Episode Book

EPISODE #6511

Why Voiceprints Are Not an Interoperable File Format

PUBLISHED	9 November 2026 at 17:25 UTC
DURATION	23:28
LISTEN	myweirdprompts.com/episode/voiceprint-biometric-interoperability-standards-95549/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	17 cited by the research (listed at the end)
BOOK GENERATED	10 November 2026 at 10:05 UTC

SYNOPSIS

Voice biometrics often promises portable voiceprints, yet no vendor-neutral standard exists to exchange speaker models. The official ISO/IEC 19794-13 biometrics standard deliberately covers only raw audio packaging, leaving extraction and speaker embeddings outside its scope due to vendor trade secrets. Looking at efforts like Brno University of Technology's Voice Biometry Standardization Initiative reveals the underlying technical friction: cross-system use requires identical frozen checkpoints, and transferring representations across different extraction pipelines demands training custom pairwise regression bridges.

TRANSCRIPT



Corn

If a word costs a hundred and thirty-five Swiss francs, you'd expect it to mean something.



Herman

That's... actually a very precise entry point, because that's the price of ISO/IEC 19794-13:2018, the standard whose title says "voice data," and the word "voiceprint" is exactly what Daniel is asking about. He keeps running into it in voice identification reading and wants to know, is a voiceprint an actual standard? A defined file format one system can export and another can read? Or is it just a loose marketing word for whatever speaker embedding a vendor happens to store? And if something real exists, who maintains it, and does anyone actually use it?



Corn

He's asking us to check whether the cheque clears.



Herman

He is. And the short answer is no, the cheque does not clear. There is no widely adopted, vendor-neutral file format for a speaker model. "Voiceprint" is a loose term, and the people who came closest to standardizing the thing behind it refuse to use the word.



Corn

Which is a strong hint before we've opened a single document.



Herman

It's the strongest hint available. So the way I'd unpack this, Corn, is three layers that the word blurs together. Layer one: the audio recording itself, the WAV file. Layer two: the representation, the speaker embedding, or the older i-vector, x-vector, the stack of numbers a system derives from your voice. Layer three: the serialization, the actual file that carries that stack of numbers around. "Voiceprint" gestures at all three, and interoperability means something completely different at each layer.



Corn

Two systems can agree perfectly on how to ship a recording and still be unable to compare a single number inside it.



Herman

That's the whole episode in one line, really. The first place to look is the standard that sounds like it's about voiceprints, ISO/IEC 19794-13, and the first thing to notice is what it says it doesn't cover. It's maintained under ISO/IEC JTC1/SC37, the biometrics committee. Published March twenty eighteen, first edition, twenty-six pages, a hundred and thirty-five Swiss francs, and it was last reviewed and confirmed in twenty twenty-four, so it's still current. On paper, that's the voiceprint standard.



Herman

Because the abstract has a sentence in it that should be carved on the door of every standards body. "This document also does not address handling of data that has been processed to the feature or voice model levels." It standardizes shipping a recording of a single speaker from a single session, with metadata. What it explicitly does not standardize is the voiceprint.



Corn

So the voiceprint standard excludes voiceprints.



Herman

For copyright reasons, essentially. Legal ones. Raw audio is hard to move across borders because of interception and privacy law in different countries, so there was real demand for a standard way to package it. But the model that comes out the other end, the thing you'd actually want to exchange between two identification engines, that was left outside the fence on purpose.



Corn

Who wanted it outside the fence?



Herman

The vendors, going back years. There's a position paper from PerSay, an Israeli voice biometrics vendor, at a W3C workshop in August two thousand eight, and it's remarkably blunt. Their position: the only biometric data that should be transferred between systems is the raw audio recordings, and it is, quote, "irrelevant to define a standard way to transfer the voiceprints themselves," because the voiceprints' data is different for any vendor and cannot be used by any other engine.



Corn

"Irrelevant." A vendor calling a standard irrelevant is a rare moment of honesty.



Herman

It's self-interested honesty, which is the best kind. Their voiceprint was their trade secret. Standardizing it would hand competitors the ability to read their enrollment data. So the audio-container camp won. Beigi and Markowitz's twenty-ten SAFE proposal, Standard Audio Format Encapsulation, standardized the audio container, and that lineage ran through ANSI/INCITS 456 into the ISO project that became 19794-13. The people who wanted portable models got nothing.



