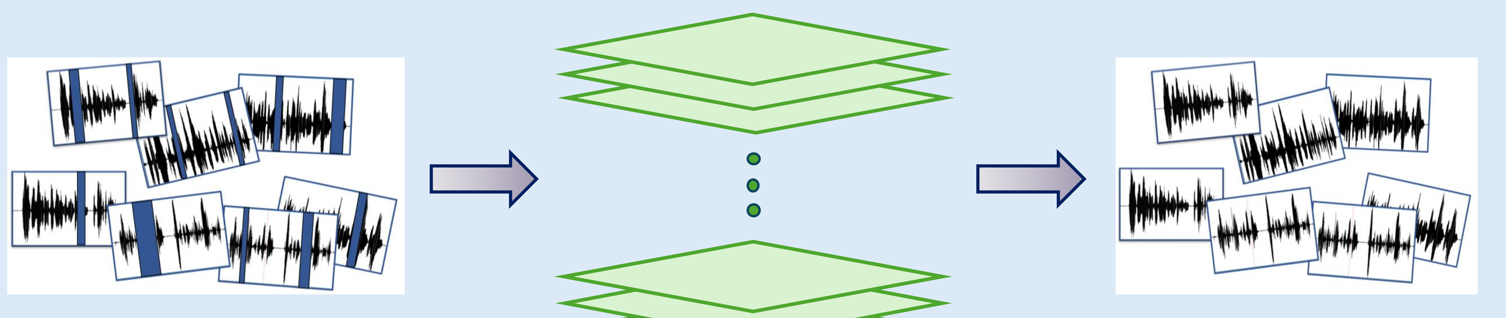
What Do Self-Supervised Speech Models Know About Words?

Ankita Pasad, Chung-Ming Chien, Shane Settle, Karen Livescu

In a nutshell

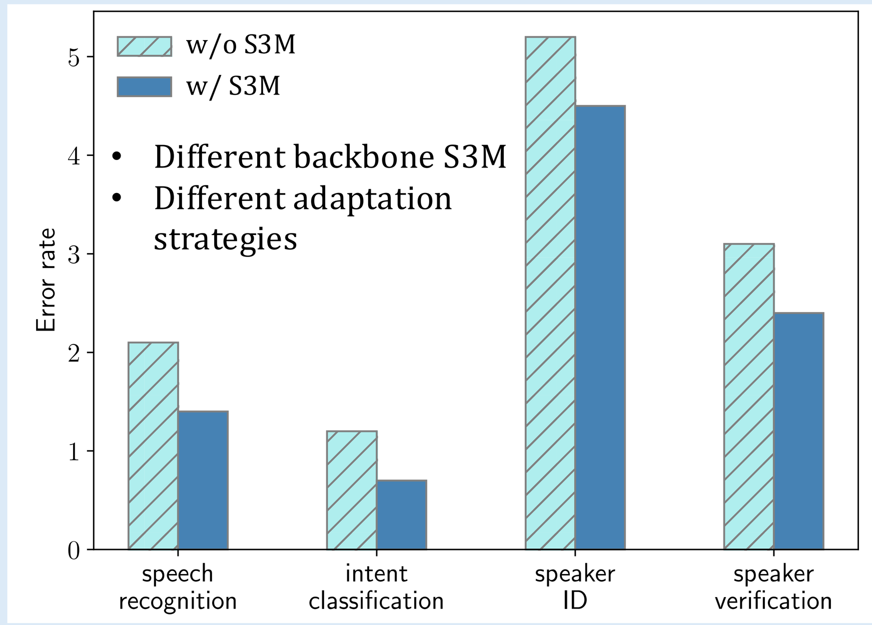
Self-supervised speech models (S3Ms) leverage unlabeled data to improve performance and data efficiency on a supervised downstream task.

Artificially corrupted audio **Recovery of clean input**



A self-supervised speech model (S3M) pre-trained with a **pretext task**

Strong empirical evidence[1]



BUT...

- Slower progress on fundamental understanding.
- Most prior analysis work has focused on phonetic and sub-word units.

In our work...

- ✓ Lightweight analytical tools for quick discovery and evaluation.
- ✓ Analysis of ten S3Ms varying in size, pre-training objective, and modality.
- ✓ Frame-wise and layer-wise analysis word-level knowledge.

