Episode Book

EPISODE #6456

The Unsung Builders of Modern Speech AI

PUBLISHED	4 November 2026 at 14:24 UTC
DURATION	26:25
LISTEN	myweirdprompts.com/episode/speech-recognition-open-source-whisper-04765/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
BOOK GENERATED	5 November 2026 at 07:03 UTC

SYNOPSIS

Speech recognition commands billions in venture funding and fills massive academic conferences, yet its foundational architects remain largely anonymous to the public. This episode maps the human ecosystem behind modern audio AI, tracing the toolkits and researchers that made ubiquitous transcription possible. From Daniel Povey's grueling work on Kaldi to Georgi Gerganov's whisper.cpp and Hervé Bredin's pyannote, explore how open-source infrastructure quietly unlocked voice models, and examine why Whisper arrived as a quiet code drop rather than a public spectacle.

TRANSCRIPT



Corn

Ask anyone on the street who built ChatGPT and you'll get a name, maybe two. Ask who built Whisper and you'll get a shrug, even from people whose working day runs on it.



Herman

Which is roughly where Daniel's head is at, I think.



Corn

He sent us this one as text. His feeling is that speech technology is "the unloved part of the AI universe" — his phrase — everyone knows the people behind ChatGPT, almost nobody can name the people behind Whisper, and yet the field must be full of people who gave years to speech recognition while the glamour went elsewhere. Plus the organisations pushing the frontier for languages speech tech never served. He wants an ecosystem map weighted toward the people. Who are the true believers, what drives them, how does it all connect.



Herman

And it's a good instinct, because the map exists, it's just never drawn with the humans on it.



Corn

Before we draw it, though, I want to test the premise. Unloved, or just unfamous? Because those are different diagnoses with different patients.



Herman

Right, so the distinction I'd make is between fame and funding, and Daniel's prompt fuses them. On fame he's completely correct. On funding, the field is not starving. The flagship speech conference, ICASSP, took about eleven thousand two hundred regular submissions this year and accepted about forty-eight hundred. That's a record, roughly fifty-eight percent above the year before.



Corn

Forty-eight hundred accepted papers is a lot of unloved.



Herman

It's an enormous field. For scale, NeurIPS's main track had around twenty-one and a half thousand valid submissions, so the top speech conference runs at roughly half of NeurIPS by volume. Half of NeurIPS is not a backwater. That's a huge field with a low public profile.



Corn

And the money side?



Herman

Also flowing. ElevenLabs raised five hundred million at an eleven billion dollar valuation in February. Deepgram raised a hundred and thirty million Series C at 1.3 billion, announced in January, and their CEO, Scott Stephenson, said they'd been cashflow positive the year before. So capital is arriving in serious amounts.



Corn

Just not evenly.



Herman

That's the catch. The money is going into voice agents and text to speech — the commercial layer. The long tail of languages, the community work we'll get to, that runs mostly on foundation grants. So the real question isn't why is speech unloved. It's why does a field this big and this well-funded produce so few nameable people.



Corn

Which is a much better question, and conveniently the one Daniel's actually asking.



Herman

One caveat before we go on — nobody's done a clean headcount comparison of speech researchers versus LLM researchers. I wouldn't want anyone walking away thinking we established that.



Corn

Noted. So the field is big and funded but faceless. Who are the faces, then?



Herman

Start with the man I'd call the closest thing ASR has to a founding figure. Daniel Povey. He created Kaldi, which came out of a two thousand nine workshop at Johns Hopkins and was first released in May twenty eleven. He actually joined the Hopkins faculty in twenty twelve specifically to concentrate on it.



Corn

A university hiring a professor to tend an open-source toolkit. That already tells you the shape of the field.



Herman

It does. And the devotion is documented, not imagined. A former student wrote that Povey spent almost all his time on Kaldi "except for eating and sleeping" and "works more than 14 hours a day." That quote is from a supporters' site, so it comes from people rooting for him, but the fourteen hours a day detail is the true-believer archetype in one sentence.



Corn

Fourteen hours a day on infrastructure nobody outside the field can pronounce.



