Self-Supervised Speech Foundation Models

Karen Livescu

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Self-introduction

About me

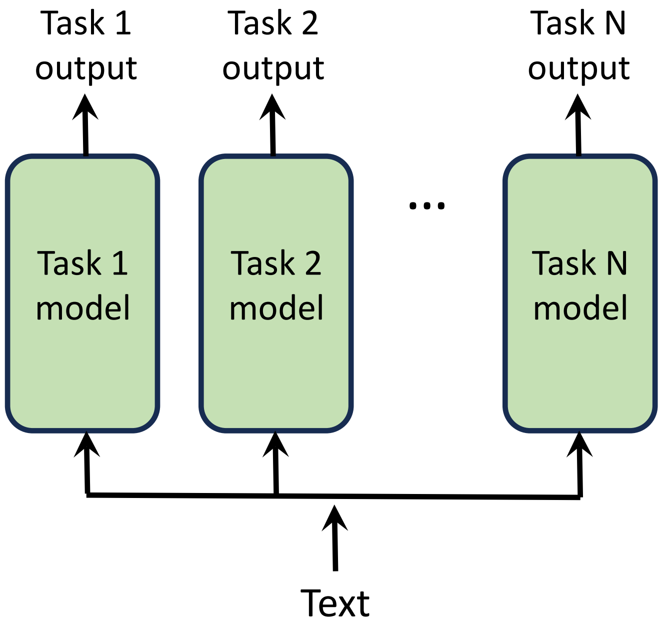
- Faculty member at TTI-Chicago since 2008
- Before that, PhD and post-doc at MIT in EECS
- Research interests: Speech, NLP, other language modalities and multi-modal language

TTI-Chicago

- Graduate institute for computer science on University of Chicago campus
- PhD program with focus on ML, algorithms, AI
- 13 tenure-track/tenured faculty, 15 research faculty (3-yr post-doctoral position)
- ~45 PhD students
- ~20 adjoint/courtesy/ visiting faculty, visiting students, ...

Evolution of (text) foundation models

The task-specific model era (- 2018)



Task 1 output

Task 2 output

Task N output

Task 1 model

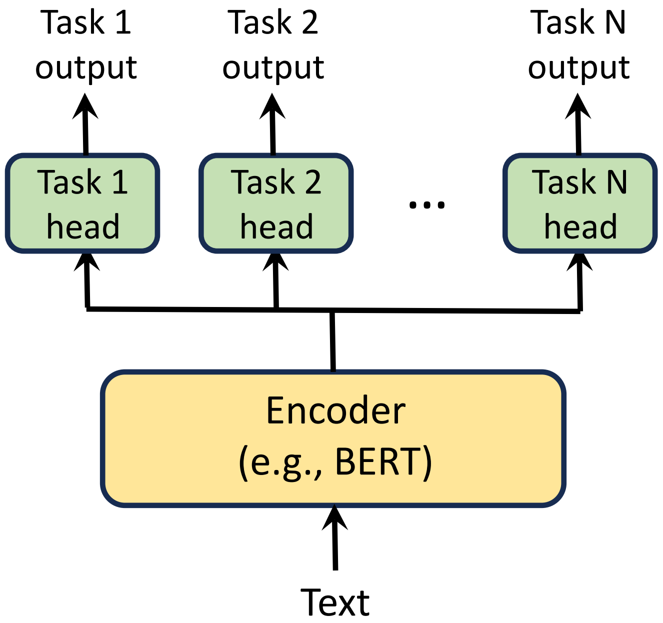
Task 2 model

...

Task N model

Text

The encoder era (2018 - 2022)



Task 1 output

Task 2 output

Task N output

Task 1 head

Task 2 head

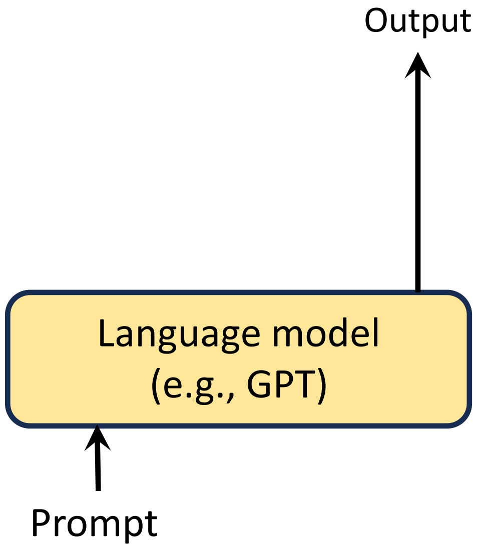
...

Task N head

Encoder (e.g., BERT)

Text

The large language model era (2022 -)



Output

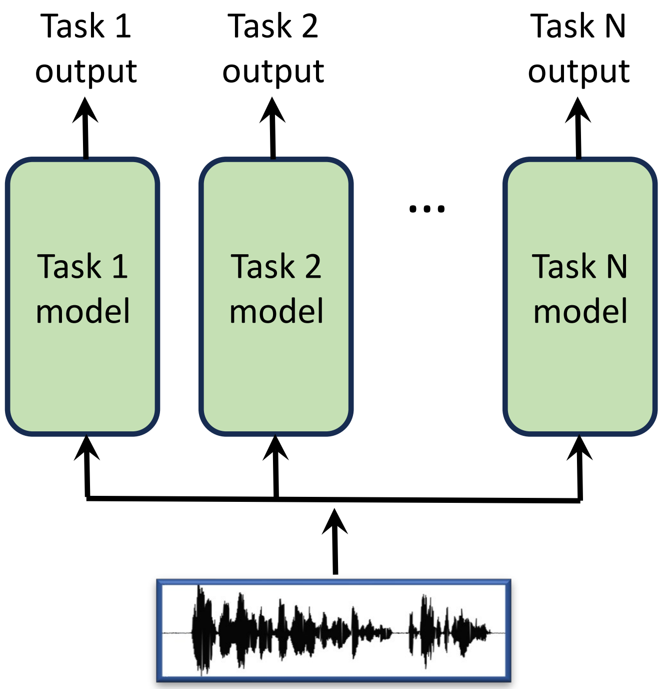
Language model (e.g., GPT)

Prompt

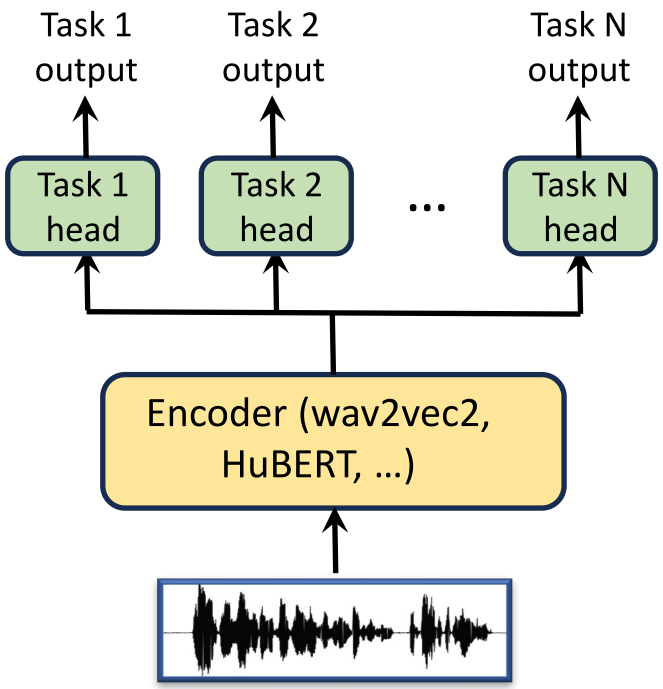
More task-universality, less human effort

Evolution of speech foundation models

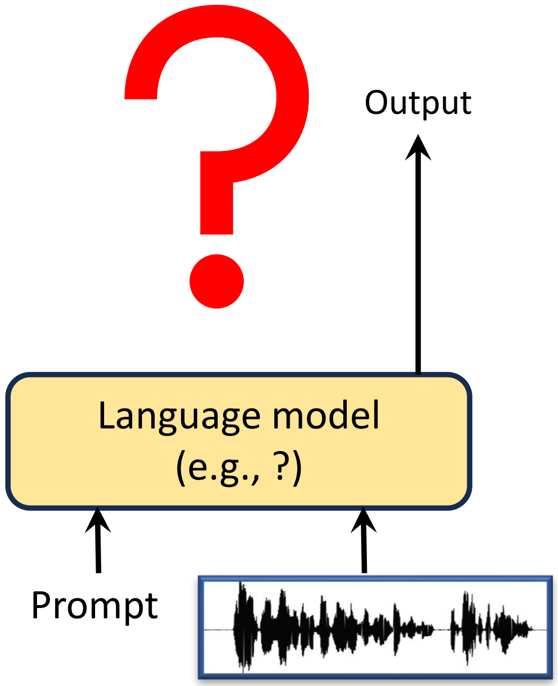
The task-specific model era (- 2020)



The speech encoder era (2020 -)

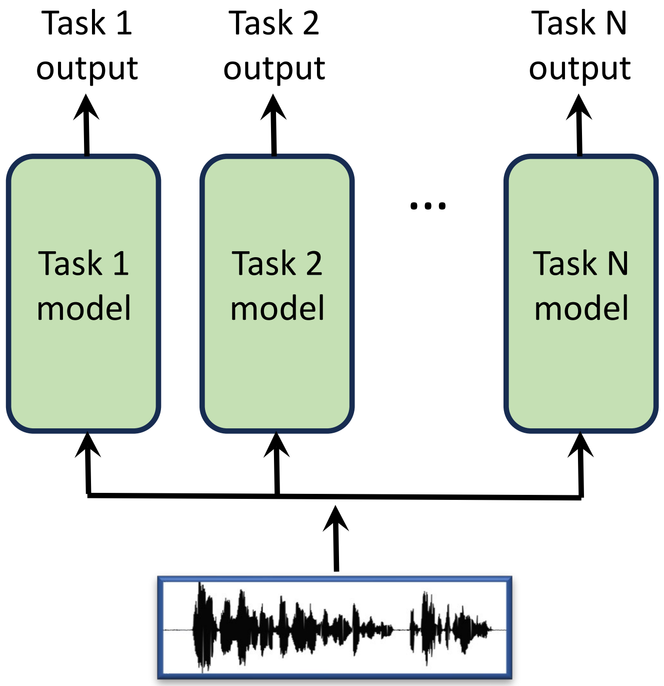


The spoken large language model era (2024? -)

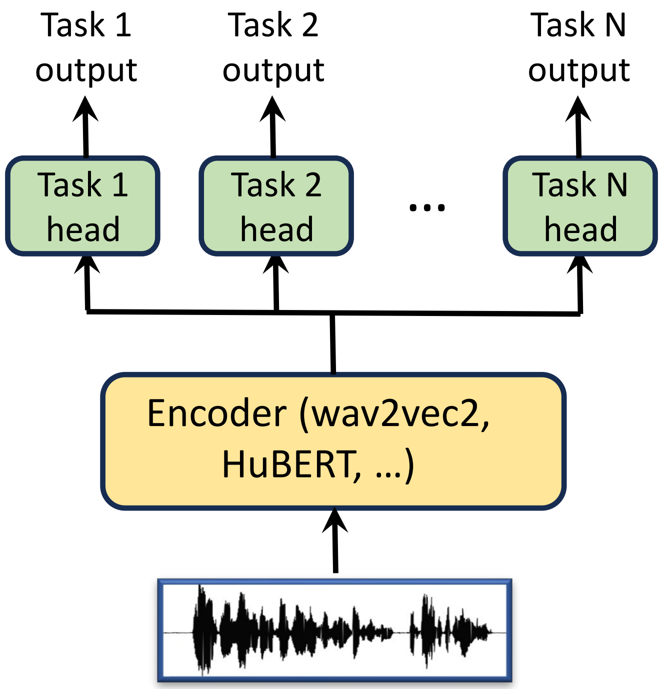


Evolution of speech foundation models

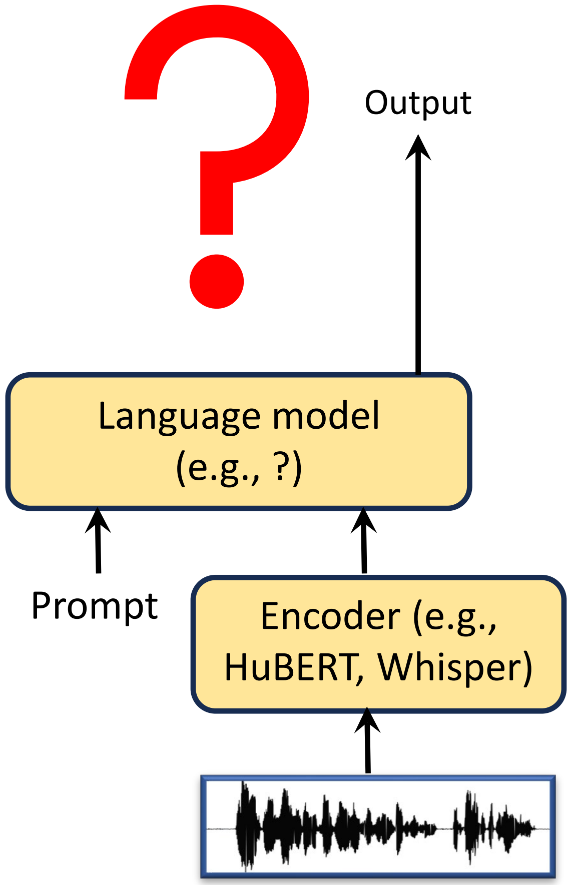
The task-specific model era (- 2020)



The speech encoder era (2020 -)



The spoken large language model era (2024? -)

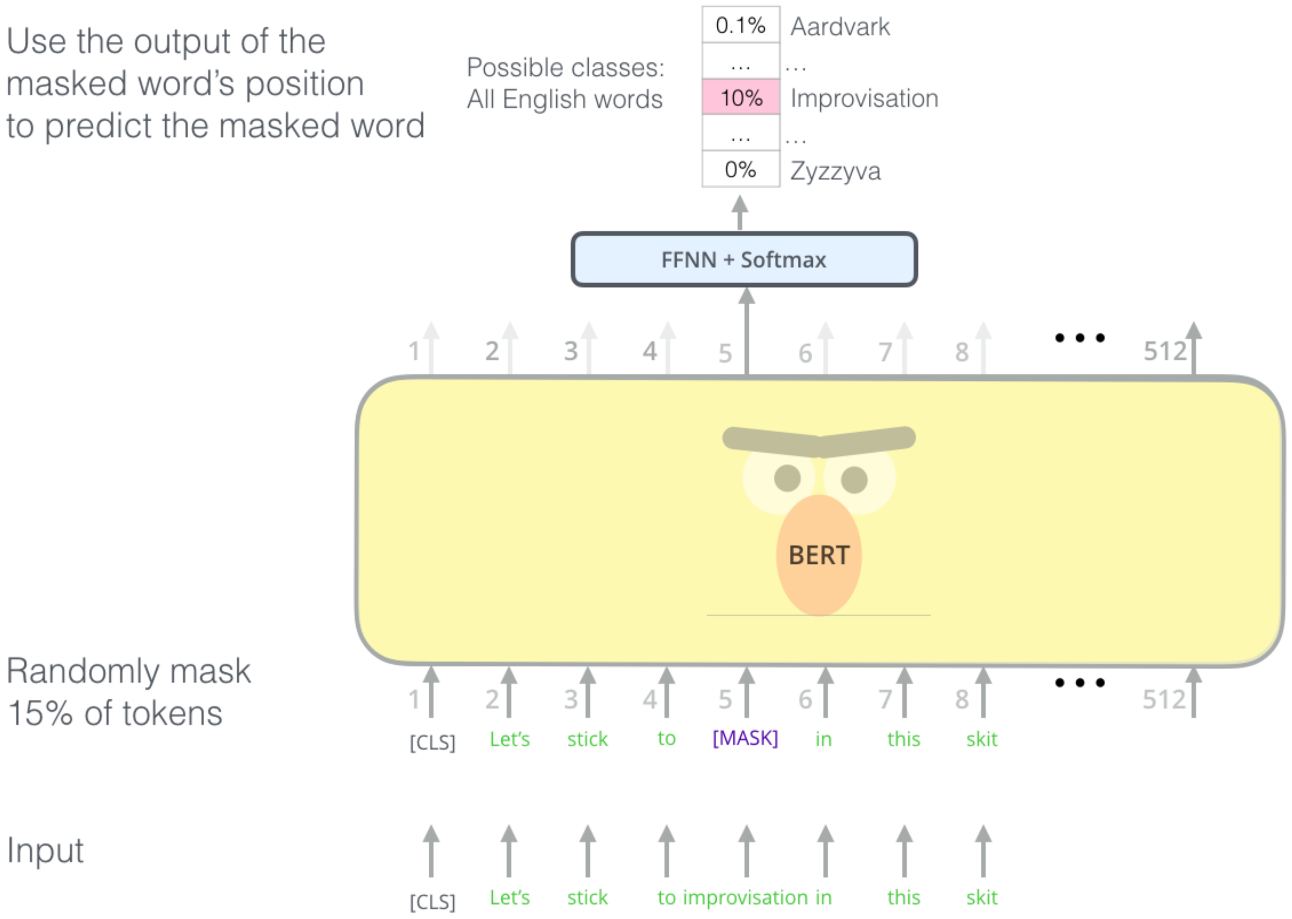


Self-supervised learning

Sometimes learners set up tasks for themselves to solve…

- Even if we only have unlabeled data, we may be able to define “pretext tasks” from the data alone
- A good pretext task is one that requires us to represent the “useful” information in the input in order to solve it well