Bob: So, what do you find from the analysis of ten S3Ms?

Alice: We use canonical correlation analysis (CCA) to study word-level pronunciation, syntax, and semantics and find that intermediate layers typically encode the most linguistic content.

Bob: Which intermediate layers?

Alice: That depends on the form of the pre-training objective. S3Ms that share pre-training objectives have similar trends, even if their pre-training data and model sizes are different.

Bob: And what about frame-wise analysis?

Alice: We find that central frames in a word segment encode the most word-identifying content, whereas edge frames contain little to none. We also propose a simple peak-detection algorithm using frame-level representations, which is effective at unsupervised word segmentation, surpassing more complex baselines.

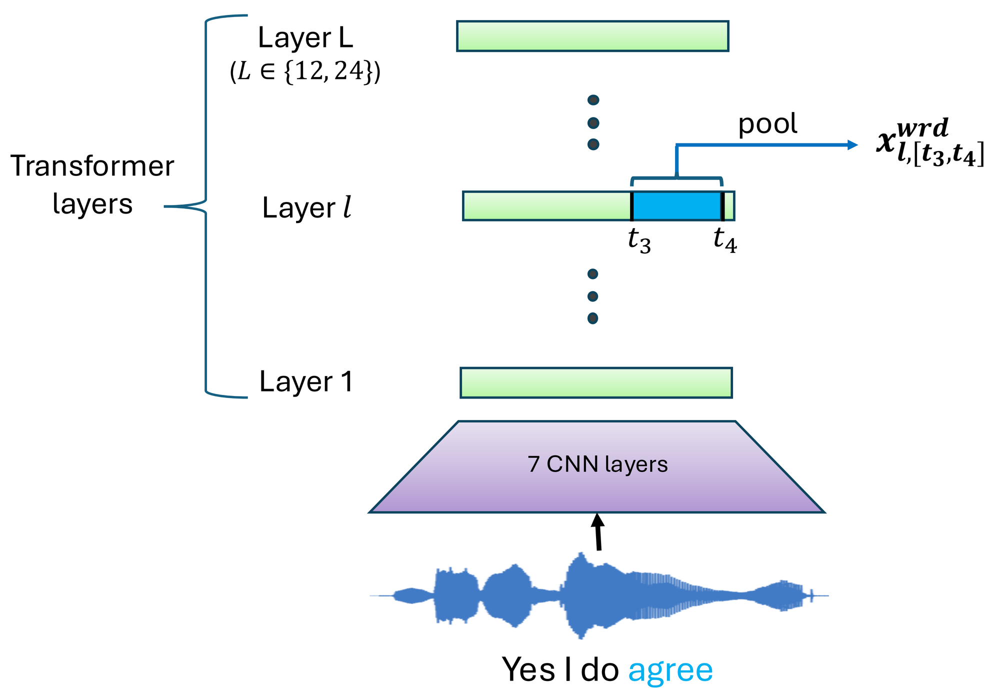
Bob: Got it, and in that case, is mean-pooling still an optimal choice?

Alice: Thanks for asking! We study that by evaluating acoustic word discrimination on S3M representations and find that different S3Ms vary in their robustness to mean-pooling.

Bob: Interesting, I am excited to read the paper! What else will I find?

Alice: You'll find our study of utterance-level representations and how they encode non-trivial semantic content. You'll find the effects of the data domain on the outcome of task-based evaluations and how the layer-wise trends from task-based studies agree with those from our task-agnostic CCA studies. You'll find many plots studying these various phenomena and maybe you can spot some interesting takeaways we might have missed!

Canonical Correlation Analysis[2,3]



- Similarity as maximum correlation between linear projections.
- Closed-form solution.
- Compare S3M representations with external word vectors.

