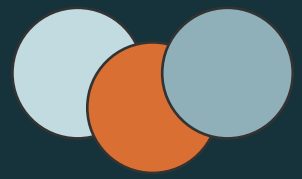


WavLM

Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu and fifteen co-authors

IEEE Journal of Selected Topics in Signal Processing (2022) · arXiv:2110.13900

Microsoft, with Shanghai Jiao Tong University and Harbin Institute of Technology

Masked prediction and full-stack pretraining

Why this paper matters

The first speech model built to be general on purpose

94k h

of pre-training audio, from three different sources

19

downstream tasks evaluated, fifteen of them from SUPERB

14 / 15

SUPERB subtasks where WavLM Large beats HuBERT Large

0.383%

equal error rate on VoxCeleb1-O speaker verification

The claim

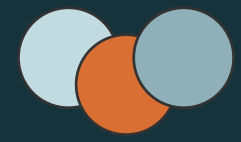
Keep HuBERT's masked prediction of cluster targets, but feed the model artificially noisy and overlapped speech while still asking it to predict the labels of the clean original. The model learns to identify and follow the main speaker — which is what speaker, diarisation and separation tasks need — without giving up any content modelling.

The result that makes the point

WavLM Base+ — a model three times smaller than HuBERT Large, with the same 95 million parameters as HuBERT Base — beats HuBERT Large on the SUPERB overall score. The gain comes from the objective and the data, not from scale.

1 The problem

What ASR-focused pre-training leaves on the table



Two concrete drawbacks

Both stated plainly in the paper's introduction

Multi-speaker tasks barely improve

Speech separation models built on top of HuBERT achieve only marginal improvement over models trained from scratch.

Two reasons the paper gives: the pre-training objective does not sufficiently enforce speaker discrimination, and the training data contains only single-speaker audio.

A model that has never heard two people at once has no reason to learn to separate them.

The audiobook problem

Libri-Light is the standard pre-training source, and its acoustic characteristics do not match real deployment conditions.

The paper notes that even the robust wav2vec 2.0 work, which trained on larger and more diverse data, still had over 90% of its audio derived from audiobooks.

So 'more data' had not yet meant 'more varied data'.

The tension the paper names

Different tasks want opposite things. Speaker verification requires the network to learn speaker characteristics regardless of content; speech recognition requires it to discard speaker characteristics and attend only to content.

That is the same tension we found in the affective-computing results last session, where ASR fine-tuning measurably damaged emotional information. WavLM is an attempt to resolve it in one model.

SUPERB, and the insight it produced

The benchmark that made 'full stack' a measurable claim

The setup

Fifteen tasks grouped into five aspects of speech: content, speaker, semantics, paralinguistics and generation.

The pre-trained model is frozen. Only a small task-specific head is trained, using a fixed downstream architecture per task, so the comparison is between representations rather than between systems.

The insight

The downstream head consumes a learned weighted sum of the hidden states from every layer, not the final layer.

Different layers carry information useful for different tasks: top layers for recognition, bottom layers for speaker verification.

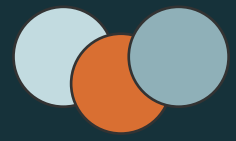
Why this matters for reading everything that follows

Every SUPERB number in this paper is a frozen-model, weighted-layer-sum number. The paper is explicit that this understates what the models can do, which is why it also runs four tasks with full fine-tuning on their own benchmarks.

This is the same practice we identified last session as essential for emotion work — and it is now the standard evaluation protocol for the whole field.

2 Three modifications

Denoising, gated relative position bias, and 94k hours



What WavLM keeps from HuBERT

Almost everything

The architecture	Seven convolutional blocks with the same strides and kernel widths, giving 25 ms of audio per frame at a 20 ms stride, then a Transformer encoder.
The objective	Masked prediction of discrete pseudo-labels, cross-entropy over masked regions only, with cosine similarity to codeword embeddings scaled by a temperature of 0.1.
The targets	K-means cluster centres on MFCCs or on latent representations, produced by the same iterative re-clustering process.
Convolutional position embedding	The kernel-128, 16-group convolution at the bottom of the Transformer stays — the gated bias is added on top of it, not instead of it.
The fine-tuning recipe	CTC with 29 tokens, frozen convolutional encoder, Transformer frozen for the first 10k steps, tri-stage learning rate, SpecAugment-style masking.

So the paper is a controlled experiment on top of HuBERT: three changes, each ablated.

