Episode Book

EPISODE #6526

Why wav2vec 2.0 Still Dominates Speech Downloads

PUBLISHED	11 November 2026 at 14:53 UTC
DURATION	25:59
LISTEN	myweirdprompts.com/episode/wav2vec-speech-recognition-persistence-29210/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	14 cited by the research (listed at the end)
BOOK GENERATED	12 November 2026 at 12:58 UTC

SYNOPSIS

Modern speech leaderboards are dominated by giant models like NVIDIA Parakeet and Whisper, yet Meta's wav2vec 2.0 base checkpoint still commands roughly 1.8 million downloads a month. The architecture turned raw speech into discrete, tokenized units using clever self-supervised pre-training with minimal labeled audio. Community fine-tunes like XLSR variants and paralinguistic encoders explain how this lightweight model became the entrenched default across edge devices, speech research, and standard developer toolchains.

TRANSCRIPT



Corn

The standard line on wav2vec is that it's been surpassed, put it on the shelf next to the flip phone. And that's true on the benchmarks and completely false on the download charts. Daniel's prompt lands right on that gap. He points out it's no longer at the cutting edge of speech recognition, and yet it ranks among the most downloaded models in its category on Hugging Face, and it's a canonical model in speech-to-text. He wants to talk about why this model and the family of fine-tunes descended from it are particularly significant. Not whether it's still good. Why it's still everywhere.



Herman

And it's a strange situation, because both halves of his observation are correct and they seem like they shouldn't be. The Open ASR Leaderboard right now is NVIDIA Parakeet, Canary, Qwen3 territory. There is no wav2vec 2.0 checkpoint anywhere in the top tier. Meanwhile the canonical base model, facebook/wav2vec2-base-960h, is still pulling about 1.8 million downloads a month.



Corn

A model nobody modified since 2022 getting 1.8 million downloads a month. That's not a graveyard, that's a commuter line.



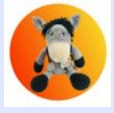
Herman

Right, something is being built with it, every day, at scale. That's the mystery Daniel's poking at.



Corn

Herman Poppleberry, take us to the beginning. What is this thing, actually?



Herman

The name covers three models over about eighteen months, and the sequence matters. The original wav2vec landed in twenty nineteen, and the idea was almost impertinent in its simplicity. You take raw audio, unlabelled, and you pre-train a fairly plain convolutional network with a contrastive task, basically teaching it to tell real speech fragments from distractors. No transcripts anywhere. Then you fine-tune on a small amount of transcribed



Herman (continued)

speech. On the Wall Street Journal corpus it cut error rates by up to thirty-six percent when you only had a few hours of transcription, and it beat Deep Speech 2, which was the best character-based system in the literature, while using a hundred times less labelled data.



Corn

A hundred times less. In a field that had been solving accuracy by throwing annotators at it.



Herman

That was the whole paradigm. Pre-wav2vec, the story of speech recognition was supervised-at-scale. Get transcripts, get more transcripts, rent the GPUs. And Meta's own framing of the problem was blunt. They said it's simply not feasible to obtain resources for each dialect and every language across all the possible domains. Read speech, telephone speech. They were right. So the move was, learn the structure of speech from the audio itself, and only pay for transcription at the very end, in small doses.



Corn

The second one, the vq one, is the bridge I always lose people on.



Herman

vq-wav2vec, late twenty nineteen. That added quantisation. You force the continuous audio into a finite set of discrete speech units, and suddenly speech looks like tokens, and everything from the NLP toolbox applies. They ran BERT-style pre-training on top of those tokens and set a new state of the art on TIMIT phoneme classification and on WSJ. That's the conceptual jump. Speech stopped being an audio problem and became a language-modelling problem.



Corn

And then 2.0 is where it clicks into the shape everyone now recognizes.



Herman

Mid-2020. Baevski, Zhou, Mohamed, Auli. The mechanism is worth walking through because the design choice is the clever part. The raw waveform goes through a multilayer CNN, which produces latent representations every twenty-five milliseconds. Those go two places at once. One copy goes to a quantizer, which maps each little chunk of audio to a unit from a finite learned inventory, short sub-phoneme units. The other copy has about half its positions masked, and then goes through a transformer, which sees the whole sequence with holes in it. The training task is a contrastive one. Given the transformer's output at a masked position, pick the correct quantized unit out of a set of candidates that includes decoys.



Corn

So it's BERT's masked-word game, except the words never existed. The model has to invent the vocabulary before it can play.



