DistilHuBERT: Speech Representation Learning by Layer-wise Distillation of Hidden-unit BERT

Heng-Jui Chang

Shu-wen Yang

Hung-yi Lee

National Taiwan University

Self-supervised Learning for Audio and Speech Processing

AAAI-22

Major Contributions

Proposed a **multi-task learning** framework to compress pre-trained speech representation models.

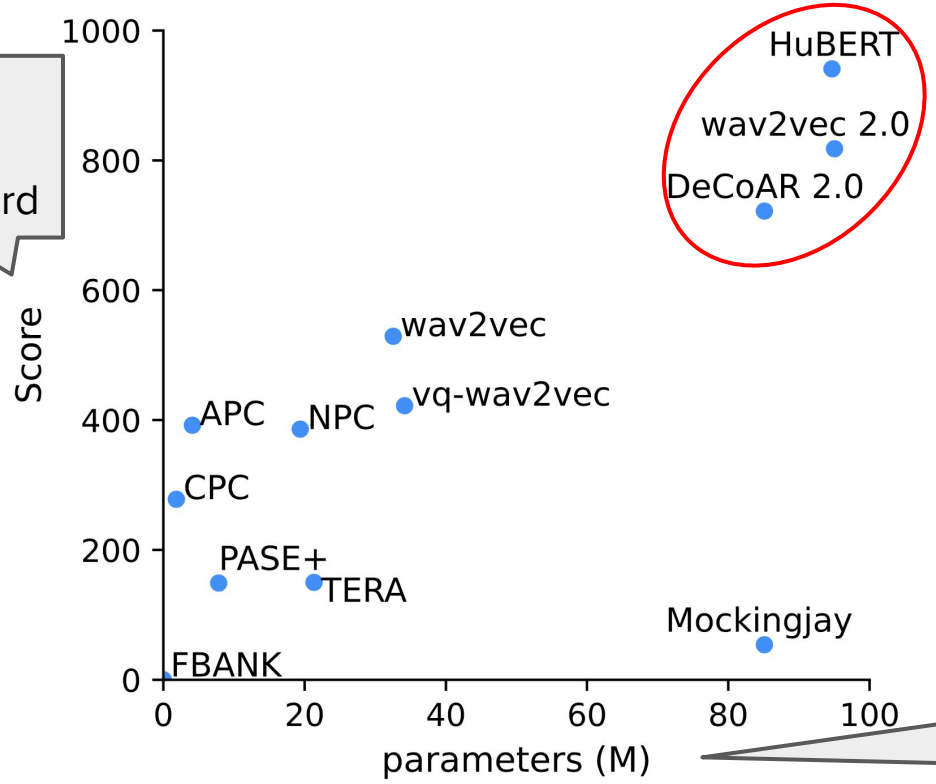
Reduced HuBERT's size by **75%** and speedup by **73%**.

Retained HuBERT's performance on **multiple speech processing tasks**.

Why compressing speech SSL models?

SSL Model Size vs. Performance Across 10 Tasks

Official SUPERB leaderboard



powerful but large

Expensive Training

Large Memory Footprint

Slow Inference Speed

# of parameters in a speech SSL model

Existing Model Compression Methods

Natural Language Processing:
DistilBERT [Sanh'19], TinyBERT [Jiao'19]
(knowledge distillation)