Self-supervised learning for text: BERT



Use the output of the masked word's position to predict the masked word

Randomly mask 15% of tokens

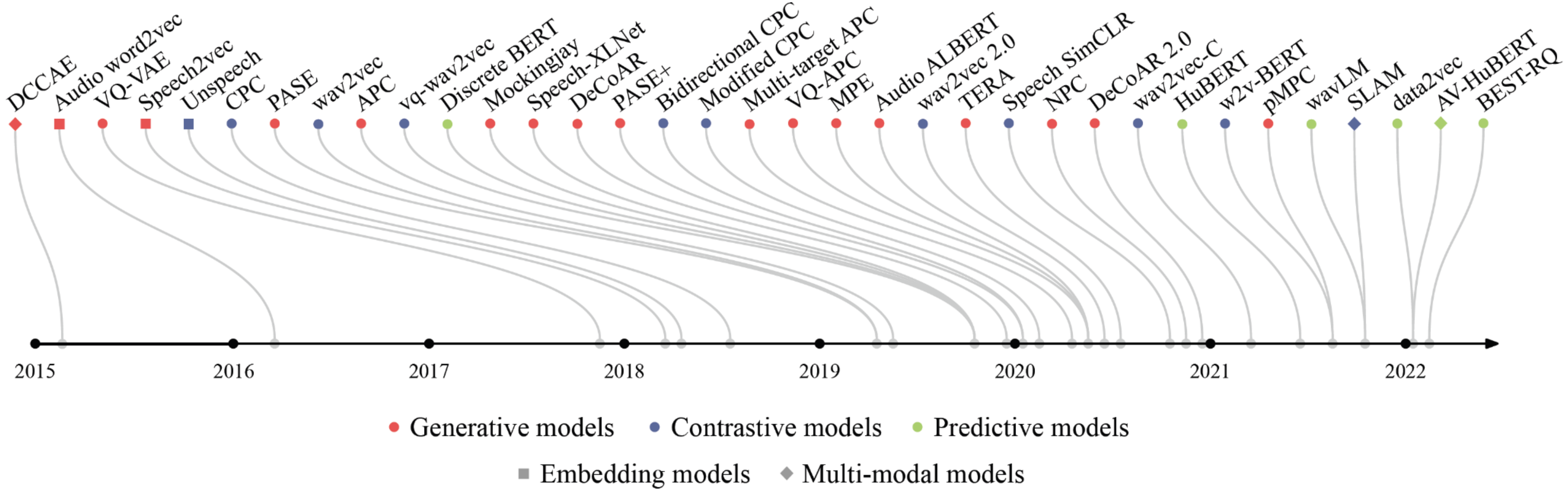
Input

[Alammar 2018]

Self-supervised learning for speech

Some speech-specific issues

- Unlike text, speech is a continuous signal with no fixed vocabulary
- Speech is continuous in time, with no boundaries between lexical units
- Speech includes informative information besides the words: speaker, accent, emotion, …
    - Which information we want to keep may depend on the downstream task
    - In practice, most common self-supervised speech models have been optimized for ASR



Mohamed et al., "Self-supervised speech representation learning: A review"
IEEE Journal of Selected Topics in Signal Processing **16**(6):1179-1210, October 2022.

wav2vec 2.0

- The task: Predict masked speech frames
- Contrastive loss
    - Predicted frame representations should be similar to quantized input features at the same frame
    - … and different from inputs at different frames

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

- FastVGS(+) Adds cross-modal retrieval loss

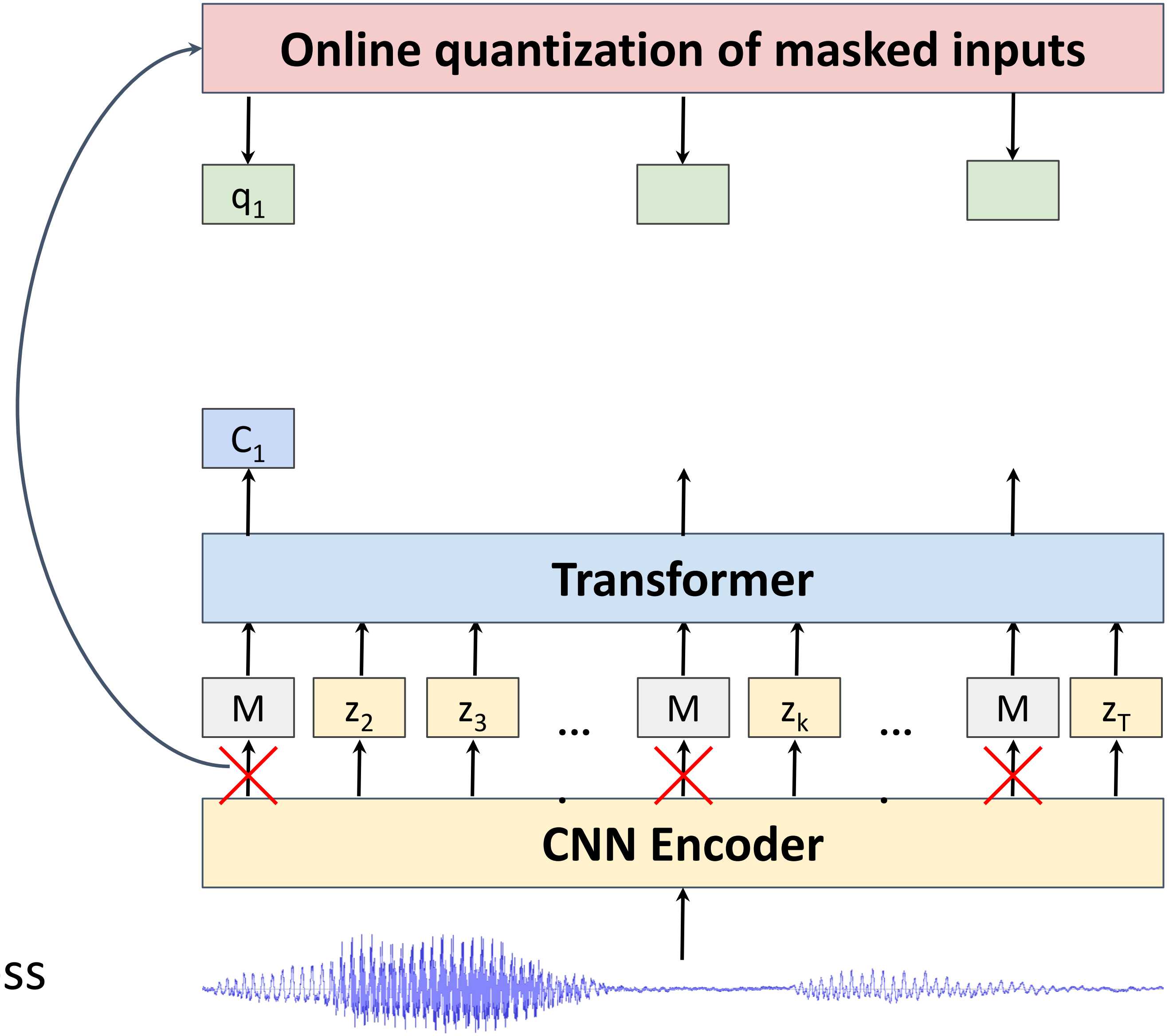


Figure credit: A. Mohamed, "Tutorial on Self-Supervised Representation Learning for Speech Processing," NAACL 2022

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

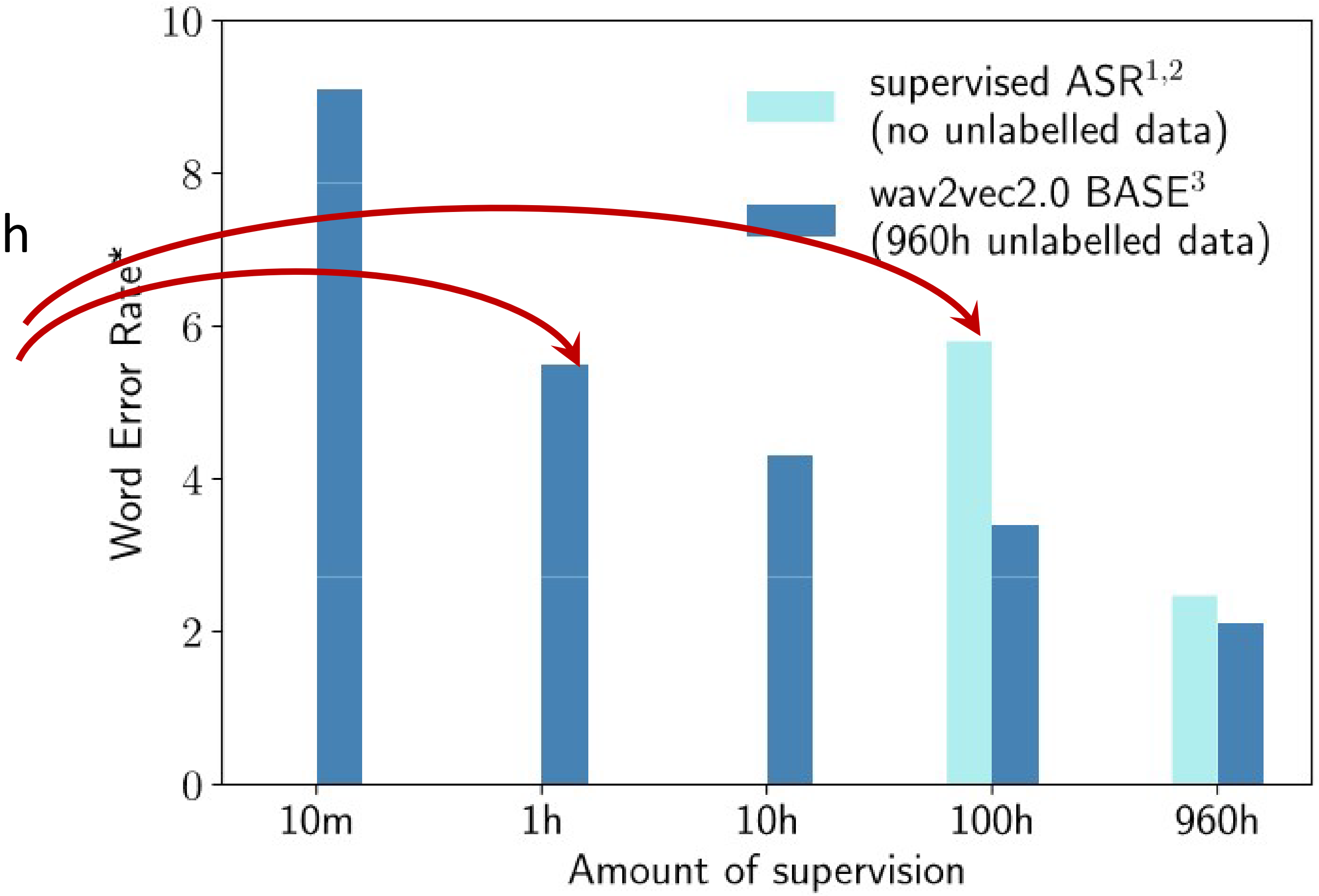
Peng & Harwath, "Self-supervised representation learning for speech using visual grounding and masked language modeling," AAAI SAS 2022

wav2vec 2.0: Some results

First major improvements on ASR using self-supervised learning

2020: wav2vec 2.0 improves performance and labeled data efficiency on the LibriSpeech benchmark

- Matches a supervised model using only 1% of the labeled data (100 hours → 1 hour)



[1] Lüscher et al., RWTH ASR Systems for LibriSpeech: Hybrid vs Attention, Interspeech, 2019

[2] Synnaeve et al., End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures, arXiv:1911.08460, 2020

[3] Baevski et al., wav2vec 2.0: A Framework for Self-supervised Learning of Speech Representations, NeurIPS, 2020

HuBERT (Hidden-unit BERT)

- Uses quantization like wav2vec 2.0, but BERT-like masked prediction loss
- Iterates quantization and re-training
- The task: Predict masked speech frames
- Log loss (cross-entropy)

$$\mathcal{L}_m = \sum_{t \in M} -\log p(y_t \mid X)$$

- First iteration uses quantized spectrogram

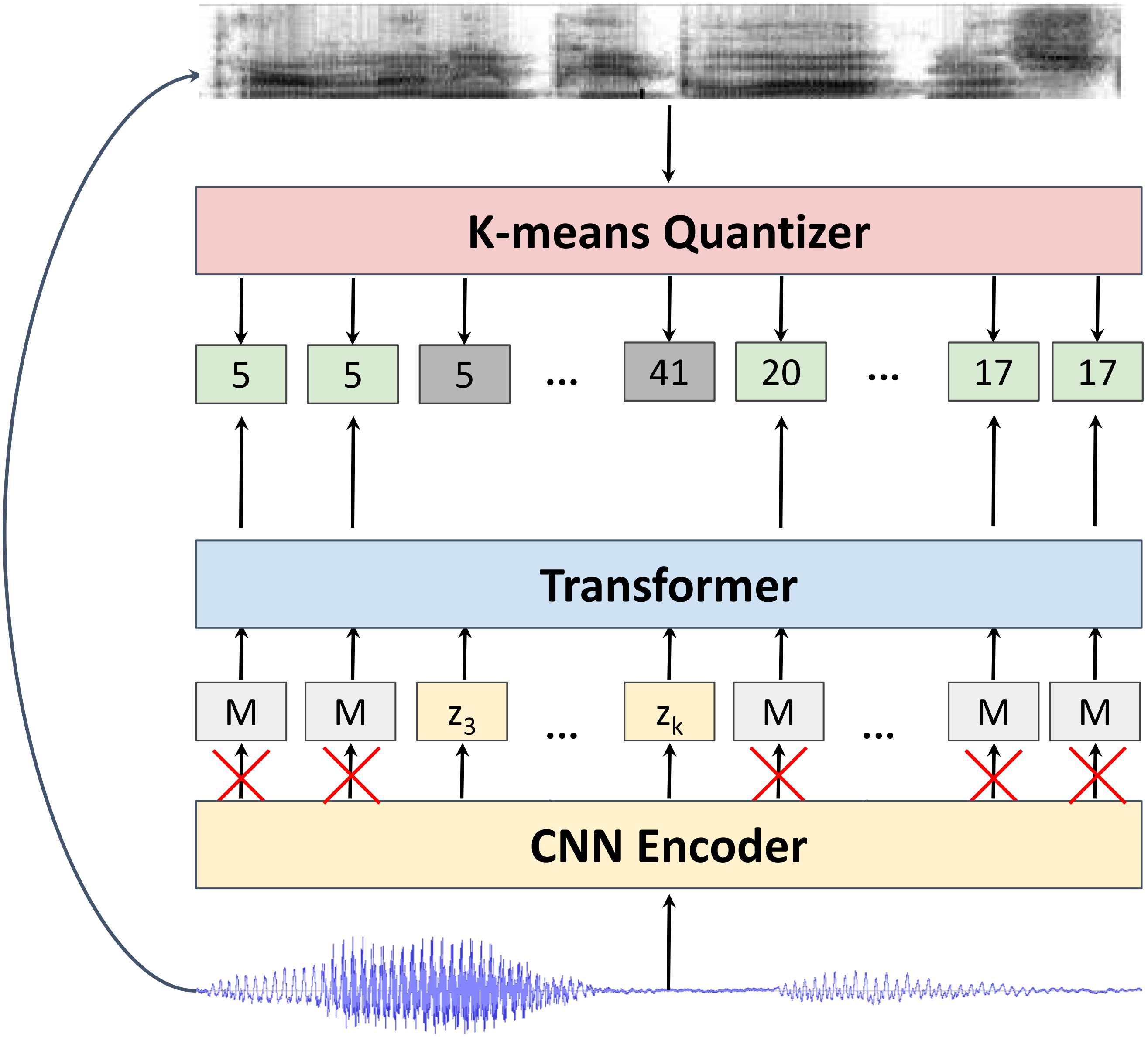


Figure credit: A. Mohamed, "Tutorial on Self-Supervised Representation Learning for Speech Processing," NAACL 2022
Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," T-ASSP 2021

HuBERT (Hidden-unit BERT)

- Subsequent iterations: Quantize HuBERT features from previous iteration
  - Which layer of previous iteration? One that is good for phonetic classification
  - (Note: not quite unsupervised anymore…)
  - In practice: Layer 6 for iteration 1, layer 9 for iteration 2
- Related models
  - WavLM: Additional denoising loss
  - AV-HuBERT: Multimodal clusters learned from speaker mouth videos