Herman

That's a better one-line version than the paper manages. And the contrast with the alternatives is the detail people skip. Other self-supervised approaches tried to reconstruct the audio signal. The problem there is that a recording contains everything. Room acoustics, microphone hiss, the speaker's timbre. If your task is to rebuild the waveform, you must represent all of it. wav2vec 2.0's quantizer can't. The inventory is finite, so the model physically cannot encode every variation. It's forced to spend its capacity on what actually matters for the linguistic content, and the noise gets rounded away by design.



Corn

The constraint is the feature. You couldn't pay me to believe that on purpose, but there it is in the results.



Herman

Which results do you want, the famous ones?



Corn

The ten minutes. Everyone quotes the ten minutes.



Herman

Ten minutes of labelled data, pre-training on fifty-three thousand hours of unlabelled audio, and it hits 4.8 word error rate on LibriSpeech clean, 8.2 on the noisy set. And with one hour of labels it beat the previous state of the art on LibriSpeech's hundred-hour subset using a hundred times less labelled data. That number is why the paper matters. Ten minutes is a coffee break. That's a transcript shorter than one of our episodes.



Corn

One of your segments. And the blog version of that same result says 5.2 clean, 8.6 noisy, which doesn't match the paper's 4.8 and 8.2.



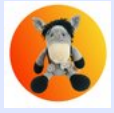
Herman

It doesn't, and I like that discrepancy. The blog post went up in September 2020 with one set of numbers, the paper went through revisions, and by the camera-ready version the same experiment reports slightly better figures. It's a tiny, harmless example of how a result firms up between announcement and publication. The blog number is the one that circulated, because the blog is what people read.



Corn

Nobody reads the revision history. They read the headline and inherit its numbers forever.



Herman

So that's the model. Then comes the part Daniel's really asking about, the family. Two releases define the trunk. XLSR in 2020 took the same recipe multilingual, pre-training one model on raw speech in fifty-three languages with the quantized units shared across languages. Seventy-two percent relative phoneme error rate reduction on CommonVoice, sixteen percent on BABEL. And the unit sharing was beautiful. Some units turn out to be used by only one language, others by languages that sit close together, English and German, Basque and Catalan. The model rediscovered linguistic geography from raw audio.



Corn

Nobody told it English and German were related.



Herman

It found out on its own. Then XLS-R, a year later, scaled it hard. Two billion parameters, roughly half a million hours of publicly available audio, a hundred and twenty-eight languages. An order of magnitude more public data than the largest prior work. Fourteen to thirty-four percent relative error reduction across four benchmarks, plus 7.4 BLEU on speech translation into English over twenty-one directions. That's the lineage. But here's the twist, and this is the fact that reframes the whole story. The most-downloaded wav2vec 2.0 model on Hugging Face is not 's.



Corn

Whose is it?



Herman

One contributor, jonatasgrosman, and his Japanese fine-tune. Wav2vec2-large-xlsr-53-japanese, sixteen point four million downloads in the last thirty days. That's roughly nine times the canonical base-960h checkpoint from. And his Portuguese, Russian, Polish, Dutch, Arabic, Greek, Chinese and Hungarian fine-tunes fill out most of the top fifteen. One person's language fine-tunes, built on 's base, out-downloading the company that built it.



Corn

The factory's proudest product is somebody else's aftermarket exhaust.



Herman

That's the community-driven long tail, and it's the answer to half of what makes the family significant. Shipped a backbone. The community shipped the world's languages.



Corn

Before you get to why it's still downloaded, I want to test the other half of the family claim, because transcription is only the visible use. The audeering models, the age and gender one.



Herman

Two point three six million downloads a month, and it's not a transcriber at all. It takes the wav2vec 2.0 encoder and puts classification heads on it for speaker age and gender. Their emotion model, based on the same backbone, does dimensional emotion recognition. There's a model doing phoneme recognition across sixty languages, and a 2025 model from a speech lab that does L2 English pronunciation assessment, aimed at language learners. Speaker verification work got a state-of-the-art result on VoxCeleb with wav2vec 2.Dysarthric speech, cerebral palsy, Parkinson's, paralysis, XLSR features cut error rates on those corpora significantly.



Corn

So it became a general speech encoder that happens to also know how to transcribe.



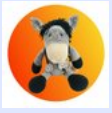
Herman

That's the pattern, encoder as feature extractor. The same way nobody uses BERT for anything directly anymore but half the field sits on transformers, wav2vec 2.0 turned into the backbone you bolt your task onto. When we talked about how speech-to-text got good, that was the Whisper side of the story. This is the other lineage. Whisper is a transcriber. wav2vec 2.0 is a set of ears that learned what speech is.



Corn

And yet the ears are no longer the best ears. Play the paradox out for me.