Modification 1: masked speech denoising and prediction

The one that matters most

Input to the model

primary

second speaker or noise mixed in

masked

Target: the pseudo-labels of the CLEAN original, at the masked positions only

What this asks the model to do

Identify which of the overlapping voices is the main speaker, follow that speaker through the interference, and predict what they were saying in the parts it cannot see.

Speaker discrimination, denoising and content modelling, all enforced by one loss.

Pseudo-labels are always generated by feeding the clean utterance to the previous-iteration network, so the target is never contaminated.

Why the overlap is capped below 50%

The mixing length is drawn uniformly from up to half the utterance, so the primary speaker is always present for longer than the secondary one.

Without that constraint, 'the main speaker' would be ambiguous and the target would be arbitrary. The constraint is what makes the task well posed.

How the mixtures are made

The simulation runs inside the training loop, on each batch

1 Select	Choose a subset of utterances in the batch by Bernoulli sampling with probability p. In the released models, denoising is applied to 20% of utterances.
2 Choose the interferer	With probability p_n mix in a noise clip from the DNS challenge noise set; otherwise mix in another utterance drawn from the same batch. p_n is 0 for Base and 10% for Base+ and Large.
3 Set the energy ratio	Sampled uniformly from −5 to 5 dB for a secondary utterance, and from −5 to 20 dB for noise — following the standard simulation practice in speech separation.
4 Choose the region	The mixing length is sampled uniformly from one frame up to half the utterance; start positions for both signals are sampled uniformly and independently.
5 Mix	Scale the secondary signal by the computed ratio and add it to the selected region of the primary. The speaker of the first utterance is the main speaker.

Note that interferers come from the same batch — no separate corpus of mixtures is needed, and the simulation costs almost nothing.

Modification 2: gated relative position bias

A small change aimed squarely at recognition

The mechanism

A learnable scalar bias is added to the attention logit for each query–key offset, on top of the usual scaled dot product.

Two sigmoid gates, computed from the current query, then modulate that bias — an update gate and a reset gate, in the style of a gated recurrent unit.

Bucketed offsets: 320 embeddings, spacing growing logarithmically up to a maximum offset of 800, shared across all layers.

The intuition the paper gives

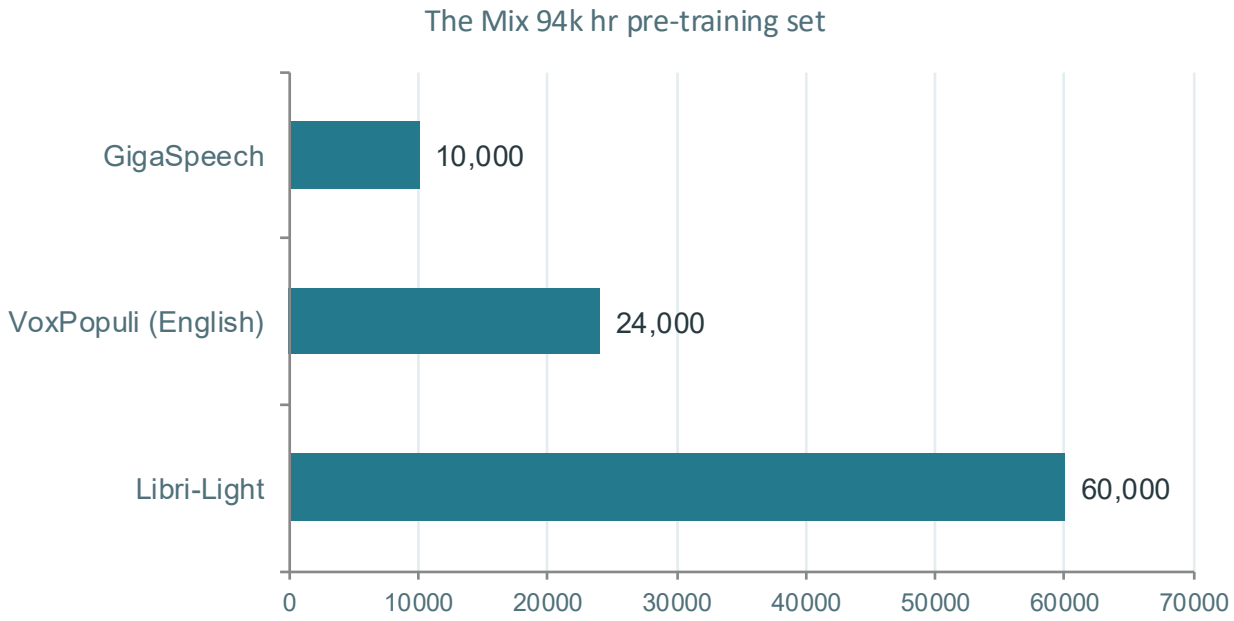
The same distance between two frames means different things depending on what those frames contain. A gap of forty frames matters differently if one end is silence than if both ends are speech.