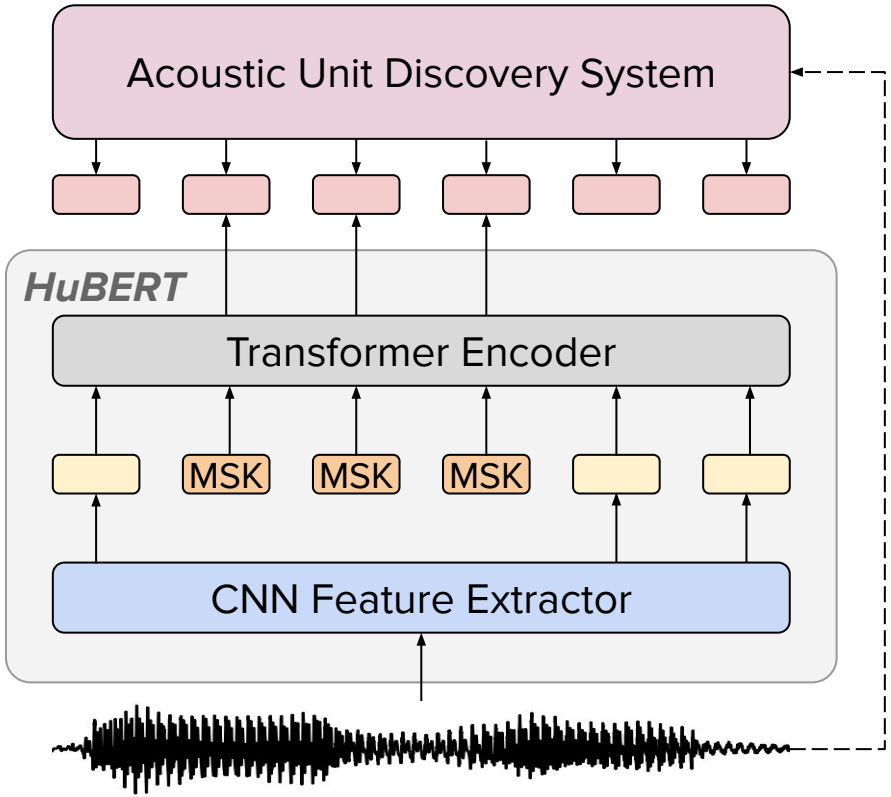
→ ineffective for speech SSL models

Speech Processing:
Prune, Adjust and Re-Prune (PARP) [Lai'20]
(parameter pruning)

→ fine-tuned with labeled data

Methods

Teacher: HuBERT [Hsu'21]

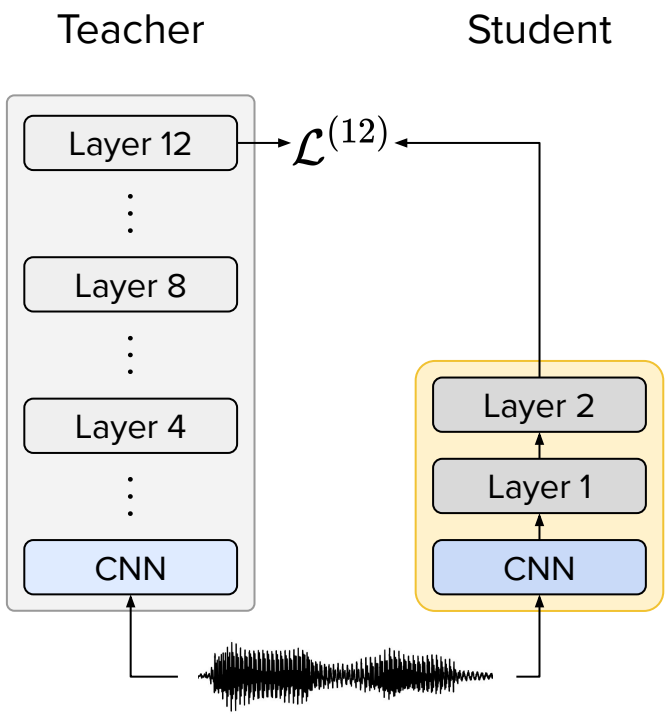


Outperforms other SSL methods on most speech processing tasks.