Herman

The Open ASR Leaderboard today, NVIDIA Parakeet at 6.5 average word error rate running at over six thousand times real time, Canary with a Qwen decoder just under that, Whisper large at 7.8. No wav2vec 2.0 anywhere near the top. For the mass market the numbers are even starker. Argmax's WhisperKit CoreML package does fifty-seven point six million downloads a month, thirty-two times the canonical wav2vec 2.0 checkpoint. And on likes, the culture metric, Whisper large has about sixteen times the likes.



Corn

So on every measure of fashion, it lost. And then every month, 1.8 million downloads of the base checkpoint alone, plus the long tail. There's no published study explaining that gap, is there.



Herman

None that I've seen. What follows is inference from model cards, course structure and the download distribution. My money is on five mechanisms, and the first is the boring one. It's the default in the tooling. Hugging Face's Transformers documentation uses wav2vec 2.0 as the canonical CTC-based example. The Audio Course has a dedicated chapter on fine-tuning the small version, pitched at any sixteen-gigabyte GPU. There's a tutorial with the model's name literally in the title. Every single person who follows the canonical path into speech technology downloads the checkpoint. That's not nostalgia, that's the syllabus.



Corn

The first model you meet is the model you keep. Ask anyone about their first car.



Herman

Second, size. Ninety-four million parameters in the base checkpoint, thirty-two bit floats. Whisper large is one and a half billion, Canary's new one is two and a half billion. The base model is a rounding error next to modern systems, and that matters for CPU inference, edge deployment, and the unglamorous one, CI pipelines. You can run this thing in a test harness without filing a budget request. And one paper makes the flip side concrete, the smallest variant at ninety-five million parameters already constitutes a challenge for edge devices. So it's small, not tiny, but small enough.



Corn

It's the biggest model that still fits through the door of a normal computer.



Herman

0 across the checkpoints and nearly all the community language fine-tunes. You can put it in a commercial product without a call to a lawyer. Whisper is MIT so it competes there, but NVIDIA's competitive stuff varies, and some of the paralinguistic fine-tunes, the audeering ones especially, are non-commercial, which is worth knowing before you wire them into anything.



Corn

Fourth. You said the download leaderboard is basically an atlas.



Herman

It is. Japanese, Thai, Indonesian with Javanese and Sundanese, Greek, Hungarian, Czech, Hindi. The most-downloaded fine-tunes map almost exactly onto the languages Whisper and the NVIDIA models serve poorly or not at all. If your product runs in Hungarian, the wav2vec 2.0 fine-tune isn't the old option, it's often the only serious open option. The long tail of languages is a durable download base because nobody's racing to displace it.



Corn

Nobody re-fights a war over a country nobody's army is in.



Herman

Fifth, the feature-extraction afterlife we already covered, which honestly may be the most durable of the five, because a generative transcriber is the wrong tool entirely if you want somebody's age from their voice. And there's a sixth, quieter signal. The ONNX community ports. The base checkpoint was re-exported for ONNX runtimes in 2025, the Chinese fine-tune too. People are still porting a 2020 architecture to new runtimes years after release. You don't do that for a dead thing.



Corn

Would you still start a new project on it, though? That's my honest question. Say somebody's building transcription today.



Herman

For English, no. I'd take Parakeet and not think twice, the speed alone is absurd. Six thousand times real time means you transcribe an hour of audio in half a second. But for a low-resource language, or for anything paralinguistic, or if I wanted a model a pre-trained encoder my students could fine-tune on one GPU, then absolutely, today, it's still the right answer. That's the ResNet comparison, isn't it. ResNet 50 stopped being state of the art years before most image pipelines finally let go of it.



Corn

ResNet 50 is exactly it. Not the best, but the reference point everyone measures against and the first thing anyone downloads. There's a paper from 2021, the History of Speech Recognition to the Year 2030 one, that predicted self-supervised pre-training would be pervasive in speech recognition by 2030.



Herman

And it's already true. Every current leaderboard model, including the ones beating wav2vec 2.0 by two points of error rate, is built on self-supervised pre-training. NVIDIA trained Parakeet that way. Qwen3's audio stack rests on it. The approach won completely, so thoroughly that it's invisible now, and wav2vec is the model that proved it could work on raw audio at all. The successors didn't replace the idea. They industrialized it.



Corn

Which does flip the framing of Daniel's question. He asked why an obsolete model is significant. The answer is that the family tree never stopped growing. HuBERT, WavLM, XLS-R, even the wav2vec 2.0-BERT variants, all of them are quantize-the-audio, mask-it, predict-the-units, in some form. The base checkpoint's 1.8 million downloads a month is the field keeping its founding document on the desk.



