Episode Book

EPISODE #6492

What Audio Models Learn Before Fine-Tuning

PUBLISHED	8 November 2026 at 14:30 UTC
DURATION	27:02
LISTEN	myweirdprompts.com/episode/speech-model-internal-representations-88381/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	9 November 2026 at 08:09 UTC

SYNOPSIS

Speech encoders arrive at fine-tuning with tens of thousands of hours of audio structure already baked into their weights. This examination of internal representations contrasts self-supervised models like wav2vec 2.0, HuBERT, and w2v-BERT against weakly supervised systems like Whisper. It explores layer-by-layer probing of tongue position and word boundaries, the resurgence of speaker identity in deep layers, the loss of lexical tone during quantization, and the trade-off between flexible latent representations and out-of-the-box robustness.

TRANSCRIPT



Corn

There's a moment Daniel's prompt lands on that I want to sit with before anything else. Before you fine-tune a model on your own voice, the model already knows things about sound. A lot of things. And none of that came from you.



Herman

None of it. The encoder shows up to your fine-tune with tens of thousands of hours of other people's audio already baked in.



Corn

That's Daniel's angle. He's asking what that pretraining actually teaches the model, what the internal representation of speech looks like once it's done, and how the choice of pretraining decides what a fine-tune can touch and what it can't. He names the two families — the self-supervised ones, wav2vec 2.0, HuBERT, w2v-BERT, that learn by masking and predicting on unlabeled audio, and Whisper, which took the weakly supervised route on hundreds of thousands of hours of transcribed web audio.



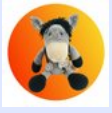
Herman

Two different bets about what a speech model even is.



Corn

So let's start with the two families and what they're actually optimizing, because I suspect it's not what most people assume.



Herman

It isn't, and the fork is load-bearing. On the self-supervised side, you take raw audio — no transcript, nothing — and you make the model solve puzzles about it. Wav2vec 2.0 runs the waveform through a stack of convolutions to get latent features, then masks spans of those latents, not the waveform itself, and hands the gappy result to a Transformer. The training task: for each masked spot, pick the correct quantized version of that latent out of a set of distractors.



Corn

A multiple-choice exam where the model writes its own answer key.



Herman

That's closer than you'd want it to be, actually. The quantization — turning the continuous latents into discrete tokens — is learned at the same time, via product quantization with a Gumbel softmax, and the paper is quite firm that learning those discrete units jointly with everything else beats fixing them in a previous step. So the answer key is being written while the exam is being graded.



Corn

And the payoff for all that?



Herman

The numbers are famous for a reason. With all nine hundred and sixty hours of LibriSpeech labels, wav2vec 2.0 hits 1.8 clean, 3.3 other. But with ten minutes of labeled audio — ten minutes — on top of fifty-three thousand hours of unlabeled pretraining, it gets 4.8 and 8.2.



Corn

Ten minutes of labels. That's the length of one of my shorter naps.



Herman

Then you'd be a viable fine-tuning dataset. The whole motivation is the label bottleneck — somewhere near seven thousand languages out there, and the overwhelming majority have nothing close to a thousand hours of transcripts. The paper even gestures at infants, who learn language by listening without any labels at all.



Corn

I have opinions about how infants eat, but I'll allow this one.



Herman

Now HuBERT comes at the same territory from a different angle, and this is where it gets philosophically interesting. Speech is awkward for masked prediction in a way text isn't. In text, BERT has a lexicon — the tokens are given. In speech, an utterance contains multiple sound units, you have no lexicon during pretraining, and the units have variable lengths with no segmentation telling you where one ends and the next begins.



Corn

No word boundaries, no phone boundaries, nothing.



Herman

Nothing. So HuBERT's fix is almost scandalously simple: run k-means clustering on the audio features first, starting with a hundred clusters, and use those cluster assignments as the targets for a BERT-style masked prediction loss. Then redo the clustering using the learned representations, twice. And the loss only applies over the masked regions, which the authors say forces the model to learn a combined acoustic and language model over continuous inputs.



