SUPERB

Speech processing Universal PERformance Benchmark

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai and sixteen co-authors

Interspeech 2021 · arXiv:2105.01051

National Taiwan University, MIT, Facebook AI Research, Johns Hopkins, Amazon AI, Carnegie Mellon

What do these representations actually encode

Why this paper matters

A four-page benchmark paper that redefined how the field reports results

10

tasks across four aspects of speech

14

pre-trained models evaluated under one protocol

frozen

upstream model, shared across every task

s3prl

the open toolkit that made the protocol reproducible

The claim

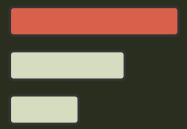
Freeze a self-supervised model, attach a lightweight prediction head per task, and you can reach competitive performance across ten different speech tasks. If that works, the field can develop one shared encoder rather than a bespoke system per application.

Why we are reading it in a session about interpretability

SUPERB is the closest thing the speech field has to a probing suite. Five of its ten tasks use linear heads, which the paper explicitly frames as a direct indication of representation quality. So the leaderboard is, in part, a claim about what information is linearly recoverable from these models — which is exactly our question.

1 Why a benchmark was needed

The state of evaluation in speech before 2021



The problem, as the paper states it

Incomparable results across incomparable setups

Different datasets	Researchers had investigated phoneme classification, speaker identification, verification, emotion, ASR, translation, spoken language understanding, voice conversion and TTS — each with its own corpus and its own splits.
Different experimental setups	Absence of a shared benchmark makes it hard to compare and draw insights across the techniques. Two papers claiming improvement may not be measuring the same thing.
Too few tasks per paper	Existing works explored a limited number of tasks, so nothing established whether a representation was generally useful or merely useful for one application.
Heavyweight downstream training	Where evaluation did happen, it often involved fine-tuning the whole model or training a large downstream system — which blurs how much credit belongs to the representation.

The fourth point is the sharpest: if the downstream model is powerful enough, it can compensate for a weak representation.

The GLUE analogy, and where it strains

Speech is not text, and a benchmark has to reflect that

What transfers

Pre-train one shared model on unlabelled data, attach a small head per task, report a table. The logic that drove NLP from ELMo and BERT to a single-backbone field.

The economic argument too: developing a deep model per use case is time- and cost-prohibitive, and a shared encoder lowers the barrier to entry.

What does not

GLUE's tasks are all classification or regression over sentences. SUPERB's span frame-level sequence labelling, utterance classification, retrieval, verification, and diarisation of overlapping speakers.

That heterogeneity means the heads cannot be uniform — and that difference will matter in section 5.

The deeper disanalogy

In text, every GLUE task is about meaning. In speech, the tasks pull in opposite directions: recognition wants the representation to discard speaker identity; verification wants it to preserve speaker identity and discard content.

So a single score across all ten tasks is asking one model to be good at mutually opposed things. That is a stronger demand than GLUE makes, and it is the tension WavLM was later built to address.

What already existed

Two prior benchmarks, and why neither was sufficient

Zero Resource Speech 2021

Evaluates representations' quality without any downstream training at all — distance-based and discrimination metrics computed directly on the features.

Clean, but it cannot tell you whether a representation is useful in a system.

NOSS

The non-semantic speech benchmark — paralinguistic and speaker tasks, with content recognition tasks deliberately excluded.

Useful for its niche, but silent on whether a representation encodes linguistic content.

SUPERB's position

Targets the direct usability of pre-trained models on popular tasks, spanning both content and non-content aspects, with a downstream training step that is real but deliberately small.

Between the two extremes, by design.

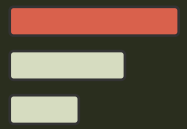
The design axis worth naming

How much downstream learning should a representation benchmark allow? Zero says none. Full fine-tuning says unlimited. Each choice measures something different, and none is obviously right.

SUPERB picks a middle point — a frozen encoder plus a small trained head — and the interesting question for us is what that specific choice makes visible and what it hides.

2 The ten tasks

Four aspects of speech, ten conventional evaluations



Ten tasks, four aspects

The taxonomy that organises the benchmark

Content

- PR — phoneme recognition
- ASR — speech recognition
- KS — keyword spotting
- QbE — query by example

Speaker

- SID — speaker identification
- ASV — speaker verification
- SD — speaker diarisation

Semantics

- IC — intent classification
- SF — slot filling

Paralinguistics

- ER — emotion recognition

Notice the shape of this table

Four tasks for content, three for speaker, two for semantics — and one for paralinguistics. The benchmark is weighted four to one toward what is said over how it is said, and the single paralinguistic task is IEMOCAP four-class emotion recognition.

