An Adapter-Based Unified Model for Multiple Spoken Language Processing Tasks

Varsha Suresh[1], Salah Ait Mokhtar[2], Caroline Brun[2], Ioan Calapodescu[2]

[1]National University of Singapore, Singapore [2]NAVER LABS Europe, France

NAVER LABS Europe

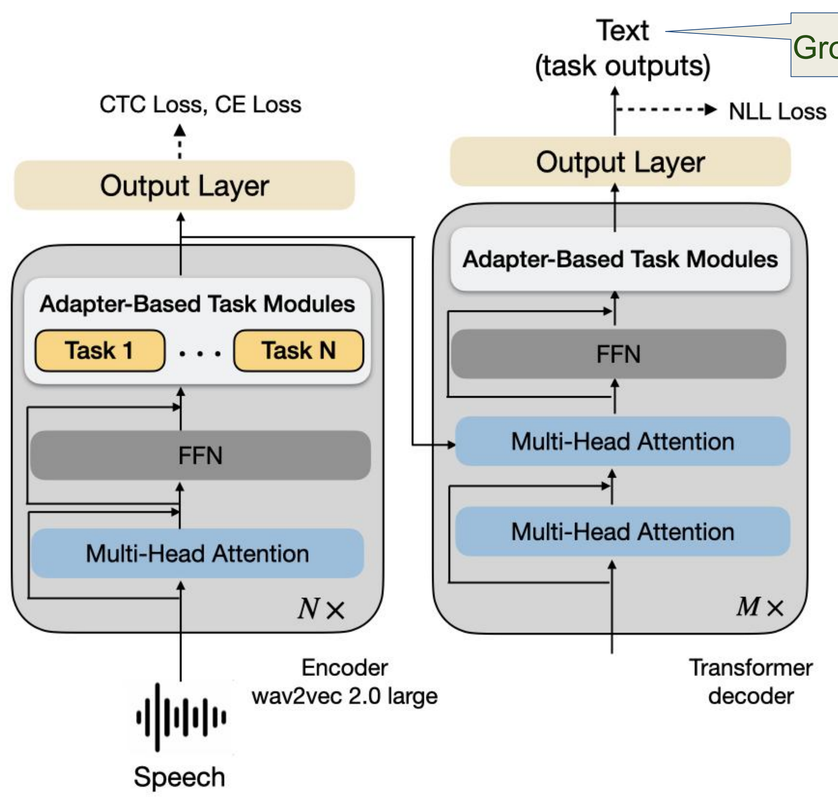
europe.naverlabs.com

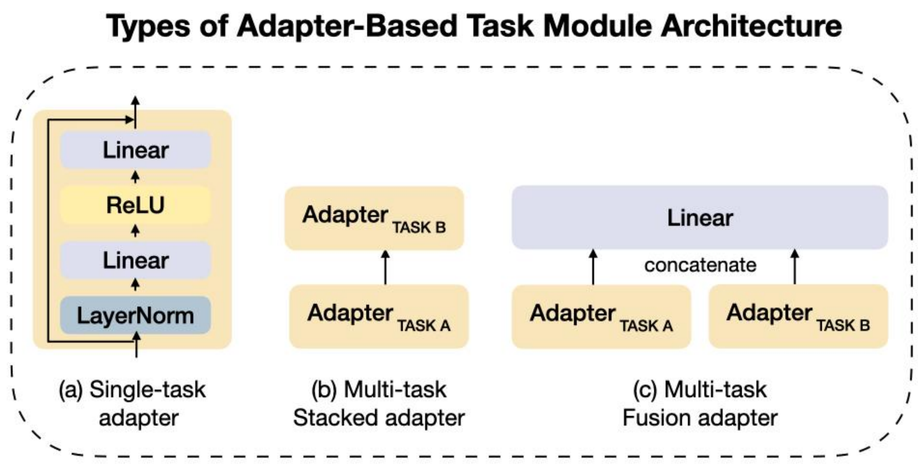
MOTIVATION

The process of fine-tuning large pre-trained speech models on downstream tasks requires substantial computational resources, particularly when dealing with multiple spoken language processing (SLP) tasks. In this work, we explore the potential of adapter-based fine-tuning in developing a unified model capable of effectively handling multiple SLP tasks.