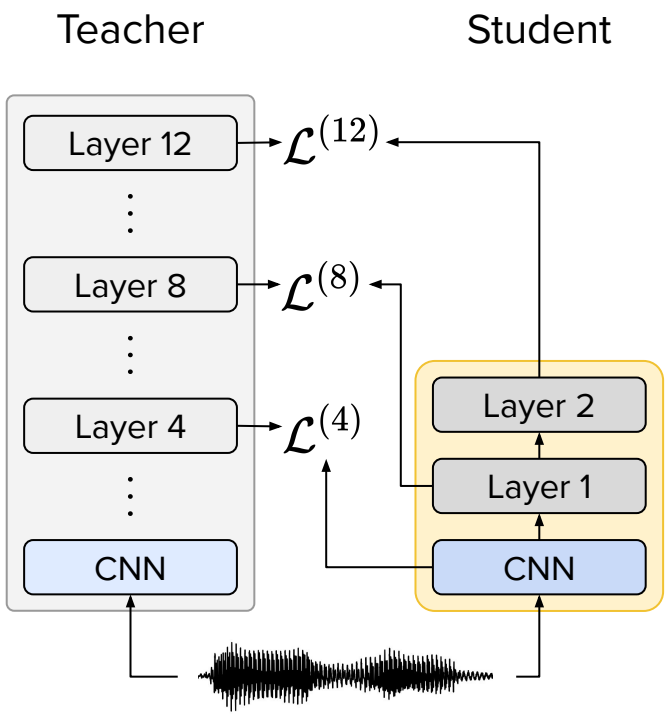
Expensive Training: 2k GPU hours
Large & Slow: 95M params