Corn

Except the Brno crowd.



Herman

And this is the part I find admirable. In July twenty fifteen, the Speech at FIT group at Brno University of Technology launched the Voice Biometry Standardization Initiative, VBS. And they didn't go the ISO route first, they just published the thing. Draft manual dated the seventeenth of July twenty fifteen, reference implementation online, presented at Interspeech in Dresden that September.



Corn

Published first, committee later.



Herman

That was the explicit philosophy, actually. The manual says, fix the core i-vector extraction algorithms, fix the necessary parameters, do the necessary minimum, and let people keep their own voice activity detection and scoring. They even had a slide about the USB analogy, the idea that a standard grows the whole market for every vendor.



Corn

How'd that work out for USB? Every vendor in a standards body, arguing.



Herman

Well, in this case there was no committee to argue in, which was both the strength and the weakness. What they standardized is a full i-vector pipeline. The reference system is a six hundred dimensional i-vector over a twenty forty-eight component Gaussian mixture model as the universal background model. They shipped the actual trained models as gzipped text files, the UBM and the six hundred dimension extractor matrix. The code is Python two point seven, numpy and scipy, Apache two licensed, patent-free, one script to extract i-vectors from audio and one to score them, with the scoring explicitly marked as not part of the standard. Output is a gzipped text file containing a single line of the i-vector, and the exchange format is defined both as binary for storage and Base64 ASCII for sending over email or HTTP.



Corn

So there is a file format. One system exports, another reads.



Herman

There is a serialization format, and that distinction is load-bearing. Here is the precondition from their own site: cross-system use of i-vectors "assumes that the i-vector extraction method, including the parameters of the method, is kept fixed, so that all i-vectors are compatible, and that their direct comparison is feasible."



Corn

Everyone must run the identical frozen checkpoint.



Herman

That is a different species of interoperability than a file format. A PDF is a PDF because any printer can render it. This is closer to, everyone must own the same printer, and the document only makes sense coming out of that one printer.



Corn

It's a standard the way a religion is. One church, universal truth.



Herman

I'd phrase it as, it's a reference model with a serialization convention bolted on. Which, to be fair, does give you something real. If two organizations both run the VBS extractor exactly as frozen, their i-vectors do compare. And there was a reason to want that, because the whole pitch was legal. A voiceprint, an i-vector, is not recoverable to speech. It can be transmitted anonymously, yet still identify the speaker. Treat it like a fingerprint, ship it across borders without the privacy law that attaches to a recording. Their listed beneficiaries are police, law enforcement, border services, intelligence agencies, banks.



Corn

A fingerprint you can't renew. Fingerprints at least have the decency not to work over the phone.



Herman

And that's the tension the initiative never addressed. Their privacy argument is that the i-vector is privacy friendly because it can't be turned back into audio. Fine. But the same property makes it a biometric identifier that travels freely between agencies with no way to revoke it. You can't get a new voice the way you get a new password.



Corn

So neither the ISO standard nor the VBS initiative really gives you a portable voiceprint. Which is a bit of a sad ending, really.

D

Dorothy

Hello dear, it is time for my grim fact.

D

Dorothy

There was a whole effort to standardise voiceprints, launched in two thousand and fifteen, and the notes say there is no evidence of it being adopted by a standards body, no version two, and no activity after that same year.

D

Dorothy

The website's last news item is the presentation they gave in September, and then it just goes quiet.

D

Dorothy

Anyway, sweetheart, I have a casserole in and your laundry is still here, when are you coming round?



Corn

Mum, we're mid-episode.



Herman

That was… I love her. Okay. The VBS site's news table ends with that September twenty fifteen Interspeech presentation. No version two, no standards-body adoption, nothing after. The site's own commercial advice to companies is to consider purchasing newer and significantly more precise systems from Phonexia, which is a Brno spinoff, so the standard's own homepage points you away from the standard. But I don't want to leave the impression the effort was pointless, because it produced the clearest demonstration I know of why voiceprints don't travel.



Corn

The Glembek work.