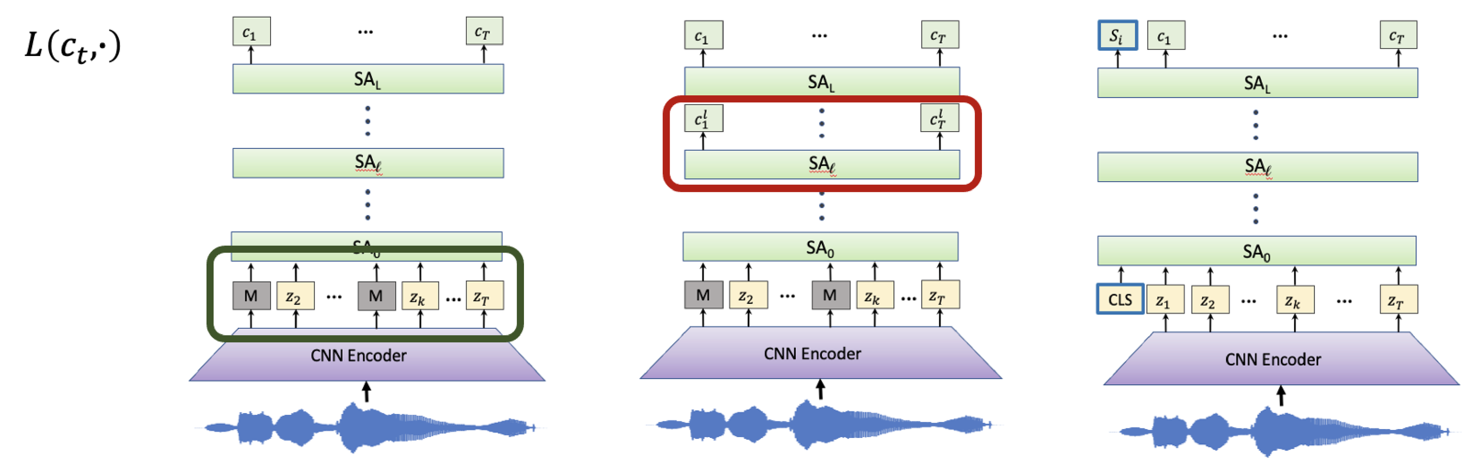
$$CCA(X,Y) = \sum_i \rho_i \, ; \rho_i = corr(v_i^T X, w_i^T Y)$$

$$v_1, w_1 = argmax_{v,w} \; corr(v^T X, w^T Y)$$

$$v_i, w_i = argmax_{v,w} \; corr(v^T X, w^T Y) \; \forall i \in [2,k] \; s.t.$$

$$corr(v_i^T X, v_j^T X) = 0 \; \forall j < i, corr(w_i^T Y, w_j^T Y) = 0 \; \forall j < i$$

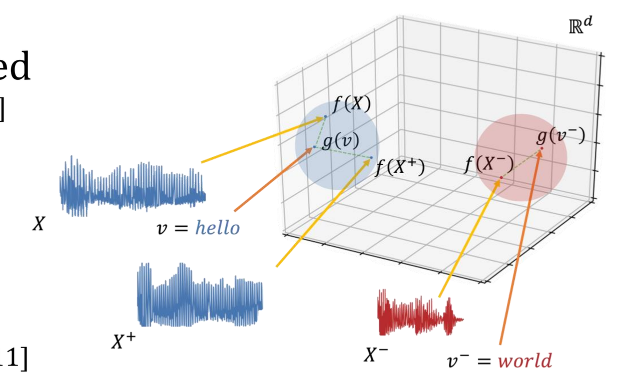
Self-supervised speech models



S3M	wav2vec 2.0[4]	HuBERT[5]	VG-HuBERT[6]
Target	$q_t = quantized(z_t)$	$y_t = discretized(c_t^l)$	$I_i = ImageEncoder(image)$
Objective	Contrastive loss	Predictive loss	Cross-modal margin contrastive loss
Similar S3Ms	FaST-VGS+[7]	WavLM[8] AV-HuBERT[9]	