Content tasks

Four tasks, two communities — transcribe, and detect

PR · phoneme recognition	Transcribe an utterance into the smallest content units. Alignment modelling is included in the task to avoid relying on potentially inaccurate forced alignment. LibriSpeech train-clean-100, dev-clean, test-clean. Metric: phone error rate.
ASR · speech recognition	Transcribe into words. Where PR analyses improvement in modelling phonetics, ASR reflects whether that improvement matters in a real scenario. Same LibriSpeech splits. Metric: word error rate.
KS · keyword spotting	Classify utterances into a predefined keyword set. Usually performed on-device, so accuracy, model size and inference time all matter. Speech Commands v1.0: ten keyword classes, a silence class, and an unknown class for false positives. Metric: accuracy.
QbE · query by example	Detect a spoken term in an audio database by deciding whether a query–document pair matches — spoken term detection without transcribing anything. English subset of QUESST 2014. Metric: maximum term weighted value, which balances misses against false alarms.

Speaker tasks

Three tasks of increasing difficulty

SID · identification	Multi-class classification of each utterance into a speaker identity, where the same speaker set appears in training and testing. VoxCeleb1. Metric: accuracy. This is the easiest of the three — the model only has to separate a closed set.
ASV · verification	Binary decision on whether two utterances share a speaker, where test speakers may never have been seen in training. VoxCeleb1 only, without VoxCeleb2 training data and without noise augmentation. Metric: equal error rate.
SD · diarisation	Predict who is speaking when, at every timestamp, with multiple speakers able to speak simultaneously. Two-speaker mixtures generated from LibriSpeech via LibriMix; time-coded labels from Kaldi alignments. Metric: diarisation error rate.

The paper's own framing

ASV is harder than SID because the speakers are open-set, and SD is harder still because the model must encode rich speaker characteristics for each frame and represent mixtures of signals. That difficulty ordering becomes an interpretive tool when we read the results.

Semantics and paralinguistics

Three tasks, and one of them should look familiar

IC · intent classification

Classify an utterance into predefined intent classes. Fluent Speech Commands, where each utterance carries three intent labels: action, object and location. Metric: accuracy.

The point is end-to-end: infer high-level semantics directly from audio rather than transcribing first and then parsing text.

SF · slot filling

Predict semantic slot types from an utterance — a type such as FromLocation for a spoken value such as Taipei. Audio SNIPS, with US-accent speakers for training and others for validation and testing.

Two metrics: slot-type F1 and slot-value character error rate. The task is reformulated as ASR with slot types encoded as special tokens.

ER · emotion recognition — the entire paralinguistics aspect

IEMOCAP, following the conventional evaluation protocol: drop the unbalanced emotion classes to leave neutral, happy, sad and angry with a similar amount of data, and cross-validate on five folds of the standard splits. Metric: accuracy.

We spent a whole session on what that protocol involves — merged excitement, discarded no-agreement turns, ten actors, Fleiss kappa of 0.27. Every paralinguistic claim made on SUPERB rests on this one column.

Three design principles

Stated explicitly, and each with a consequence

Conventional evaluation protocols

Each task uses the metric its own community already uses — PER, WER, accuracy, MTWV, EER, DER, F1.

Consequence: the numbers are interpretable to specialists, but the ten metrics have different scales, directions and meanings, so they cannot simply be averaged.

Publicly available datasets

Everything is downloadable so anyone can participate.

Consequence: the benchmark inherits the biases of the public corpora available in 2021 — which is to say LibriSpeech, VoxCeleb and IEMOCAP.

Limited labelled data

Small training sets, to benchmark generalisability rather than downstream data volume.

Consequence: this is what makes it a representation benchmark. With enough labels, a weak representation catches up.

The second principle is doing more work than it appears to. 'Publicly available' in 2021 meant read audiobooks, celebrity interview clips and acted emotional dialogues — and most of the models being evaluated were pre-trained on the first of those.

The corpora behind the ten tasks

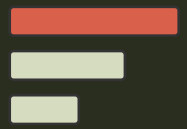
Worth listing, because the pattern matters

Task	Dataset	Split or protocol	Metric
PR	LibriSpeech	train-clean-100 / dev-clean / test-clean	PER ↓
ASR	LibriSpeech	train-clean-100 / dev-clean / test-clean, 4-gram LM	WER ↓
KS	Speech Commands v1.0	10 keywords + silence + unknown	Accuracy ↑
QbE	QUESST 2014	English subset only	MTWV ↑
SID	VoxCeleb1	Closed speaker set	Accuracy ↑
ASV	VoxCeleb1	No VoxCeleb2, no noise augmentation	EER ↓
SD	LibriMix	Two-speaker mixtures from LibriSpeech	DER ↓
IC	Fluent Speech Commands	Action, object and location labels	Accuracy ↑
SF	Audio SNIPS	US-accent speakers train, others test	F1 ↑ / CER ↓
ER	IEMOCAP	Four classes, five-fold cross-validation	Accuracy ↑

Three of the ten tasks — PR, ASR and SD — are built on LibriSpeech or on material generated from it.

3 The framework

Frozen upstream, lightweight head, weighted layers



The evaluation framework

Three constraints that define what is being measured

- Freezing is not only about fairness. The paper argues that fine-tuning requires huge resources and hinders re-usability — the whole point is a shared encoder serving many applications at once.
- The explicit constraint is that downstream models be as lightweight as possible for all tasks, because their parameter size and training cost are part of what makes the framework usable.
- A model that succeeds under all three constraints is what the paper calls a universal representation encoder.

