WORKSHOP ON WORLD MODELS · UNIVERSITY OF CHICAGO BOOTH · 31 AUG TO 2 SEP 2026

S-JEPA

Soft Clustering Anchors for Self-Supervised Speech Representation Learning

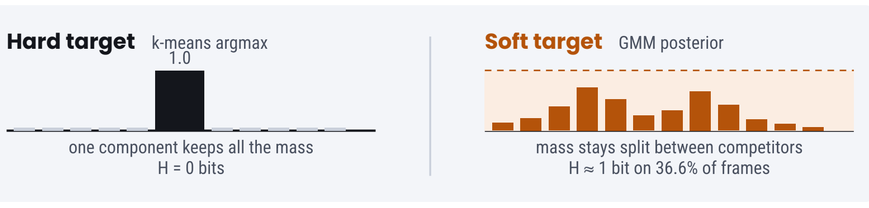
Georgios Ioannides[1,3,7] · Adrian Kieback[3,8,*] · Judah Goldfeder[4,*] · Linsey Pang[5] · Aman Chadha[3,6] · Aaron Elkins[3] · Yann LeCun[2] · Ravid Shwartz-Ziv[2]

[1]Carnegie Mellon University · [2]New York University · [3]James Silberrad Brown Center for AI · [4]Columbia University · [5]Northeastern University · [6]Stanford University · [7]Amazon GenAI[†] · [8]Discern Science International[†]

* Equal contribution · † Work does not relate to position at Amazon or Discern · gioannid@alumni.cmu.edu

James Silberrad Brown Center for Artificial Intelligence

CODE ON GITHUB



CONTRIBUTION

HuBERT and WavLM supervise the encoder against a **one-hot cluster index**. S-JEPA supervises against the **full GMM posterior**, matched by a single KL divergence at masked positions.

On held-out speech, **36.6% of frames** carry more than 1 bit of predictor entropy. An argmax at those frames selects one component and discards the rest of the mass.

S-JEPA recasts HuBERT-style clustering as a continuous soft-target JEPA objective, which retains boundary uncertainty and removes the offline re-clustering step.

At **51.8M parameters** S-JEPA sets the sub-90M WER frontier and matches HuBERT-Base on emotion recognition at just 55% of its parameter count, without teacher distillation

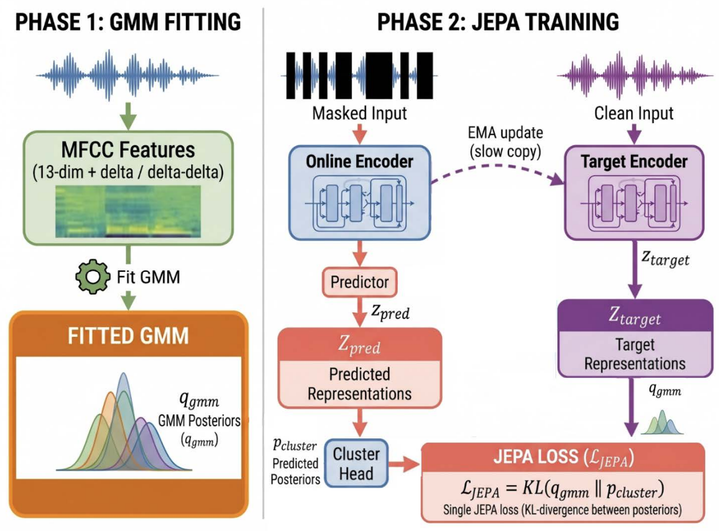
12.10%	64.83%	51.8M	1 pass
WER on LibriSpeech test-clean. Lowest among sub-90M SSL encoders.	Emotion recognition accuracy. Matches HuBERT-Base at 55% of its size.	Encoder parameters, 6 Transformer layers.	No offline re-clustering, no pre-trained teacher.

01 Motivation

HuBERT-style pre-training clusters acoustic features offline, then trains the encoder to predict cluster IDs at masked positions under cross-entropy. The approach works well and carries two costs.

- **Hard assignment discards ambiguity.** Frames at phone boundaries, transitions and silence shifts carry genuine acoustic ambiguity, and one ID cannot represent it.
- **Training stops and restarts.** Each iteration refits the centers over the whole corpus, then relabels every frame.

02 Method



Phase 1 fits a GMM over MFCC features and freezes it. **Phase 2** updates a GMM online over EMA encoder features. The predictor, cluster head and GMM are discarded after pre-training, and only the encoder is retained for downstream use.