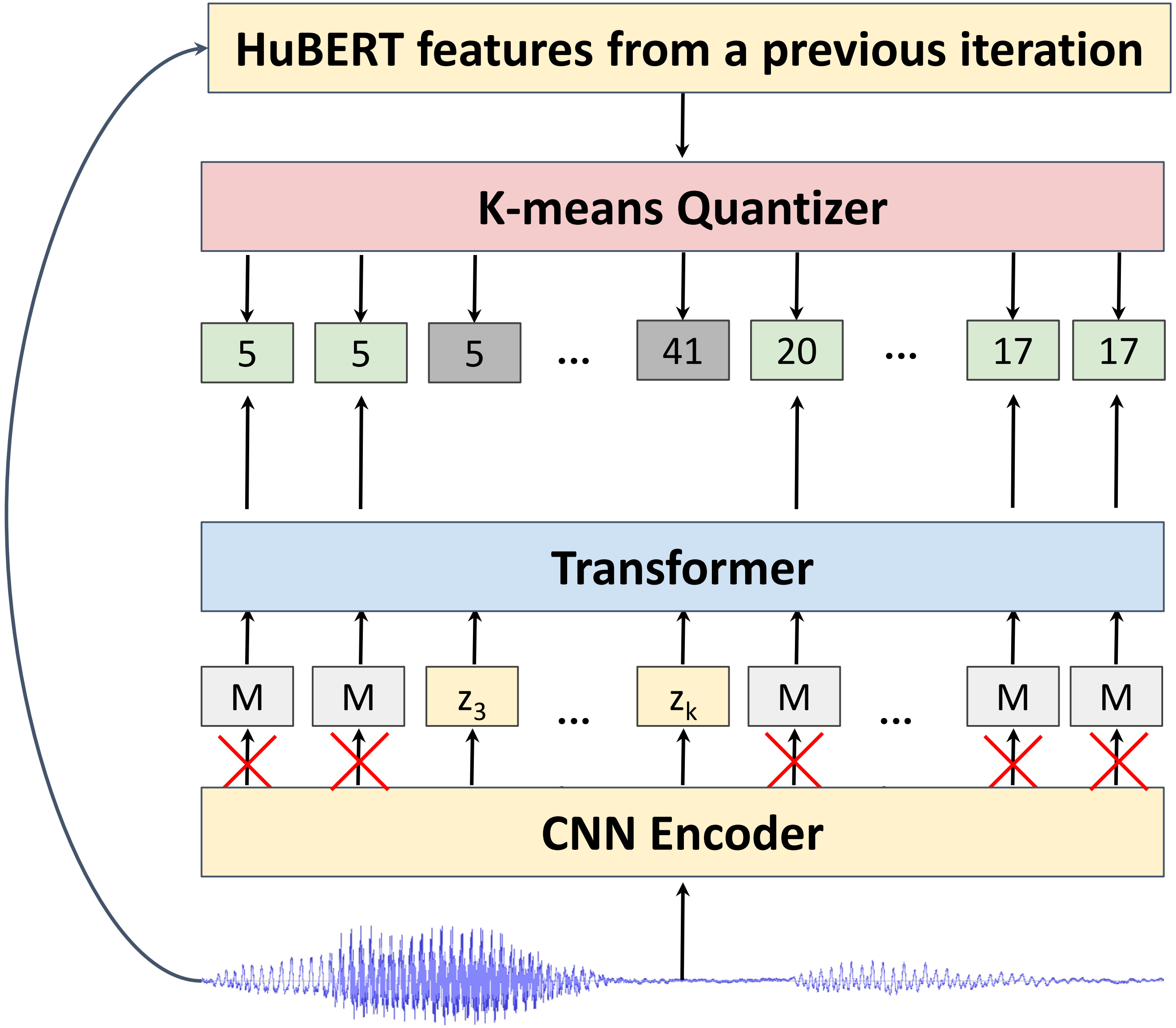
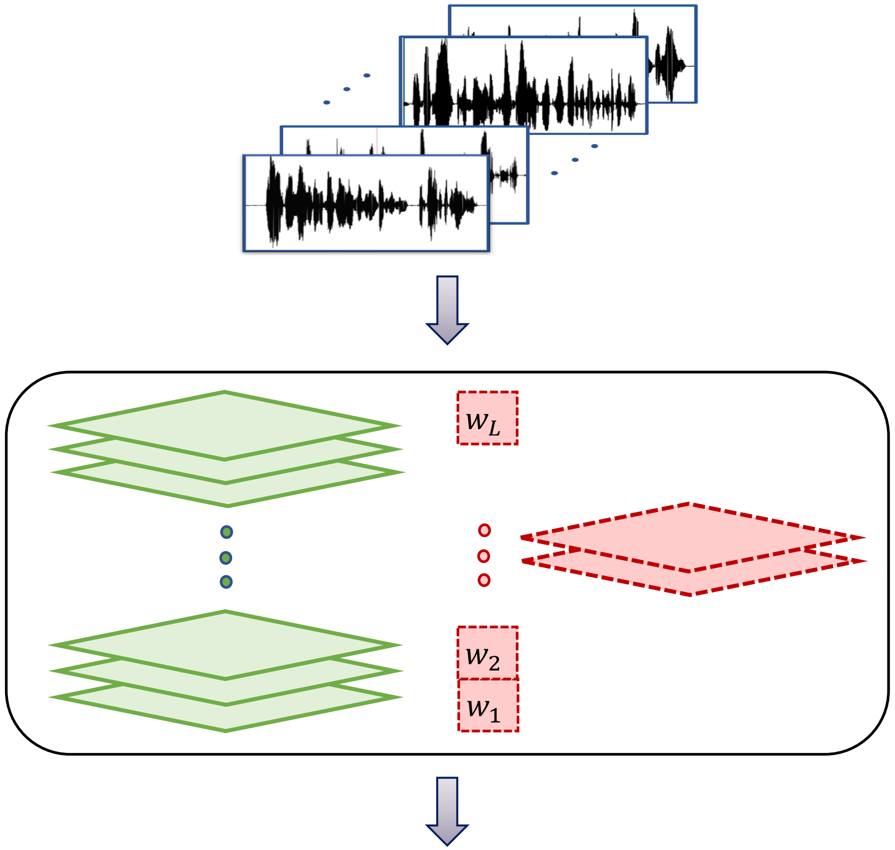
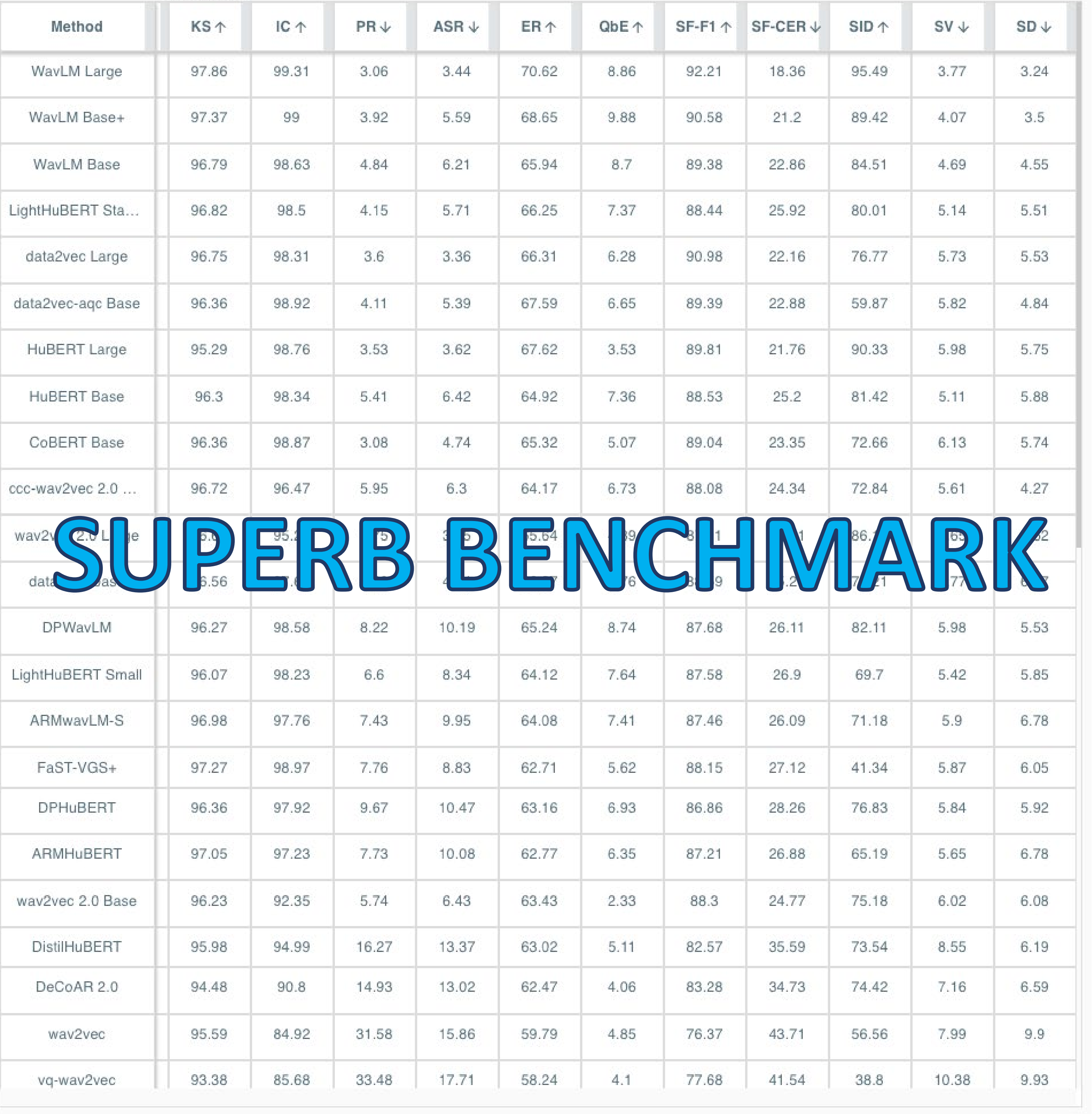


Figure credit: A. Mohamed, “Tutorial on Self-Supervised Representation Learning for Speech Processing,” NAACL 2022
Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” T-ASSP 2021

What can speech encoders *do*?



Task-specific labels

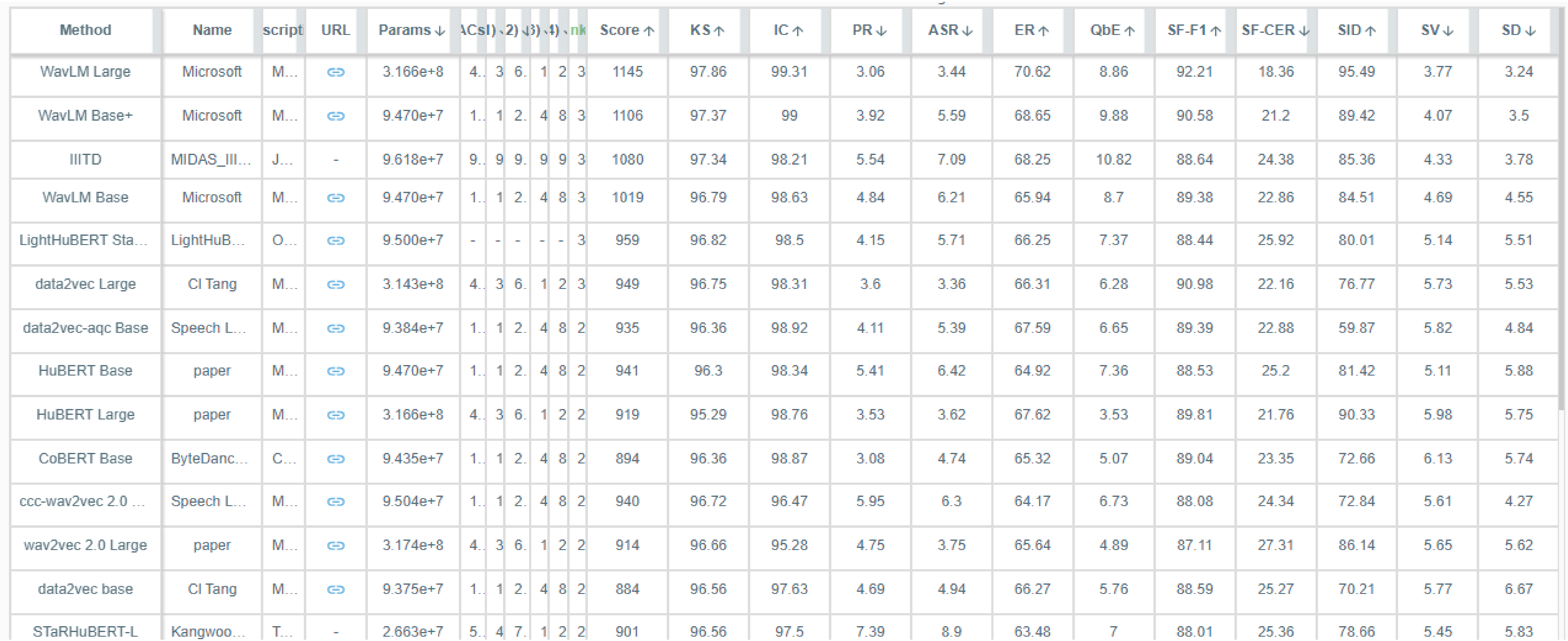


SUPERB BENCHMARK

Method	KS ↑	IC ↑	PR ↓	ASR ↓	ER ↑	QbE ↑	SF-F1 ↑	SF-CER ↓	SID ↑	SV ↓	SD ↓
WavLM Large	97.86	99.31	3.06	3.44	70.62	8.86	92.21	18.36	95.49	3.77	3.24
WavLM Base+	97.37	99	3.92	5.59	68.65	9.88	90.58	21.2	89.42	4.07	3.5
WavLM Base	96.79	98.63	4.84	6.21	65.94	8.7	89.38	22.86	84.51	4.69	4.55
LightHuBERT Sta…	96.82	98.5	4.15	5.71	66.25	7.37	88.44	25.92	80.01	5.14	5.51
data2vec Large	96.75	98.31	3.6	3.36	66.31	6.28	90.98	22.16	76.77	5.73	5.53
data2vec-aqc Base	96.36	98.92	4.11	5.39	67.59	6.65	89.39	22.88	59.87	5.82	4.84
HuBERT Large	95.29	98.76	3.53	3.62	67.62	3.53	89.81	21.76	90.33	5.98	5.75
HuBERT Base	96.3	98.34	5.41	6.42	64.92	7.36	88.53	25.2	81.42	5.11	5.88
CoBERT Base	96.36	98.87	3.08	4.74	65.32	5.07	89.04	23.35	72.66	6.13	5.74
ccc-wav2vec 2.0 …	96.72	96.47	5.95	6.3	64.17	6.73	88.08	24.34	72.84	5.61	4.27
wav2vec 2.0 Large	[illegible]	[illegible]	[illegible]	[illegible]	65.64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
data2vec Base	96.56	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DPWavLM	96.27	98.58	8.22	10.19	65.24	8.74	87.68	26.11	82.11	5.98	5.53
LightHuBERT Small	96.07	98.23	6.6	8.34	64.12	7.64	87.58	26.9	69.7	5.42	5.85
ARMwavLM-S	96.98	97.76	7.43	9.95	64.08	7.41	87.46	26.09	71.18	5.9	6.78
FaST-VGS+	97.27	98.97	7.76	8.83	62.71	5.62	88.15	27.12	41.34	5.87	6.05
DPHuBERT	96.36	97.92	9.67	10.47	63.16	6.93	86.86	28.26	76.83	5.84	5.92
ARMHuBERT	97.05	97.23	7.73	10.08	62.77	6.35	87.21	26.88	65.19	5.65	6.78
wav2vec 2.0 Base	96.23	92.35	5.74	6.43	63.43	2.33	88.3	24.77	75.18	6.02	6.08
DistilHuBERT	95.98	94.99	16.27	13.37	63.02	5.11	82.57	35.59	73.54	8.55	6.19
DeCoAR 2.0	94.48	90.8	14.93	13.02	62.47	4.06	83.28	34.73	74.42	7.16	6.59
wav2vec	95.59	84.92	31.58	15.86	59.79	4.85	76.37	43.71	56.56	7.99	9.9
vq-wav2vec	93.38	85.68	33.48	17.71	58.24	4.1	77.68	41.54	38.8	10.38	9.93

https://github.com/s3prl/s3prl
https://superbbenchmark.org/leaderboard
Yang et al., "SUPERB: Speech processing Universal PERformance Benchmark", Interspeech, 2021

What can speech encoders *do*?



Method	Name	script	URL	Params ↓	ACsI) ↓	2) ↓	3) ↓	4) ↓	nk	Score ↑	KS ↑	IC ↑	PR ↓	ASR ↓	ER ↑	QbE ↑	SF-F1 ↑	SF-CER ↓	SID ↑	SV ↓	SD ↓
WavLM Large	Microsoft	M...	🔗	3.166e+8	4..	3	6.	1	2	3	1145	97.86	99.31	3.06	3.44	70.62	8.86	92.21	18.36	95.49	3.77	3.24
WavLM Base+	Microsoft	M...	🔗	9.470e+7	1..	1	2.	4	8	3	1106	97.37	99	3.92	5.59	68.65	9.88	90.58	21.2	89.42	4.07	3.5
IIITD	MIDAS_III...	J...	-	9.618e+7	9..	9	9.	9	9	3	1080	97.34	98.21	5.54	7.09	68.25	10.82	88.64	24.38	85.36	4.33	3.78
WavLM Base	Microsoft	M...	🔗	9.470e+7	1..	1	2.	4	8	3	1019	96.79	98.63	4.84	6.21	65.94	8.7	89.38	22.86	84.51	4.69	4.55
LightHuBERT Sta...	LightHuB...	O...	🔗	9.500e+7	-	-	-	-	-	3	959	96.82	98.5	4.15	5.71	66.25	7.37	88.44	25.92	80.01	5.14	5.51
data2vec Large	CI Tang	M...	🔗	3.143e+8	4..	3	6.	1	2	3	949	96.75	98.31	3.6	3.36	66.31	6.28	90.98	22.16	76.77	5.73	5.53
data2vec-aqc Base	Speech L...	M...	🔗	9.384e+7	1..	1	2.	4	8	2	935	96.36	98.92	4.11	5.39	67.59	6.65	89.39	22.88	59.87	5.82	4.84
HuBERT Base	paper	M...	🔗	9.470e+7	1..	1	2.	4	8	2	941	96.3	98.34	5.41	6.42	64.92	7.36	88.53	25.2	81.42	5.11	5.88
HuBERT Large	paper	M...	🔗	3.166e+8	4..	3	6.	1	2	2	919	95.29	98.76	3.53	3.62	67.62	3.53	89.81	21.76	90.33	5.98	5.75
CoBERT Base	ByteDanc...	C...	🔗	9.435e+7	1..	1	2.	4	8	2	894	96.36	98.87	3.08	4.74	65.32	5.07	89.04	23.35	72.66	6.13	5.74
ccc-wav2vec 2.0 ...	Speech L...	M...	🔗	9.504e+7	1..	1	2.	4	8	2	940	96.72	96.47	5.95	6.3	64.17	6.73	88.08	24.34	72.84	5.61	4.27
wav2vec 2.0 Large	paper	M...	🔗	3.174e+8	4..	3	6.	1	2	2	914	96.66	95.28	4.75	3.75	65.64	4.89	87.11	27.31	86.14	5.65	5.62
data2vec base	CI Tang	M...	🔗	9.375e+7	1..	1	2.	4	8	2	884	96.56	97.63	4.69	4.94	66.27	5.76	88.59	25.27	70.21	5.77	6.67
STaRHuBERT-L	Kangwoo...	T...	-	2.663e+7	5..	4	7.	1	2	2	901	96.56	97.5	7.39	8.9	63.48	7	88.01	25.36	78.66	5.45	5.83

https://github.com/s3prl/s3prl
https://superbbenchmark.org/leaderboard
Yang et al., “SUPERB: Speech processing Universal PERformance Benchmark”, Interspeech, 2021

What do speech encoders “know”?

What we know

+ Self-supervised models are great! Most state-of-the-art speech systems use them

+ Some layers seem more important than others, depending on the task

☹ Adapting a self-supervised model for a task takes trial and error: which model to use, how to fine-tune, …

☹ We have little guidance for designing pretext tasks

What we want to know

- What kind of linguistic information is encoded in each model, and in each layer?
- How is linguistic information distributed across time?
- How does the pretext task affect what is learned?
- Can the results guide how we use models for downstream tasks?