academia / small corps.: difficult to reproduce or applying to products

Typical Knowledge Distillation

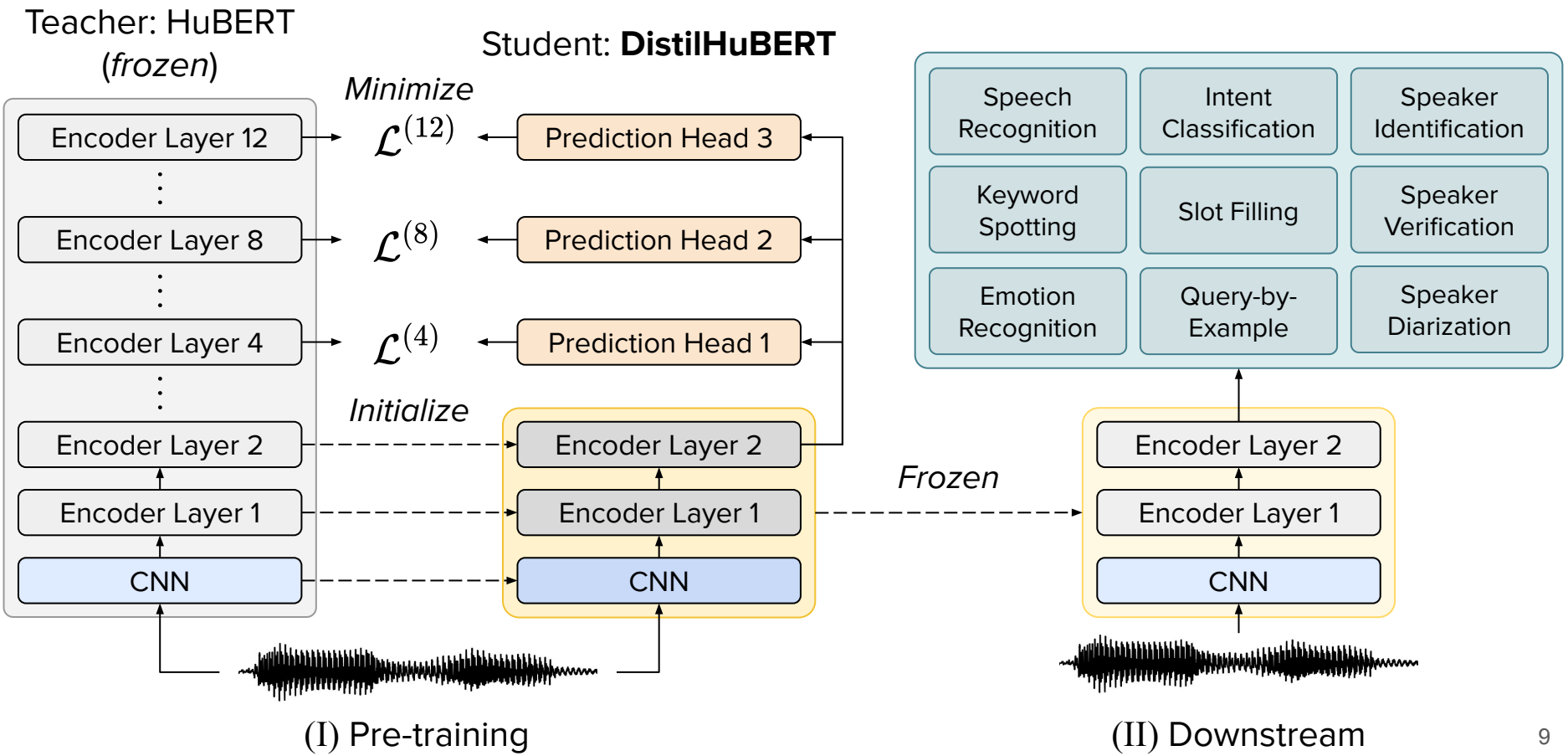


Distill Last



Distill w/ Hidden

Proposed DistilHuBERT Framework



Experiments