18.4 % average improvement achieved across **5 SLP tasks.**

MULTI-TASK ADAPTERS





SLP Tasks

(subset of SUPERB)

1. Automatic Speech Recognition (ASR)
2. Phoneme Recognition (PR)
3. Spoken Emotion Recognition (SER)
4. Slot Filling (SF)
5. Intent Classification (IC)

EVALUATION & RESULTS

- We show that task-specific adapters can perform multiple SLP tasks within a single encoder-decoder model making efficient use of encoder representations and improving computational efficiency as compared to having a frozen encoder and task-specific decoders.

TABLE 1: Overall Performance Comparison

	LibriSpeech		IEMOCAP	SNIPS	FSC	Avg
	ASR	PR	SER	SF	IC	
	WER↓	PER↓	Acc↑	F1↑ , CER↓	Acc↑	
WavLM large SUPERB [3]	3.4	3.1	**70.6**	92.2, 18.4	99.0	89.5
wav2vec2.0 large SUPERB [3]	3.1	4.7	65.6	87.1, 27.3	95.2	85.5
wav2vec2.0 large (Ours)	**3.5**	**2.4**	68.2	**95.4, 11.8**	**99.5**	**90.9**

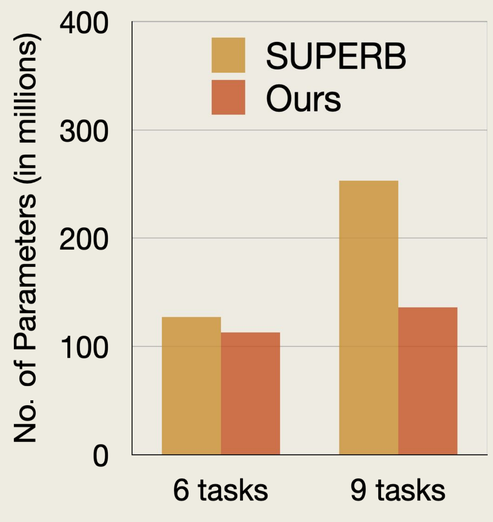
- As adapters naturally support MTL, we also consider adapter stacking [1] and adapter fusion [2] architectures to perform positively correlated tasks together, further improving performance over single task adapter settings.

TABLE 2: Single Task vs. Multi-Task Learning

wav2vec2.0 large (Ours)	IEMOCAP	SNIPS	FSC
	SER	SF	IC
	Acc↑	F1↑ , CER↓	Acc↑
Single	65.6	94.7, 12.9	99.4
Stacked	**68.2**	94.4, 13.5	**99.5**
Fusion	65.4	**95.4, 11.8**	99.3

- Our unified model enables multi-task capabilities in a parameter-efficient and scalable manner as seen in Graph 1.

No. of Additional Trainable Parameters vs. No. of Tasks



Graph 1: Our approach requires less trainable parameters. Even as the no. of tasks increases, the increase in the parameter count remains significantly lower with the ratio dropping to 53.6%.

FUTURE DIRECTIONS

1. Broaden the scope of our approach to add more tasks such as Speaker Identification, Speaker Diarization and other datasets.
2. Evaluate for different choices of SSL models such as HuBERT and WavLM and also explore different adapter architectures.

REFERENCES

1. Le, Hang, et al. "Lightweight Adapter Tuning for Multilingual Speech Translation." Proceedings of the 59th ACL, 2021.
2. Zhao, Yuting, and Ioan Calapodescu. "Multimodal robustness for neural machine translation." Proceedings of EMNLP. 2022.
3. Yang, Shu-wen, et al. "SUPERB: Speech Processing Universal PERformance Benchmark." Interspeech 2021 (2021).

Scan for the pdf