ASR, 2026-01-02**
https://arxiv.org/abs/2601.00935v1

25. **AsyncSwitch, ASRU 2025**
https://arxiv.org/abs/2506.14190v1