Experimental Setup

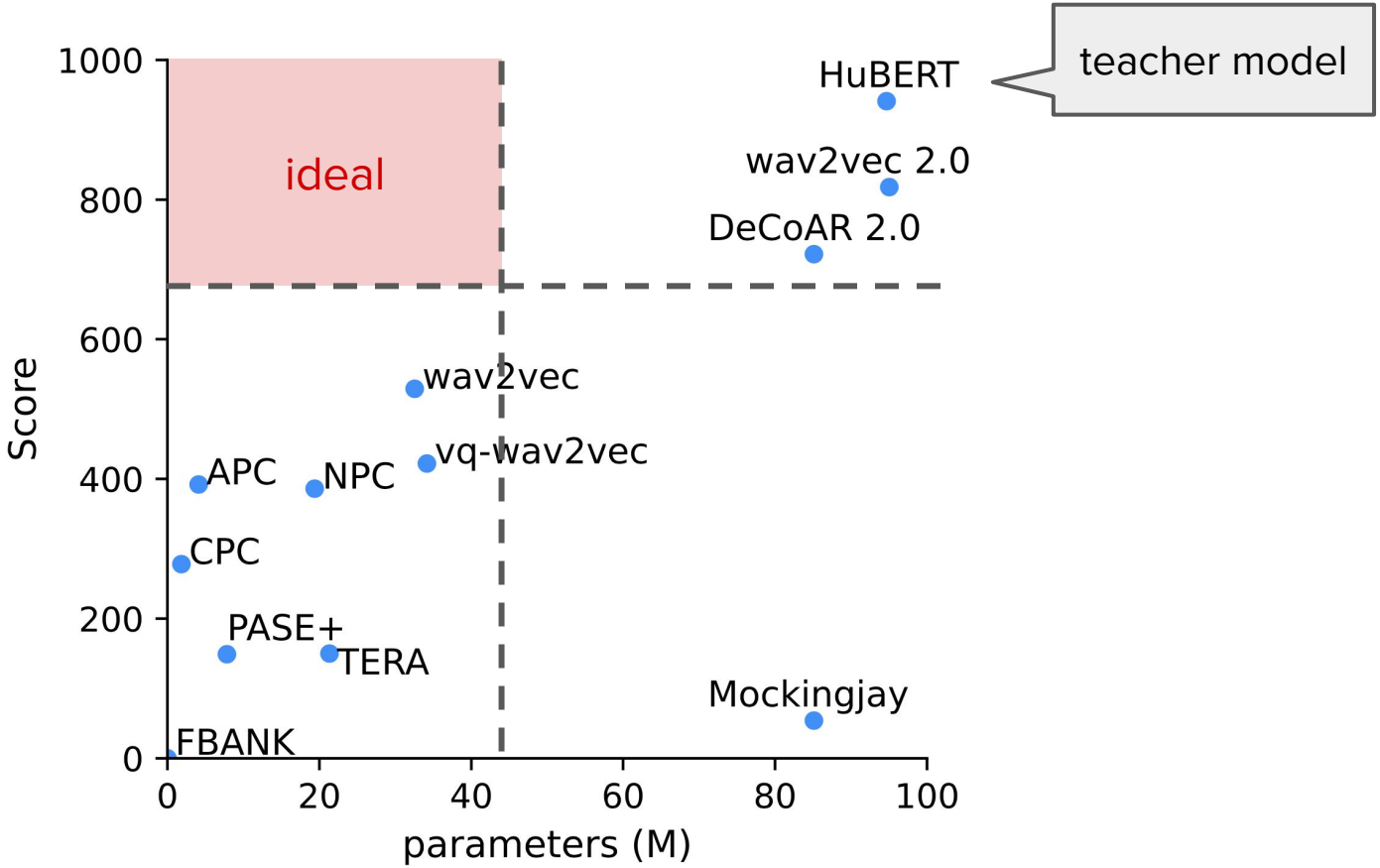
Data: LibriSpeech 960h

Implementation: S3PRL

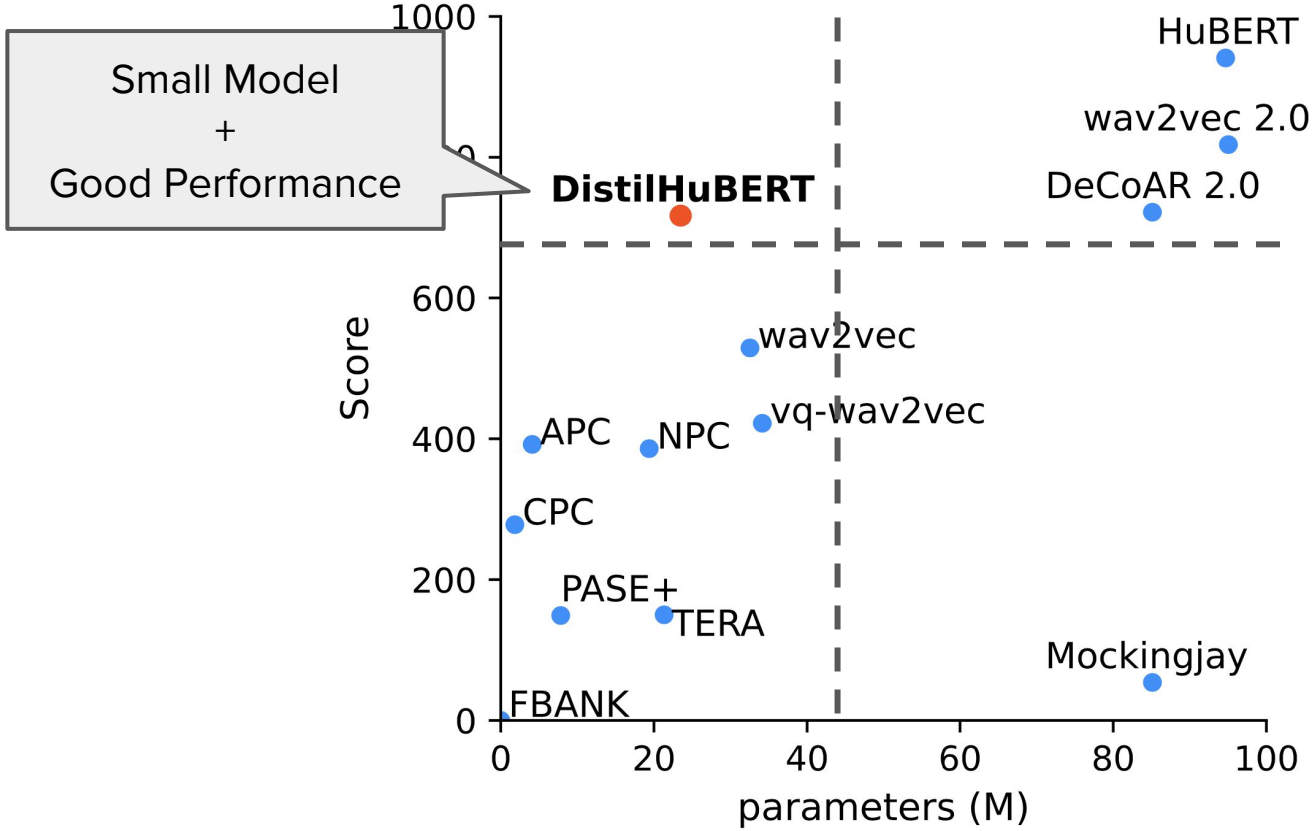
Benchmark: Speech processing Universal PERformance Benchmark (SUPERB) [Yang'21]