Hold on to the middle box. It is the least discussed part of the protocol and, for our question, the most consequential.

The weighted sum across layers

One sentence in the paper; a large consequence

What the paper says

Since the last-layer representation is not always the best, the framework collects multiple hidden states from the pre-trained model and weighted-sums them as the final representation.

The weights are learned per task, jointly with the downstream head.

Why this was necessary

We saw the reason two sessions ago. Masked-prediction models behave like autoencoders: the deepest layers drift back toward the encoder output, and content and speaker information live at different depths.

Taking the last layer would systematically understate every model.

But notice what is now being benchmarked

Not a representation — a family of representations, with the benchmark selecting the best convex combination for each task separately.

A twenty-four-layer model offers a richer family than a twelve-layer one, independent of how good any individual layer is. And a model whose layers are highly specialised can score well on all ten tasks without any single layer being generally useful.

How lightweight is lightweight?

The heads are not equally small, and that matters

Task	Downstream model	Loss or scoring
PR	Frame-wise linear transformation	CTC
KS, SID, IC, ER	Mean pooling, then a linear transformation	Cross-entropy
ASR	Two-layer 1024-unit BLSTM, decoded with the official 4-gram LM	CTC on characters
SF	Same as ASR, with slot types encoded as special tokens in the transcript	CTC on characters
ASV	x-vector, with softmax replaced by AMSoftmax	Cosine-similarity backend
SD	Single-layer 512-unit LSTM, end-to-end	Permutation-invariant training
QbE	None — dynamic time warping over hidden states	Best distance function and layer reported

Five tasks get a linear probe. Two get a recurrent network plus a 4-gram language model. One gets a purpose-built speaker architecture. One gets no training at all. 'Lightweight' covers a wide range.

The tuning policy

What is allowed to vary, and what is held fixed

Held fixed

The upstream model is frozen. The downstream architecture per task is identical across all candidate models. The datasets, splits and metrics are fixed.

The constrained track of the challenge enforces all of this; an unconstrained track allowing fine-tuning was planned for later.

Allowed to vary

The learning rate is searched from 1e-1 to 1e-7 on a log scale, for every combination of representation and task.

That is seven orders of magnitude, tuned per cell of the results table. The paper says the space for downstream hyper-parameter tuning is limited, and by the standards of the field it is — but it is not nothing.

How to read a SUPERB number, precisely

It is the best result achievable by: this frozen model, with the best learned convex combination of its layers, feeding a fixed downstream architecture, at the best learning rate in a seven-decade sweep.

That is a fair and well-specified protocol. It is also several maximisations away from 'the quality of this representation', and the gap is worth keeping in view.

Fourteen models, three learning approaches

Plus filter banks as the control

Generative

APC and VQ-APC — language-model-style generation of future frames from past ones, over filter bank input.

Mockingjay and TERA — BERT-style masking of acoustic features in time, and in TERA also in frequency, then regenerating the masked parts.

NPC — masked reconstruction with convolutions for speed.

DeCoAR 2.0 — Mockingjay plus a vector-quantisation layer, larger masks, more data.

Discriminative

Modified CPC — contrastive InfoNCE prediction of future latents.

wav2vec and vq-wav2vec — CPC for ASR, then with a quantisation module producing pseudo-text for a BERT.

wav2vec 2.0 Base and Large — masking in latent space with contrastive targets, end to end.

HuBERT Base and Large — masked prediction of offline cluster labels.

Multi-task and control

PASE+ — many objectives at once: waveform generation, prosody regression, contrastive InfoMax, with reverberation and additive noise applied to the input.

FBANK — log mel filter banks, zero parameters, no pre-training. The control condition, and a surprisingly strong one.

Parameter counts span four orders of magnitude, from 1.84 million for modified CPC to 317 million for wav2vec 2.0 Large. Pre-training data ranges from 50 hours to 60,000. The table compares models that differ on almost every axis at once.

Part I in four sentences

Before the break

1. SUPERB collects ten conventional speech tasks across content, speaker, semantics and paralinguistics, using public datasets and deliberately small labelled training sets.

2. Every candidate model is frozen; only a per-task head is trained, plus a learned weighted sum over the model's hidden layers.

3. Five tasks use linear heads and are framed as a direct indication of representation quality; the other five allow recurrent networks, an x-vector, a language model, or no training at all.

