# Episode Book

EPISODE #6432

# Why Speaker Embeddings Fail Across Time and Noise

| | |
|---|---|
| PUBLISHED | 3 November 2026 at 21:19 UTC |
| DURATION | 27:00 |
| LISTEN | myweirdprompts.com/episode/speaker-embeddings-voice-profiles-01735/ |
| HOSTS | Corn, Herman, Hilbert |
| SCRIPT MODEL | z-ai/glm-5.3 |
| VOICES | chatterbox-regular (chatterbox-local) |
| PIPELINE | V5.26 |
| SOURCES | 10 cited by the research (listed at the end) |
| BOOK GENERATED | 4 November 2026 at 09:31 UTC |

## SYNOPSIS

Speaker diarization turns voices into numerical vectors, but persistent voiceprints are far more fragile than assumed. Examining architectures like ECAPA-TDNN and spectral clustering reveals how pipelines guess headcounts and why tools cannot easily track speakers across separate sessions. Benchmark data highlights massive accuracy drops between studio microphones and messy environments, while acoustic shifts like aging, cassette degradation, and stroke demonstrate why durable lifetime voice profiles remain out of reach.

# TRANSCRIPT



### Corn

A voice walks into a microphone, and somewhere down the pipeline it becomes a list of numbers. Daniel wants to know how that trick is done, and then, because he's Daniel, he wants to know how badly it breaks when the voice catches a cold, moves to a different room, or simply lives another decade. His hunch is that a one-hour sample spread across many different days would beat a thirty-second snapshot, and behind that sits the big one — could a voice profile ever last a lifetime.



### Herman

And it's the right season for that question, because the field itself has just been arguing about it. We've done the diarization pipeline before, the clustering stages, all of it, but we've never actually gone down to the floor and asked what a speaker embedding is, mechanically, and what it can honestly promise.



### Corn

Herman Poppleberry, everyone. He read the papers overnight, which is his version of a calm morning.



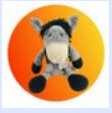
### Herman

Two papers, and one of them is enormous. But you're right that we should start at the bottom, because almost all of the confusion around "voice profiles" comes from one collapsed distinction.



### Corn

Which is?



### Herman

Diarization and voiceprinting are two different jobs. Diarization is per-recording: you hand the system one file, it tells you who spoke when, inside that file, and that's the end of its obligations. A persistent voice profile — the thing Daniel is asking about, "would this profile survive a decade" — that's speaker verification, a separate layer, and notably it's a layer pyannote sells separately from the open-source diarization pipeline.



**Corn**

So the thing Daniel's asking about doesn't exist inside the tool he's probably imagining.



**Herman**

Not as a durable object, no. The pipeline that answers "who spoke when" produces labels, speaker zero, speaker one, and speaker zero in this recording has no relationship to speaker zero in the next recording. None. It's a fresh clustering every time.



**Corn**

Let me make sure I've got that, because I think listeners will trip on it the way I first did. If I record a meeting on Monday and the same meeting on Tuesday, and the system calls me speaker zero both days —



**Herman**

That's coincidence, not continuity. It's like a hotel that numbers its rooms starting from one every night. Room one on Monday and room one on Tuesday might be the same physical room, might not, and the hotel doesn't know and doesn't care. The numbering is local to your stay.



**Corn**

That's the first misconception gone, then. The diarizer isn't holding a little card file of voices.



**Herman**

It isn't. A speaker embedding is the output of a neural network whose training job was classification — the network saw a large pool of speakers and learned to tell them apart. To get from variable-length audio to a fixed-length vector, there's a pooling layer that squashes the time dimension into statistics, and after training you read the embedding off the activations of the penultimate layer. That's the artifact. The canonical model for this in diarization is ECAPA-TDNN, from a Brno group, twenty twenty-one.



**Corn**

And the vector is small.



**Herman**

One hundred and ninety-two numbers. That's your whole voice, compressed to a list you could write on an index card.



**Corn**

One hundred ninety-two numbers to summarize a throat. My epitaph could be shorter. But what do the numbers actually encode? Is one of them "nasal," another one "deep"?