Gating lets the position bias be adjusted by conditioning on the current speech content, rather than being a fixed function of distance alone.

- Because the position embeddings are shared across layers, the parameter count barely moves — 94.70M for WavLM Base against 94.68M for HuBERT Base.
- The ablation is specific: removing it costs mainly phoneme recognition and speech recognition, the content-related tasks. Speaker tasks barely notice.
- So the two modifications are complementary by design: denoising targets speaker and multi-speaker tasks, gating targets content tasks.
- A separate practical fix accompanies it: rescaling the attention logits before the softmax to prevent 16-bit overflow, which was destabilising large-model training.

Modification 3: 94,000 hours from three sources

Deliberately not all audiobooks



- GigaSpeech contributes audiobooks, podcasts and YouTube — read and spontaneous styles across many topics.
- VoxPopuli contributes European Parliament recordings from 2009 to 2020: plenary sessions, committee meetings and other events.
- Only 10k of GigaSpeech's 40k hours are used; the authors say the other 30k are poorly processed, with long silences and some utterances containing only background noise.
- So a substantial share is still audiobook. Libri-Light alone is 64% of the total.

Where the extra data shows up

The paper is specific: the combined dataset especially boosts test sets that are not drawn from audiobooks — speaker verification, out-of-domain ASR, intent classification, slot filling and emotion recognition. Domain match, not volume, is doing the work.

Model configurations

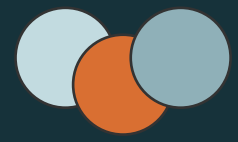
And one detail that is easy to miss

	WavLM Base	WavLM Base+	WavLM Large
Transformer layers	12	12	24
Hidden dimension	768	768	1,024
Attention heads	8	8	12
Parameters	94.70M	94.70M	316.62M
Pre-training data	LibriSpeech 960h	Mix 94k hr	Mix 94k hr
Update steps	400k	1.2M	700k
Peak learning rate	5×10^{-4}	5×10^{-4}	1.5×10^{-3}
Warm-up steps	32k	96k	32k
Batch size (audio)	350 s	350 s	720 s
GPUs	32 V100	32 V100	64 V100
Denoising applied to	20% of utterances	20% of utterances	20% of utterances
Noise-mixing probability	0	10%	10%

WavLM is trained on labels from HuBERT: Base uses the 6th layer of iteration-1 HuBERT Base; Base+ and Large use the 9th layer of iteration-2 HuBERT Base.