Herman

Kaldi became the substrate. If you did speech research for a decade, you either used it or you were arguing with it. And then twenty nineteen, his Hopkins position ended — supporters campaigned for him — and in November of that year he moved to Beijing as chief speech scientist at Xiaomi.



Corn

That's a bruising arc for the open-source purists. The patron saint of free toolkits going to a phone manufacturer.



Herman

I'd frame it differently, and this is the part I find moving. At Xiaomi he built next-generation Kaldi — k2, Lhotse, Icefall, Sherpa — all of it Apache licensed. He took the corporate job and kept the commons. He was elected a Fellow of the International Speech Communication Association this year, which is the field's own hall of fame.



Corn

Does the new stack actually perform?



Herman

Xiaomi claims their Zipformer encoder beats Google's Conformer. Note the verb — claims. That's the company saying it, not a neutral benchmark.



Corn

I'll hold my applause.



Herman

The honest gap in the Povey story is motivation. Nobody's found a first-person account of why he cares. We know he moved continents twice, worked fourteen-hour days, and gave everything away — the motivation has to be inferred from the behaviour. Which, to be fair, the behaviour states pretty loudly.



Corn

Actions fourteen hours a day speak louder than interviews. What's the academic version of him?



Herman

Mirco Ravanelli. He founded and leads SpeechBrain, he's an assistant professor at Concordia and associated with Mila in Montreal. Same impulse, different institution — build the shared toolkit rather than the product. SpeechBrain is the PyTorch-native successor a lot of labs moved to when Kaldi's recipe style started feeling its age.



Corn

We did a whole episode once on how ASR went from frustration to Whisper, and Kaldi was the long winter before it. What I remember is that Kaldi was famous for being powerful and brutal to install.



Herman

Brutal is unfair. Powerful and uninviting. The people who mastered it formed a priesthood, and the field is partly faceless because its central tool demanded vows.



Corn

Then Whisper arrives and the priesthood dissolves. Who's the solo-operator figure there?



Herman

Georgi Gerganov. Whisper comes out in September twenty twenty-two, and within weeks he's built whisper.cpp — the version that runs on ordinary CPUs, no GPU required. His own explanation for why it was fast is wonderfully unglamorous. He said the model was "like 90% I have the code already written for the GPT-2" and that "It might be interesting to have it running on a CPU."



Corn

Ninety percent pre-written. That's the most honest origin story in the field — most "breakthroughs" are a Tuesday for someone with the right drawer open.



Herman

Gerganov is the solo open-source engineer as a distinct species. One person, no lab, no launch event, and the artifact becomes infrastructure that commercial products quietly stand on. His project, ggml, joined Hugging Face in February this year — though whether whisper.cpp itself is covered by that arrangement, nobody's confirmed.



Corn

So even his happy ending has an asterisk.



Herman

They usually do. Next one — Hervé Bredin, who built pyannote. That's the who-spoke-when layer, the diarization that tells you which of the words belonged to whom. And here's the dependency story: pyannote took off precisely because Whisper supplied the words. The two tools locked together like halves of a sandwich and half the transcription industry ate it.



Corn

And he stayed in research?



Herman

Until March twenty twenty-five, when he went on leave from France's CNRS — their national research agency — to work full time on pyannoteAI, which he'd co-founded in twenty twenty-four with Vincent Molina and Juan Coria. Reported seed of about eight point one million euros.



Corn

There's the pattern. Open component becomes load-bearing, load-bearing becomes a company.



Herman

It's the only reliable exit route for a toolkit author. The toolkit itself is free, so the value migrates to hosting, support, the enterprise version. Whether that preserves the commons or drains it depends entirely on the licence, and pyannote's stayed open.



Corn

You said Gerganov's status is unconfirmed. Is Hannun's the same?



Herman

Awni Hannun, yes. His arc is the lab-to-toolchain route. Technical lead on Baidu's Deep Speech back in the mid twenty-tens, then he co-created Apple's MLX in December twenty twenty-three and maintains mlx-whisper, which is the Apple-silicon-native path. One secondary report says he left Apple at the end of February this year — where he went, unconfirmed. I'd treat the departure as probable and the destination as unknown.