Herman

What do you make of the jonatasgrosman numbers, though, nine to one against 's own checkpoint? Does that surprise you?



Corn

Not once you've met developers. The base model is a component, and components get downloaded once per pipeline, silently, by a lockfile. The language fine-tune is the thing a human chose, for a language they care about, often because there was no alternative. Nine to one is the ratio of things people want to things machines fetch. If anything I'd expect it higher.



Herman

You've clearly been reading the same download charts I have.



Corn

Slowly. Very slowly.



Herman

One more piece of the family significance, then. The downstream research is where you see the representations were real. The dysarthric speech results are the ones I find moving, from a clinician's point of view. Speakers whose speech is impaired enough that conventional systems fail on them, and XLSR-based features cut errors by up to twenty-two percent on one corpus. Nobody fine-tuned for that population specifically. The encoder had simply learned speech generally enough that it transferred to speech that doesn't sound like the training data.



Corn

That's the actual test of a representation. Not clean benchmark audio. Whether it survives the parts of speech that don't match the textbook.



Herman

The Flemish Dutch result is the tidy small version of the same thing. Fine-tune XLSR-53 on Flemish, thirty percent error reduction, and the paper's phrasing was that all pre-trained models improve phone separability but not all improve recognition. wav2vec 2.0 did both. That's the difference between learning sounds and learning speech.



Corn

Fourteen years of medicine and the thing you bring to a model conversation is dysarthric speech.



Herman

Fourteen years of listening to people talk is exactly the qualification, Corn. That's the whole job.



Corn

Fair. So where does that leave the paradox? My tally is, it's the tutorial default, it's small, it's permissively licensed, it's the only option for a long tail of languages, it's a backbone for tasks that aren't transcription, and it's still being ported to new runtimes. Do all those together feel like they explain sixteen million downloads a month for one Japanese fine-tune, or is something still missing?



Herman

Mostly, I think. The Japanese number I'd attribute to a very specific thing, which is that Japanese is a huge market, Whisper's Japanese has always been its weakest major language, and the fine-tune is free and good enough. One dominant option in a big market collects everyone's downloads. But I'll concede the honest edge of it, we're inferring from the shape of the numbers, not from anyone's survey of why people download.



Corn

I'd add a seventh, unglamorous one. Inertia is a legitimate engineering strategy. If your system works, on a model you understand, with an architecture simple enough that you can actually debug it, replacing it to chase two points of error rate is a project with a budget and a risk. Plenty of those 1.8 million downloads are the sound of people not breaking things that work.



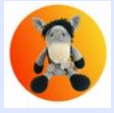
Herman

That's the take you hold for everything with a power switch, and it's usually correct.



Corn

Usually. Correct today.



Herman

Fine. Today.



Corn

There's one more wrinkle worth naming. The base checkpoint's own model card claims 3.4 clean, 8.6 other on LibriSpeech, and that 8.6 for noisy is not a great number against the modern field. Six and a half average for Parakeet across much harder data. The gap on clean read speech is small, the gap everywhere else is not. The people still downloading it mostly aren't confused about that.



Herman

They know exactly what they're downloading. A known quantity with a five-year paper trail.



Hilbert

It runs in real time on a CPU, that's the whole of it.



Herman

Well, there's a voice we haven't factored in.



Hilbert

The ONNX export of the base 960 hour model, Corn. I use it weekly. I had a run of old cassette tapes, a local radio drama that went off the air years ago, and I wanted the lot transcribed before the tape sheds. Tried the big Whisper first. My laptop, a 2014 thing I see no reason to replace, needed eleven hours for one side of one tape. The wav2vec 2.0 export did it at speaking speed. Not eleven hours. Eighty minutes, roughly, for a C-90.



Corn

You transcribed the whole run on that machine?



Hilbert

Most of it. The last fourteen tapes are still with Molcho, the man who services my deck. He replaces the pinch rollers, and he keeps the tapes until he's done a batch, which is taking him a while. I stopped asking. I keep a spreadsheet of the download counts, the checkpoints I use. The Japanese fine-tune is nine times the base model, and has been for



Hilbert (continued)

as long as I've tracked it. I don't find that insulting to the way you'd expect. I find it correct. They made the base, other people made it useful.



Herman

You track the counts how often?



Hilbert

Sundays. It's not a lot of work. I'll say this for the tutorial you two keep crediting, the one on fine-tuning wav2vec 2.0. I did it in 2022 with the cassettes in mind, and they've since rewritten it around a different model. That annoyed me more than it should have. The old version was fine. I also wrote to a podcast host, some fellow who'd called wav2vec dead in an episode, a short note, polite, explaining that dead and not state of the art are different things. That was over a year ago. No reply.