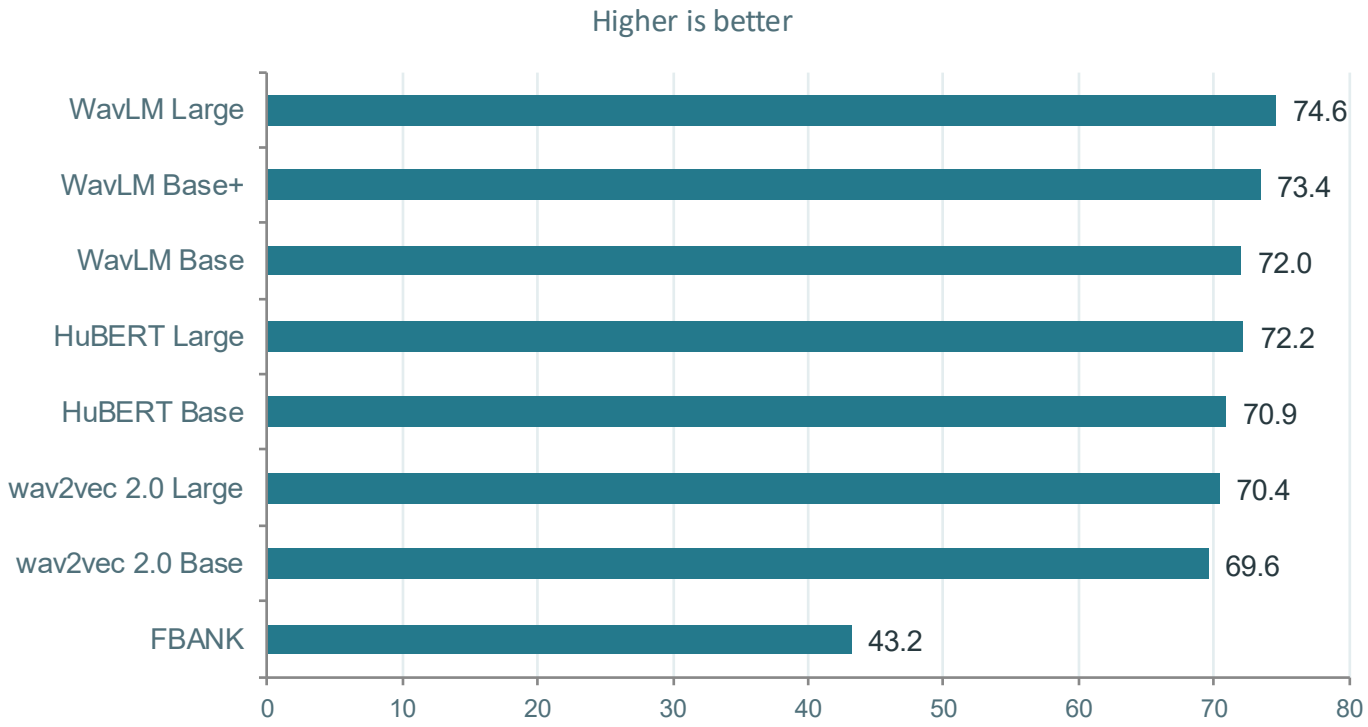
3 Results

Fifteen frozen tasks, and four with full fine-tuning



SUPERB overall score

Frozen models, weighted sum across layers, fifteen tasks

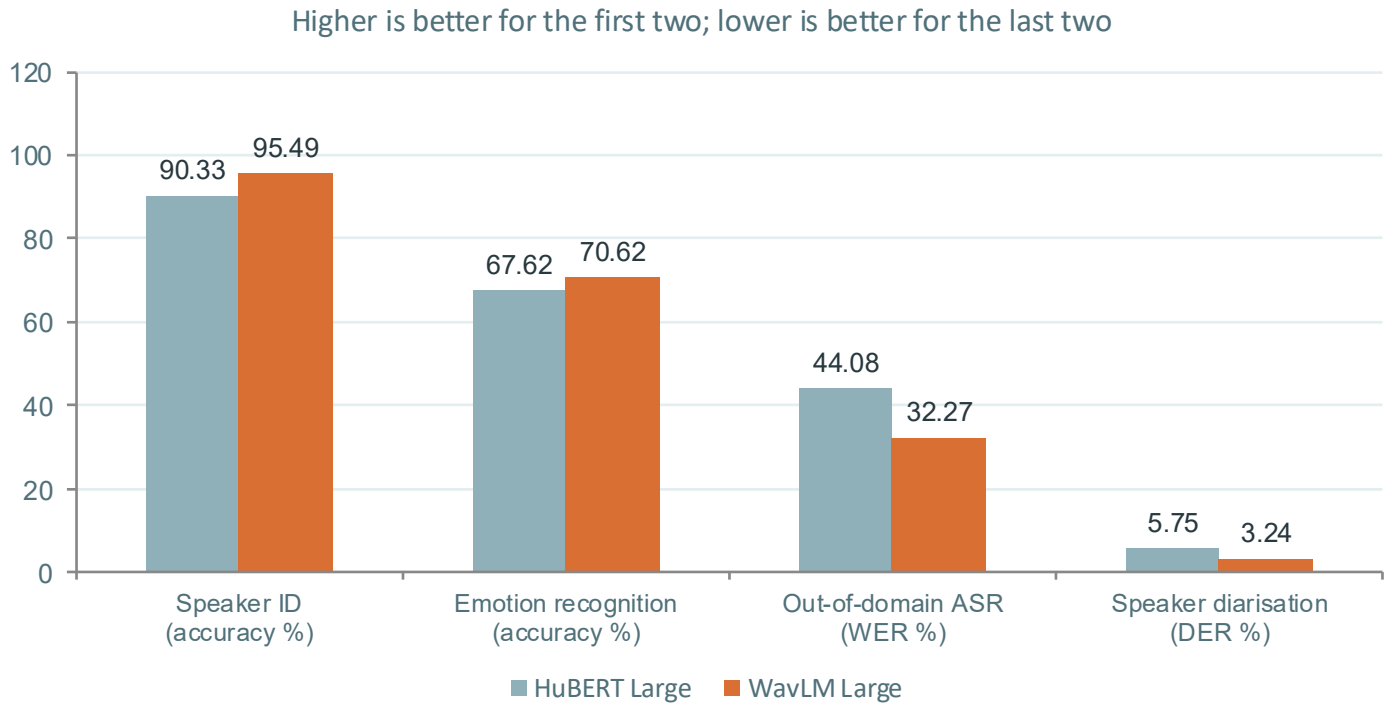


- WavLM Base beats both wav2vec 2.0 Base and HuBERT Base on every one of the fifteen subtasks — a fair comparison, since all three share data and parameter count.
- WavLM Base+ at 95M parameters beats HuBERT Large at 317M.
- WavLM Large improves on HuBERT Large by 2.4 points overall and is better on 14 of the 15 subtasks.
- The FBANK row is the reference point: hand-crafted filter banks with the same downstream heads score 43.2.