Linguistic features

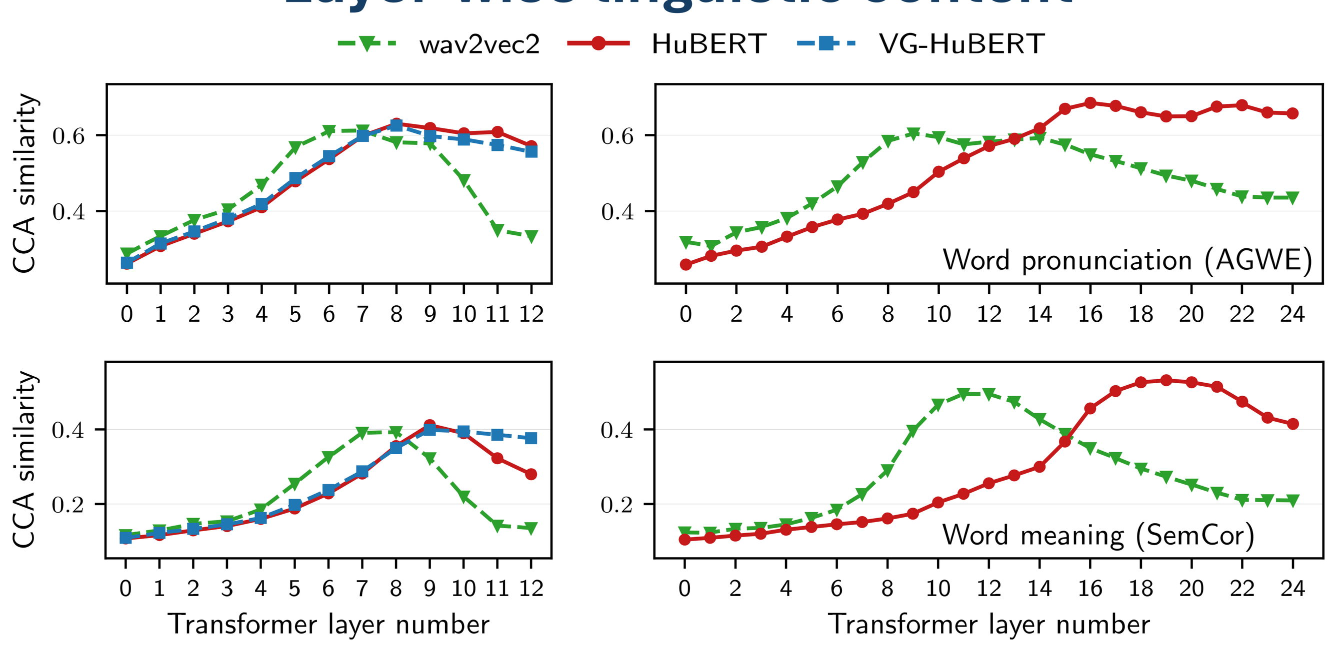
Acoustically grounded word embeddings[10]



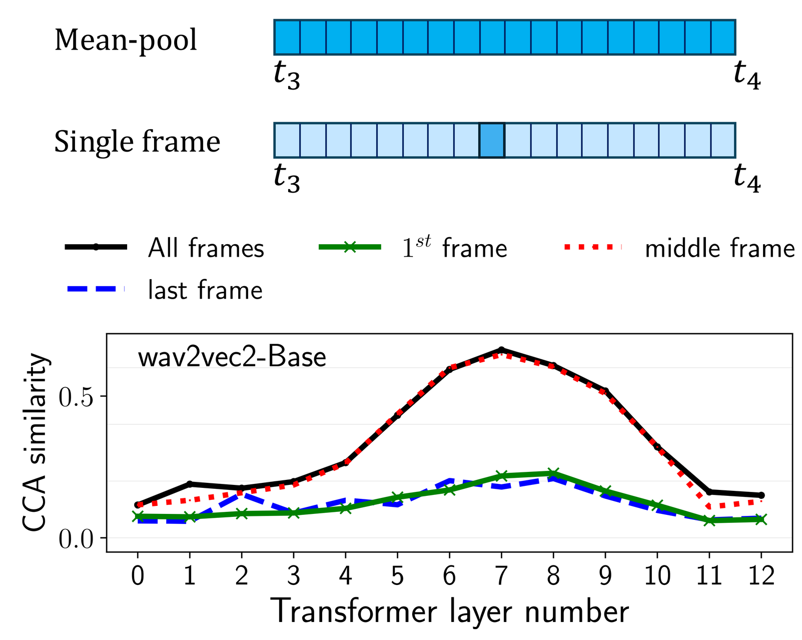
Semantic attributes[11]

WORD	NN.GROUP	NN.ACT	...	NN.ARTIFACT	VB.CHANGE
family	0.96	0.04	...	0.00	0.00
mix	0.00	0.00	...	0.00	0.91
industry	0.79	0.21	...	0.00	0.00

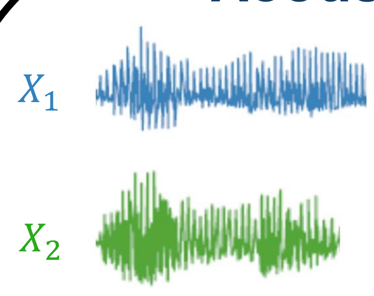
Layer-wise linguistic content



Distribution across frames



Acoustic word discrimination



Do X_1 and X_2 correspond to the same word?