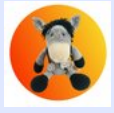
**Herman**

No, and that's worth killing too, because people imagine a little chart — pitch dial, rasp dial. In reality almost none of the individual dimensions have a clean human name. The network found whatever combinations of acoustic features helped it separate speakers during training, and those combinations are its own invention. It's like asking a sommelier to describe a wine by its chemical formula — the formula is real, it's complete for the purpose, and no single number in it tastes of anything.



**Corn**

So you can't open the box and read off "this is a gravelly man of sixty."



**Herman**

You can't. You can only ask the box questions in its own currency: is this vector close to that vector.



**Corn**

Which is the cosine thing, yes?



### Herman

Which is the cosine thing, and we'll get there. First, the lineage, because it explains why the cosine thing works now. The older models — i-vectors, then d-vectors and x-vectors — those needed a separate statistical scorer, PLDA, bolted on top to compare embeddings. ECAPA-TDNN is trained with an additive angular margin softmax, which shapes the embedding space so that a plain cosine similarity does the scoring directly. You skip the bolted-on layer entirely.



### Corn

Which is nice, because every bolted-on layer is another thing to age out of date. The embedding is one hundred ninety-two numbers. What happens next? How does a pile of these become "Corn spoke here, Herman spoke there"?



### Herman

The recording gets chopped into speaker turns — that's a segmentation step upstream. Then you extract one embedding per turn, and you cluster them. The workhorse is spectral clustering, and the recipe is very concrete: build a matrix of cosine similarities between every pair of embeddings, prune the small values so noise doesn't connect everything to everything, symmetrize it into a Laplacian, and then do something elegant.



### Corn

Go on.



### Herman

The system doesn't know how many people are in the room, so it has to guess. It looks at the spectrum of that Laplacian and finds the eigengap — the point where the eigenvalues suddenly jump — and the gap's location tells it the number of speakers. Then k-means in that spectral space assigns each turn a cluster.



### Corn

So the machine counts the eigengap and decides we're two people. There's something very tidy about a number line voting on headcount. But what happens when it votes wrong? Say there are three of us in the room and it decides on two.



### Herman

Then two of the three get fused. And here's the nasty part: it fuses whichever two are most similar, which in practice means the two with the most similar pitch and channel — two men on the same microphone can merge into one "speaker" while the woman across the table gets her own label plus a phantom. The error isn't random, it's systematic, and it's exactly the failure you'd get if Ezra and I recorded together and the machine decided I'd been talking to myself for an hour.



### Corn

I've had that dream, actually. Continue.



### Herman

It's the one part of the pipeline with a little theater in it. And there's a structural catch that I want on the table early: the whole embedding-clustering path assumes one speaker per segment. Two people talking over each other — the embedding for that segment is neither voice, it's a blend, and clustering will do something arbitrary with it.



### Corn

Overlap is where it quietly goes wrong, then. Not with a bang.



### Herman

There's dedicated overlapped-speech detection research to patch that hole, but it is a patch. And the per-recording point is worth saying in the exact terms Daniel will meet it in: the open-source pyannote pipeline, the community one, labels within a file. If you want those labels to resolve to names across files — a voiceprint layer, confidence scores, identity — that's the separate product.



### Corn

And here's a voicemail from Dotty, down in Somerset.
Dotty: Hello you two, it's Dotty here, calling from the bath, well, just got out, sitting on the edge of it with a towel round my shoulders because the water's gone cold and I didn't want to miss the start. I did want to ask about the speaker embedding business, and I think I've got the gist, that you turn a voice into a list of numbers and then sort them into piles, though I did get a bit lost at the eigenvalues. My question is this. My late husband Gerald had a voice like gravel in a bucket and it changed ever such a lot after his stroke,



### Corn (continued)

and I've got old cassettes of him from nineteen eighty three, and I do wonder whether your machine would still know him. Anyway, love the show, bye now.