Corn

A hundred clusters is a weirdly small number. English has more phones than that, before you've added a single accent.



Herman

It's deliberately crude, and the paper's core claim is that it doesn't matter. The line I love — HuBERT, quote, relies primarily on the consistency of the unsupervised clustering step rather than the intrinsic quality of the assigned cluster labels. The targets don't have to be right. They have to be consistent. If the same sound keeps getting the same wrong label, the model still has to learn the sequential structure of speech to predict that label from context.



Corn

So the teacher is an idiot, but a reliably consistent idiot.



Herman

A consistent idiot who never gets tired. And that's the argument against pseudo-labeling as the alternative — a student mimicking a teacher is capped by the teacher's supervised data and annotation quality, while a pretext task forces the model to compress the entire signal into its latents. Everything in the audio, not just what a transcript bothered to mention. HuBERT's stated ambition is a high-fidelity representation that includes speaker identity, emotion, hesitation, interruptions — plus laughter, coughing, lip-smacking, engines, birds, sizzling food.



Corn

Sizzling food.



Herman

In the paper.



Corn

HuBERT has heard more sizzling food than most line cooks.



Herman

And at a billion parameters it takes up to nineteen percent relative WER off the hardest LibriSpeech test sets. So the crude teacher works.



Corn

w2v-BERT, since Daniel named it — the one that supposedly combines both.



Herman

I'd like to be careful there, because I'm working from how it's used downstream more than from its own paper. The short version people give is that it pairs wav2vec 2.0-style contrastive learning with HuBERT-style masked prediction in one model, and I've no reason to doubt it, but I can't claim I've sat with the original. What I can say is that on the Whisper language ID table, a w2v-BERT model of roughly six hundred million parameters scores over seventy percent while Whisper's own much larger model gets about sixty-five.



Corn

The runt out-identifying the big one. Noted.



Herman

Hold that thought, it comes back later. The question of what actually comes out of these models — the representation — is where the probing literature gets rich. Start with articulation. There's work using electromagnetic articulography, sensors on the tongue and jaw tracking the vocal tract while people speak, then asking whether you can decode those physical traces from the model's hidden states with just a linear layer.



Corn

And?



Herman

Best correlation around 0.81. And to train that linear probe you need about five minutes of data to hit 0.77.



Corn

Five minutes. The model's latents track tongue position, linearly.



Herman

The tongue is in there. Then the layer-by-layer picture — this is the intellectual core of the whole episode, Corn, so let me have it. In wav2vec 2.0 Base, phonetic information peaks around layers six and seven. In the Large model trained on sixty thousand hours, you get phonetics strong at layers eleven and around eighteen, nineteen, with a dip at thirteen to seventeen where word-level, contextual content rises. Word identity, across wav2vec 2.0, HuBERT and WavLM, concentrates in intermediate layers, and — this detail kills me — word identity information clusters near the center of each word segment. The model has never been told words exist, and it has learned where the middle of a word is.



Corn

It invented the notion of a word's center without a dictionary.



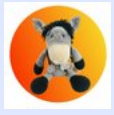
Herman

From raw audio alone. Acoustics at the bottom, phonetics in the middle, words and meaning above them — that's the standard picture. But the standard picture has a crack in it. There's a probing study covering eleven self-supervised speech models, and the finding that challenges the consensus is that larger models recover speaker-discriminative information in their deep layers. The top is supposed to be purely linguistic — abstract away who's talking, keep what was said. Not so. The bigger the model, the more speaker identity comes back near the top, and the intermediate layers of these SSL models beat specialized speaker embeddings on several attributes.



Corn

Which is a nice callback to a question we did a long while back — whether your fine-tune goes useless as your voice ages. The answer hides exactly here, doesn't it. If speaker identity survives to the deep layers, then a fine-tune that captures your voice isn't grabbing something fragile at the bottom. It's grabbing something the model carries all the way up.