And, we want to do this analysis in a lightweight way (without tuning lots of downstream models)

Layer-wise analysis of speech encoders

Speaker ID

- [Chen+ 2022, Fan+ 2021, van Niekerk+ 2021]

Phones

- [Abdullah+ 2023, Hsu+ 2021, Ma+ 2021, Pasad+ 2021]

Words

- [Sanabria+ 2023, Pasad+ 2021, Choi+ 2024]

Prosodic features, tone

- [Ji+ 2022, Kim+ 2022, Shen+ 2024, de la Fuentes & Jurafsky 2024]

Dialect variation

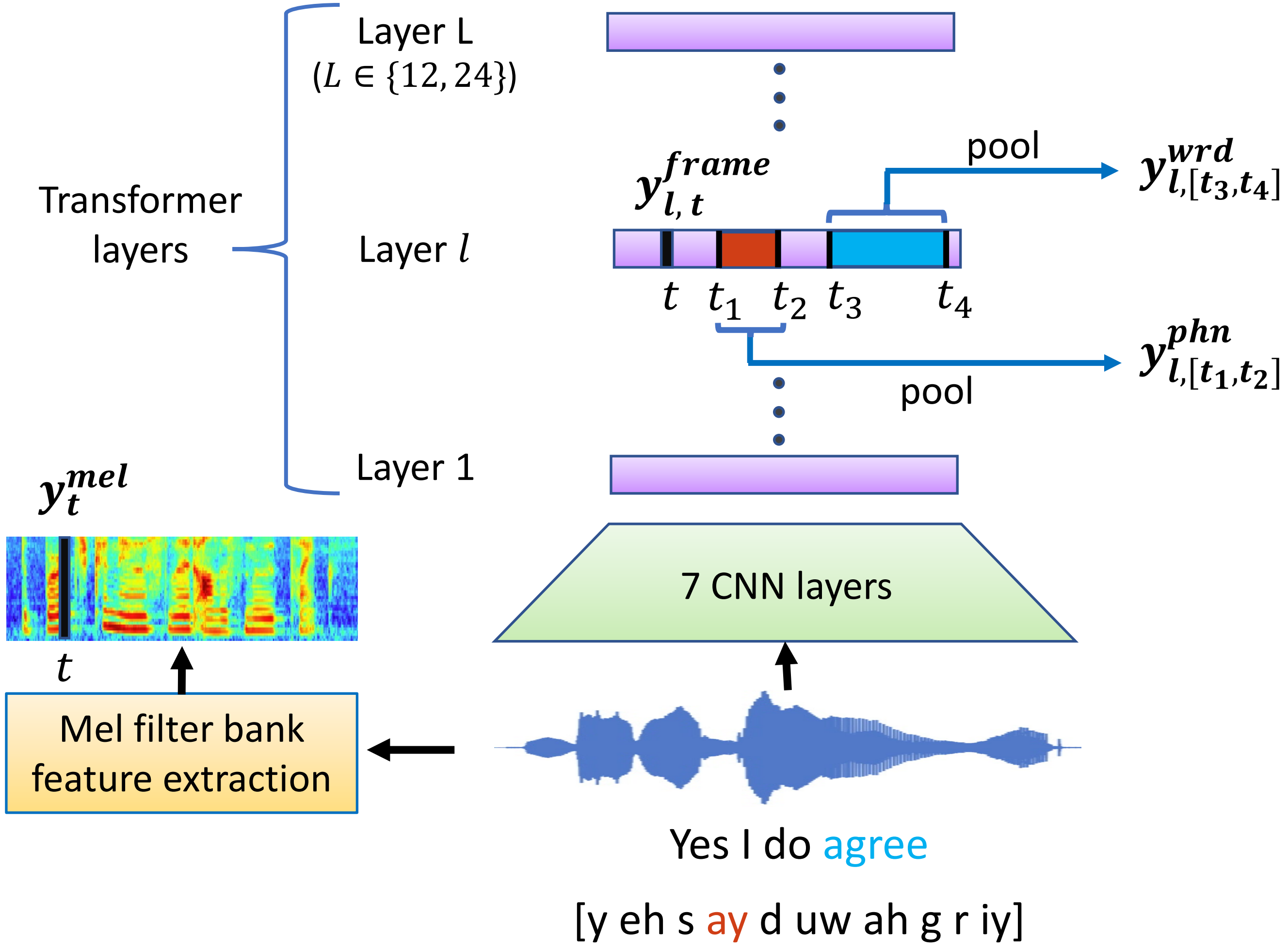
- [Bartelds+ 2022]

Cross-model analysis

- [Pasad+ 2023, 2024]

Layer-wise analysis

Method: Extract layer-wise representations, and measure their “similarity” to external linguistic variables



Measuring similarity via canonical correlation analysis (CCA)

- Defines a similarity between two random vectors (X, Y) via correlations between their projections
- We use it to measure similarity between layer representations and some external vector variable

$$\rho_1 = \max_{a_1, b_1} corr(a_1^T X, b_1^T Y)$$

$$\rho_\mathrm{k} = \max_{a_k, b_k} corr\left(a_k^T X, b_k^T Y\right) \quad s.t. \quad a_k^T C_{xx} a_i = 0, \quad b_k^T C_{yy} b_i = 0 \quad \forall\, i < \ k$$

$$\mathrm{score} = \frac{1}{d} \sum_k \rho_k$$

- Score = 1 if the two “views” are linearly related, score = 0 if they have no correlated components
- Has a closed-form solution via an SVD
- We use a variant, projection-weighted CCA (PWCCA) [Morcos+ 2018]
- Some nice properties of CCA: lightweight; scale-invariant; applicable to any linguistic variable

Layer-wise analysis: Some notes

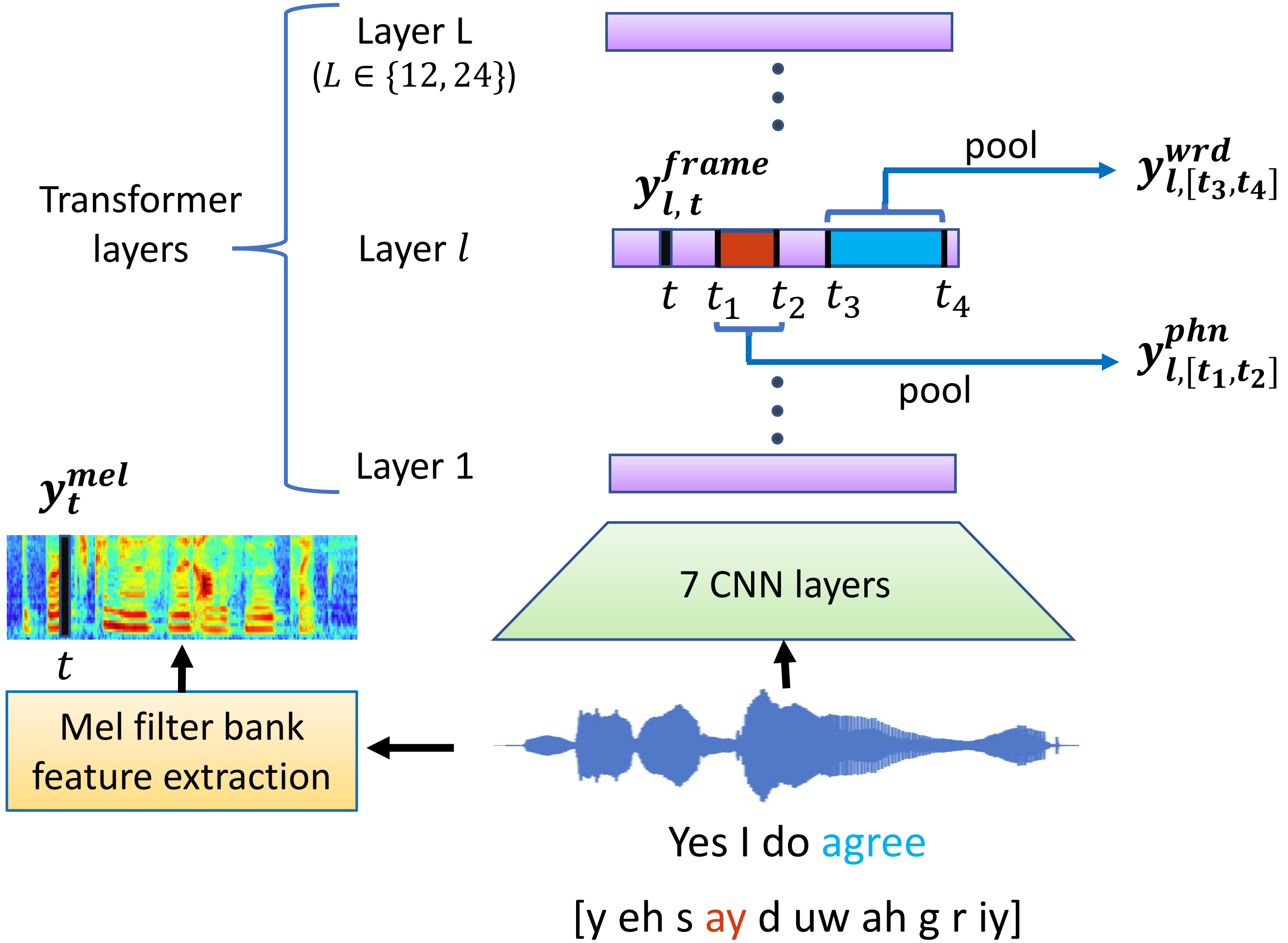
This analysis does not:

- Study the information content in an entire layer, only over short segments (frame/phone/word segments)
- Reveal precisely how a model will perform on some task
- Measure mutual information (even when we say “information”…)

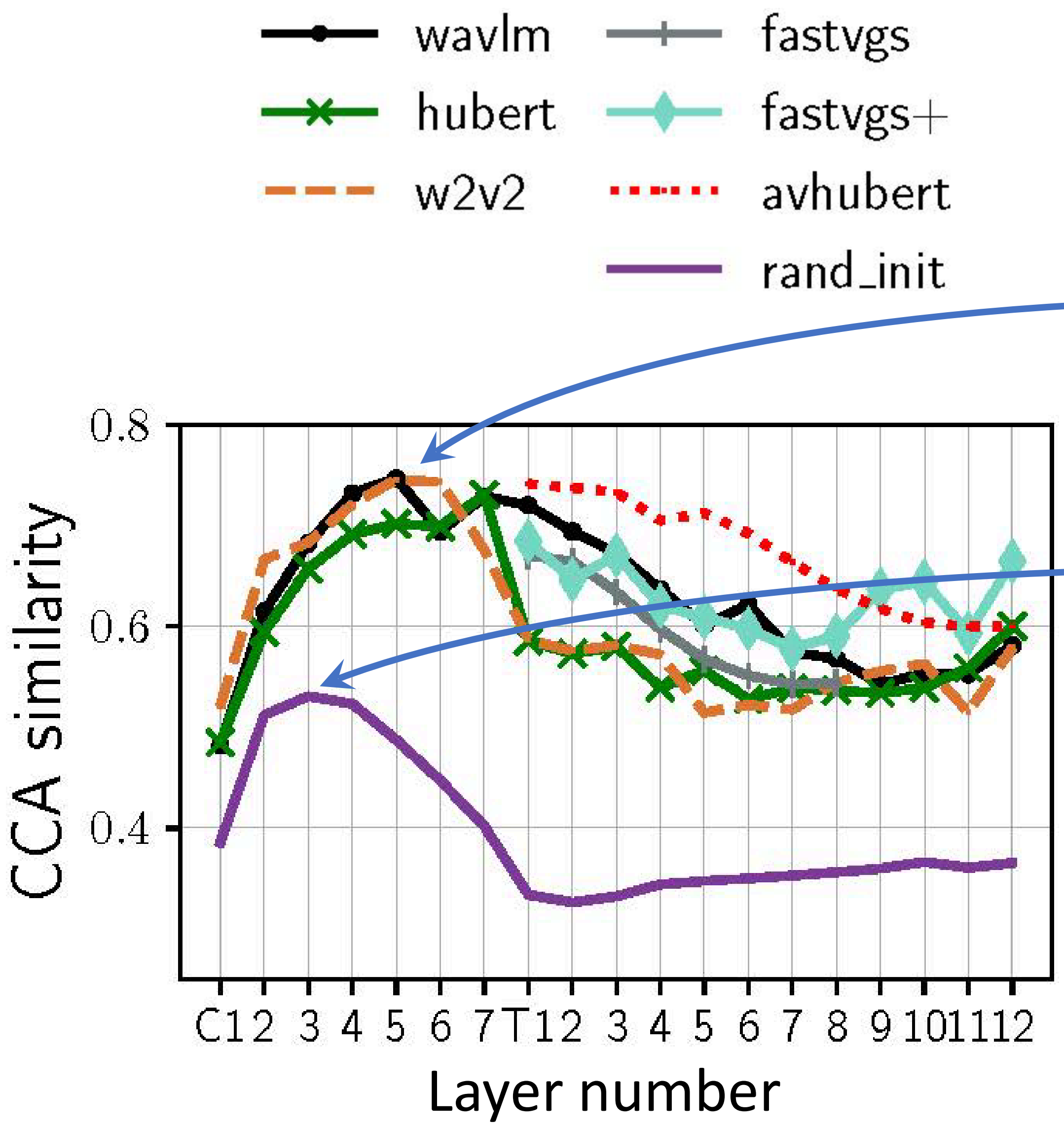
This analysis is linear

- To emulate typical use cases

Every form of analysis has its quirks
➔ important to corroborate findings in multiple ways



Similarity with spectral features



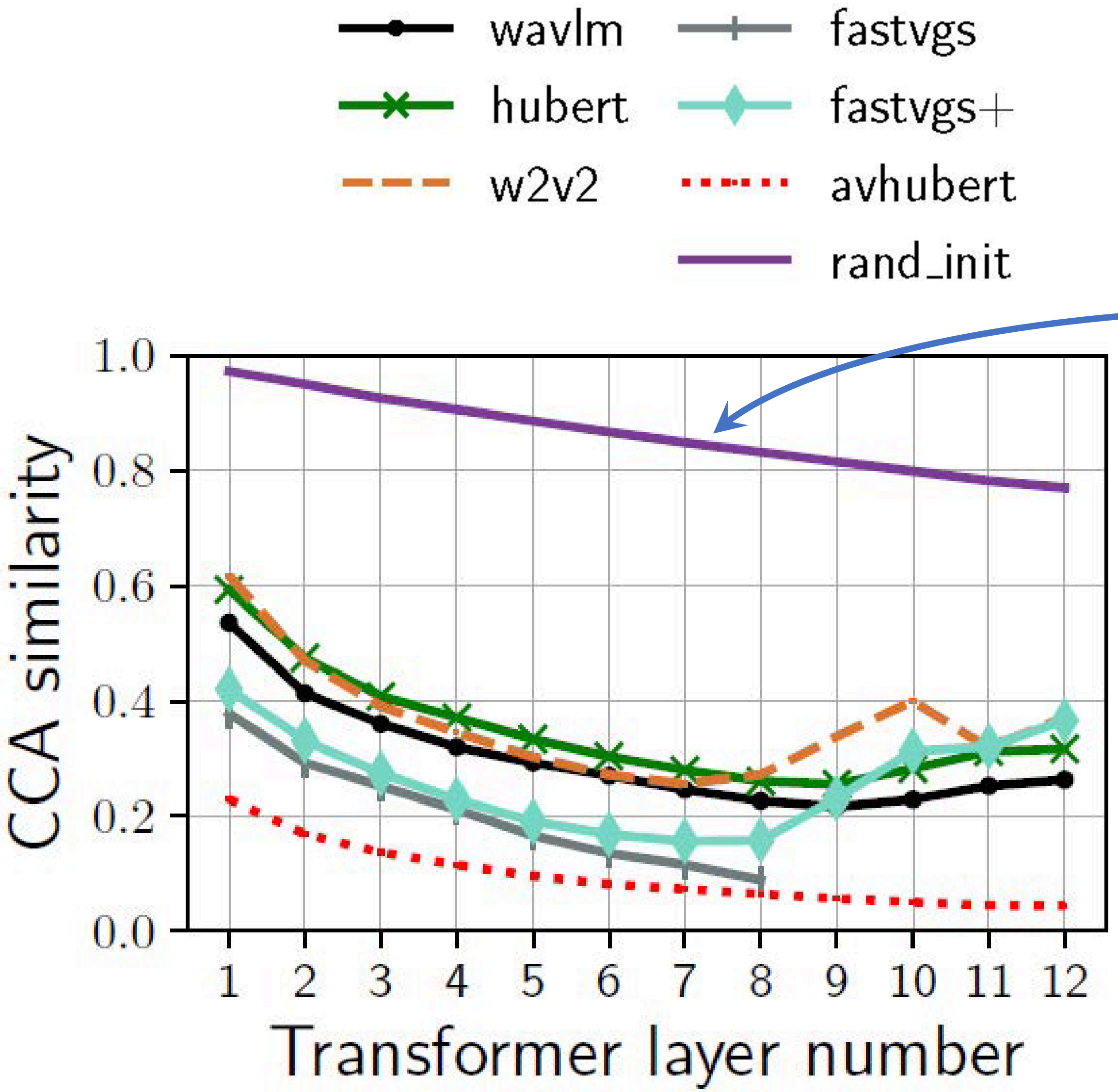
High correlation with spectral features around layers C5-C7: Do we need the raw audio? [1-2]

Random convolutional layers also somewhat correlated with spectral features, but less so

[1] Wu et al., “Performance-efficiency trade-offs in unsupervised pre-training for speech recognition,” ICASSP 2022

[2] Lin et al., “MelHuBERT: A simplified HuBERT on Mel spectrogram,” ASRU 2023

Similarity with "local" features (output of CNN)

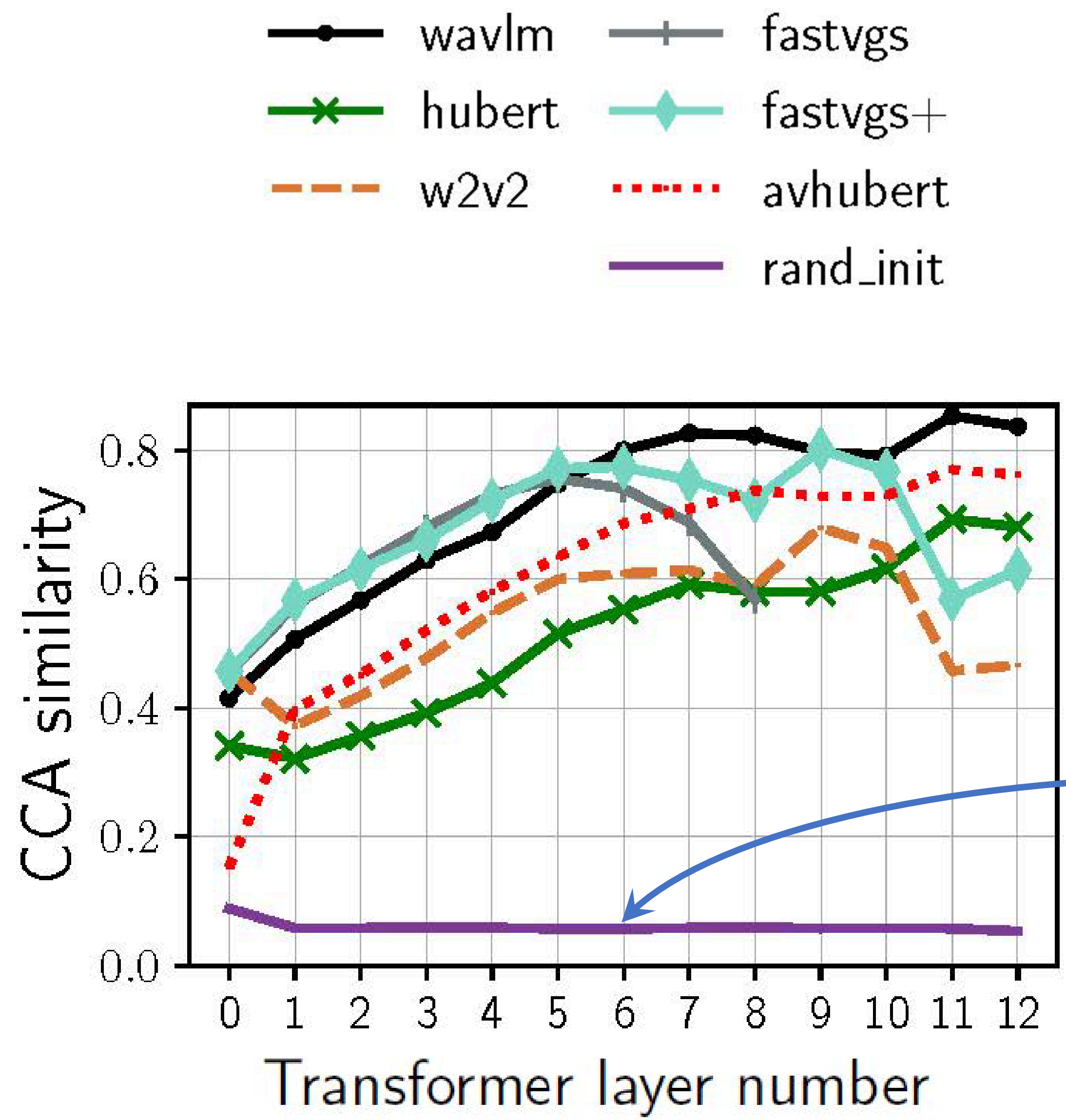


Random model just slowly forgets the input

Audio-only models have autoencoder-like behavior

Multi-modal models don't attempt to reconstruct the input

Phonetic content (similarity with phone 1-hot vectors)