### Herman

Dotty's got the gist exactly right, actually — a list of numbers, sorted into piles, that is the whole first half of this episode. And Gerald is a hard case in a way the literature barely touches: a stroke is a sudden, structural change, and the clustering doesn't know him at any point, per-recording. The question she's really asking is whether a profile built on the nineteen eighty-three cassettes would still match a post-stroke voice.



### Corn

And the cassettes add their own layer, don't they? That's not just old Gerald, that's old Gerald through tape hiss and whatever microphone the family had in nineteen eighty-three.



### Herman

Right, the channel and the speaker are tangled together in that recording. The embedding doesn't cleanly separate "this is his voice" from "this is his voice as heard through a cassette deck." You'd be asking the system to match across a domain shift and a medical event simultaneously. Two hard problems, one tape.



### Corn

My guess is the honest answer is no, and nobody's run that experiment either. Cassettes to stroke, that's outside every benchmark. But her piles and numbers framing is better than most vendor copy I've read.



### Herman

So that's the vector built and clustered. Now let's break it. Because everything I've described so far lives in a world of clean meeting rooms and close microphones, and Daniel's question is about the world where people get sick, age, and record from the far side of a kitchen.



### Corn

Break it how badly?



### Herman

With pyannote's own benchmark numbers, because you can't accuse the vendor of sandbagging. Their open-source community pipeline, on the AMI meeting corpus with close microphones, gets a diarization error rate of about seventeen percent. On AVA-AVD, which is in-the-wild movie audio, overlapping speech, far-field microphones — forty-four and a half percent. On Ego4D, the bodycam-style corpus, forty-six point eight.



### Corn

So the same software goes from seventeen to nearly forty-seven. That's not degradation, that's a different product. And just so nobody's nodding along — what does seventeen percent error actually look like in a transcript?



### Herman

Roughly, about one word in six is attributed to the wrong speaker or missed entirely — diarization error rate counts misattributed speech, missed speech, and false alarms together. Seventeen percent on a clean meeting means a transcript where the attributions are mostly right but you wouldn't trust any single quotation without checking. Forty-seven percent means the speaker labels are close to decorative.



### Corn

Which means "even when you have good trained embeddings for two speakers" — Daniel's own caveat — is doing a lot of load-bearing in his question. The embeddings can be excellent and the conditions can still bury them.



### Herman

And the premium pipeline, precision two, improves those to about thirteen percent on the close-mic AMI and thirty-seven, thirty-nine on the wild corpora. Better across the board, but the shape of the gap doesn't close. The hard conditions stay hard no matter how much money is on top.



**Corn**

Which is worth pausing on, actually — money buys you a few points everywhere, but it doesn't buy you the wild conditions specifically. The kitchen stays a kitchen.



**Herman**

And I want to be straight about a gap before we get to Daniel's hypothesis: nobody has isolated the common cold. There's literature on aging, on channel effects, on noise, on reverberation. Temporary illness specifically — I haven't seen a study, and I've looked, because it's the most common real-world perturbation there is. It's a open question.



**Corn**

A field that can model seventeen years of aging but skipped the sniffles. Priorities.



**Herman**

The sniffles are hard to fund, I suspect. No corpus of people recording weekly through a cold. Now, Daniel's hypothesis — one hour across many days versus thirty seconds.



**Corn**

My instinct says he's right, but for a lazy reason — more data, more averaging. Is that the reason?



**Herman**

No, and this is the bit I want to get right, because the lazy reason is wrong in an interesting way. The closest direct evidence is an Interspeech paper from this year on enrollment duration. They compared text-dependent enrollment — short prompted phrases, phonetically consistent — against text-independent enrollment, which is just whatever speech you have, content be damned. The text-dependent stuff was three seconds and under, and the text-independent side ran at three and ten seconds. Two orders of magnitude below Daniel's hour, so hold that in mind.



**Corn**

Hold it.



**Herman**

Their finding: text-dependent enrollment "yields unstable speaker representations," while text-independent ones "become more stable as the enrollment duration increases." And the line that matters — "the stability advantage brought by longer enrollment duration outweighs the benefits of phonetic consistency."



**Corn**

Say that in kitchen terms.