Herman

That's exactly the connection. And it cuts both ways — if identity is woven through layers people assumed were purely about words, then "just take the last layer" is bad advice for anyone doing anything with voice.



Corn

One more piece before we leave the SSL side. The discrete units.



Herman

Right, because a lot of modern pipelines — speech language models, voice conversion — run on discrete tokens, not the continuous latents. And here's the quiet loss. Osakuade and King probed Mandarin and Yorùbá, both lexical tone languages, and found the continuous SSL latents do encode tone. The discrete units obtained by quantizing those latents tend to prioritize phonetic structure, and lexical tone becomes less reliably encoded. Across quantization methods, not just k-means.



Corn

Quantization is where prosody goes to die.



Herman

It's a structural loss of exactly the stuff that makes a voice sound like a person. There's a proposed fix — cluster once for phonetics, then again on the residual — but the headline finding stands.



Corn

So summarize the SSL side for me. What's actually in the box?



Herman

A layered representation where acoustics, articulation, phonetics, words, speaker identity and delivery style all coexist — with the top layer messier than the textbooks say. And one more finding that decides the whole shape of the discussion: a similarity study by Chung, Belinkov and Glass found the training objective has more impact on the representation than architectural choices — RNN versus Transformer versus CNN, directionality, all of it. And the word-knowledge work found HuBERT and WavLM consistently outperform wav2vec 2.0 on acoustic word discrimination, word segmentation and spoken semantic similarity.



Corn

The objective, not the architecture. You could build it out of anything, but what you ask it to do during pretraining is what it becomes.



Herman

What it's asked to do is what it learns.



Corn

So that's what the SSL models learn. Now the other bet — Whisper looked at all of this and went somewhere else entirely.



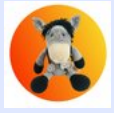
Herman

Whisper's authors diagnosed the SSL paradigm pretty bluntly. The encoders learn high-quality representations, they said, but they lack an equivalently performant decoder to map those representations to output — so you need a fine-tuning stage, and fine-tuning is a complex process requiring a skilled practitioner, with a risk of exploiting brittle, spurious, dataset-specific patterns. Their number for that risk comes from vision: fine-tuning a computer vision model on ImageNet produced a 9.2 percent increase in object classification accuracy on ImageNet itself, with no improvement in average accuracy across seven other natural image datasets.



Corn

9.2 percent of nothing, in other words.



Herman

Spurious gain. Memorizing the exam. So Whisper's bet: skip fine-tuning entirely, train on six hundred and eighty thousand hours of weakly supervised web audio with transcripts, and ship a system that works reliably out of the box without a decoder fine-tune for every deployment.



Corn

Six hundred eighty thousand against wav2vec's fifty-three thousand unlabeled. Twelve times the audio, and all of it with imperfect transcripts found lying around the internet.



Herman

Of which a hundred and seventeen thousand hours cover ninety-six languages, plus a hundred twenty-five thousand hours of translation into English. And the weak labels are weak — they don't even normalize the text, they just let the seq2seq model learn the mess.



Corn

Which buys what, exactly?



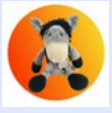
Herman

Robustness to conditions it saw. In the paper's noise sweep, LibriSpeech-trained models degrade faster than the best Whisper under white noise and pub noise, and models that lead under clean conditions get overtaken once the signal-to-noise drops below ten decibels. Hundreds of thousands of hours of messy web audio — bad microphones, music, background chatter — buys you a model that doesn't fall apart in a café.



Corn

You'd expect the mirror to be true too. Conditions it didn't see.