Corn

You wrote fan mail to a model.



Hilbert

I wrote a correction. There's a difference, Corn. He was wrong in a way a person with a spreadsheet wouldn't be.



Herman

The C-90 figure, by the way, checks out against the speed you'd expect from a CPU run of that checkpoint. For what it's worth, that's the quiet argument for the whole family. The new leaderboard models are faster than real time on a GPU. On his laptop, nothing else in the top tier even runs.



Hilbert

It's not a laptop for throwing out. He had a good fifteen hundred hours left in him. Anyway. Sixteen million a month for the Japanese one. I'd want to see it stay there.



Corn

The open question I'd leave with Daniel's framing is whether the download dominance holds. Because the direction of travel is visible. WhisperKit at fifty-seven million a month tells you where the mass market went, and every year the efficient new models get smaller. When a Parakeet-class transcriber fits on a phone, the tutorial default stops being the small option and starts being the big one.



Herman

The legacy holds either way, though. wav2vec's real significance isn't any single checkpoint. It's the template. The proof that self-supervised pre-training on raw audio works, which every current leaderboard model now assumes. The canonical model isn't the one that wins the benchmark. It's the one that teaches the next one.



Corn

If you take one thing from today, take that distinction. State of the art is a property of a leaderboard, canonical is a property of a culture. wav2vec 2.0 lost the first and owns the second, and the second is worth more.



Herman

One thing to watch, whether jonatasgrosman's Japanese fine-tune keeps its nine-to-one lead as the newer multilingual models spread. That ratio is the clearest live signal of whether the long tail migrates or stays put.



Corn

Thanks as always to our producer, Hilbert Flumingtop, at the desk through all of that. If today held your attention, subscribe and leave us a review, and tell a friend who still thinks state of the art is the only metric that matters. For more along these lines, there's episode ten, How ASR Went From Frustration To ... Whisper Magic; episode five, Fine-Tuning ASR For Maximal Usability; and episode nine, Benchmarking Custom ASR Tools - Beyond The WER. This has been My Weird Prompts, where no question is too obscure and no rabbit hole too deep. Find everything at my weird prompts dot com.



Herman

See you tomorrow, Corn.



Corn

See you tomorrow.

SOURCES

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

1. **Hugging Face model card , created 2022-03-02, last modified 2022-11-14, retrieved 2026-11-11. [primary]**
   https://huggingface.co/facebook/wav2vec2-base-960h

2. **Meta AI blog , published 2020-09-24, retrieved 2026-11-11. [primary]**
   https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

3. **Baevski, Zhou, Mohamed, Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, submitted 2020-06-20, revised 2020-10-22, retrieved 2026-11-11.**
   https://arxiv.org/abs/2006.11477v3

4. **Schneider, Baevski, Collobert et al., wav2vec: Unsupervised Pre-training for Speech Recognition, submitted 2019-04-11, revised 2019-09-11, retrieved 2026-11-11.**
   https://arxiv.org/abs/1904.05862v4

5. **Baevski, Schneider, Auli, vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations, submitted 2019-10-12, revised 2020-02-16, retrieved 2026-11-11.**
   https://arxiv.org/abs/1910.05453v3

6. **Conneau, Baevski, Collobert et al., Unsupervised Cross-lingual Representation Learning for Speech Recognition (XLSR), submitted 2020-06-24, revised 2020-12-15, retrieved 2026-11-11.**
   https://arxiv.org/abs/2006.13979v2

7. **Babu, Wang, Tjandra et al., XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale, submitted 2021-11-17, revised 2021-12-16, retrieved 2026-11-11.**
   https://arxiv.org/abs/2111.09296v3

8. **Hugging Face model card, created 2022-03-02, retrieved 2026-11-11.**
   https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-japanese

9. **Hugging Face model card, created 2023-09-04, retrieved 2026-11-11.**
   https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender

10. **Hugging Face model card, created 2022-04-06, retrieved 2026-11-11.**
    https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

11. **Hugging Face model card, created 2022-03-02, retrieved 2026-11-11.**
https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft

12. **Hugging Face model card, created 2024-02-28, retrieved 2026-11-11.**
https://huggingface.co/argmaxinc/whisperkit-coreml

13. **Hugging Face model card, created 2023-11-07, retrieved 2026-11-11.**
https://huggingface.co/openai/whisper-large-v3

14. **Open ASR Leaderboard results dataset, last commit 4 days before 2026-11-11, retrieved 2026-11-11.**
https://huggingface.co/datasets/hf-audio/open-asr-leaderboard-results/blob/main/english_short_latest.csv