Herman

Glembek and colleagues, twenty fifteen, same group. The question was, can we use i-vectors produced by one system for scoring on another system? And their presentation has two slides I think about a lot. One is titled "It works." The other is titled "It does not work," annotated in caps, "MISMATCH."



Corn

Peer review by slide title.



Herman

The finding is that the interpretation of i-vectors changes from system to system, depending on feature extraction, on how the background model was trained, the initialization, the EM algorithm, the Gaussian splitting protocol, on how the extractor was trained. Two systems both output six hundred numbers per speaker, and the numbers mean different things.



Corn

Six hundred numbers that agree on nothing except the count.



Herman

Their fix was to learn a transformation between each pair of systems. They started with one-hidden-layer neural networks, tried multi-layer networks, and concluded that no hidden layer at all, plain linear regression, worked best. Trained on about nine thousand hours of audio, using THEANO, this was twenty fifteen.



Corn

A regression between every pair. So four systems is six bridges.



Herman

And the four systems in that study tell you the fragmentation story on their own. A reference system, nineteen MFCCs with deltas over a twenty forty-eight component model, six hundred dimensions, nine thousand hours. A second one trained on two thousand hours. A third with five hundred twelve components and a four hundred dimensional vector. A fourth called Perseus using twenty Perseus coefficients, which is its own front end entirely. All four called their output an i-vector. None directly comparable. "



Corn

Meaning, we got it working for the pairs we tried.



Herman

Meaning it's a migration project every time, not a file format. And you asked earlier whether anyone actually uses the VBS standard in practice, which is the other half of Daniel's question. There's one documented research use I know of: a Speakers in the Wild challenge system at Interspeech twenty sixteen by Kudashev and colleagues used the VBS i-vector extractor as one of their two systems, the other built in-house. A lab using it as an off-the-shelf baseline, which is exactly what a reference implementation is for.



Corn

Not a bank shipping a voiceprint to a border agency.



Herman

I haven't seen any evidence of commercial or agency deployment. And the thing is, the underlying cause hasn't gone away, it's gotten worse. The i-vector era at least had a shared architecture. Today's speaker embeddings are neural checkpoints, every one with its own architecture, dimension, training data and license. On Hugging Face, searching speaker embedding gets you a long tail of community uploads. The most downloaded one in a recent search had a hundred and five downloads in thirty days. Another had fifteen. These are models with tens of downloads, not millions.



Corn

The most popular speaker model in the world has the audience of a decent birthday party.



Herman

And they're mutually incompatible by construction. One is thirty million parameters trained on low-quality speech under a non-commercial license, another is a SpeechBrain checkpoint from twenty twenty-three, another is a gated Apache-licensed five hundred twelve dimensional model. There's no common dimension, let alone a common meaning. This is the exact fragmentation the VBS Initiative was trying to fix in twenty fifteen, and nobody is proposing the modern equivalent.



Corn

We did an episode on safetensors and STT formats, and the lesson there was the same one wearing different clothes. A file format guarantees you can move bytes. It says nothing about whether the other end can do anything with them. Speaker models are that lesson with the volume turned up.



Herman

Because STT weights at least mostly agree on the task. Speaker embeddings don't even agree on the coordinate system. Meanwhile the word itself has gone fully feral. On GitHub this year, "VoicePrint" as a repo name attaches to a forensic tool for attributing AI voice clones, matching a synthetic voice back to its source. Another April project uses it for authentication with ECAPA-TDNN embeddings, reporting ninety-four point six percent accuracy and a three point eight percent error rate, which is at least a real model. And then there are two near-duplicate watermarking repos that embed a "sixty-four bit voiceprint hash" as an inaudible watermark in audio.



Corn

Sixty-four bits. That's not a speaker model, that's a fortune cookie with ambitions.



Herman

It's a hash. A sixty-four bit hash of a speaker embedding has thrown away essentially all of the information that makes the embedding work. But it's called a voiceprint, because the word is available and sounds technical. Even Wikipedia can't pick one meaning. The disambiguation page lists a spectrogram, the stored speaker template, a Canadian



Herman (continued)

broadcast reading service that got renamed in twenty twelve, and an English record label. Four unrelated things, one word.