ASR	PR	Intent Classification	Speaker Identification
Keyword Spotting		Slot Filling	Speaker Verification
Emotion Recognition		Query-by-Example	Speaker Diarization

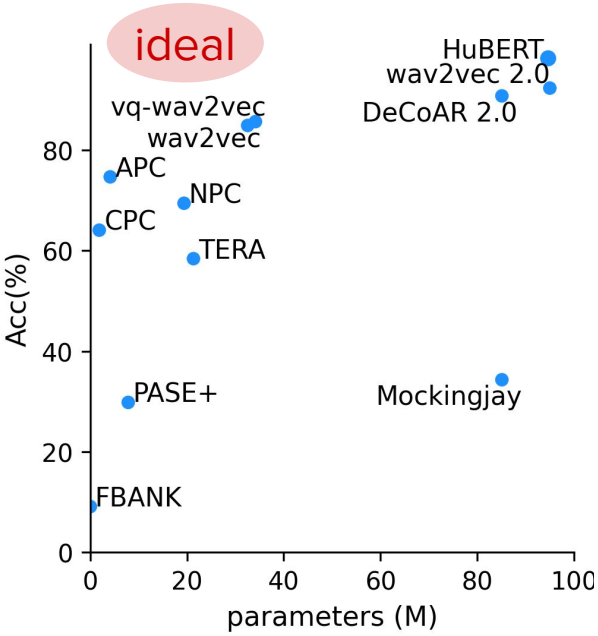
Size vs. Overall Performance on SUPERB



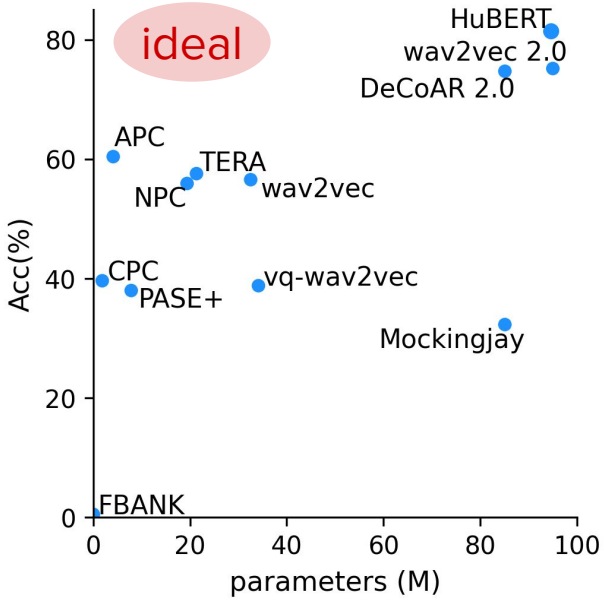
Size vs. Overall Performance on SUPERB



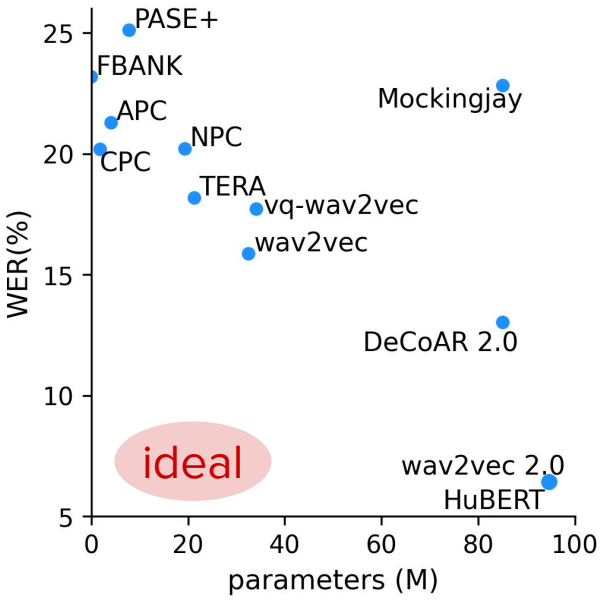
Size vs. Performance on 3 SUPERB Tasks



Intent Classification (*semantics*)