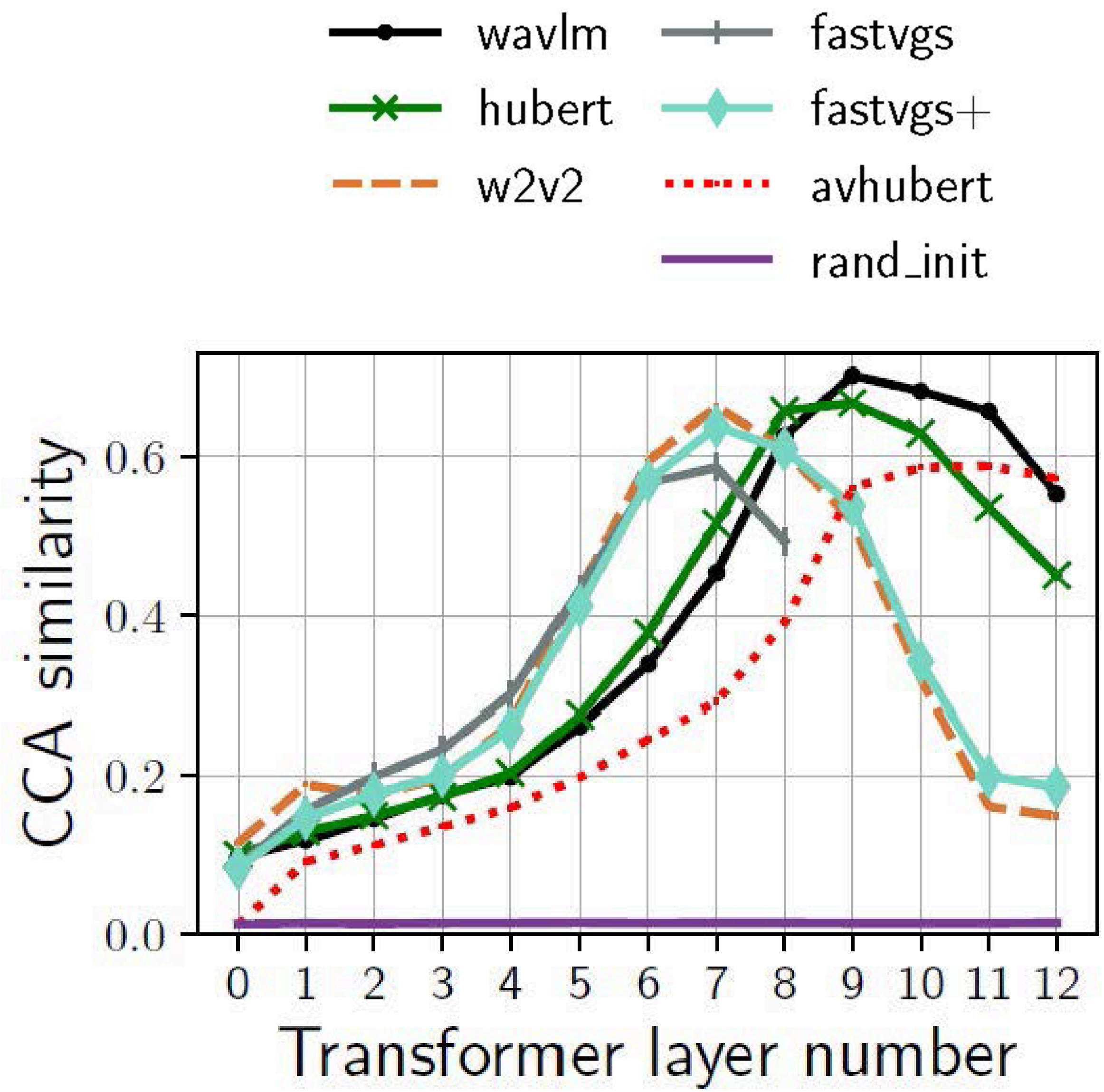
Models trained to predict phone-like clusters retain phonetic information in highest layers

Models that predict more local features "forget" phonetic information in highest layers*

Random model contains no phonetic information

* "Forget" = "transform into a less linearly accessible form"

Word content (similarity with word 1-hot vectors)



Each model has a peak in word content at some intermediate layer

- Higher layer for certain pretext tasks than others

Implications: Improved fine-tuning?

The final few layers are less stable, and encode less phone and word identity information
➔ Re-initialize final layers before fine-tuning?

Results with wav2vec 2.0 on LibriSpeech test-other set:

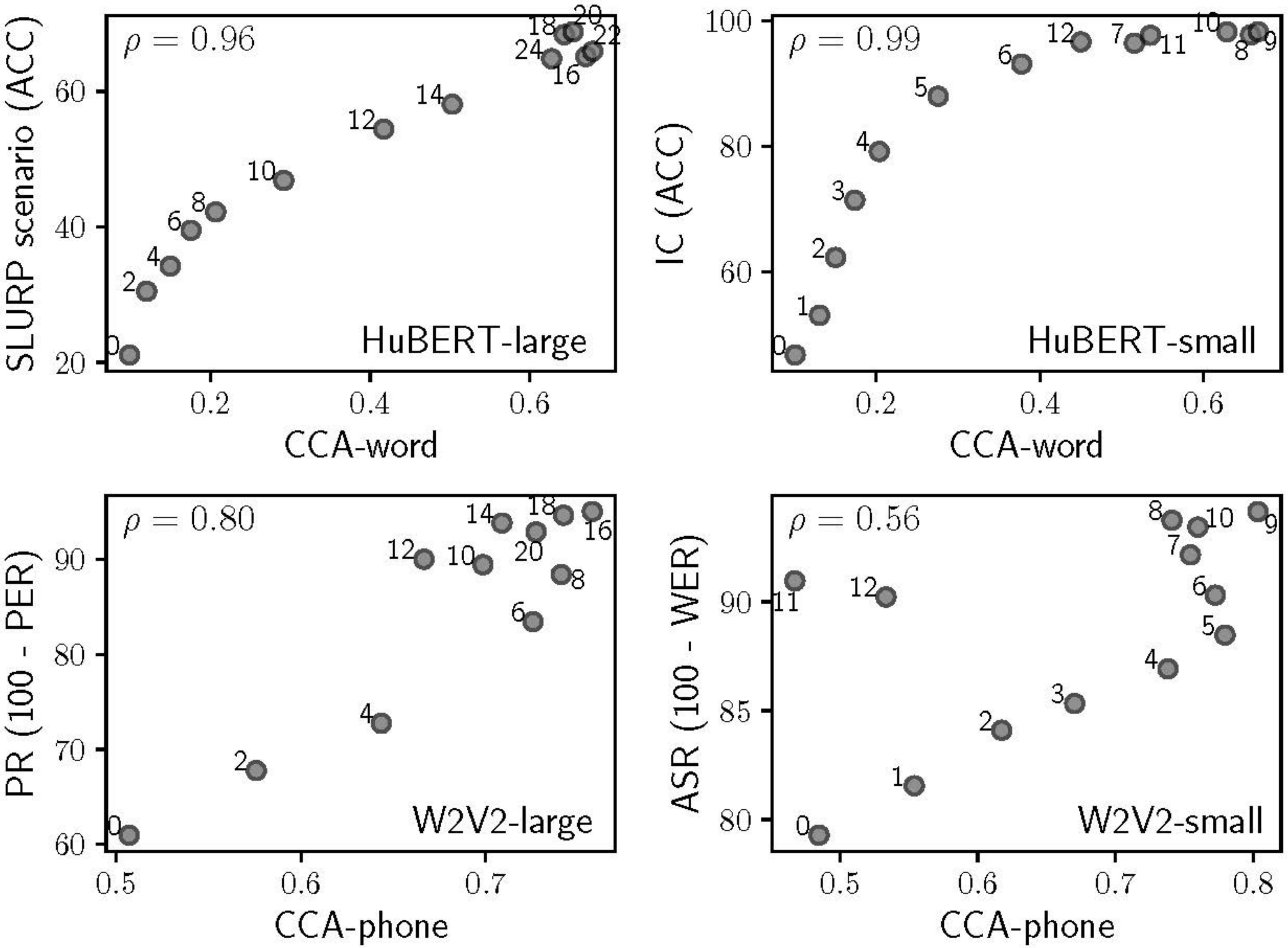
Fine-tuning set size	*k*	Standard fine-tuning → Re-init last *k* layers (WER, ↓)
10m	3	56.7 → 51.8
1h	1	29.9 → 29.3
10h	1	20.6 → 19.4

Re-initialization of final layers improves performance, especially in the lowest-resource setting

Implications: Correlation with downstream performance?

Yes

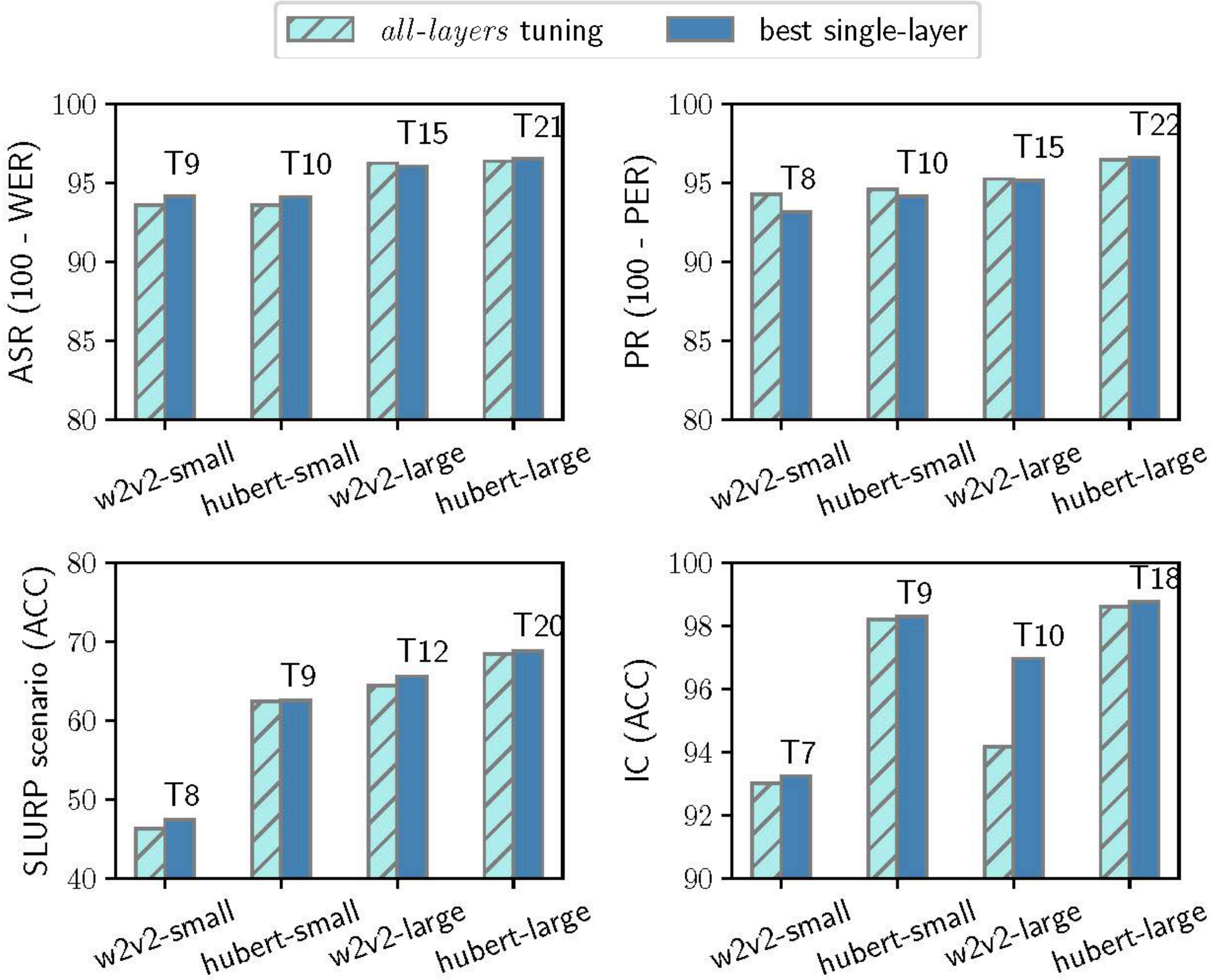
- CCA-word correlates well with speech recognition, spoken language understanding tasks
- CCA-phone correlates well with recognition performance
- Note: Layer weights in fine-tuning do not correlate nearly as well with task performance
  - ρ = 0.66 for layer weights, 0.90 for CCA-word



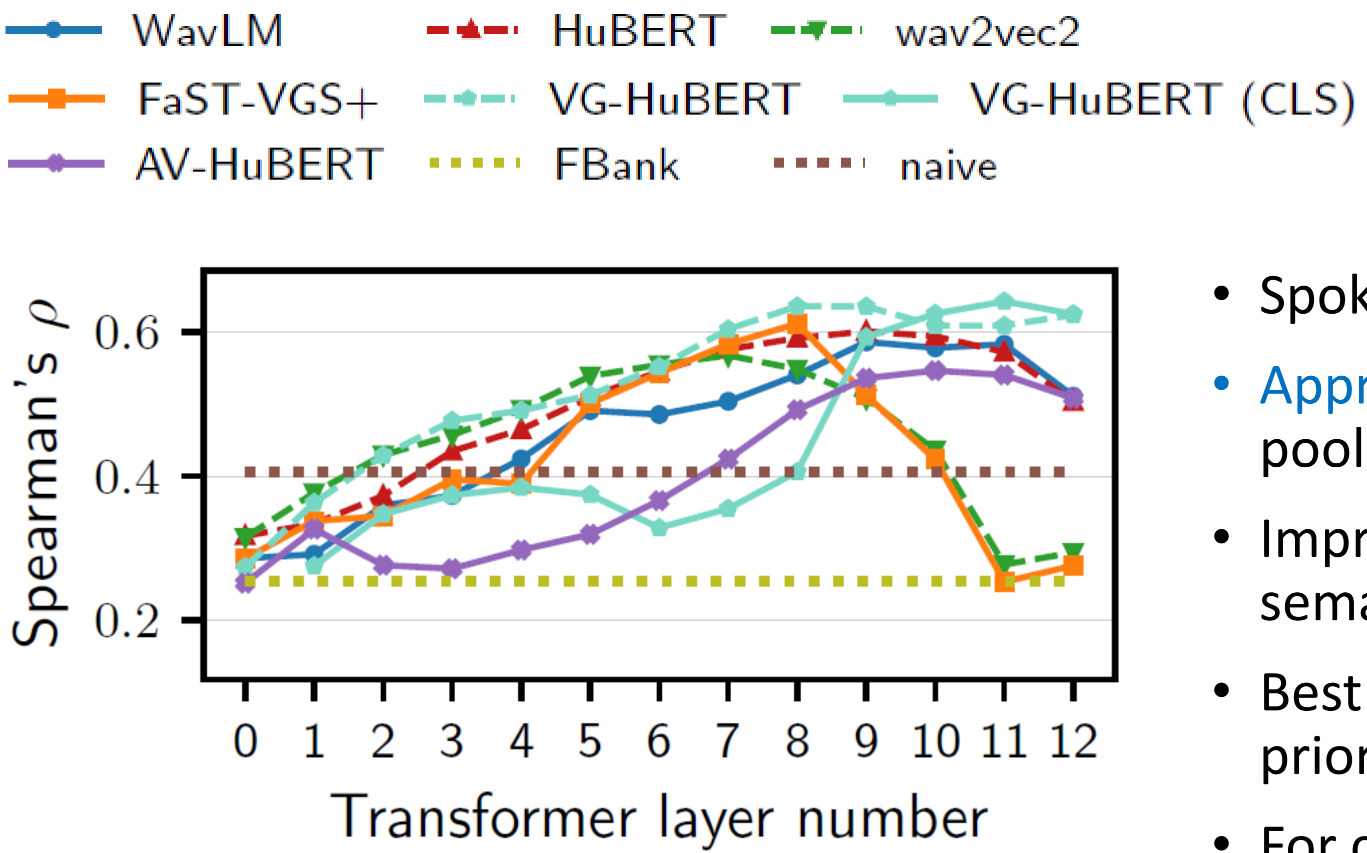
Implications: Do we need all layers for downstream tasks?

The higher layers often contain less linguistic information
→ Do we really need them?

- No (compared to using weighted sum of layers, as in SUPERB)



How much “semantics” have we learned?



- Spoken sentence similarity task
- Approach: Cosine similarity between mean-pooled representations
- Improvement over naïve baseline suggests some semantics is encoded
- Best results (with VG-HuBERT) improve on best prior results (Merkx et al. 2021, Zhu et al. 2022)
- For comparison, a text oracle (based on RoBERTa) has ρ = 0.77 (Zhu et al. 2022)

Discussion

We are starting to understand what is inside self-supervised speech models…

- Layer-specific information depends on the pretext task
- Intermediate layers often contain the "deepest" linguistic information
  - Much like some text language models [Voita et al. 2019, Blevins et al. 2018]
- Speech encoders seem to learn *some* semantics, not as much as text models
- Analysis suggests ways to improve fine-tuning & save some compute

What about larger models?