Corn

And the people who built the closest thing to a standard refuse to say it.



Herman

The VBS site says i-vectors are "occasionally referred to as audio voice-prints, although this term should be used with caution and therefore we avoid using it at all." The standardizers avoid the word. So when Daniel keeps running into it in vendor literature, that's not a standard he's failing to find the spec for. It's a word vendors use because "speaker embedding" doesn't fit on a slide.



Corn

"Voiceprint" sounds like it was matched by a fingerprint. Nobody ever bought the fingerprint equivalent of a sixty-four bit hash.



Hilbert

It's forty, by the way.



Corn

Forty what?



Hilbert

The Brno extractor was built on forty hours of me not getting an answer. Different thing, same year. I had two speaker systems in a building I won't name, and the second machine could not read the first one's numbers, and the man from the supplier said it was a compatibility issue, which it was, in the sense that they were incompatible.



Herman

So what did you actually do?



Hilbert

Made a notebook. A paper one, Herman, before you say the word spreadsheet. One row per pair of systems, a column of correction factors, tuned by ear over about six weeks. Room four was the one that ran hot, its factors drifted every summer and I'd re-tune them in June. I numbered the rows by the room, not the system name, because the system names meant nothing and the rooms didn't lie.



Corn

You tuned speaker verification by ear.



Hilbert

By result, Corn, the ear came second. And the word, "voiceprint," I never used it. It was a stack of numbers that only meant anything next to the exact machine that made them, same as a lathe setting. If you moved the machine, the numbers were scrap. Nobody wanted to hear that in a meeting, so the notebook stayed in my pocket. When I did the sound desk in Shepherds Bush the same rule held, the numbers belonged to the box.



Herman

The correction factors drifting with the temperature, that's a real thing, the room acoustics changing the front end's behavior.



Hilbert

Room four had a wall of glass on the south side, if you want the physics. June, the factors moved. I stopped apologizing for it. Graham, who ran the second system, took eleven minutes to do a swap I could do in four, and he was paid more than me. That's not relevant to your standard, but I remember it.



Corn

The notebook is the most honest interoperability document this field has produced.



Hilbert

It's in a drawer. The June factors are still in it.



Herman

I want to come back to something in that, actually, because the June drift is the Glembek finding happening in a building instead of a lab. The interpretation of the numbers depends on everything around them, including the glass wall. What breaks when people assume "voiceprint" means "portable" is that the numbers were never coordinates in a shared space to begin with.



Corn

And Hilbert's answer to non-shared coordinates was a pocket notebook and a grudge about Graham.



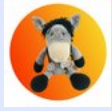
Herman

Which is arguably more scalable than what the field produced, because at least it was maintained. The VBS site has been quiet for eleven years.



Corn

Here's what I keep landing on, though. The embedding era should have made this easier. The models are all public checkpoints on Hugging Face now, Apache licenses, no vendor vault. The raw ingredients for a standard are lying around free, and the result is more fragmentation, not less.



Herman

Because openness of the weights doesn't buy agreement on the representation. Every team trains its own embedding for its own data and publishes it, and nobody has an incentive to converge. The i-vector world had fewer actors and a shared architecture, and



Herman (continued)

even they couldn't get past "everyone run my frozen extractor."



Corn

PerSay saw this in two thousand eight and called the whole idea irrelevant, and eleven years of trying since has mostly proved them right.



Herman

The misconception I'd want to correct before we go, because it's the one Daniel's question is really aimed at. People hear "voiceprint" and picture a fingerprint card, a standard artifact any qualified examiner can read.



Corn

It's the opposite. The closest real standard excludes the model, the closest real model standard excludes everyone else's extractor, and the word itself is avoided by the people who studied it hardest. A voiceprint is whatever the vendor says it is, which is the fingerprint of nothing.



Herman

Though to be fair to the word, it's earned one thing. Eleven years of the field not solving the problem, and the word still sounds like it was solved. That's durability.