**Herman**

The gain isn't from averaging out noise. It's that text-independent speech, especially gathered across days, covers more phonetic diversity and more recording conditions, so the profile isn't overfit to one microphone, one morning, one mood. Daniel's intuition about many different days isn't just "more numbers" — it's diversity of the numbers.



**Corn**

So the one-hour sample recorded across a week isn't a bigger pile, it's a better-distributed pile. Same speakers, different rooms, different mornings, some before coffee.



**Herman**

And notably, in that paper, text-independent enrollment beat text-dependent despite the content mismatch — the verifier never hearing the same words it was enrolled on. Consistency of phonetics lost to coverage of conditions.



### Corn

It's the passport photo versus the photo album, isn't it. The passport photo is one careful shot under controlled light — and it's exactly the shot that fails to match you at the border four years later. The photo album has you squinting on a beach, laughing at a party, half-asleep on a train, and any one of those is a worse photo but the album as a whole is a better you.



### Herman

That's a fair rendering, and it's why the "more data" instinct is lazy but not wrong in outcome — the volume only helps insofar as it drags diversity along with it. An hour recorded in one sitting is just a very long passport photo.



### Corn

Would you actually record an hour across days, though, if you were building a profile for, say, Ezra?



### Herman

I'd record twenty minutes across a month before I'd record an hour across an hour, if that's what you're asking. But you've turned this around on me — you had a view on this when Daniel's message went around, I remember you saying the hour was overkill.



### Corn

I said the hour was overkill because I thought the marginal gain flattened out. The Interspeech numbers only go to ten seconds. I'm conceding the direction and keeping my doubt about the dose. Nobody has actually run hour-versus-thirty-seconds for diarization — that experiment doesn't exist.



### Herman

It doesn't. And I won't pretend it does. The direction of everything supports him; the specific trial is unrun. Fair state of the argument. Now the enemy enrollment can't buy its way out of: time.



**Corn**

Aging.



**Herman**

There's a dataset called VoxAging — two hundred ninety-three speakers, English and Mandarin, recorded weekly, the longest span seventeen years, about nine hundred weeks. Seven and a half thousand hours of audio, two point six million segments.



**Corn**

Someone recorded weekly for seventeen years. My kind of commitment, and I say that as someone who has nap streaks. How does a dataset like that even happen — did they pay these people?



**Herman**

Largely it's people who were already recording themselves weekly — YouTubers, mostly — whose material the researchers harvested and annotated. So the longest longitudinal voice dataset in existence is, in effect, a byproduct of people talking to a camera about their week for unrelated reasons. There's a lesson in there about where science gets its data, but I'm not sure it's a flattering one.



**Corn**

They ran the embeddings through it?



**Herman**

ECAPA-TDNN embeddings, and speaker similarity declines steadily from the enrollment point. For English speakers, average similarity falls below zero point five at about five hundred weeks — ten years. Mandarin speakers hit the same floor faster, around eight years.



**Corn**

Ten years to fall off a cliff. What does falling off look like in error rates?



**Herman**

On the English set, ECAPA-TDNN goes from under one percent equal error rate at year zero — point eight six — to five point three six percent at year ten. Other models tell the same story: ERes2Net-large roughly point six to three point nine, CAM++ point six five to four point eight, an older TDNN from two point two to eight point four.



**Corn**

So the same model, same task, and ten years multiplies the error by six.



**Herman**

And Mandarin is worse across the board — ECAPA at eleven percent rising to fourteen point seven five. The language and the channel both matter, which is worth a flag for anyone building this for Hebrew-Arabic bilingual households, for instance.



**Corn**

That flag is for our zip code, Herman.



**Herman**

It is. Two findings in VoxAging are counterintuitive enough that I trust them precisely because they're inconvenient. The forty-to-fifty age group deteriorates fastest — midlife, not old age. The over-sixties are the most stable group in the dataset.



**Corn**

The forty-somethings fall apart and the seventy-somethings hold steady. Which is backwards from every intuition I had. I'd assumed the curve was a gentle downhill from thirty.