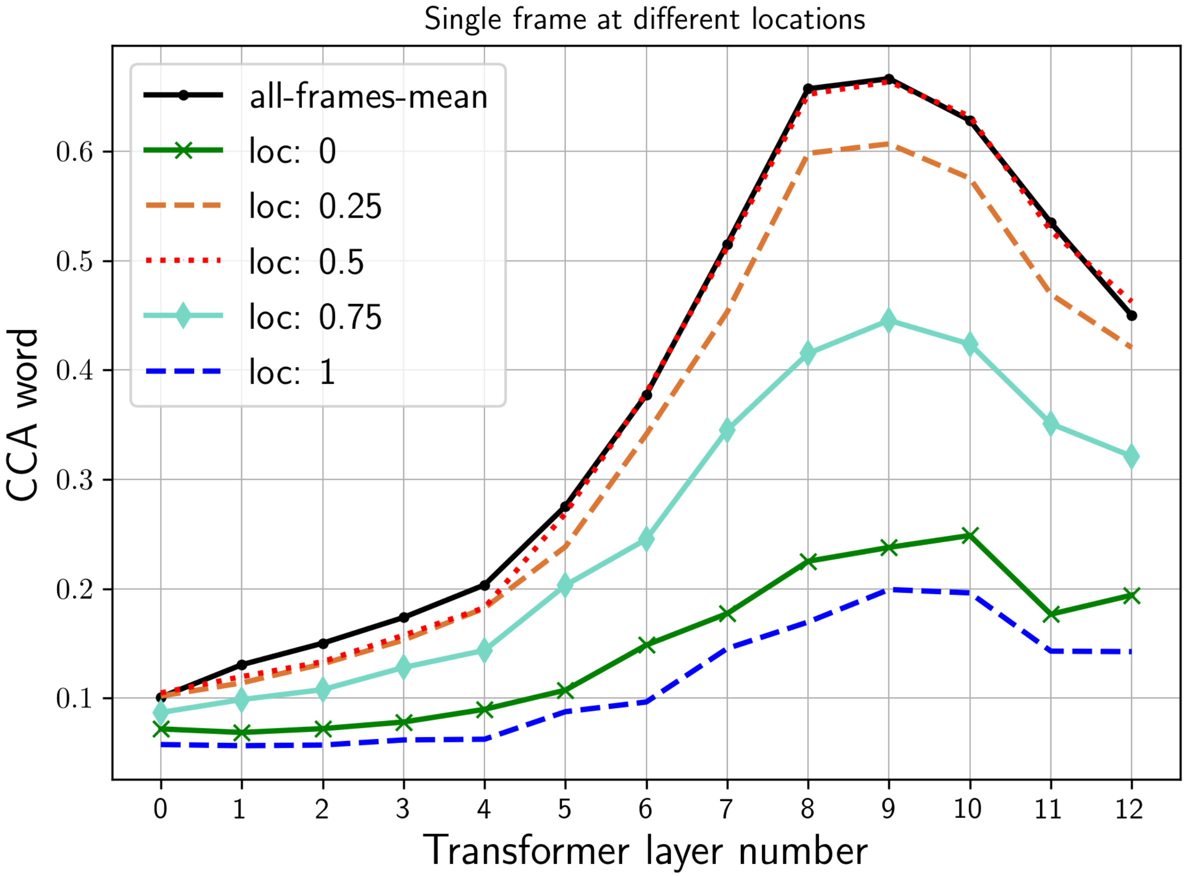
- Layer-wise behavior similar between Small and Large models
- For downstream tasks, Large models are *not always better* than Small

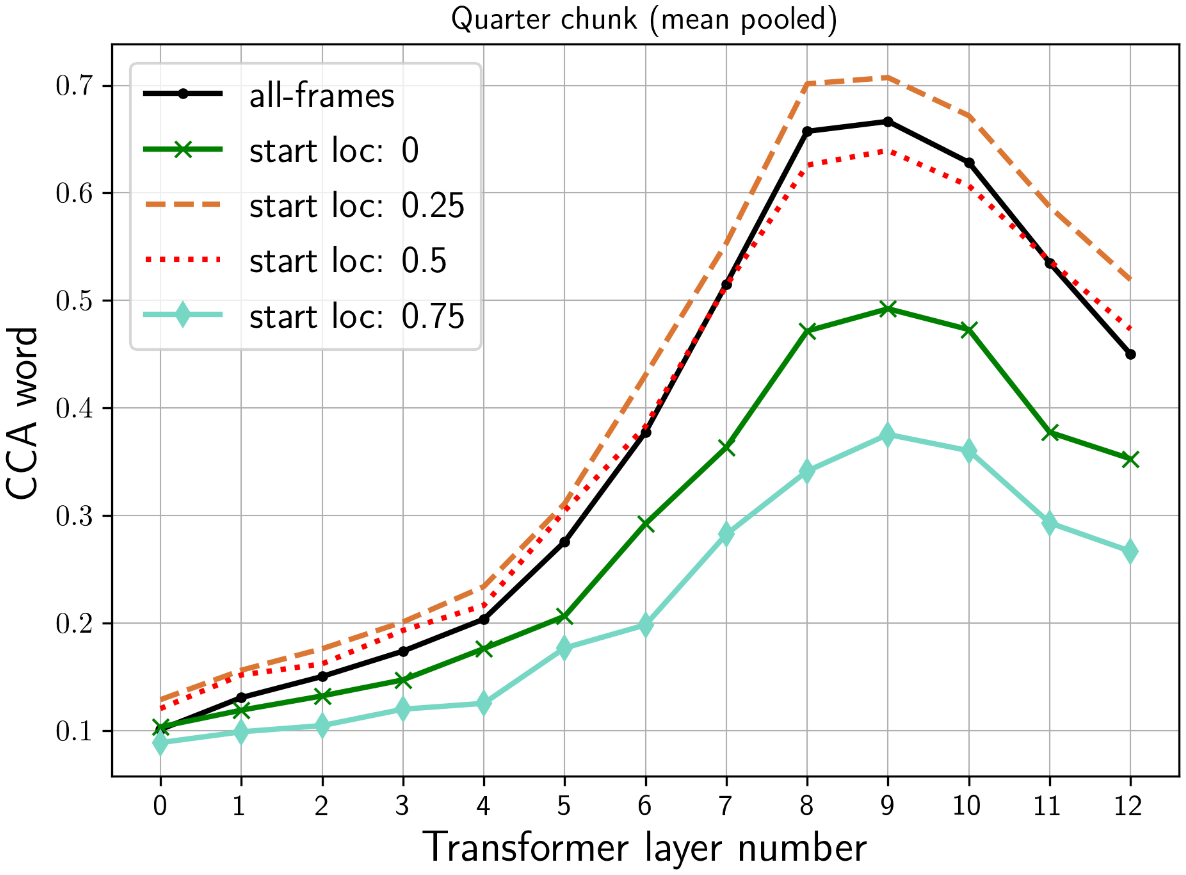
There is a lot more to do…

- What about supervised models (e.g. Whisper)?
- Should we be trying to make pre-trained models more semantic? Connect them to text LLMs? Or learn large autoregressive speech LLMs from scratch?

Try your own analysis! https://github.com/ankitapasad/layerwise-analysis

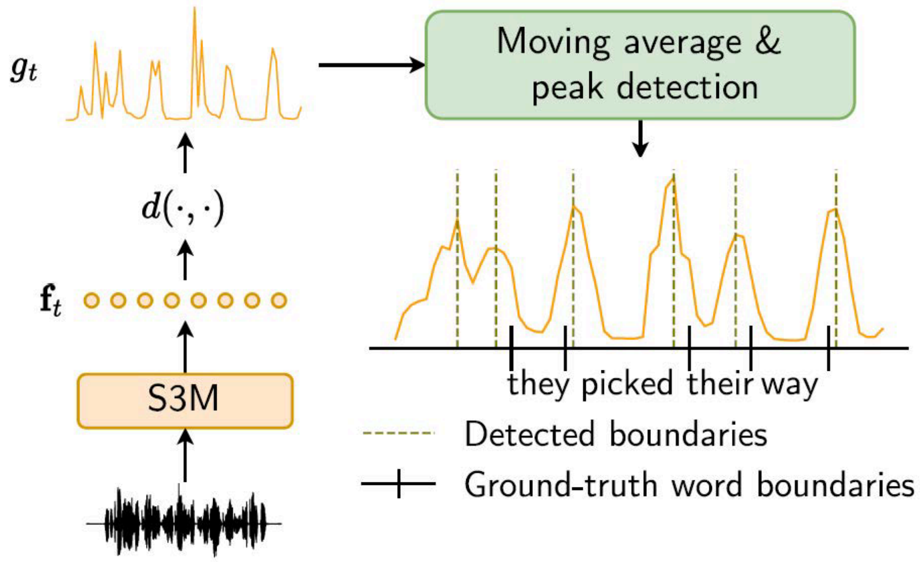
Digging deeper: How local is word information?

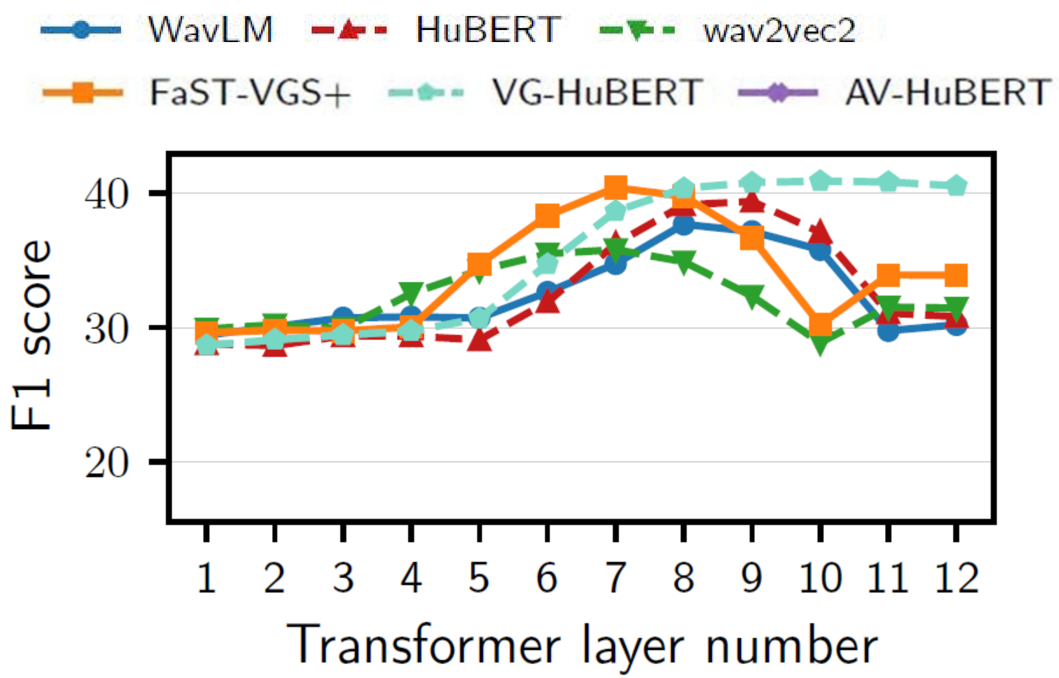


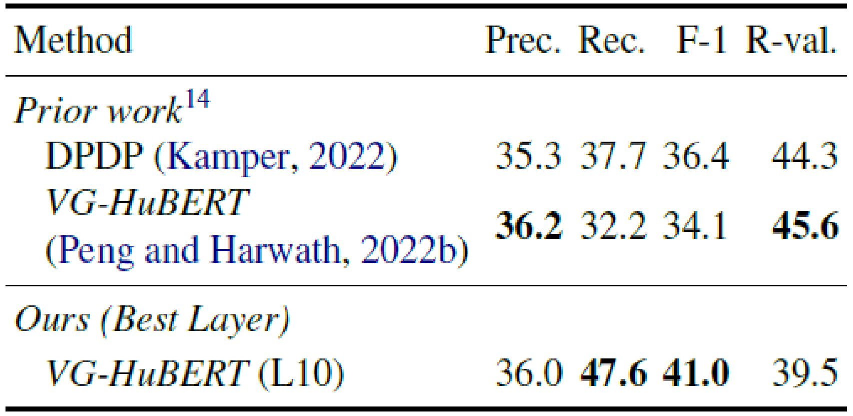


- Central frame of each word contains as much word information as mean-pooled embedding
- Word information concentrated earlier in the segment (loc: 0.25) rather than later (loc: 0.75)
- Whole-word pooling is close to best segment representation, but could do slightly better

Results on simple tasks with nonparametric models: Word segmentation



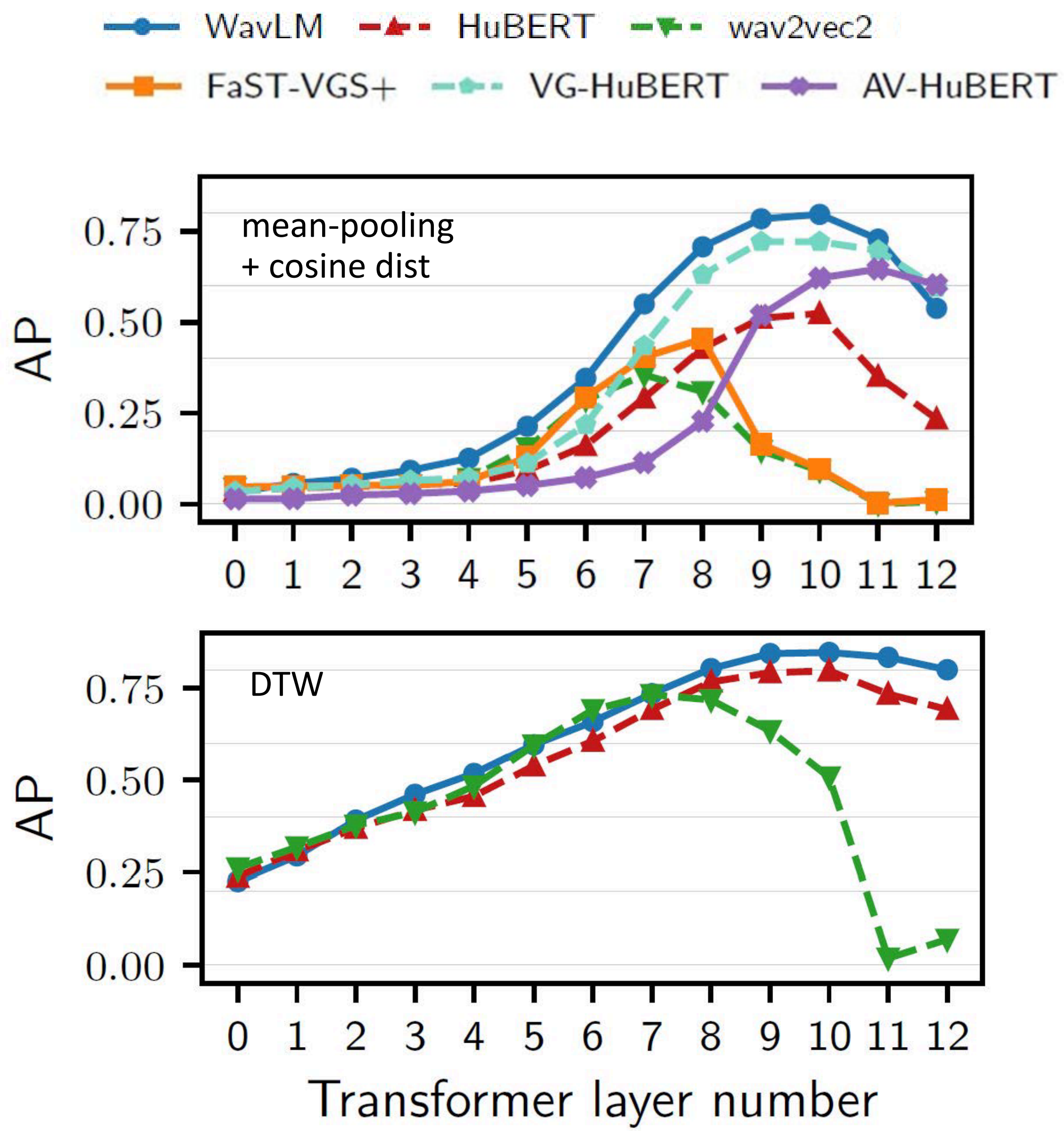




Method	Prec.	Rec.	F-1	R-val.
Prior work[14]				
DPDP (Kamper, 2022)	35.3	37.7	36.4	44.3
VG-HuBERT (Peng and Harwath, 2022b)	**36.2**	32.2	34.1	**45.6**
Ours (Best Layer)				
VG-HuBERT (L10)	36.0	**47.6**	**41.0**	39.5

- Out of the box, self-supervised models compete with or outperform more complex methods
- (But the problem is far from solved)

Results on simple tasks with nonparametric models: Acoustic word discrimination

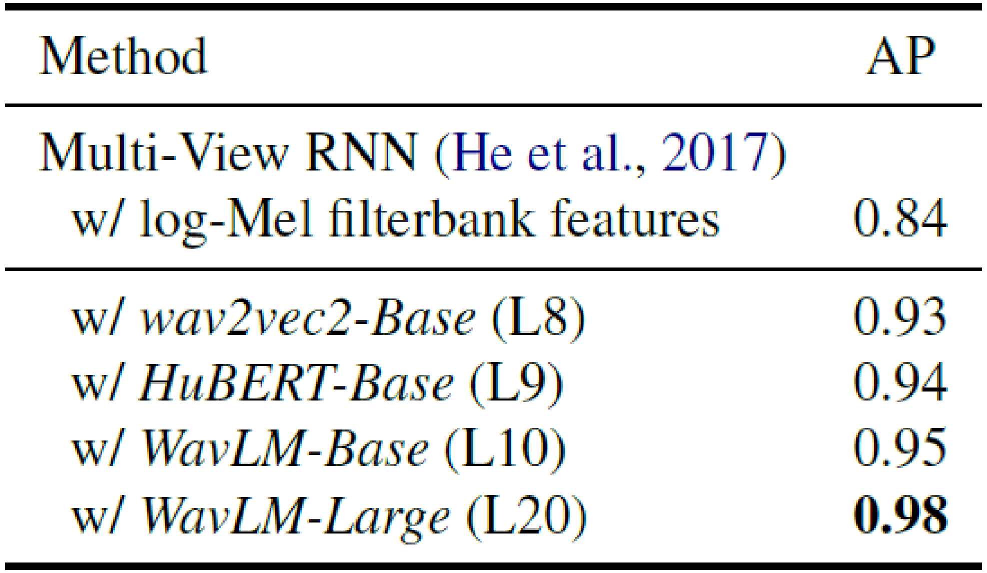


- Self-supervised models do quite well out of the box!
- DTW helps a lot

  ➔ mean-pooling is not optimal (similar to findings of Sanabria et al. 2023)

Results on simple tasks: Acoustic word discrimination (2)

- What if we have a small amount (~100 minutes) of labeled speech?
- Results (average precision) on Switchboard word discrimination benchmark:

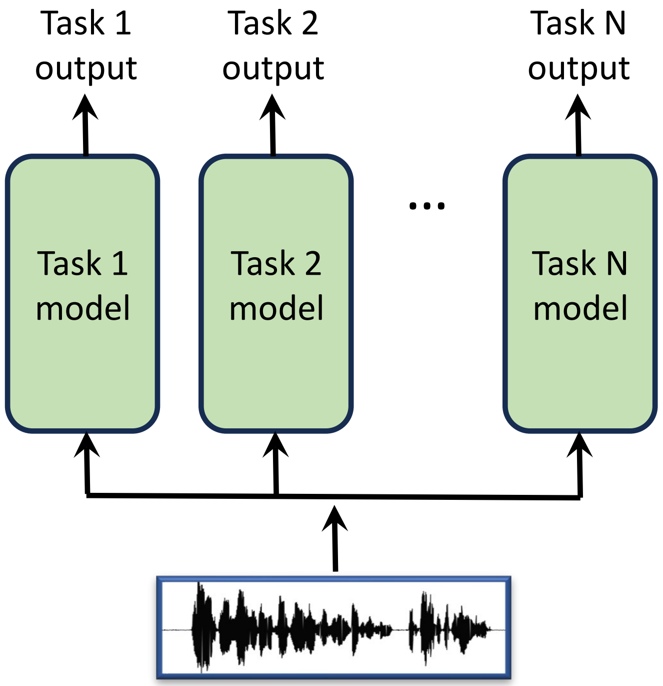


Method	AP
Multi-View RNN (He et al., 2017)	
w/ log-Mel filterbank features	0.84
w/ *wav2vec2-Base* (L8)	0.93
w/ *HuBERT-Base* (L9)	0.94
w/ *WavLM-Base* (L10)	0.95
w/ *WavLM-Large* (L20)	**0.98**

- Self-supervised models + minimal supervision achieve close to perfect performance!