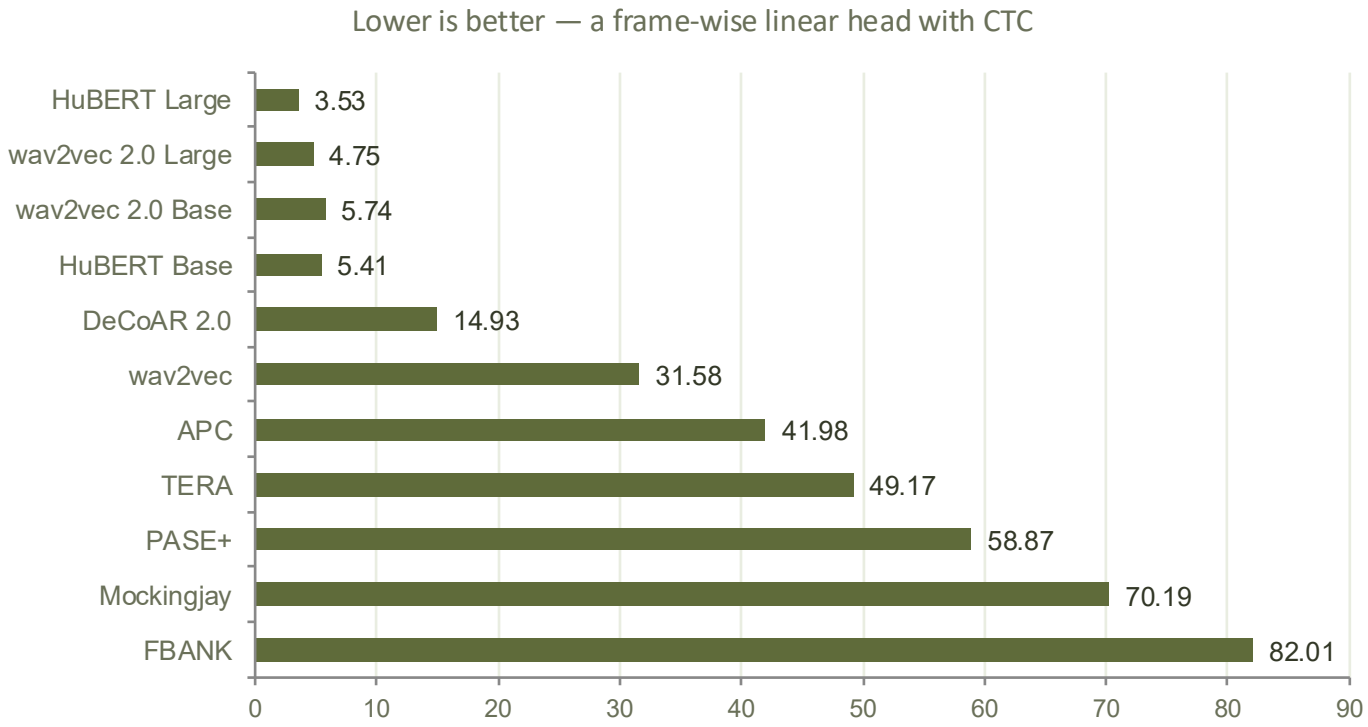
4. A reported number is therefore the best achievable under a fixed head, the best layer combination, and the best learning rate in a seven-decade sweep.