The overall score is computed by the authors, not by SUPERB — error rates inverted, the keyword-detection score rescaled, then averaged. Reasonable, but it is their aggregation.

Where the gains actually are

HuBERT Large against WavLM Large on selected subtasks

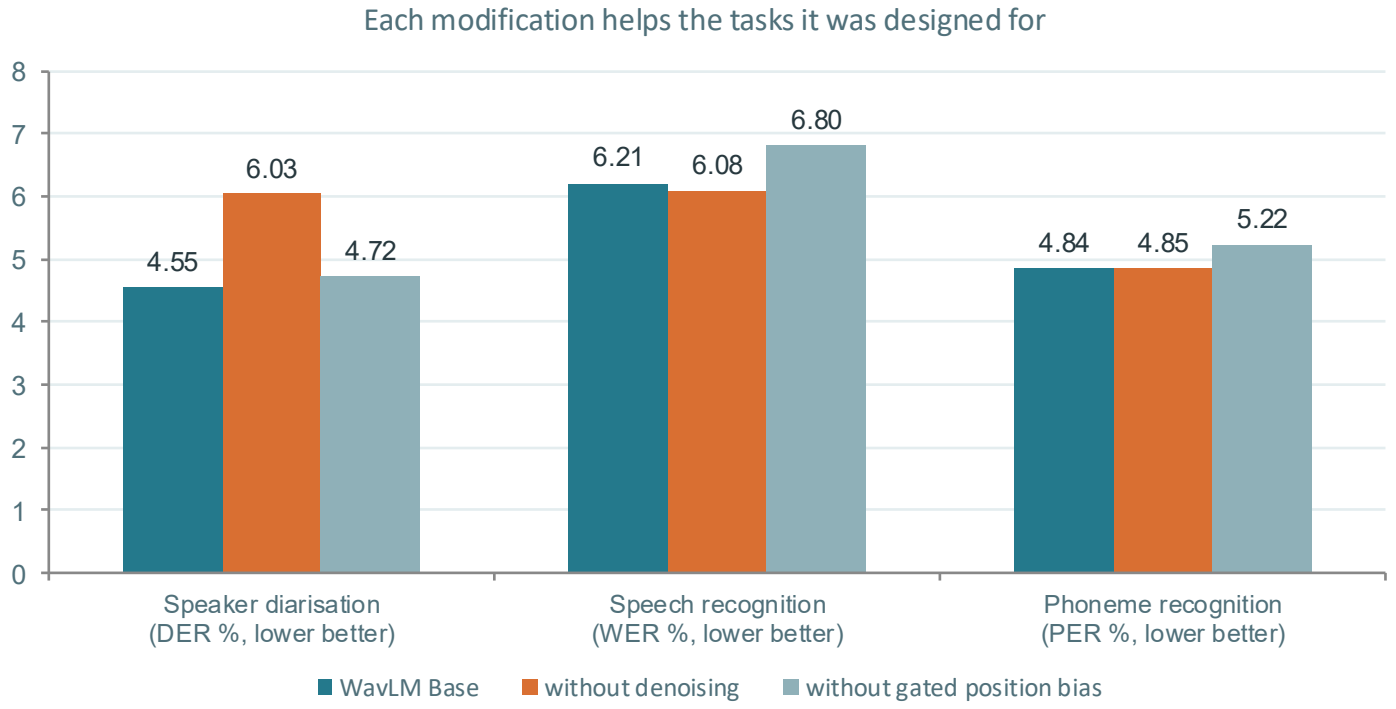


- Speaker identification rises by more than five points; diarisation error nearly halves.
- Out-of-domain ASR falls from 44.1 to 32.3 — the single largest relative gain on the benchmark, and it is a domain-shift result, not a modelling one.
- Emotion recognition rises from 67.6 to 70.6 accuracy. For reference, filter banks score 35.4 and wav2vec 2.0 Base scores 63.4.
- In-domain ASR barely moves — 3.62 to 3.44 — which is exactly what you would expect if the modifications target non-content tasks.