Beyond speech encoders

The task-specific model era (- 2020)



Task 1 output

Task 2 output

Task N output

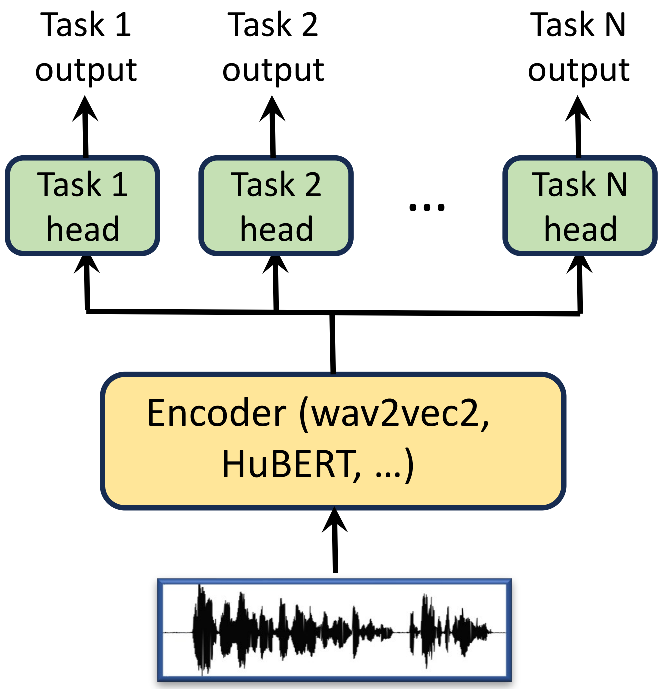
Task 1 model

Task 2 model

…

Task N model

The speech encoder era (2020 -)



Task 1 output

Task 2 output

Task N output

Task 1 head

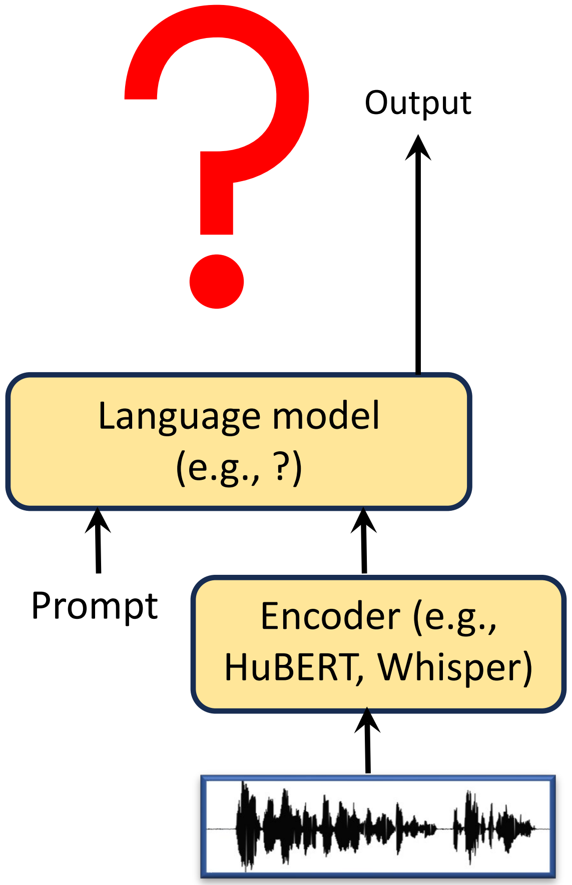
Task 2 head

…

Task N head

Encoder (wav2vec2, HuBERT, …)

The spoken large language model era (2024? -)



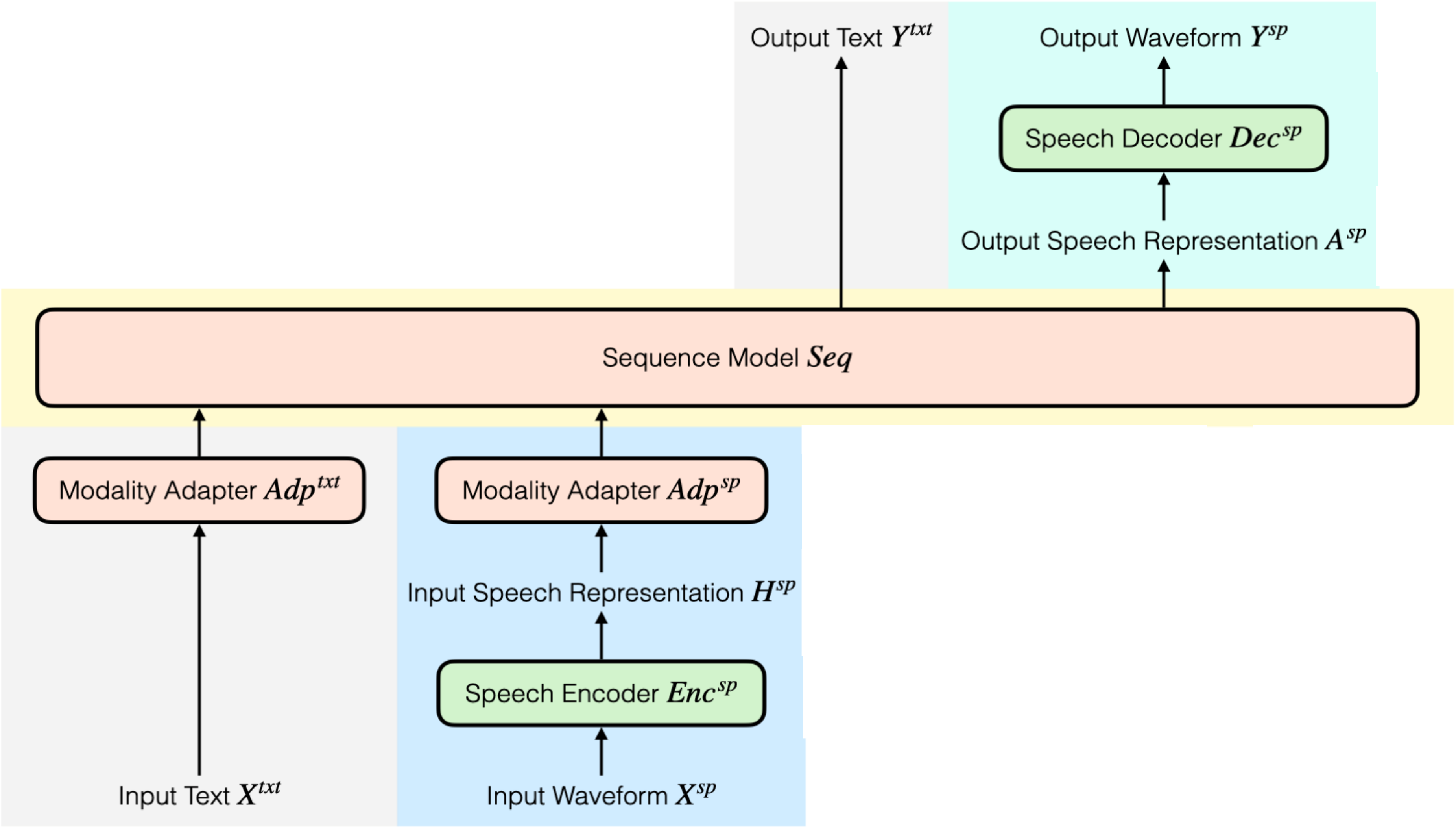
Output

Language model (e.g., ?)

Prompt

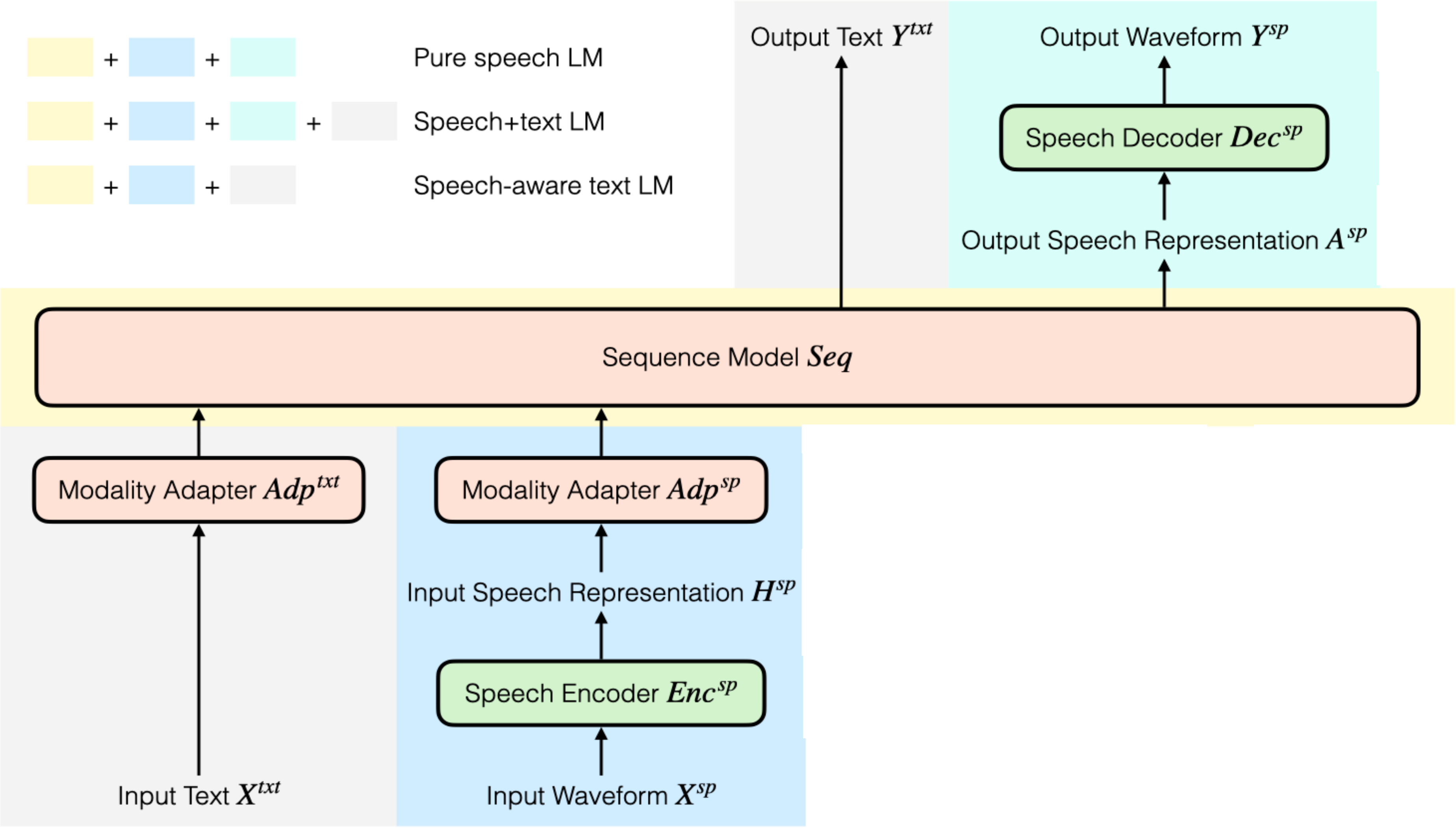
Encoder (e.g., HuBERT, Whisper)

Spoken language models


Output Text Y^{txt}
Output Waveform Y^{sp}
Speech Decoder Dec^{sp}
Output Speech Representation A^{sp}
Sequence Model Seq
Modality Adapter Adp^{txt}
Modality Adapter Adp^{sp}
Input Speech Representation H^{sp}
Speech Encoder Enc^{sp}
Input Text X^{txt}
Input Waveform X^{sp}

Arora et al., "On the landscape of spoken language models: A comprehensive survey," arXiv:2504.08528.