4 Results

What the table shows

Phoneme recognition

The clearest column in the table, and the widest spread

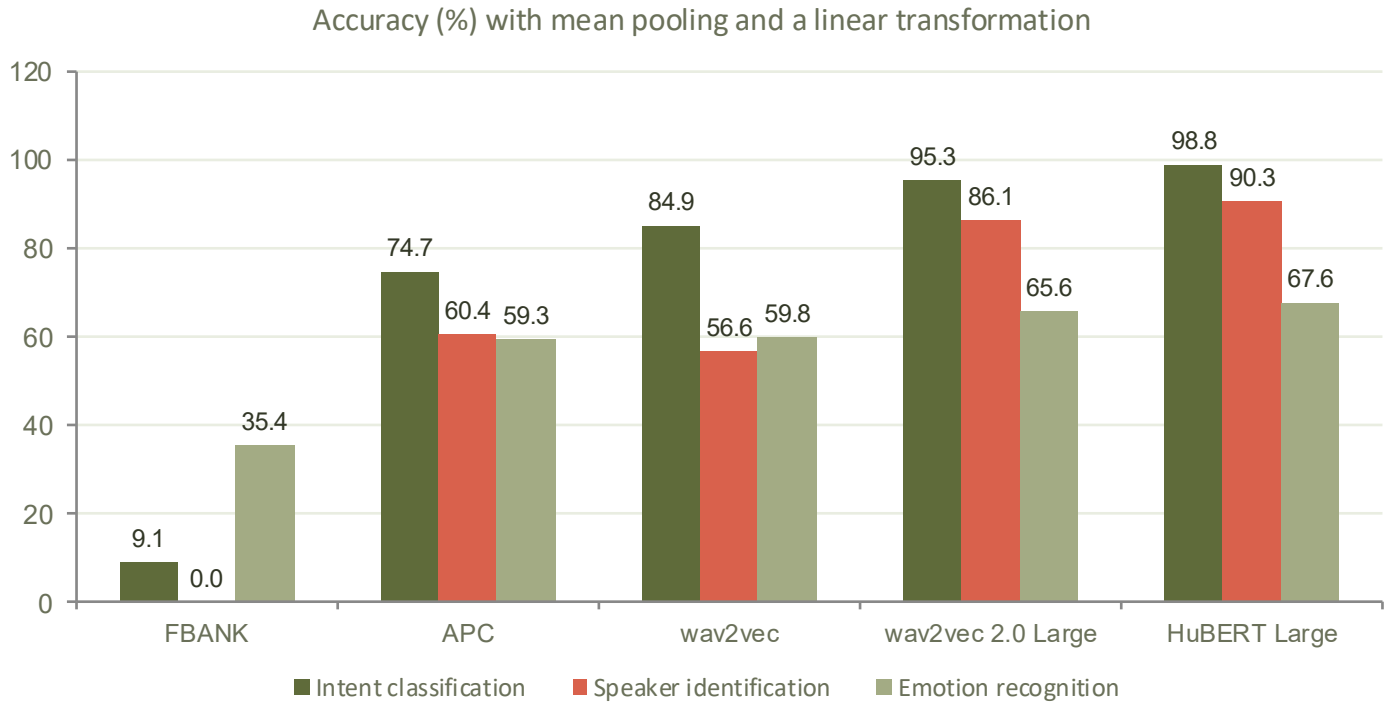


- Filter banks reach 82% phone error with a linear head — essentially unusable, as expected.
- wav2vec 2.0 and HuBERT reach single digits with nothing but a linear transformation. The paper calls this a surprise, and it is.
- That is the strongest evidence in the table for the probing claim: phonetic identity is close to linearly decodable from these representations.
- The jump from DeCoAR 2.0 at 14.93 to wav2vec 2.0 Base at 5.74 is the transition from filter-bank inputs to raw waveform with latent masking.

If you want one number for 'does this model encode phonetics', this column is it — and it is the closest thing in the suite to a clean probing result.

The five linear-probe tasks

Where the paper's probing claim actually applies

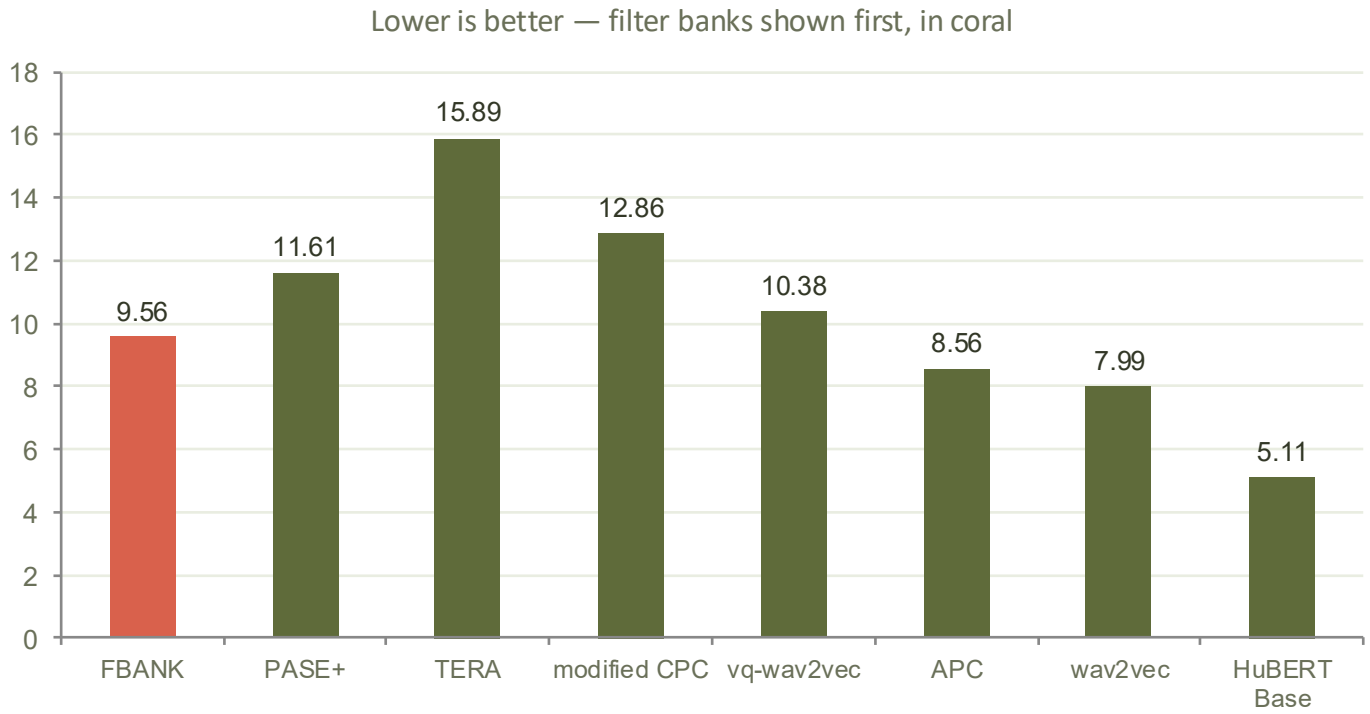


- Filter banks score essentially zero on speaker identification with a linear head — 8.5×10^{-4} accuracy on a 1,251-way problem.
- Intent classification is close to solved: 98.76% from a linear layer on frozen HuBERT features.
- Emotion recognition is the flattest column in the whole table. From filter banks to the best model is 32 points; between the four SSL models shown, about eight.
- So three different kinds of information — phonetic, speaker, semantic — are all linearly recoverable, and the paralinguistic one least so.