The emotion result is the one for us. A model whose only changes were mixing in interfering speakers and diversifying the pre-training audio gains three points on emotion recognition, with no emotional supervision anywhere.

The ablations

WavLM Base, with each modification removed in turn



- Remove denoising and diarisation error rises from 4.55 to 6.03 — a third worse — while recognition is unaffected and even fractionally better.
- Remove the gated position bias and recognition and phoneme error both worsen, while diarisation barely changes.
- This is a clean double dissociation: each change moves the tasks it targets and leaves the others alone.
- It also means the small ASR cost of denoising is real, and is being paid deliberately.

Ablations like this are what make a follow-up paper worth reading. Two modifications, two distinct effects, each measured against the model that lacks it.

Speaker verification

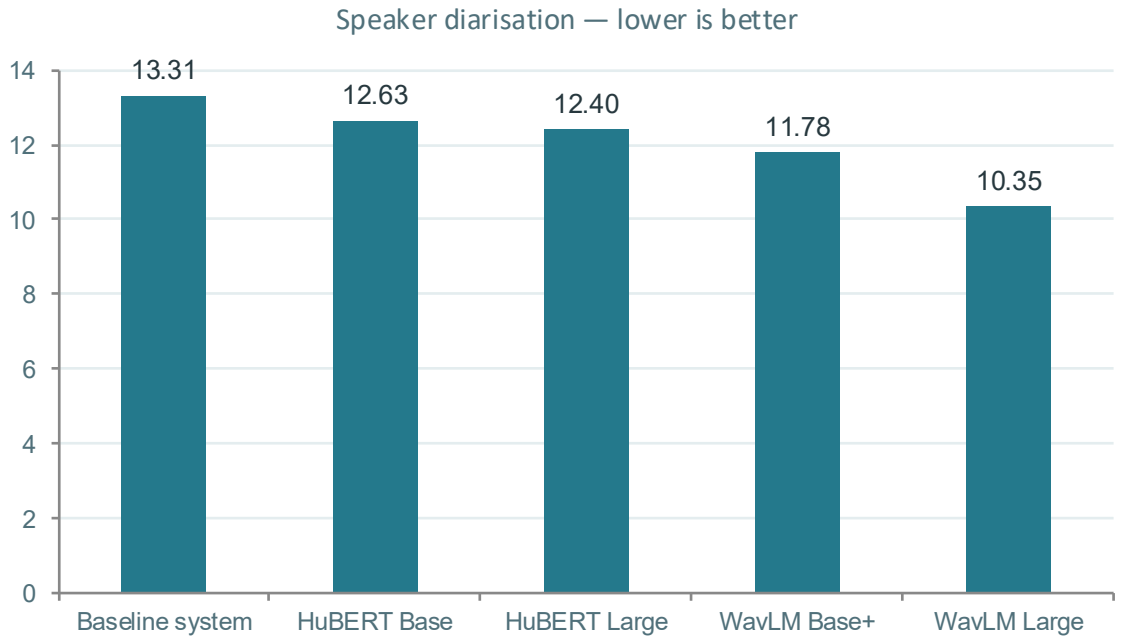
Equal error rate on VoxCeleb1, with a fine-tuned downstream model