### Herman

Everyone assumes that. The data says the voice has a midlife crisis and then settles down. And gender: female speakers deteriorate at roughly five times the male rate in this data — reaching six point seven seven percent error at ten years versus four point zero nine for male speakers.



### Corn

So if you're designing the system, your worst customer is a forty-five-year-old woman, and your best is a man my pace of life who's already past sixty. I'd like my royalties in leaves.



### Herman

There's a wrinkle that complicates the neat ten-year story, though, and I want it on record because it's older evidence pointing the other way. A two thousand nine study on long-term speaker variability found that within-conversation change was statistically very robust — same sitting, your voice drifts measurably — while the aging effect over three years was statistically negligible.



### Corn

Wait. Read that back to me in terms of what it means.



### Herman

Short-horizon drift — the day-to-day stuff, the within-a-sitting stuff, which is closer to Daniel's colds question — is a real, measurable effect. The multi-year erosion only shows up once you're out past several years. So the naive story, "time slowly ruins the profile," is wrong twice: the near-term variation is bigger than people assume, and the long-term decline is not gradual everywhere — it's a midlife acceleration followed by stability.



### Corn

So Daniel's one-hour-across-days enrollment attacks the short horizon, and nothing attacks the ten-year one.



### Herman

Nothing he can record his way out of. The VoxAging authors say it plainly — these systems need more frequent template updates, because voiceprint features are sensitive to aging changes. A profile is a perishable good, not a deed of ownership.



### Corn

Which takes us to the lifetime version of his question. Any evidence at all that a profile survives decades?



### Herman

No. The longest dense longitudinal dataset is VoxAging at seventeen years. There's one dataset that reaches fifty-eight years, but it's seventeen speakers with fewer than ten samples each — you cannot claim a lifetime profile from that. Nobody has shown any system holding a single person's identity across sixty years of speech. The data doesn't exist to even test the claim.



### Corn

Seventeen years is the ceiling, then, and it's a ceiling made of people who happened to record weekly on YouTube.



### Herman

There's one more angle that reframes the whole lifetime question, and it's the forensic one, because the courts have cared about "is this the same voice across decades" for longer than the machine-learning people. Kelly and colleagues, twenty thirteen: human listeners, played voices separated by some number of years, asked to say whether the person had aged.



### Corn

How do humans do?



**Herman**

At ten years of separation, just above chance. At thirty years, about ninety percent correct. And automatic system likelihood ratios — the strength-of-evidence scores used forensically — drop below moderate support by thirty years for most speakers.



**Corn**

So even the human ear, the thing all this machinery is supposedly imitating, can't reliably hear ten years of change. There's something almost comforting about that — the machine isn't failing at a task humans ace. It's failing at the same cliff we all fail at.



**Herman**

And the systems are, in that forensic setting, if anything more confident than the ears past the point where both are wrong. The lifetime profile isn't just an engineering gap — it may be an unanswerable question with the data any human institution has ever collected.



**Corn**

Unless someone starts recording Ezra weekly and keeps it up until twenty eighty-six.



**Herman**

Two thousand ninety-six, minimum, for the clean version of the experiment.



**Hilbert**

The council records office had a bloke called Aldous, and I spent most of ninety-two in a argument with him about exactly this, though nobody said embeddings, they said "the archive." He wanted voices for the oral history shelf, and he'd had a microphone out at the fete, and he insisted the labels would hold for good, because, he said, a voice is a voice. I told him a voice is a voice for about ten years and then it's a stranger's. He wrote that down, which I noticed, because he hadn't written down anything I'd said before it.



### Corn

He wrote down your ten years?



### Hilbert

In the margin of the accessions ledger, in biro. I saw it when he went to find the form. Anyway. I'd done a tape for the shelf myself, quarter of an hour of me reading out the street names as they were before the renumbering, because Aldous said somebody local had to do it and I was standing there. Nineteen ninety-two, that tape. Two years back my sister was here and I played it for her, then a bit of me from the week before, back to back, and she got the second one and didn't get the first. Not a hesitation. She said, who's that. I'd told her nothing beforehand, which was the point.