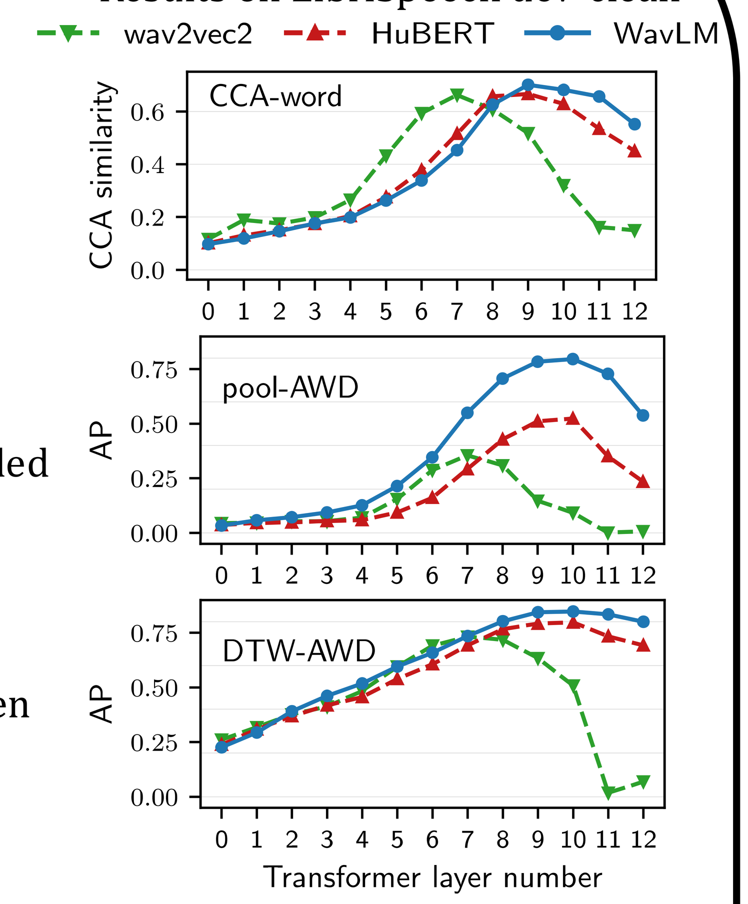
pool-AWD

Cosine similarity of mean-pooled representations

DTW-AWD

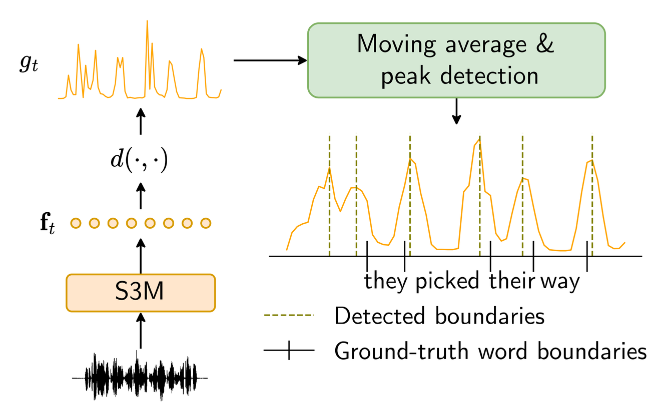
Dynamic time warping between frame-level representations