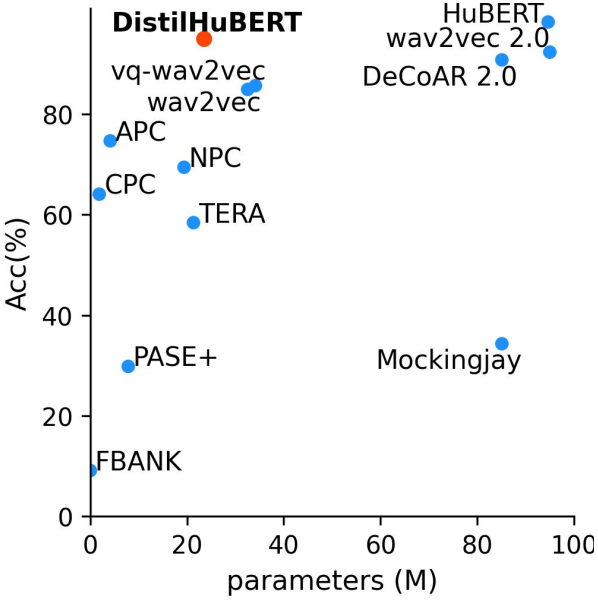


Speaker Identification (*speaker*)

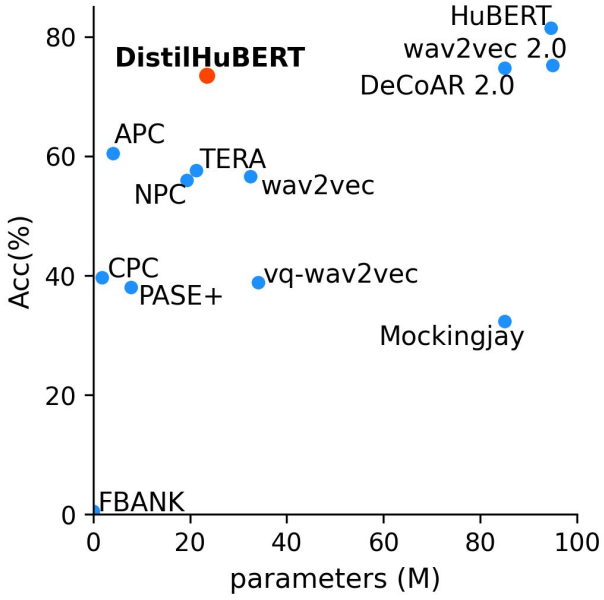


Automatic Speech Recognition (*content*)

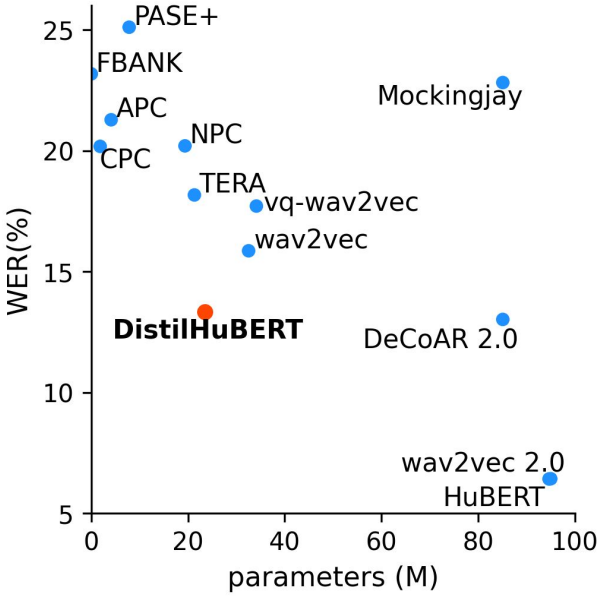
Size vs. Performance on 3 SUPERB Tasks



Intent Classification (semantics)