Herman

Exactly true, and the paper is honest about it, which I respect. Language ID, zero-shot on Fleurs: Whisper gets sixty-four and a half percent. A two-billion-parameter mSLAM model gets seventy-seven and a half, and that six-hundred-million-parameter w2v-BERT gets seventy-one. Whisper has no training data at all for twenty of Fleurs' hundred and two languages, so it's capped around eighty percent even in principle — on the eighty-two it has, it reaches eighty-three.



Corn

Held back by its own menu.



Herman

And then there's Welsh. This is my favorite failure in the whole paper. The dataset appears to contain about nine thousand hours of Welsh translation data — fourth largest language in the training set, ahead of French, Spanish and Russian. It scores thirteen BLEU. Because the majority of that supposedly Welsh data is English audio with English captions, mis-classified as Welsh. Nine thousand hours of the wrong language, sitting in fourth place, fooling everyone who reads the table.



Corn

Fourth place in Welsh, by volume. There's a joke about the internet in that, but the fact is funnier.



Herman

Weak supervision's strength and weakness are the same mechanism. Nobody checked, because checking is the whole thing you were trying to avoid.



Corn

We did say, back when we did the episode on how fine-tuning works mechanically, that the fine-tune is the last mile of a very long road. Whisper's authors are essentially arguing the last mile is where the accidents happen.



Herman

They are. Theirs is a foundation-model argument: don't specialize, generalize. The SSL community's argument is the opposite — specialize, but specialize carefully. And then there's the behavioral evidence on Whisper's own encoder, which the Whisper paper itself doesn't probe at all — that comes from third parties. The Trinity College Dublin phonemic restoration work perturbed segments in real words and in pseudorandom non-words, then probed wav2vec 2.0 large and Whisper large-v3 layer by layer for articulatory features — place, manner, voicing. wav2vec 2.0 was better at restoring the perturbed segments.



Corn

The masked-prediction model is better at filling in what was masked.



Herman

It's almost tautological until you remember Whisper saw ten times more data. But the training objective, not the data volume, decided that skill. Same conclusion as Chung's similarity study, arrived at behaviorally. The Whisper probing also shows a different layer profile — some models show a convex relationship between how separable a task is and depth, while Whisper's models converge, with weak tasks like stress detection and intent classification improving all the way down the stack while strong tasks like multi-speaker detection stay high. And on paralinguistic style, a probe on Whisper-large-v2 goes from about fifty-one percent accuracy at layer zero to about eighty-three at layer thirty-one.



Corn

Style rising with depth in Whisper. Speaker identity returning near the top in SSL. Both models carry the person all the way up, just differently.



Herman

Which sets up the question Daniel actually asked at the end — what a fine-tune can and can't change. The short, uncomfortable answer: fine-tuning both does more than people think and less. Zaiem, Parcollet and Essid looked at this directly. Using SSL encoders as frozen feature extractors severely limits how much of the encoder you can exploit — but fine-tuning hurts the robustness acquired during pretraining, especially in low-resource scenarios. You pay for the adaptation out of the robustness account. Their fix is continual-learning methods that slow the forgetting, and with XLSR they get relative gains of about sixteen percent on English and twenty-two and a half on Danish, and the probing shows the gains line up with less forgetting.



Corn

So there's a measurable exchange rate between adapting and forgetting.



Herman

A real one. The children's-speech work fills in the other half. When they analyzed why fine-tuned SSL models do better on kids' phoneme recognition, the finding was that the superiority comes largely from the models having learned features for older children's



Herman (continued)

speech and then adapting those to younger children. Fine-tuning reweights an existing capability. It doesn't create one.



Corn

Which is Daniel's question answered sideways. A fine-tune on my voice can't add a me that wasn't already representable.



Herman

If you're a younger child, evidently you're a reweighting of an older one. And the transfer question has boundaries people skip. Poncelet and Van hamme, on Flemish Dutch — every pretrained model improved linear phone separability in Flemish, but not all of them improved actual ASR. What mattered was a substantial amount of data and a matching pretraining domain. Fine-tuning multilingual XLSR on Flemish Dutch took thirty percent off the WER, which is enormous, but that's because the domain matched.