Results on LibriSpeech dev-clean

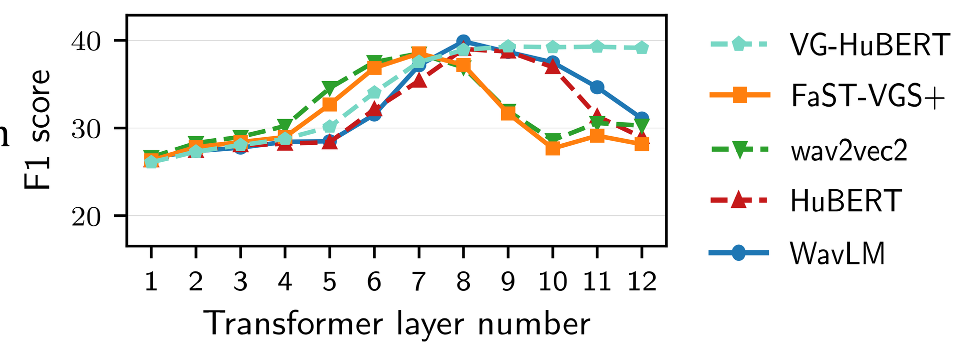


- All three models have similarly high CCA scores.
- AWD has similar trends as CCA.
- pool-AWD has drastic differences in relative AWD scores.
- DTW-AWD closes the performance gap with improved scores.

Unsupervised word segmentation



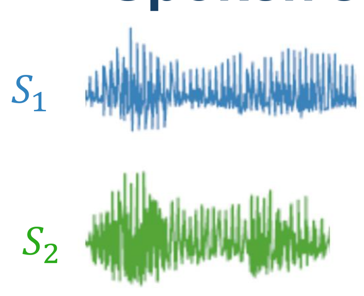
Results on LibriSpeech dev-clean



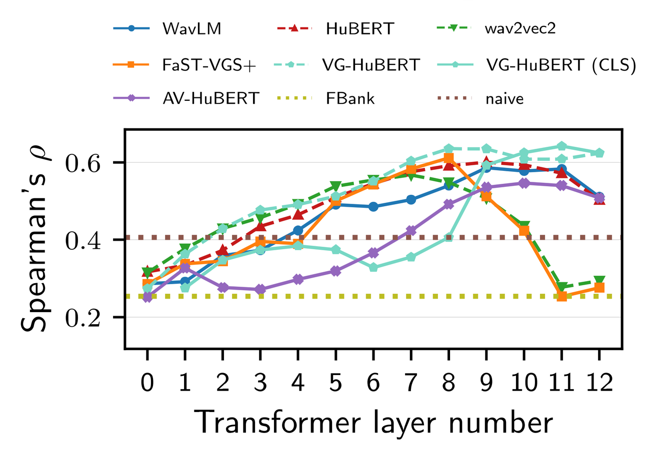
Results on Buckeye test

Method	Precision	Recall	F1	R-val
DPDP[12]	35.3	37.7	36.4	44.3
VG-HuBERT[6]	**36.2**	32.2	34.1	**45.6**
Ours (Best Layer)				
HuBERT-Base (L9)	33.8	46.6	39.2	34.9
VG-HuBERT (L10)	36.0	**47.6**	**41.0**	39.5

Spoken sentence similarity[13]



Are S_1 and S_2 semantically similar sentences?



[1] S. Yang et al., "SUPERB: Speech processing universal performance benchmark", Interspeech, 2021

[2] Hotelling, "Relations between two sets of variates. Biometrika", 1936.

[3] Morcos et al., "Insights on representational similarity in neural networks with canonical correlation," NeurIPS 2018

[4] Baevski et al., "wav2vec 2.0: A Framework for self-supervised learning of speech representations", NeurIPS, 2020

[5] Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units", TASLP, 2021

[6] Peng and Harwath, "Word discovery in visually grounded, self-supervised speech models", Interspeech, 2022

[7] Peng and Harwath, "Fast-slow transformer for visually grounding speech", ICASSP, 2022

[8] Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing", JSTSP, 2022

[9] Shi et al., Learning audio-visual speech representation by masked multimodal cluster prediction, ICLR, 2022

[10] Shi et al., "Whole-word segmental speech recognition with acoustic word embeddings," SLT 2021.

[11] Tsvetkov et al., "Evaluation of word vector representations by subspace alignment," EMNLP 2015.

[12] Kamper, "Word segmentation on discovered phone units with dynamic programming and self-supervised scoring", TASLP, 2022.

[13] Merkx et al., "Semantic sentence similarity: size does not always matter", Interspeech, 2021

Codebase