Corn

And that gets us to the actual question Daniel opened with. The people behind Whisper. Name them.



Herman

They're not anonymous, that's the thing. The paper lists six authors, in order: Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever.



Corn

Brockman and Sutskever are household names. In this household, anyway.



Herman

Two of the six are famous! Just not for Whisper. And Radford — first author, the one whose name would be on the marquee — left OpenAI in December twenty twenty-four for independent research and became an adviser to Thinking Machines Lab. But here's the thing I find structurally interesting: nobody's turned up an interview with the authors about how Whisper was made. Maybe one exists and it hasn't surfaced — I can't prove a negative — but the canonical account isn't out there the way it is for ChatGPT.



Corn

ChatGPT shipped with a company narrative attached. Faces, blog posts, a launch moment. Whisper shipped as a technical report and a set of weights.



Herman

And no personality attached to the product. There was no demo day, no naming ceremony, no founder interview circuit. You went to a repository, you downloaded a model, it worked. The most consequential speech release of the decade arrived by paper.



Corn

It's the difference between a premiere and a delivery.



Herman

Daniel's right that the credit gap is real, and the answer is boring: it's about how the thing entered the world, not about anyone hiding the names. The authors are on page one.



Corn

And the current roles of the middle three — Kim, Xu, McLeavey?



Herman

Couldn't confirm. Which is itself a small data point for Daniel's thesis, when you think about it. The people behind the tool that changed everyone's audio workflow can't be located in the present tense.



Corn

Those are the individuals. But individuals don't float free — they sit in an ecosystem. And I suspect the ecosystem has a fault line.



Herman

It does, and drawing it is the second half of Daniel's map. The layers: a shared substrate, which is Mozilla's Common Voice — the crowd-sourced corpus anyone can contribute to. I'll resist giving a language count because sources disagree, somewhere between a hundred and thirty and over two hundred fifty depending who you ask and when. Then the open toolkits — Kaldi and its successors, SpeechBrain, whisper.cpp, pyannote. Then the frontier labs at scale. And then community organisations building for themselves. The interesting friction is between the last two.



Corn

Frontier labs meaning whom, concretely?



Herman

Meta's Omnilingual ASR, out November last year, covers over sixteen hundred languages, more than five hundred never served before, with models from three hundred million up to seven billion parameters. Google's USM is two billion parameters, trained on twelve million hours across more than three hundred languages, with a stated goal of a thousand. These are open releases from giants, and they're real coverage.



Corn

Sixteen hundred languages in one release. What's the objection — that's the gift horse department, isn't it?



Herman

The objection isn't the model. It's the data. Go to New Zealand. Te Hiku Media, a Māori organisation up in the far north. Their chief technology officer is Keoni Mahelona, a Native Hawaiian engineer. They collected over three hundred labelled hours of te reo Māori in about ten days — twenty-five hundred volunteers reading over two hundred thousand phrases. They built a recogniser reporting ninety-two percent accuracy in te reo, eighty-two bilingual. And when big tech came asking for the data, they said no. Mahelona's explanation is worth hearing in his words: "To have our language stolen from



Herman (continued)

us by the colonizers through generations of abuse and residential schools, and then for them to turn around and sell it back to us?... We couldn't do this. It wasn't right."



Corn

That's not a licensing dispute. That's someone who knows the history of the resource.



Herman

He's also called for "sovereignty for our languages and our data." So they license it under their own Kaitiakitanga licence — guardianship — with terms they set, and they built on DeepSpeech and later NeMo, themselves open tools. The stack is borrowed, the data is theirs.



Corn

So the fault line isn't open versus closed. It's sixteen hundred languages released by a giant, versus a community that could have had coverage tomorrow and chose ownership.



Herman

And you can hold both positions honestly. 's release serves languages nobody was serving. Te Hiku's refusal is the only reason those three hundred hours exist on Māori terms — a community that gives away its language data may never get it back on its own terms. Sovereignty isn't nostalgia, it's leverage.



Corn

And now it's time for another of our regular segments. Herman, your mother's waiting, I believe.