The flatness of the emotion column is the finding this room should take away, and we will come back to why it might be flat.

The most interesting result in the paper

Filter banks beat many self-supervised models on the harder speaker tasks



- Four of the eight self-supervised models shown are worse than filter banks at speaker verification.
- The paper states it plainly: many SSL representations are worse than FBANK when it comes to real-world speaker problems beyond SID.
- The same pattern appears in diarisation, and filter banks are also competitive on ASR and slot filling once non-linear heads are allowed.
- HuBERT Base is the exception, taking verification error from 9.56 to 5.11 with no VoxCeleb2 data and no augmentation.

A benchmark that lets the trivial baseline beat half the field is doing its job. This is the single most useful thing SUPERB contributed in 2021.

The rankings disagree with each other

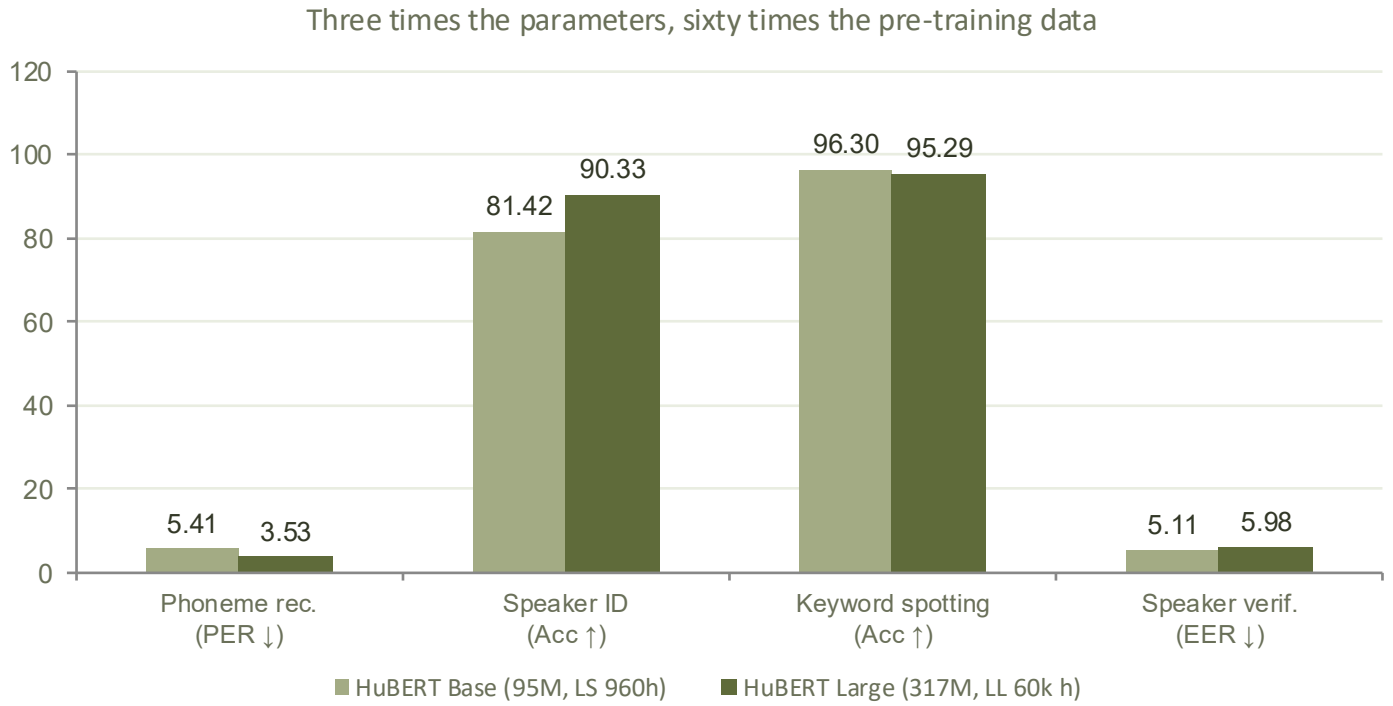
There is no single ordering of these models

Best on phoneme recognition	HuBERT Large · 3.53 PER
Best on keyword spotting	wav2vec 2.0 Large · 96.66% — HuBERT Large is 8th
Best on query by example	HuBERT Base · 0.0736 MTWV — HuBERT Large is 0.0353
Best on speaker verification	HuBERT Base · 5.11 EER — ahead of both Large models
Best on speaker identification	HuBERT Large · 90.33%
Best on diarisation	wav2vec 2.0 Large · 5.62 DER

The paper also notes that the ranking on phoneme recognition aligns only weakly with the ranking on speech recognition.