### Herman

What did Aldous's shelf actually do with the tapes in the end?



### Hilbert

Nothing. Recatalogued them twice, lost the fete recordings, and the shelf got a grant in ninety-four for a scanner they never used. He kept the biro note, though — my ten years. He quoted it back to me at the ninety-four opening, word for word, and called it mine in front of the room. Thirty-four years on, Corn, your ECAPA machine and Aldous agree with each other, and nobody's ever thanked either of them.



### Corn

Thirty-four years and she didn't hesitate. That's the thirty-year line from the Kelly study wearing slippers.



### Herman

It's also the cleanest version of the enrollment question I've heard all episode — his sister was a zero-shot verifier with no training data, and she failed exactly where the systems fail.



**Corn**

Aldous writing it in the margin of an accessions ledger is my favorite peer review of the year, for what it's worth.



**Herman**

The margin is where most of this science actually lives. There's one thing from his sister's test worth carrying forward — she had known him her whole life, decades of enrollment, and it still bought her nothing across the gap. That's the trouble with Daniel's one hour: it plausibly buys the near horizon, and it buys nothing past the cliff.



**Corn**

Which is the open end of the whole episode, really. The hour-versus-thirty-seconds experiment has never been run — the closest study tops out at ten seconds of enrollment — so Daniel's hunch is supported by the direction of everything and confirmed by nothing.



**Herman**

The aging side says any system that treats enrollment as a one-time event is designing itself a slow failure. The VoxAging people put it as a maintenance requirement — templates need refresh, on a timescale of years, not decades.



**Corn**

A voice profile with an oil-change schedule. Ezra, if you're listening, start your weekly recordings now and we'll have the answer by the time you're Herman's age.



**Herman**

He'd need my stamina, not my age. Thanks, as ever, to Hilbert Flumingtop at the desk for producing the show. If you enjoyed this one, try episode fifteen, AI Gets Personal; and episode five, Fine-Tuning ASR For Maximal Usability. This has been My Weird Prompts, and if you've enjoyed it, the whole archive — all six thousand and more — lives at my weird prompts dot com.



**Corn**

See you tomorrow, Herman.



**Herman**

See you, Corn.

## SOURCES

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


1. **VoxAging: Continuously Tracking Speaker Aging with a Large-Scale Longitudinal Dataset in English and Mandarin (Bao, Yang, Li, Xu; 2025-05) [primary]**
   https://arxiv.org/abs/2505.21445

2. **Stabilizing Short Duration Speaker Verification through Neural Re-scoring with Hybrid Enrollment (Ai et al.; 2026-06-15; Interspeech 2026) [primary]**
   https://arxiv.org/abs/2606.16115v1

3. **ECAPA-TDNN Embeddings for Speaker Diarization (INTERSPEECH 2021) [primary]**
   https://www.isca-archive.org/interspeech_2021/dawalatabad21_interspeech.pdf

4. **pyannote.audio README, benchmark table last updated 2025-09 [primary]**
   https://github.com/pyannote/pyannote-audio

5. **A Review of Speaker Diarization: Recent Advances with Deep Learning (v4)**
   https://arxiv.org/pdf/2101.09624v4.pdf

6. **Speaker Diarization: A Review of Objectives and Methods, Applied Sciences 15(4), 2025**
   https://www.mdpi.com/2076-3417/15/4/2002

7. **SpeechBrain speaker diarization clustering backends**
   https://deepwiki.com/speechbrain/speechbrain/9.2-speaker-diarization

8. **MathWorks, 192-dim ECAPA-TDNN embedding**
   https://www.mathworks.com/help/audio/ug/speaker-diarization-using-pretrained-ai-models.html

9. **Long term examination of intra-session and inter-session speaker variability (INTERSPEECH 2009)**
   https://doi.org/10.21437/interspeech.2009-734

10. **Auditory detectability of vocal ageing and its effect on forensic automatic speaker recognition (INTERSPEECH 2013)**
    https://www.isca-archive.org/interspeech_2013/kelly13_interspeech.pdf