Herman

Mum! Always a pleasure.

D

Dorothy

Hello dear, it's your mother, with Dorothy's History of Toast.

D

Dorothy

For about ten years, Wikipedia said the electric toaster was invented by a Scotsman called Alan MacMasters. He never existed, sweetheart. It was a hoax, and it was only taken down in twenty twenty-two.

D

Dorothy

Ten years of everyone believing a man who was never there. That's the history of toast for you, dear, things just go along wrong for a long while.

D

Dorothy

Anyway. I've got a bit of bread going stale on the side, so I shall have that with my tea. Bye now.



Corn

Ten years of a man who never existed. There's a lesson in there about citations, Herman.



Herman

There is, and I refuse to learn it before lunch. Back to the map — the labour layer. Because speech for underserved languages runs on recorded voices, and somebody pays those somebodies.



Corn

This is the annotation end. Who's the figure there?



Herman

Manu Chopra. In twenty seventeen, at twenty-one, he visits a data company in Mumbai and learns the workers are earning forty cents an hour. His reaction, quoted at the time: "This cannot be the only way this work can happen." He founded Karya, which pays at least five dollars an hour — roughly twenty times India's minimum wage — and, this is the structural innovation, pays workers again every time their data is licensed. Not once, repeatedly.



Corn

So the dataset becomes an annuity rather than a transaction.



Herman

That's the whole idea. The recording of your voice keeps earning because it keeps getting used. Karya does collection for AI4Bharat, which is the India-scale story — IIT Madras, Mitesh Khapra as corresponding author, and their IndicVoices release: seven thousand three hundred forty-eight hours from over sixteen thousand speakers across a hundred forty-five districts and twenty-two languages at the twenty twenty-four launch, since reported at twelve thousand hours and growing speakers.



Corn

A hundred and forty-five districts is the detail that isn't accidental. That's deliberate geographic spread, not whoever's near the campus.



Herman

Exactly the point. If you collect in three cities you get three accents. And the pattern repeats across the African continent. Lelapa AI — CEO and co-founder Pelonomi Moiloa, started twenty twenty-two after the founders met at the twenty seventeen Deep Learning Indaba. Her line: "It's important for us as Africans to own that narrative." Their product is Vulavula, which is Xitsongo for "speak" — Xitsongoa for speak, forgive me, Xitsonga —



Corn

You've earned the etymology stumble of the day.



Herman

It offers recognition for isiZulu, isiXhosa, Sesotho and Afrikaans, on roughly two and a half million dollars of seed. Which, set against ElevenLabs' eleven billion valuation, tells you everything about where the money sits.



Corn

Five orders of magnitude. That's the map in one comparison.



Herman

Then there's the movement layer. Rwanda — Digital Umuganda came out of a twenty eighteen Mozilla hackathon, and an engineer there, Didier Mugwaneza, with a Mozilla Fellow named Josh Meyer, built the first Kinyarwanda speech recogniser. By February twenty twenty-one, Kinyarwanda was second only to English in contributed hours on Common Voice — that ranking is as of twenty twenty-one, to be clear, later figures cite over two thousand four hundred hours. A language most of the industry had never heard of, out-contributing everything but English.



Corn

That's what happens when the infrastructure is open and one stubborn engineer shows up.



Herman

Masakhane is the continental version — founded twenty nineteen, credited to Jade Abbott and Vukosi Marivate, though that's from secondary sources. Their African Languages Hub, directed by Chenai Chair, opened a call in January covering fifty languages, described as "a movement toward a more equitable digital future." Backers include Google dot org, the UK's development arm, Canada's IDRC and the Gates Foundation.



Corn

So the giants fund the movement whose entire premise is independence from the giants.



Herman

They fund the parts of it they can be seen at. I don't doubt the money is welcome — this work doesn't happen without it — but it does mean the sovereignty work is permanently grant-shaped. Te Hiku can refuse 's terms because its authority comes from its community. Masakhane's next year depends on a renewing a foundation's priorities.



Corn

One more frontier before we close the loop — accessibility.