Bigger is not reliably better

HuBERT Base against HuBERT Large, on four tasks



- Scale helps enormously on phoneme recognition and speaker identification.
- It slightly hurts keyword spotting, and it clearly hurts speaker verification — 5.11 becomes 5.98.
- Query by example is the starkest reversal: 0.0736 for Base against 0.0353 for Large, a halving.
- One reading: the larger model's layers are more specialised, so a single weighted sum serves some tasks better and others worse.

Useful corrective to the scaling narrative: on a benchmark spanning several kinds of information, more parameters and more data do not move every column in the same direction.

5

What is actually being measured?

The session's question, applied to the instrument

The probing claim

What a linear head does and does not establish

What the paper asserts

That PR, KS, SID, IC and ER serve as the direct indication of representations' quality, following the conventional linear evaluation protocol.

One sentence, no further discussion.

What linear separability shows

That the information is present and encoded in a linearly accessible geometry.

It does not show that the model uses that information, that the information is represented in a form the model itself relies on, or that a non-linear probe would not find far more.

The standard objection from the probing literature

A probe's accuracy conflates how much information is present with how easy it is to extract. Two models can encode the same information and score differently because one happens to encode it more linearly.

The usual remedies — control tasks, probe-complexity comparisons, selectivity measures — are absent here, which is fair for a four-page benchmark paper but matters for anyone reading SUPERB as evidence about representational content.

The domain confound

Most of the models and several of the tasks share a corpus

Where the models were pre-trained

APC, VQ-APC, NPC, Mockingjay on LibriSpeech 360 hours. TERA, DeCoAR 2.0, wav2vec, vq-wav2vec, wav2vec 2.0 Base, HuBERT Base on LibriSpeech 960. PASE+ on LibriSpeech 50.

Modified CPC, wav2vec 2.0 Large and HuBERT Large on Libri-Light — which is also LibriVox audio.

Every model in the table is pre-trained on read audiobooks.

Where the tasks are evaluated

PR and ASR on LibriSpeech test-clean. SD on LibriMix, which is generated from LibriSpeech. SF on synthesised multi-speaker audio.

Genuinely out of domain: KS on Speech Commands, QbE on QUESST, SID and ASV on VoxCeleb, IC on Fluent Speech Commands, ER on IEMOCAP.

Why this complicates the generalisability claim

Roughly a third of the benchmark evaluates models on the same corpus they were pre-trained on. Those columns measure in-domain quality; the others measure transfer. The table does not distinguish them, and any aggregate score mixes the two.

Note where the models do worst: speaker verification and emotion — both on corpora nothing was pre-trained on. That correlation is suggestive, and SUPERB provides no way to disentangle it from the difficulty of the tasks themselves.

Ten columns, ten different lenses

Why the table resists aggregation

Different head capacity	A linear layer for phoneme recognition; a two-layer BLSTM plus a 4-gram language model for speech recognition. The second can compensate for weaknesses the first exposes.
Different metric direction and scale	Four metrics go up, four go down, one is a probability-like MTWV in the hundredths, and slot filling reports two numbers at once.
Different label reliability	Word error rate is measured against transcripts. Emotion accuracy is measured against a majority vote among three annotators who agreed at a kappa of 0.27.
Different domain relationship	Some columns are in-domain for nearly every model; others are transfer.
No overall score in this paper	The single SUPERB number that later papers quote does not appear here. It was constructed afterwards by aggregating these columns — an operation this table does not license on its own.

So: what do these representations encode?

What the benchmark can and cannot tell us

What it can answer

Is phonetic content linearly decodable from frozen features? Yes, strikingly so.

Is speaker identity present in a closed-set sense? Yes. In an open-set sense? Much less than you would assume.

Is high-level intent recoverable without transcription? Yes, almost perfectly.

Does any one model dominate across aspects? No.

What it cannot answer

Where in the model the information lives — though the learned layer weights would say, and are not reported.

Whether information absent from a linear probe is absent from the model.

Whether a low score means weak encoding, non-linear encoding, or a noisy label set.

Whether performance reflects representation quality or domain match.

The honest formulation

SUPERB measures the downstream utility of frozen representations under a fixed, well-specified protocol. That is a valuable and previously missing measurement, and it is not the same as characterising what the representations encode.

The probing and layer-analysis literature answers the second question; SUPERB answers the first. The field has often read the first as evidence for the second.

6 Appraisal and legacy

What SUPERB achieved, and what it became

Critical appraisal

What the design buys, and what it costs

Strengths

It made results comparable. Before this, no two speech SSL papers were measuring the same thing.

Ten tasks across four aspects, with conventional metrics each community already trusted.

The frozen-encoder constraint is the right one for a representation benchmark, and the weighted-layer-sum mechanism became standard practice across the field.

It revealed a real weakness — the filter-bank result on open-set speaker tasks — that directly motivated later models.

An open toolkit, s3prl, without which none of the above would have been adopted.