Corn

Matching domain as a precondition. So the Bangla study —



Herman

Yes, the mirror result. Thirteen-domain Bangla benchmark, BanSpeech, about six and a half hours, eight thousand utterances. They compared wav2vec 2.0 pretraining, Whisper, and fully supervised CNNs, and self-supervised cross-lingual pretraining won — better than weak supervision, better than full supervision, on that hard multi-domain task.



Corn

On a benchmark Whisper should love. Messy, small, real-world.



Herman

Small is the whole point — the domain where weak supervision's magic requires data the domain doesn't have. And the Arabic dialect result completes the pattern: Whisper zero-shot beats fully fine-tuned XLS-R on every Arabic dataset they tested, and then deteriorates significantly on five dialects it hadn't seen — Algeria, Jordan, Palestine, the UAE, Yemen.



Corn

Coverage, not comprehension.



Herman

Which is the honest summary of the whole weak-supervision approach. Conditions it saw, it survives. Conditions it didn't, it's just another model.



Corn

One more thing I want on the table, because it undermines half of what we've cited. The benchmark rankings themselves.



Herman

Zaiem and colleagues again, and this one deserves more attention than it gets. Alter the downstream architecture — change the probe head you attach — and the performance ranking of the evaluated models fluctuates significantly. Which model "wins" depends partly on the plumbing you measured it through.



Corn

So every claim we've made today that model A beat model B is partly a claim about B's hat.



Herman

Partly. It's the evaluation paradox of this whole literature — the rankings are real enough to act on and fragile enough to distrust. Both at once.



Corn

Hm. On that cheerful note, I feel like the person who fine-tunes on his own voice should hear the practical shape of it. Two papers, stitched together: fine-tuning costs you robustness, and it reweights rather than invents. What's that leave Daniel with, concretely?



Herman

A voice fine-tune that makes the encoder better at him and measurably worse at the pub he doesn't record in. The evidence says you can shift which layers carry which load — the infant-vocalization work found the model leveraging phonetic representations at middle layers — but you're redistributing an inheritance, not writing a new one.



Corn

You inherit the fifty-three thousand hours, and you spend your ten minutes — well, we record more than ten minutes.



Herman

You record considerably more than ten minutes.



Corn

I'd want the option of napping through it.



Herman

You'd nap through the recording and the fine-tune both, and claim the compute.



Corn

The compute naps whether I do or not. That's what I like about it.



Hilbert

—that's where the ten minutes is wrong, by the way.



Hilbert

It's not the ten minutes that's doing the work. Herman said it himself, Herman, you said fifty-three thousand hours of unlabeled audio, then the labels, then the win. Nobody asks what's in the fifty-three thousand. I'll tell you what's in mine. Everything. Every dictation tape since nineteen eighty-eight. I read the shipping news aloud, four minutes a morning. Weather, tonnage, delays at the port.



Herman

You've been recording your own voice for decades?



Hilbert

Unlabeled. Nobody's transcribed any of it. That's the point. It's pretraining, not fine-tuning. A man on Jaffa Road has been transferring it for me for eleven years — cassettes first, then minidiscs, then the digital recorders when the minidisc deck died. He keeps the masters. I asked him once what he does with them and he said, nothing, they're yours. Two hundred and forty tapes now.



Corn

You labeled them by date?



Hilbert

By weather. The date's on the tape but the weather's on the label. It's how I find things. The deck's at my mother's. There was no room here after the second minidisc batch, and she doesn't use the porch.



Herman

Two hundred and forty tapes of your own voice, unlabeled, at your mother's, indexed by weather. That is — structurally, that's a corpus.



Hilbert

It's forty years of pretraining on one speaker. Corn will get the joke eventually. The tonnage figures are the interesting part. August nineteen ninety-four, the delays at Ashdod ran six days. I said so on tape. Nobody else did.



