Few-Shot Spoken Language Understanding Via Joint Speech-Text Models

Chung-Ming Chien[1], Mingjiamei Zhang[2], Ju-Chieh Chou[1], Karen Livescu[1]
TTIC[1], The University of Chicago[2]

Overview

Few-shot Spoken Language Understanding (SLU)

- Assumes limited labeled speech data access, alongside more readily obtainable text data.

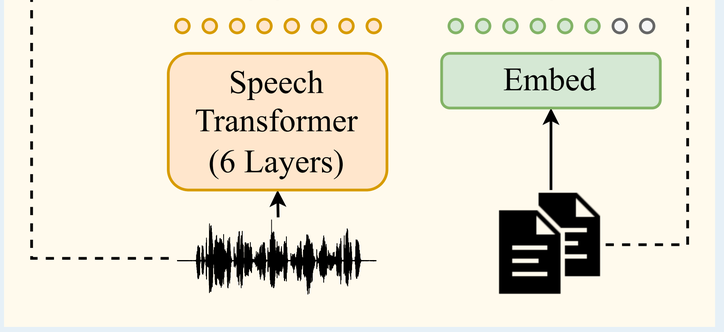
Speech-text models for **few-shot & zero-shot SLU**

- Match the performance of previous models with **0-20% of speech data**.

Analysis of hidden representations

- Explains the zero-shot text-to-speech transferability.
- Suggests **fine-tuning with bottom layers frozen**, which improves zero-shot performance.

Speech-Text Models

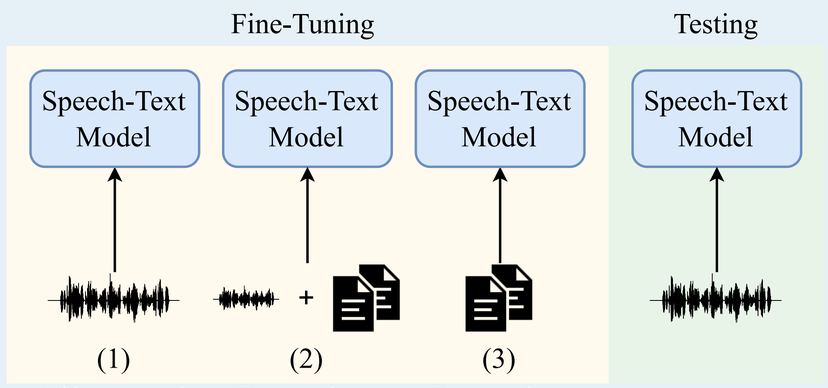


- Learn shared representations for speech & text.
- Improve ASR, speech translation, etc.

Conclusion

- **Speech-text models exhibit zero-shot transferability from text to speech** in SLU.
- Few-shot performance matches previous work trained with 5+ times more speech data.
- Bottom layers are task-agnostic and top layers are task-specific.
- **Freezing bottom layers enhances zero-shot performance.**

Few-Shot SLU

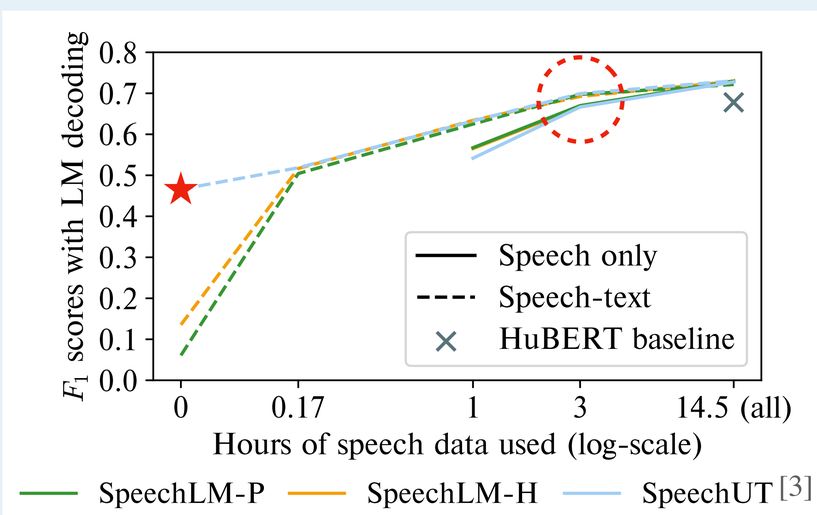


SLUE Benchmark [1]

Sentiment Analysis (Classification)

	Labeled Data		Speech-Only	Speech-Text	
	Speech	Text	HuBERT	Speech-LM-P	Speech-LM-H[2]
Baselines	1 hr	-		36.9	37.7
	12.8 hrs	-	43.0	45.6	45.3
Proposed	-	full		45.2	45.2
	10 mins	full		45.2	38.3
	1 hr	full		46.4	43.4

NER (sequence labeling)