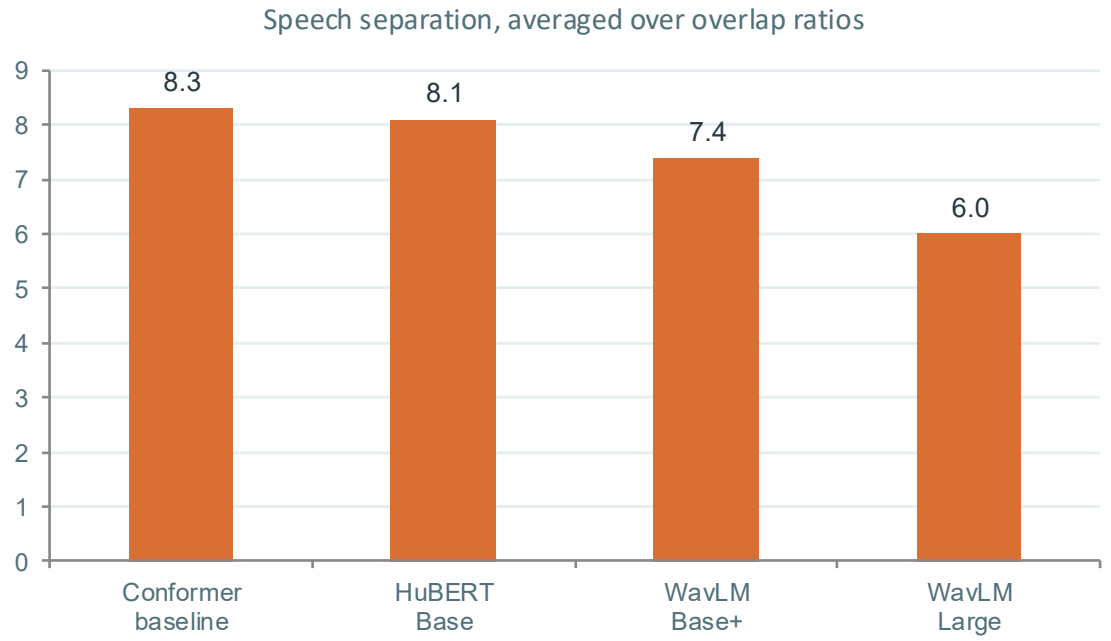
- The downstream model is held constant — a small ECAPA-TDNN — so only the input representation changes.
- HuBERT Base is roughly level with hand-crafted filter banks. Self-supervised features are not automatically good speaker features.
- WavLM Large more than halves the error against the purpose-built baseline, and the final system beats the winner of the 2021 VoxSRC challenge on all three trial lists.
- The last bar adds large-margin fine-tuning and quality-aware score calibration — engineering on top of the representation, not part of it.

This is the clearest evidence for the paper's thesis: a general self-supervised model beating a specialised architecture at the task that architecture exists for.

The multi-speaker tasks

Where the denoising objective was supposed to pay off





- HuBERT features give a separation result comparable to the baseline — the marginal improvement the paper predicted in its introduction, now measured.
- WavLM Large cuts separation word error by 27.7% relative on average, and by 32.5% on the hardest 40% overlap condition.
- On diarisation, WavLM Large sets a new best result on CALLHOME, and WavLM Base+ again beats HuBERT Large.
- A counter-intuitive detail: for separation, freezing the pre-trained model works better than fine-tuning it — the fine-tuned model overfits simulated mixtures and is tested on real meeting recordings.

And speech recognition

The honest result

With all 960 labelled hours

WavLM Large reaches 1.8 on test-clean and 3.2 on test-other.

wav2vec 2.0 Large and HuBERT Large are at 1.8/3.3 and 1.9/3.3. Essentially a tie.

Without a language model

This is where WavLM Base clearly separates from wav2vec 2.0: 9.8 against 11.1 on test-clean with 10 hours of labels, and 5.7 against 6.1 with 100 hours.

A strong language model masks acoustic differences.

Out of domain, the gap is large

On SUPERB's out-of-domain ASR task, WavLM Large reaches 32.3 word error against HuBERT Large's 44.1 — a 27% relative reduction.

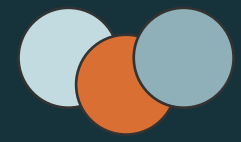
How to read this

On the benchmark the field had been optimising for four years — LibriSpeech with a Transformer language model — three quite different pre-training objectives all land within a tenth of each other. That benchmark has stopped discriminating between methods.

Where the models still differ is everywhere else: no language model, out of domain, multi-speaker, speaker identity, paralinguistics. Which is precisely the argument for evaluating on nineteen tasks instead of one.

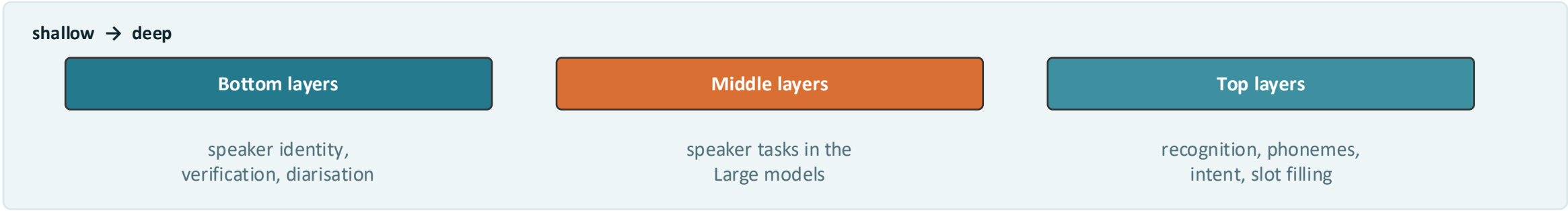
4 Layer analysis

Where in the network each task lives



What the learned layer weights show

The weights the downstream heads assign to each layer, normalised

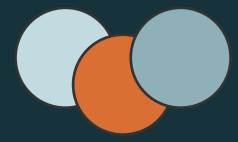


- The pattern is essentially the same for HuBERT and WavLM, which tells you it is a property of masked-prediction pre-training rather than of these particular modifications.
- In the Base models, speaker information concentrates in the bottom layers; in the Large models it shifts to the middle.
- Content and semantic information sits at the top in both.
- The paper's own conclusion: how to leverage the hidden states of intermediate layers is the key to success on speaker-related tasks.
- For speaker verification, diarisation and separation specifically, almost all the weight comes from the bottom layers.