A K-component diagonal-covariance GMM supplies the posterior q_t at each masked frame. Training minimizes one loss with no auxiliary terms.

$$L = (1/|S|)\, \Sigma_{t \in S}\, \mathrm{KL}(\, q_t \,\|\, p_t\,)$$

Both arguments are distributions over the same K components. S is the masked positions, or masked plus visible early in training.

03 Design choices

Online GMM updates

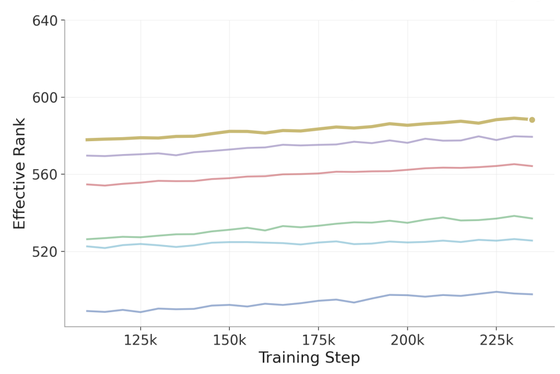
After each minibatch we EMA-update the GMM from responsibility-weighted statistics of the EMA encoder features. No gradient flows through it, so the mixture tracks the encoder as it evolves rather than freezing at a re-cluster boundary.

Offline re-cluster (HuBERT, WavLM) plus a full corpus relabel pass	$O(N_s D + I N_s K D)$
Online update (S-JEPA, per step) reuses the training forward pass	$O(BKD)$

Adaptive layer selection by effective rank

A label-free signal selects the layer to cluster on. We track the effective rank of each layer's embeddings and anchor the GMM at the current argmax.

$$\mathrm{erank}(Z) = \exp(\, -\Sigma_i\, p_i \log\, p_i\,), \quad p_i = \sigma_i \,/\, \Sigma_j\, \sigma_j$$



Phase 2. Layer 4 dominates from the first measured checkpoint and its gap over layer 2 widens. The active layer moved from ℓ^* = 2 to ℓ^* = 4 midway through Phase 2 and remained there. We report this as a correlation with continued WER progress; the experiment does not establish causation.

Periodically switched EMA decay

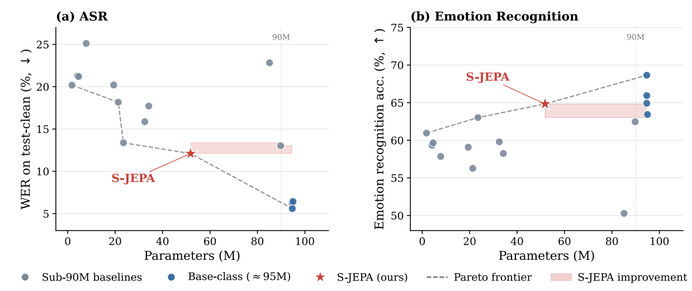
A fixed slow decay keeps recent encoder improvements out of the GMM, and a fixed fast decay moves the target at every step. Alternating between the two rates separates the requirements in time.

$$a_t \in \{\, a_{\text{fast}},\, a_{\text{slow}} \,\} = \{\, 0.999,\, 0.9999 \,\}$$

Flipped on a fixed cadence, averaging every 20,000 steps.

04 SUPERB benchmark results

We evaluate three of the SUPERB tasks rather than the full suite. ASR, emotion recognition and slot filling cover the content, paralinguistic and semantic categories of the benchmark, the coverage required to test whether a soft target helps beyond phonetic content. The compute budget that limits this study to one model at one scale also excluded the remaining tasks.



WER and emotion recognition accuracy against encoder size. S-JEPA (star) reaches the lowest WER below 90M parameters and matches HuBERT-Base emotion accuracy at 55% of its parameter count. The frontier is over parameter count only: pre-training data is not matched, since HuBERT, WavLM Base, wav2vec 2.0 and DeCoAR 2.0 use LibriSpeech's 960 hours, WavLM Base+ roughly 94,000, and S-JEPA roughly 83,000.

Full leaderboard figures behind the plot above. Encoder frozen, small task heads trained on top, greedy CTC decoding for ASR. Baselines from the public SUPERB leaderboard, parameter counts inference-time encoder only. With 4-gram LM rescoring S-JEPA reaches 8.50 WER and 2.90 CER; the greedy figure is quoted for comparability.