Herman

Google's Project Euphonia, which by late twenty twenty had gathered over a thousand hours from about a thousand participants, mostly people with impaired speech. And the figure I find quietly extraordinary is Dimitri Kanevsky, a Google researcher who is himself deaf — he trained a model on his own speech, about twenty-five hours of recordings of one person's voice, and it understood him where nothing general-purpose could.



Corn

Twenty-five hours to be understood by a machine after decades of being misunderstood by them.



Herman

That's the frontier that has no commercial gradient at all. There's no voice-agent market at scale for atypical speech. It happens because somebody at a lab argued for it, or it doesn't happen.



Corn

So the loop closes like this. The money sits in voice agents and text to speech — ElevenLabs, Deepgram. The mission sits in foundation-funded language work and one-off accessibility projects. And the load-bearing infrastructure, the toolkits everything stands on, is maintained by a handful of people on research salaries and caffeine.



Herman

Daniel asked for an ecosystem map weighted toward the people. The honest version has the shapes of continents but the density in the wrong places — crowded where the money is, inhabited where the belief is.



Corn

You've been sitting on the wrong shape of that map yourself, haven't you.



Herman

What do you mean?



Hilbert

They're right about the believers, Corn.



Hilbert

I've kept a notebook. Small one, green cover, one line a day. Every time the voice assistant mishears the same word of mine, I write the date. Been doing it four years. Total stands at one thousand one hundred and sixty-two.



Herman

One word?



Hilbert

Flumingtop. Surname. It offers me "flume in top", "plumbing top", and once, in March, "Florence Stoppard", which I have not corrected because it was the most interesting thing it ever said.



Corn

You never wrote to anyone about this.



Hilbert

No. That's the point. Herman spent twenty minutes on people who gave fourteen hours a day, and not one of them has ever heard me say my own name. I stopped saying it to the thing about two years ago. I type it now. Nobody told me to. It's a matter of principle and I've kept it to myself until this morning.



Herman

Eleven sixty-two. That's, what, roughly one mishearing per working day for four years.



Hilbert

Some weeks clean. Worst week was last November, six of them, because I was booking a delivery and had to give the name to a phone tree that then read it back to me, wrong, three separate times before it would let me through to a person. That is the only line in the notebook that goes on for two lines.



Corn

The phone tree confirmed the mishearing back to you as security. Identity verified against the wrong word.



Hilbert

Verified I was a man who says plumbing top, yes. That's the system working as built, Corn. Nobody's fault and nobody's problem. Same as you said about the grants.



Herman

The believers built the thing that transcribes te reo and Kinyarwanda and it still can't take one man's surname at a phone tree. I don't think those are in tension, actually. The Kinyarwanda recogniser took one stubborn engineer too. The failure you're describing is a commercial deployment that never met its user.



Hilbert

Hm. The notebook says otherwise, but the notebook is one side of it.



Corn

Fair witness, though. A thousand lines of it.



Herman

So here's the one thing to take from this. If anything's unloved in speech, it isn't the field — a record-breaking conference and an eleven billion dollar valuation don't read as neglect. What's unloved is the connective tissue: the toolkit maintainers, the community data collectors, the sovereignty arguments, the twenty-five-hour accessibility models. Famous nowhere, load-bearing everywhere.



Corn

And the money sits in the part that talks back — voice agents, generated speech — while the long tail of languages runs on grants that renew.



Herman

Which leaves the question I'd watch: what happens to the community-built layer when a grant cycle turns? Te Hiku owns its data outright, so Te Hiku endures. Masakhane's next fifty languages depend on a foundation's priorities holding. Whether that layer survives on ownership or on patronage is undecided.



Corn

Producer credit where it's due — Hilbert Flumingtop at the desk, notebook and all. If this was your kind of episode, go back for episode ten, How ASR Went From Frustration To ... Whisper Magic; episode seven, Building Custom ASR Tools; and episode eight, Building Your Own Whisper. This has been My Weird Prompts. If today's map earned its keep, leave us a review, and the archive lives at my weird prompts dot com.



Herman

See you tomorrow, everyone.



Corn

See you tomorrow.