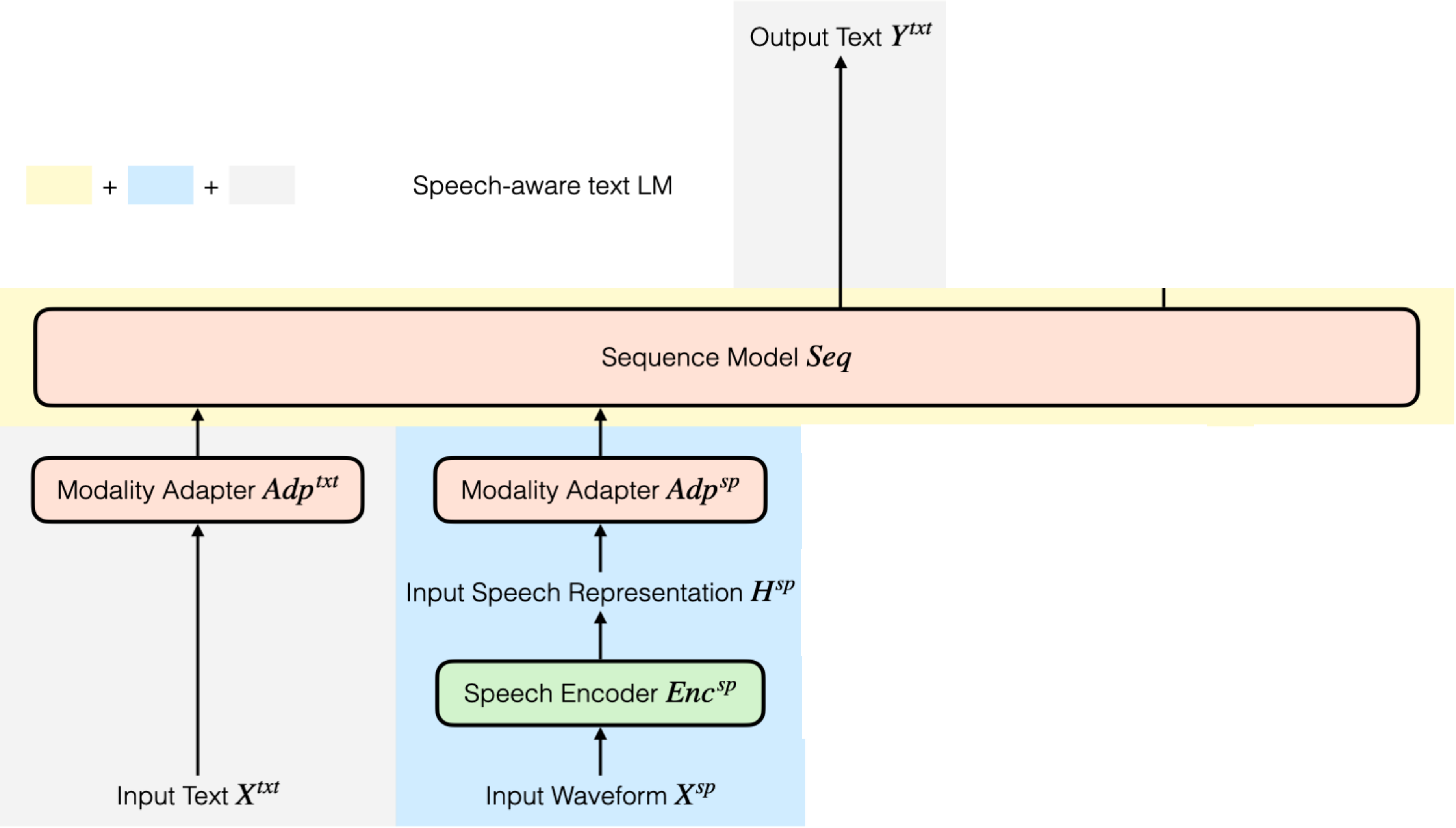
Spoken language models



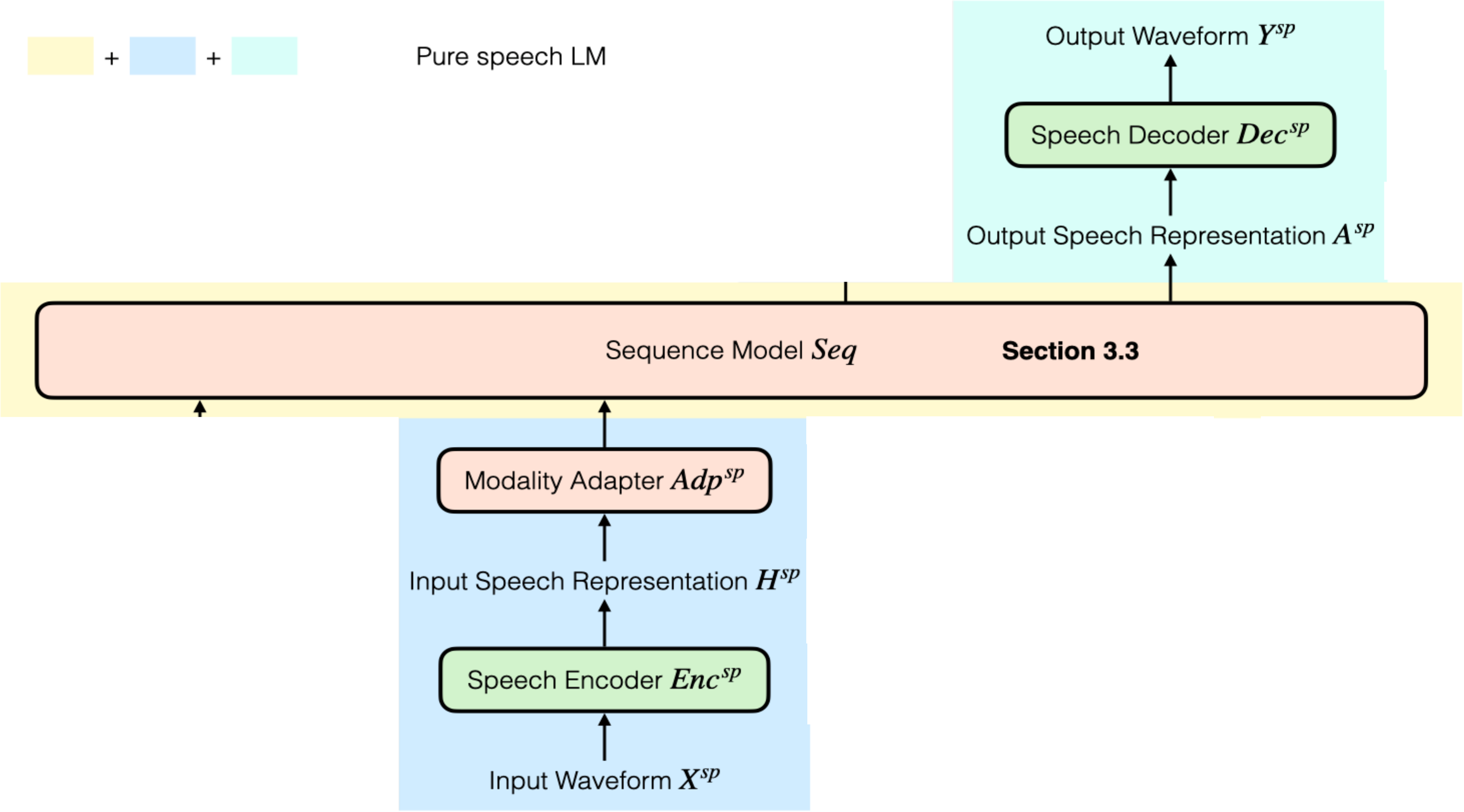
🟨 + 🟦 + 🟩 Pure speech LM
🟨 + 🟦 + 🟩 + ⬜ Speech+text LM
🟨 + 🟦 + ⬜ Speech-aware text LM
Output Text Y^{txt}
Output Waveform Y^{sp}
Speech Decoder Dec^{sp}
Output Speech Representation A^{sp}
Sequence Model Seq
Modality Adapter Adp^{txt}
Modality Adapter Adp^{sp}
Input Speech Representation H^{sp}
Speech Encoder Enc^{sp}
Input Text X^{txt}
Input Waveform X^{sp}

Arora et al., "On the landscape of spoken language models: A comprehensive survey," arXiv:2504.08528.

Spoken language models


🟨 + 🟦 + ⬜
Speech-aware text LM
Output Text $\boldsymbol{Y}^{txt}$
Sequence Model $\boldsymbol{Seq}$
Modality Adapter $\boldsymbol{Adp}^{txt}$
Modality Adapter $\boldsymbol{Adp}^{sp}$
Input Speech Representation $\boldsymbol{H}^{sp}$
Speech Encoder $\boldsymbol{Enc}^{sp}$
Input Text $\boldsymbol{X}^{txt}$
Input Waveform $\boldsymbol{X}^{sp}$

Spoken language models



🟨 + 🟦 + 🟦 Pure speech LM

Output Waveform Y^{sp}

Speech Decoder Dec^{sp}

Output Speech Representation A^{sp}

Sequence Model Seq

Section 3.3

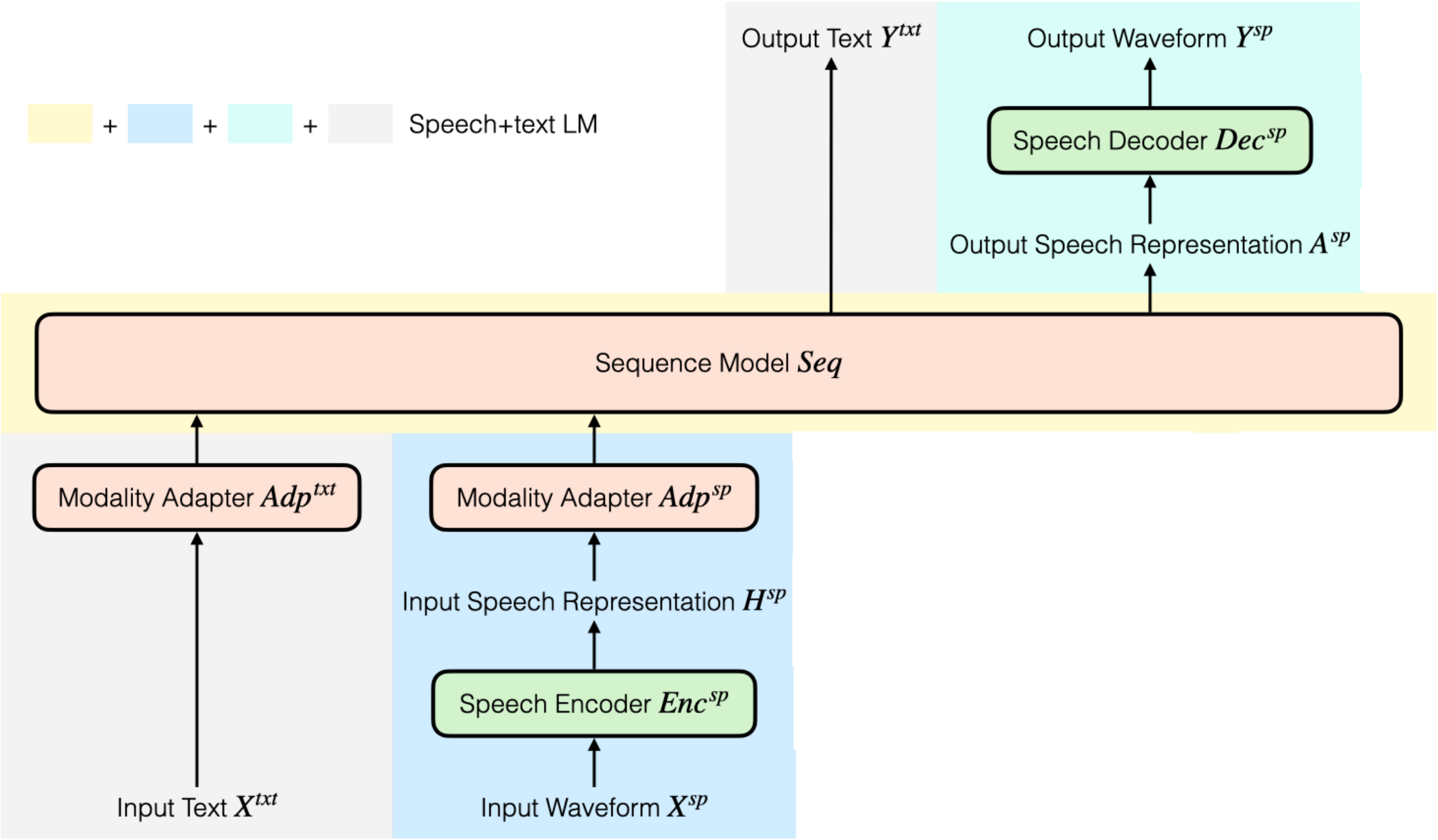
Modality Adapter Adp^{sp}

Input Speech Representation H^{sp}

Speech Encoder Enc^{sp}

Input Waveform X^{sp}

Spoken language models


+ + + Speech+text LM
Output Text Y^{txt}
Output Waveform Y^{sp}
Speech Decoder Dec^{sp}
Output Speech Representation A^{sp}
Sequence Model Seq
Modality Adapter Adp^{txt}
Modality Adapter Adp^{sp}
Input Speech Representation H^{sp}
Speech Encoder Enc^{sp}
Input Text X^{txt}
Input Waveform X^{sp}

Spoken language models: Where is the state of the art?

Example benchmark: Dynamic-SUPERB

- An evolving collection of ~180 speech + audio tasks contributed by 76 academic + industry researchers
- Only "understanding" tasks: audio in, text out
- Each task = a text instruction + an audio input + a desired output
- Evaluated by GPT-4o as a judge

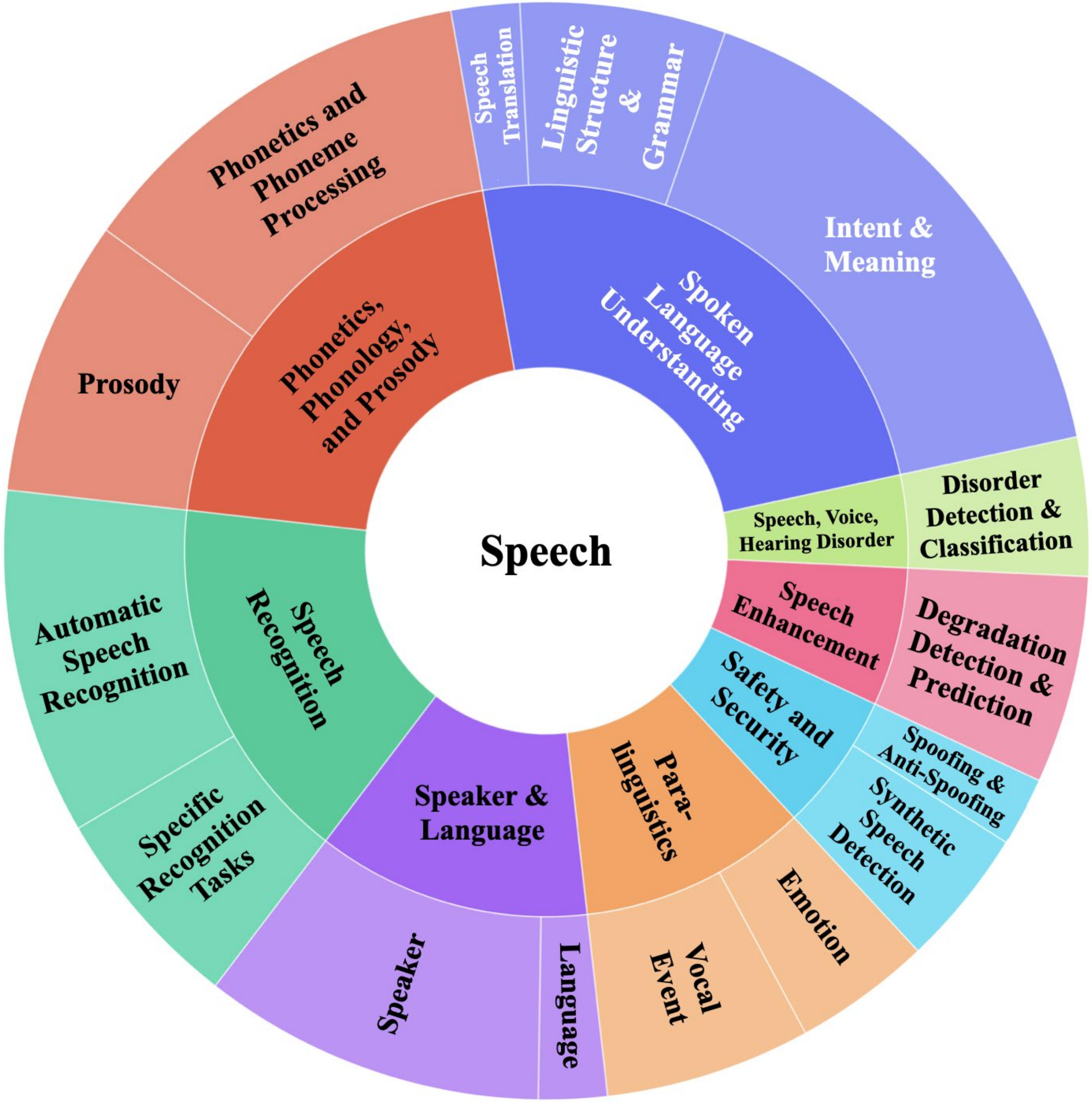
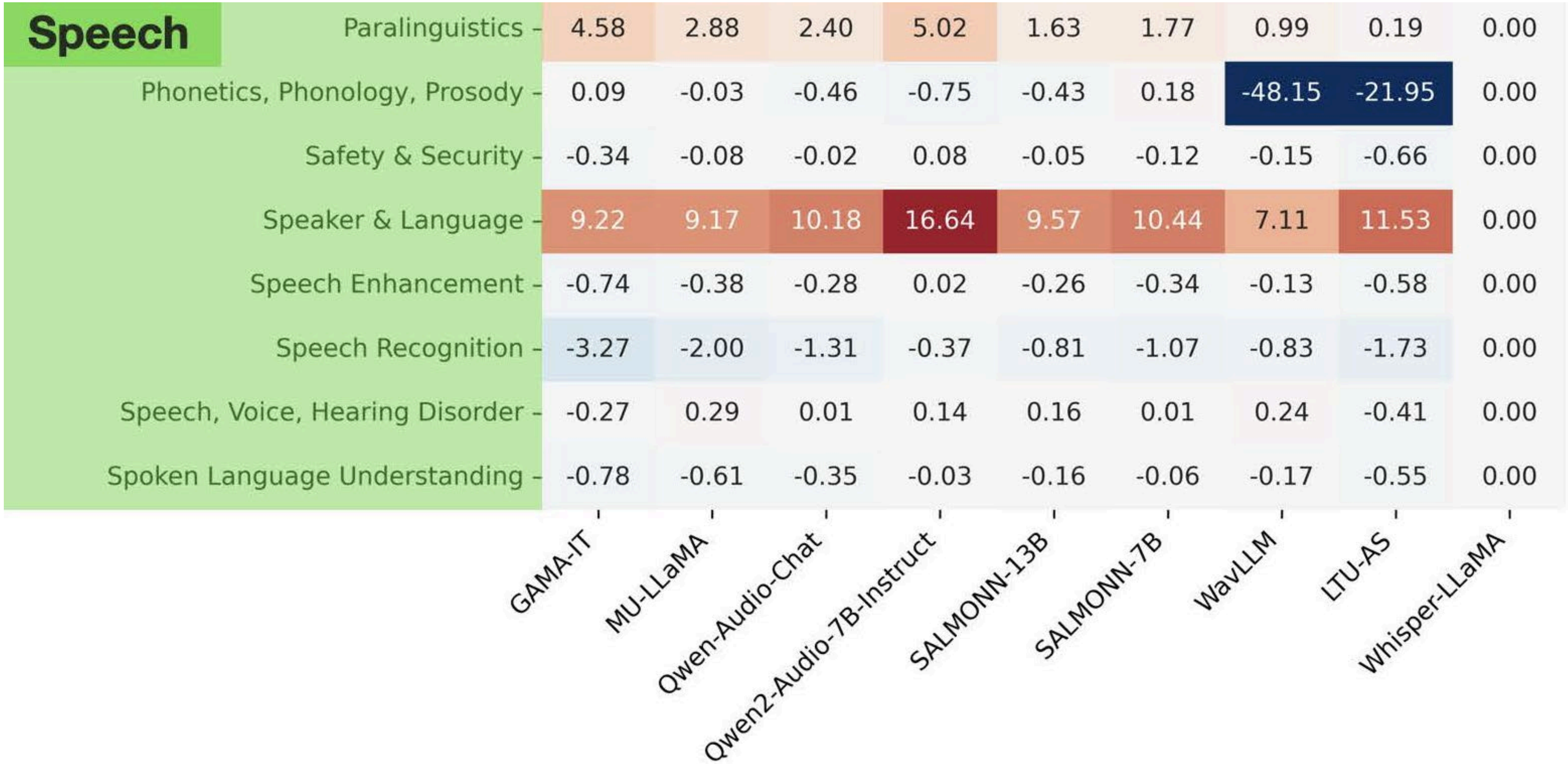

Speech
Phonetics, Phonology, and Prosody
Phonetics and Phoneme Processing
Prosody
Spoken Language Understanding
Speech Translation
Linguistic Structure & Grammar
Intent & Meaning
Speech, Voice, Hearing Disorder
Disorder Detection & Classification
Speech Enhancement
Degradation Detection & Prediction
Safety and Security
Spoofing & Anti-Spoofing
Synthetic Speech Detection
Para-linguistics
Emotion
Vocal Event
Speaker & Language
Language
Speaker
Speech Recognition
Automatic Speech Recognition
Specific Recognition Tasks

Fig. from C.-y. Huang et al., "Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities of Spoken Language Models with 180 Tasks," ICLR 2025.

Spoken language models: Where is the state of the art?

Main takeaway: Current SLMs do worse than a Whisper+Llama baseline on most “linguistic content” tasks, better on paralinguistics, speaker, or language tasks



Speech	GAMA-IT	MU-LLaMA	Qwen-Audio-Chat	Qwen2-Audio-7B-Instruct	SALMONN-13B	SALMONN-7B	WavLLM	LTU-AS	Whisper-LLaMA
Paralinguistics	4.58	2.88	2.40	5.02	1.63	1.77	0.99	0.19	0.00
Phonetics, Phonology, Prosody	0.09	-0.03	-0.46	-0.75	-0.43	0.18	-48.15	-21.95	0.00
Safety & Security	-0.34	-0.08	-0.02	0.08	-0.05	-0.12	-0.15	-0.66	0.00
Speaker & Language	9.22	9.17	10.18	16.64	9.57	10.44	7.11	11.53	0.00
Speech Enhancement	-0.74	-0.38	-0.28	0.02	-0.26	-0.34	-0.13	-0.58	0.00
Speech Recognition	-3.27	-2.00	-1.31	-0.37	-0.81	-1.07	-0.83	-1.73	0.00
Speech, Voice, Hearing Disorder	-0.27	0.29	0.01	0.14	0.16	0.01	0.24	-0.41	0.00
Spoken Language Understanding	-0.78	-0.61	-0.35	-0.03	-0.16	-0.06	-0.17	-0.55	0.00

Spoken language models

Spoken language models are now…

- The most active focus of the research community, in both industry and academia
- On their way to taking on the same role for spoken language that text LLMs have for written language

Spoken language models are not (yet)…

- Truly task-universal or language-universal

State of the art on most tasks is still self-supervised encoder + prediction head + fine-tuning

Some other driving forces

- Complementary contributions from both industry and academia
- The need to study (and improve) inclusiveness and safety

The end

Questions? Comments?