Method	Params	WER ↓	ER acc ↑	SF F1 ↑	SF CER ↓
SUB-90M					
modified CPC	1.8M	20.18	60.96	71.19	49.91
APC	4.1M	21.28	59.33	70.46	50.89
VQ-APC	4.6M	21.20	59.66	68.53	52.91
PASE+	7.8M	25.11	57.86	62.14	60.17
NPC	19.4M	20.20	59.08	72.79	48.44
TERA	21.3M	18.17	56.27	67.50	54.17
DistilHuBERT	23.5M	13.37	63.02	82.57	35.59
wav2vec	32.5M	15.86	59.79	76.37	43.71
vq-wav2vec	34.2M	17.71	58.24	77.68	41.54
Mockingjay	85.1M	22.82	50.28	61.59	58.89
DeCoAR 2.0	89.8M	13.02	62.47	83.28	34.73
S-JEPA (ours)	**51.8M**	**12.10**	**64.83**	**83.05**	**33.17**
BASE-CLASS (≈95M)					
HuBERT Base	94.7M	6.42	64.92	88.53	25.20
WavLM Base	94.7M	6.21	65.94	89.38	22.86
WavLM Base+	94.7M	5.59	68.65	90.58	21.20
wav2vec 2.0 Base	95.0M	6.43	63.43	88.30	24.77
DinoSR	94.7M	4.71	–	88.83	23.57

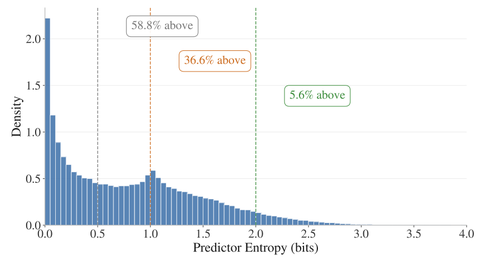
05 Representation probing

Accuracy on LibriSpeech, mean ± s.d. across 3 splits. Chapter ID is the hardest of the five, since chapters within a book share a speaker and the available cue is prosodic and recording variation.

Task	WavLM	HuBERT	S-JEPA
Speaker ID	91.1 ±1.0	99.1 ±0.3	**99.7** ±0.3
Gender	96.3 ±0.5	99.5 ±0.3	**99.6** ±0.3
Chapter ID	59.5 ±1.0	88.2 ±1.0	**93.2** ±1.0
Phoneme (linear)	**86.1** ±0.2	84.8 ±0.1	82.8 ±0.1
Phoneme (MLP)	87.5 ±0.1	86.5 ±0.1	**88.0** ±0.1

A linear probe favors WavLM on phonemes, but an MLP probe reverses the ranking. The gain from the MLP is far larger for S-JEPA (+5.2) than for WavLM (+1.4) or HuBERT (+1.7). S-JEPA therefore appears to encode phonetic information in a form that a nonlinear probe recovers more readily.

06 Structure of predictor uncertainty



Per-frame predictor entropy over ≈153K held-out frames, 500 dev-clean utterances, masking disabled.

The distribution is bimodal. A confident regime sits below 0.3 bits, and a second mode sits near 1 bit, the entropy of an exact two-way tie. Diffuse softness would produce a unimodal slide from low to high entropy, and softness degenerated to the uniform prior would concentrate near $\log_2 K = 9$ bits. The observed shape identifies a specific competing pair of components at each uncertain frame.

07 Scope and limitations

- Parameter count is the only quantity held fixed. Pre-training data and compute differ across the table, so this is not a data-matched or compute-matched claim.
- The online GMM, adaptive layer selection and switched EMA decay are evaluated **as a package**, and are not ablated individually.
- Results cover one model at one scale on three SUPERB tasks. Scaling into the Base regime remains untested, and all training is in English.

CLOSEST PRIOR WORK

DinoSR (Liu et al. 2023) also removes the offline re-cluster step, but assigns each frame to one centroid and trains by cross-entropy against a one-hot index, so the inter-cluster mass is still discarded. It clusters the top eight layers at once with a codebook each; we anchor a single GMM at the effective-rank argmax. data2vec shares the single-pass EMA structure but regresses continuous targets.

SETUP

Pre-trained on roughly 83,000 hours combining LibriLight and an English subset of Granary. Encoder is 6 Transformer layers, 768-dim, 12 heads. K=100 in Phase 1 over 39-dim MFCC, K=500 in Phase 2 over EMA encoder features. AdamW, global batch 192, 8 GPUs, block masking at ratio 0.65.

REFERENCES

Hsu et al. 2021 (HuBERT) · Chen et al. 2022 (WavLM) · LeCun 2022, Assran et al. 2023 (JEPA) · Baevski et al. 2022 (data2vec) · Liu et al. 2023 (DinoSR) · Chiu et al. 2022 (BEST-RQ) · Skean et al. 2025, Garrido et al. 2023 (effective rank) · Chang et al. 2022 (DistilHuBERT) · wen Yang et al. 2021 (SUPERB)