Corn

Forward-looking thought, then. If a standardization effort started tomorrow, what would it even standardize? You can't freeze an embedding model the way Brno froze an i-vector pipeline, the field moves too fast now, and a frozen twenty twenty-six checkpoint would be Phonexia's advice page all over again within two years. Per-pair regressions, notebooks, migration projects forever. The question I'd leave with Daniel is whether interoperability is even the right goal, or whether the honest architecture is shipping audio and keeping the models where they were trained.



Herman

And the word will keep drifting regardless. Watermark hashes, forensic clone attribution, whatever attaches next. The thing Daniel actually wants, a portable speaker model, stays a research problem wearing a marketing word.



Corn

Thanks as always to our producer, Hilbert Flumingtop, at the desk with the notebook in the drawer.



Herman

If you enjoyed this one, try episode three, Safetensors or something else; and episode seven, Building Custom ASR Tools. This has been My Weird Prompts. If you got something out of this one, the reviews help, wherever you listen.



Corn

Find us at my weird prompts dot com, and we'll be back soon. See you then, Herman.



Herman

See you, Corn.

SOURCES

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

1. **Voice Biometry Standardization Initiative homepage , retrieved 2026-11-09; news table last entry 2015-09. [primary]**
   https://voicebiometry.org/

2. **Voice Biometry Standard [primary]**
   Draft, Glembek, Burget, Matějka, dated 2015-07-17 , retrieved 2026-11-09.
   https://www.voicebiometry.org/download/vbs_manual.pdf

3. **ISO/IEC 19794-13:2018, Biometric data interchange formats [primary]**
   Part 13: Voice data, Edition 1, published 2018-03, last reviewed and confirmed 2024 , retrieved 2026-11-09.
   https://www.iso.org/standard/72276.html

4. **Glembek et al., Voice-print transformation for migration between automatic speaker identification systems, 2015-09-10 , retrieved 2026-11-09. [primary]**
https://www.fit.vut.cz/research/group/speech/public/publi/2015/glembek_eafs_2015-09-10.pdf.en

5. **PerSay position paper, W3C Speaker biometrics and VoiceXML 3.0 Workshop, 2008-08 , retrieved 2026-11-09. [primary]**
https://www.w3.org/2008/08/siv/Papers/PerSay/PerSay_position_paper.pdf

6. **Beigi & Markowitz, Standard audio format encapsulation (SAFE), Springer, 2010 , retrieved 2026-11-09. [primary]**
https://www.recotechnologies.com/~beigi/ps/beigi_safe_2010.pdf

7. **IEC webstore listing for ISO/IEC 19794-13:2018, retrieved 2026-11-09.**
https://webstore.iec.ch/en/publication/62740

8. **NEN catalogue entry, 26 pages, publication date 2018-03-01, retrieved 2026-11-09.**
https://www.nen.nl/nen-iso-iec-19794-13-2018-en-244341

9. **Beigi & Markowitz, A Standard Audio Encapsulation Method, W3C SIV workshop position paper, 2008, retrieved 2026-11-09.**
https://www.w3.org/2008/08/siv/Papers/Recognition_Technologies/Position_paper-Beigi_and_Markowitz.pdf

10. **Kudashev et al., A Speaker Recognition System for the SITW Challenge, Interspeech 2016, retrieved 2026-11-09.**
https://www.isca-archive.org/interspeech_2016/kudashev16_interspeech.pdf

11. **disambiguation page, last edited 2013-11-09, retrieved 2026-11-09.**
https://en.wikipedia.org/wiki/Voiceprint

12. **disambiguation page, last edited 2016-10-03, retrieved 2026-11-09.**
https://en.wikipedia.org/wiki/Voice_print

13. **created 2026-08-01, retrieved 2026-11-09.**
https://github.com/wobushannes/VoicePrint

14. **created 2026-04-24, retrieved 2026-11-09.**
https://github.com/Swathilekha/Voice_print_authentication

15. **created 2025-11-08, retrieved 2026-11-09.**
https://github.com/LilRag/Hybrid-Audio-Verification-via-Biometric-Watermarking

16. **created 2026-09-25, CC-BY-NC-4.0, retrieved 2026-11-09.**
https://huggingface.co/hynt/Speaker-Embedding-30M-Low-Quality-Speech

17. **created 2024-10-14, retrieved 2026-11-09.**
https://huggingface.co/csukuangfj/speaker-embedding-models