Deliberately limited labelled data, which is what makes it a representation benchmark at all.

Limitations

Four of ten tasks are content; one is paralinguistics, and it is IEMOCAP four-class.

Head capacity varies enormously, so the columns are not commensurable as representation measures.

Roughly a third of the benchmark is in-domain for the models being ranked.

The probing claim is asserted without control tasks, non-linear baselines or selectivity measures.

The learned layer weights — a free interpretability result — are never reported.

English only, discriminative tasks only, and the models compared differ in size, data and objective simultaneously.

What SUPERB became

A benchmark that grew into a family

SUPERB-SG	Adds the generative and semantic tasks the first release deferred — speech translation, voice conversion, speech enhancement and separation. The fifteen-task version that WavLM reports on.
ML-SUPERB	The multilingual extension, covering more than a hundred languages, and a direct answer to the English-only limitation acknowledged here.
Dynamic-SUPERB	An instruction-tuning benchmark for speech, reframing evaluation around a model that follows natural-language task descriptions rather than a fixed set of heads.
The protocol itself	The frozen-encoder, weighted-layer-sum, lightweight-head recipe became the default way to evaluate any new speech representation — including in papers that never cite SUPERB.
The leaderboard effect	Models began to be designed against these ten columns. WavLM's introduction reads as a direct response to two of them, and its denoising objective targets the speaker tasks this benchmark showed were weak.

The ER column, revisited

What this benchmark means for affective computing

- One task out of ten. Emotion recognition is the entire paralinguistics aspect of SUPERB.
- It is IEMOCAP four-class with the conventional protocol — which we established means merged excitement, discarded no-agreement turns, ten actors, and a Fleiss kappa of 0.27 among the original annotators.
- The spread between self-supervised models on this column is small: about eight points between APC and HuBERT Large, against seventy-eight points on phoneme recognition.
- Credit where due: SUPERB specifies five-fold cross-validation on standard splits, which is more than most IEMOCAP papers do, and it is why this column is comparable across papers at all.
- But a score of 67.6% on a label set that noisy is not the same kind of measurement as a 3.5% phone error rate, and averaging them into one number treats them as if it were.

A concrete opportunity

Everything we would need for a better paralinguistic benchmark exists: MSP-Podcast with defined speaker-independent splits and a non-retrieved control partition, CREMA-D with per-modality perceptual ratings, and rater distributions rather than majority votes.

What is missing is the aggregation into a protocol with the rigour SUPERB brought to the content tasks — and the willingness to report against a noise-aware ceiling rather than against 100%.

Where this sits in the series

Five sessions, one recurring problem

The corpora sessions

IEMOCAP and CREMA-D gave us labels, and showed that emotional labels are noisy, expensive and shaped by design choices most users never examine.

MSP-Podcast and AffectNet showed that harvesting moves the bias rather than removing it.

The model sessions

wav2vec 2.0, HuBERT and WavLM showed that learning speech structure from unlabelled audio is now close to a solved and downloadable problem.

And that what the objective withholds determines what the representation generalises over.

This session

The instrument that connects the two. SUPERB is how the field decides whether a representation is any good — and it inherits the label problems from the first half while measuring the models from the second.

The recurring problem, in one sentence

In every session the bottleneck turned out to be not the model but the measurement: no official splits, majority votes over disagreeing raters, selection mechanisms nobody could inspect, and benchmarks whose aggregate scores hide which columns are trustworthy.

Better representations are now cheap. Better evaluation is not, and that is where the open work is.

Take-aways

Five things to carry out of this session

1	**A benchmark is a theory of what matters**	Ten columns, four of them content and one paralinguistics. That ratio became the field's priorities.
2	**Linear decodability is a lower bound**	A low probe score means the information is not linearly accessible. It does not mean the information is absent.
3	**Check the domain relationship**	About a third of SUPERB is in-domain for the models it ranks. In-domain and transfer columns measure different things and should not be averaged.
4	**The trivial baseline is worth running**	Filter banks beat half the field on open-set speaker tasks, and that single result motivated the next generation of models.
5	**Report the layer weights**	They are computed for free in every SUPERB run, they say where each task's information lives, and almost nobody publishes them.

Questions for discussion

1. The ER column is flat across every self-supervised model. Is that because these representations encode emotion similarly, because the information is non-linearly encoded, or because IEMOCAP's label noise has already saturated the task — and what experiment would separate the three?

2. SUPERB benchmarks the best convex combination of a model's layers, not any single representation. Is a model whose layers are ten specialists a universal representation encoder, and does the title's claim survive that reading?

3. Speech tasks pull in opposite directions: recognition wants speaker identity discarded, verification wants it preserved. Is one score across ten such tasks the right target, or should we expect a family of specialised encoders?

4. Roughly a third of the benchmark is in-domain for the models being ranked. What would a version of SUPERB look like that reported in-domain and transfer columns separately?

5. If you were designing the paralinguistic half of a benchmark today, with MSP-Podcast, CREMA-D and rater distributions available: what would the tasks be, and what would you evaluate against instead of a majority vote?