Corn

The one thing I can't get past — you find a tape by looking up what the weather was.



Hilbert

You remember the weather better than the date. Everybody does. Nobody admits it.



Herman

Forty years of one voice, no labels. If anyone ever ran a real speaker-adaptation study, you're the control group nobody's using.



Hilbert

I'm not a control group, I'm a person with a porch's worth of tapes. Ashdod, six days, nineteen ninety-four. It's on tape twenty-one.



Corn

Cutting-room floor, before we close. The one I couldn't fit: someone measured how big the smallest wav2vec 2.0 and HuBERT variants actually are — ninety-five million parameters — and the same work got a nearly seventy-five percent size reduction through distillation for about a tenth of a percent of accuracy. Ninety-five million is already too fat for a phone, and that's the small one.



Herman

The other shelf item I'll sneak in: the same probing family found HuBERT's fixed twenty-millisecond resolution isn't right for everything, because speaker characteristics and semantics live at different time scales than phonemes. One clock, several tempos.



Corn

Which loops back to the open question. Nobody has directly measured what a fine-tune on one person's voice can and cannot change inside an SSL encoder. The closest evidence is sideways — continual learning showing the robustness drain, the children's-speech work showing adaptation rather than creation, the layer probes showing identity woven all the way to the top. Two families of papers orbiting an answer nobody's written down.



Herman

And underneath that, the philosophical split, which is the thing I'd leave anyone with. The two paradigms disagree about what fine-tuning is for. Whisper's authors treat it as a bug — the place where spurious patterns sneak in. The SSL community treats it as the entire point, then writes papers about doing it less destructively. If you're fine-tuning on your own voice, you're not just picking a model. You're picking a side in that argument.



Corn

Weigh it by weather, like everything else.



Herman

Producer credit — Hilbert Flumingtop at the desk, as ever.



Corn

If this was your kind of episode, go back for episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; episode eleven, How Does Fine Tuning Work Anyway; and episode nine, Benchmarking Custom ASR Tools - Beyond The WER. This has been My Weird Prompts. If the show earns its forty minutes in your week, a review wherever you listen helps more than you'd think — it's how the archive finds new ears.



Herman

See you soon.

SOURCES

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

1. **Baevski, Zhou, Mohamed et al., wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, arXiv v3 (submitted 2020-06-20, revised 2020-10-22). Retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2006.11477v3

2. **Hsu, Bolte, Tsai et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, arXiv v1 (submitted 2021-06-14). Retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2106.07447v1

3. **Radford et al., Robust Speech Recognition via Large-Scale Weak Supervision (Whisper), arXiv, 28pp. Retrieved 2026-11-08. [primary]**

   https://arxiv.org/abs/2212.04356

4. **Cho, Wu, Mohamed et al., Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech, arXiv v3 (submitted 2022-10-21, revised 2023-07-21). Retrieved 2026-11-08.**

   https://arxiv.org/abs/2210.11723v3

5. **What do self-supervised speech models know about words? (TTIC), arXiv v1. Retrieved 2026-11-08.**

   https://arxiv.org/html/2307.00162v1

6. **Layer-Wise Analysis of a Self-Supervised Speech Representation Model, ASRU 2021. Retrieved 2026-11-08.**

   https://doi.org/10.1109/asru51503.2021.9688093

7. **Comparative Layer-Wise Analysis of Self-Supervised Speech Models, ICASSP 2023. Retrieved 2026-11-08.**

   https://doi.org/10.1109/icassp49357.2023.10096149

8. **Chiu, Fung, Li et al., A Large-Scale Probing Analysis of Speaker-Specific Attributes in Self-Supervised Speech Representations, arXiv v4 (submitted 2025-01-09, revised 2026-09-05; accepted ISCSLP 2026). Retrieved 2026-11-08.**

   https://arxiv.org/abs/2501.05310v4

9. **Osakuade & King, Lexical Tone is Hard to Quantize: Probing Discrete Speech Units in Mandarin and Yorùbá, arXiv v1 (submitted 2026-04-08; Proc. Speech Prosody 2026, 38**