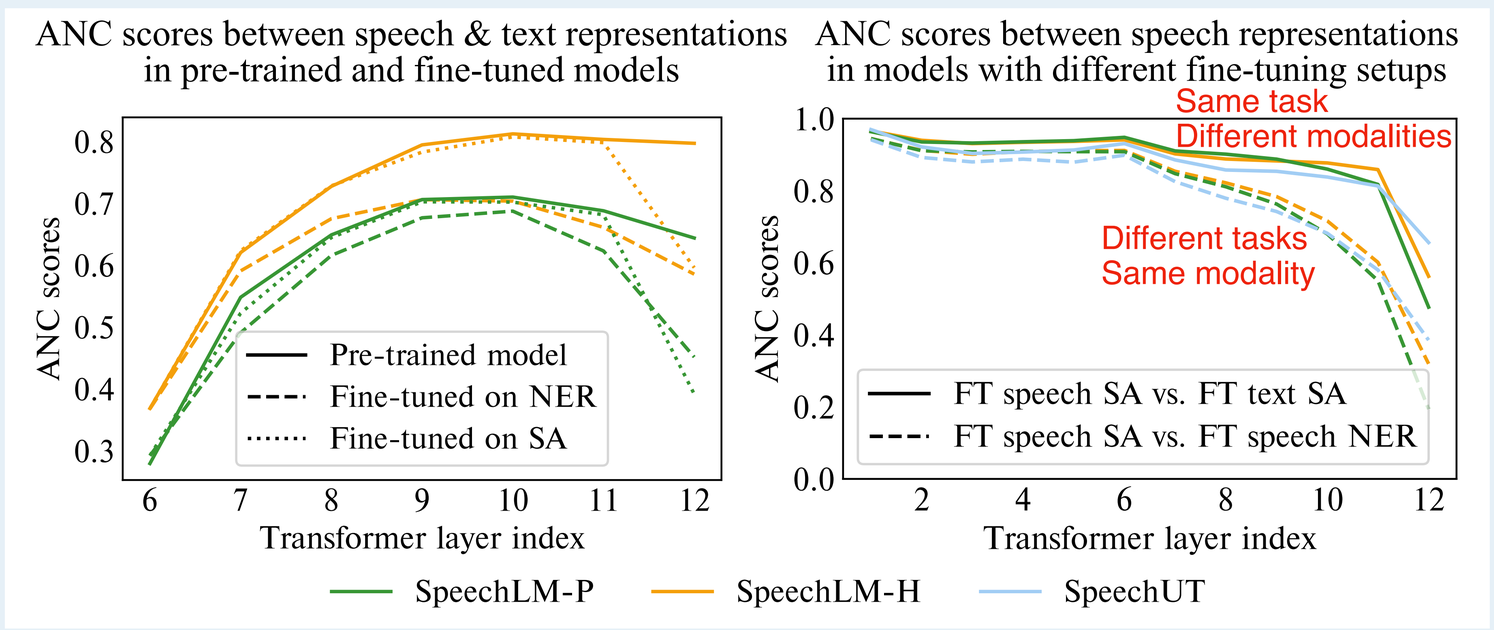


Analysis

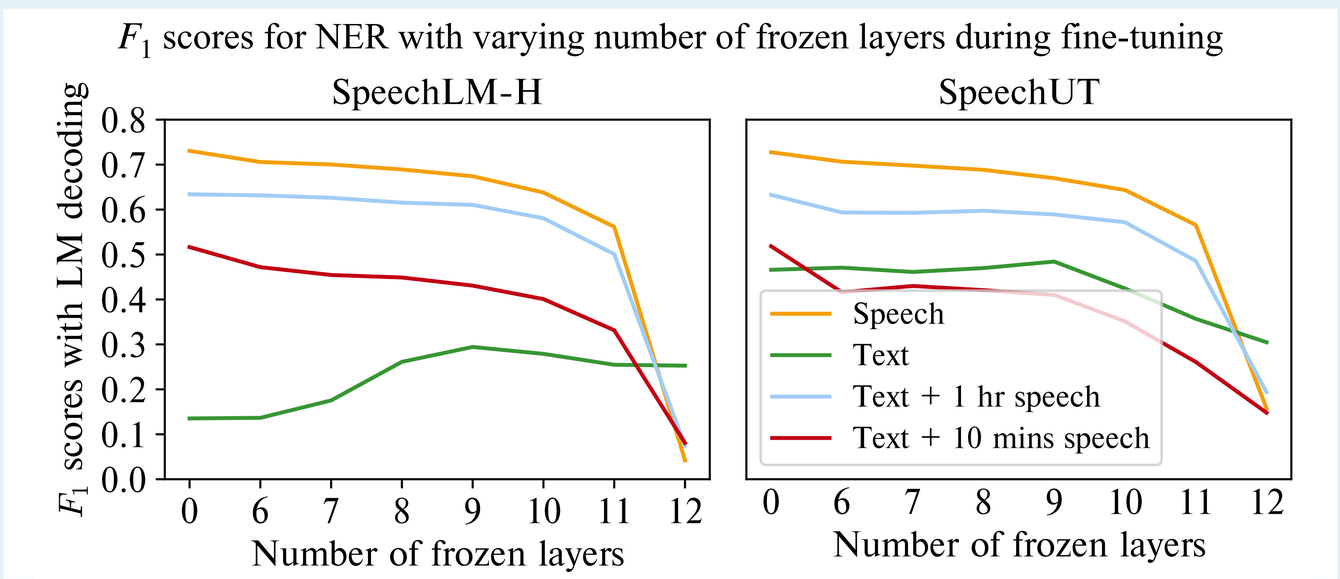
Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

$X, Y \in \mathbb{R}^d$: different views (e.g. text & speech) of the same data instance.



- Bottom layers align speech & text into a shared space.
- Fine-tuning only influences top layers.
- Tasks affects top layers more than input modalities
  → **top layers are task-specific**.



- Fine-tuning with **frozen bottom layers** leads to a slight performance reduction but improves **zero-shot cross-modal transfer**.

[1] S. Shon, et al, "SLUE: New benchmark tasks for spoken language understanding evaluation on natural speech," in ICASSP, 2022.

[2] Z. Zhang, et al, "SpeechLM: Enhanced speech pre-training with unpaired textual data," preprint arXiv:2209.15329, 2023.

[3] Z. Zhang, et al, "SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training," in EMNLP, 2022.

[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.