Which is why every result in this paper uses a weighted sum across layers — and why, for any paralinguistic task, you should too.

5 Appraisal and synthesis

The two papers together



Critical appraisal

What the design buys, and what it costs

Strengths

Nineteen tasks, fifteen of them on a standard frozen-model benchmark. The breadth is the contribution and it is done properly.

Clean ablations showing a double dissociation between the two modifications.

The denoising simulation is cheap: interferers come from the batch, and only 20% of utterances are corrupted.

A 95M-parameter model beating a 317M one is strong evidence that the objective, not scale, is responsible.

Honest reporting of the small ASR cost, and of the odd finding that HuBERT's pseudo-labels work better than WavLM's own.

Code and checkpoints released.

Limitations

Still two-thirds audiobook, in a paper whose motivation is escaping audiobook bias.

English only, despite VoxPopuli being a 23-language corpus from which only the English portion is used.

The mixing is simulated: additive mixtures at sampled energy ratios, not real overlapped conversation with its own room acoustics and turn-taking.

The overall SUPERB score is the authors' own aggregation of fifteen heterogeneous metrics.

Depends on HuBERT for its targets, so the clustering bootstrap problem is inherited rather than solved.

No emotion-specific or health-related evaluation beyond SUPERB's single four-class emotion task.

The two papers side by side

One framework, two questions

	HuBERT (2021)	WavLM (2022)
The question asked	Can masked prediction work without a real lexicon?	Can one model serve every speech task?
Core mechanism	Offline k-means targets, loss on masked frames only	Corrupted input, clean targets, gated position bias
Input during pre-training	Clean single-speaker audio	20% corrupted with a second speaker or noise
Pre-training data	LibriSpeech 960h or Libri-Light 60k h	Mix 94k hr from three sources
Targets	Bootstrapped from MFCC k-means, refined twice	Taken from released HuBERT models
Evaluation	Speech recognition only	Nineteen tasks across five aspects of speech
Best ASR result	1.8 / 2.9 with a 1B-parameter model	1.8 / 3.2 with 317M
Key ablation	α: masked-only prediction versus all frames	Each modification removed in turn
What it left open	Anything beyond ASR; single-phase training	Multilingual, real overlap, non-audiobook data

WavLM changes what goes in and how position is encoded. Everything else is HuBERT.

Take-aways

Four things to carry out of this session

1 Corrupt the input, keep the target clean

That single asymmetry turns a content model into a speaker model without a second loss term, a second head, or a second training stage.

2 Design the objective for the task you want

Denoising helps speaker tasks; gated position bias helps content tasks. The ablation shows each moving only what it targets.

3 A saturated benchmark hides progress

Three quite different objectives land within a tenth of each other on LibriSpeech, and differ by twelve points on out-of-domain ASR.

4 Use the whole layer stack

Both papers, and the interpretability work behind them, agree: speaker information is low, content is high, and the last layer is rarely the one you want.

Questions for discussion

1. WavLM's emotion gain is a side effect of an objective designed for speaker tasks. What would a pre-training corruption designed specifically for paralinguistics look like — and what would you corrupt?

2. Adding denoising costs a little ASR accuracy and buys a lot of speaker capability. Is a single general model the right goal, or should we accept a family of objective-specialised backbones?

3. The paper reports that HuBERT's pseudo-labels produce a better second-iteration WavLM than WavLM's own. What would explain that, and what does it say about the iterative refinement story?

4. Every emotion claim in this paper rests on SUPERB's IEMOCAP four-class task — the benchmark we spent a session criticising. What would a trustworthy paralinguistic entry in a benchmark like SUPERB require?

5. Three objectives now tie on LibriSpeech. What should replace it as the benchmark that tells us whether speech representation learning is still making progress?

Chen et al. (2022), IEEE JSTSP · arXiv:2110.13900 · Hsu et al. (2021), IEEE/ACM TASLP 29, 3451–3460