   42). Retrieved 2026-11-08.
   https://arxiv.org/abs/2604.07467v1

10. **What Do Speech Foundation Models Not Learn About Speech? arXiv. Retrieved 2026-11-08.**

    https://arxiv.org/html/2410.12948

11. **Heard but Not Heeded: Paralinguistic Information Encoding and Loss in Audio-Language Models, arXiv. Retrieved 2026-11-08.**

https://arxiv.org/html/2609.00727v1

12. **On the Interpretability of Whisper Encodings Using Sparse Autoencoders, arXiv v2. Retrieved 2026-11-08.**

https://arxiv.org/html/2605.12225v2

13. **Under the hood: Phonemic Restoration in transformer-based automatic speech recognition, Trinity College Dublin repository. Retrieved 2026-11-08.**

https://www.tara.tcd.ie/items/5f9494a6-ea0d-48c6-97c1-7cecbe55f4d2

14. **Zaiem, Parcollet & Essid, Less Forgetting for Better Generalization: Exploring Continual-learning Fine-tuning Methods for Speech Self-supervised Representations, arXiv v1 (submitted 2024-06-30). Retrieved 2026-11-08.**

https://arxiv.org/abs/2407.00756v1

15. **Kerpicci et al., Application of Knowledge Distillation to Multi-task Speech Representation Learning, arXiv v2. Retrieved 2026-11-08.**

https://arxiv.org/abs/2210.16611v2

16. **Wang, Boumadane & Heba, A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding, arXiv v3. Retrieved 2026-11-08.**

https://arxiv.org/abs/2111.02735v3

17. **Fan, Li, Zhou et al., Exploring wav2vec 2.0 on speaker verification and language identification, arXiv v2. Retrieved 2026-11-08.**

https://arxiv.org/abs/2012.06185v2

18. **Poncelet & Van hamme, Comparison of Self-Supervised Speech Pre-Training Methods on Flemish Dutch, arXiv v1 (ASRU 2021). Retrieved 2026-11-08.**

https://arxiv.org/abs/2109.14357v1

19. **Samin et al., Investigating self-supervised, weakly supervised and fully supervised training approaches for multi-domain automatic speech recognition: a study on Bangladeshi Bangla, arXiv v3. Retrieved 2026-11-08.**

https://arxiv.org/abs/2210.12921v3

20. **Talafha, Waheed & Abdul-Mageed, N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition, arXiv v1 (Interspeech 2023). Retrieved 2026-11-08.**

https://arxiv.org/abs/2306.02902v1

21. **Chemudupati et al., On the Transferability of Whisper-based Representations for 'In-the-Wild' Cross-Task Downstream Speech Applications, arXiv v1. Retrieved 2026-11-08.**

https://arxiv.org/abs/2305.14546v1

22. **Becerra, Ragano & Hines, Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction, arXiv v1. Retrieved 2026-11-08.**

https://arxiv.org/abs/2204.02135v1

23. **Li, Hasegawa-Johnson & McElwain, Analysis of Self-Supervised Speech Models on Children's Speech and Infant Vocalizations, arXiv v2. Retrieved 2026-11-08.**
https://arxiv.org/abs/2402.06888v2

24. **Zaiem, Kemiche, Parcollet et al., Speech Self-Supervised Representations Benchmarking: a Case for Larger Probing Heads, arXiv v2. Retrieved 2026-11-08.**
https://arxiv.org/abs/2308.14456v2

25. **Chung, Belinkov & Glass, Similarity Analysis of Self-Supervised Speech Representations, arXiv v2 (ICASSP 2021). Retrieved 2026-11-08.**
https://arxiv.org/abs/2010.11481v2