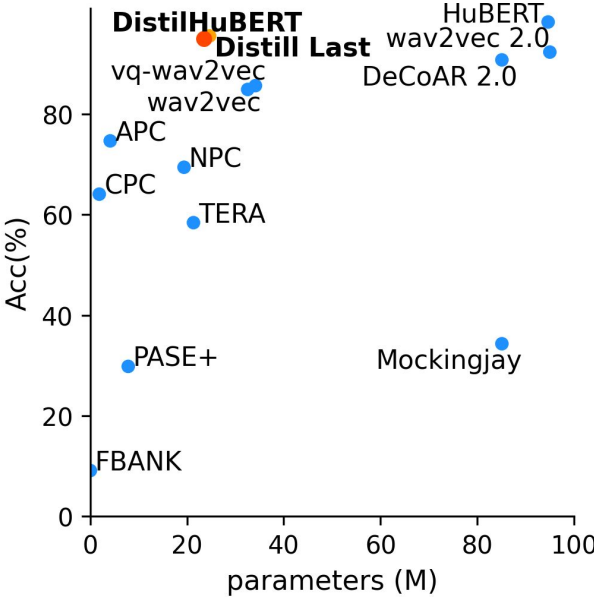


Speaker Identification (speaker)

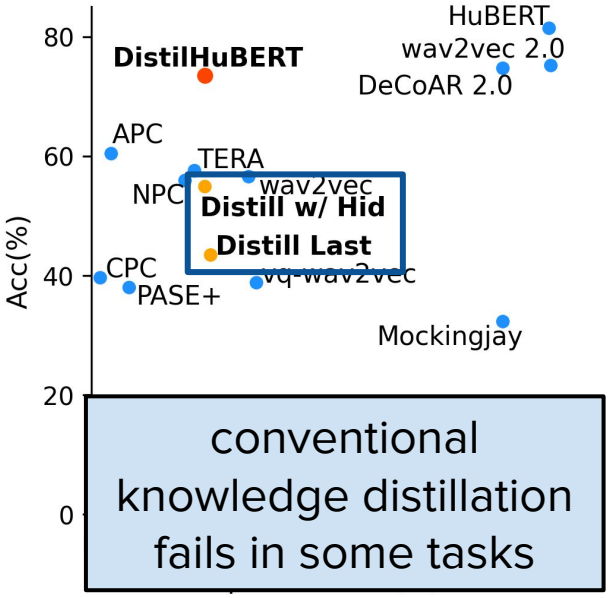


Automatic Speech Recognition (content)

Size vs. Performance on 3 SUPERB Tasks

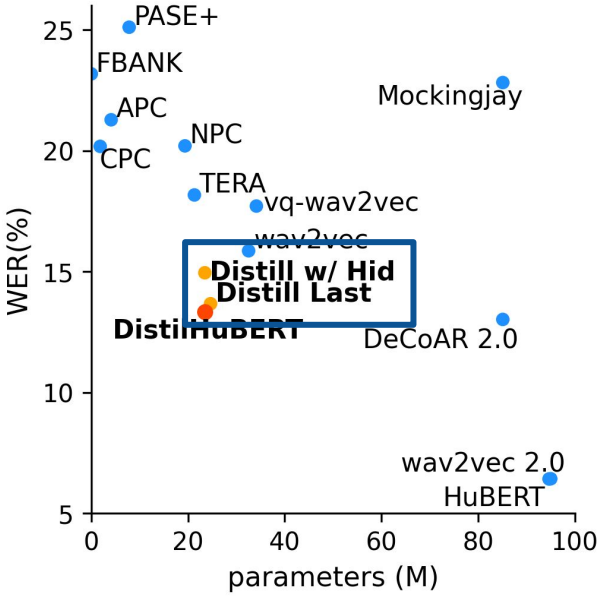


Intent Classification (semantics)



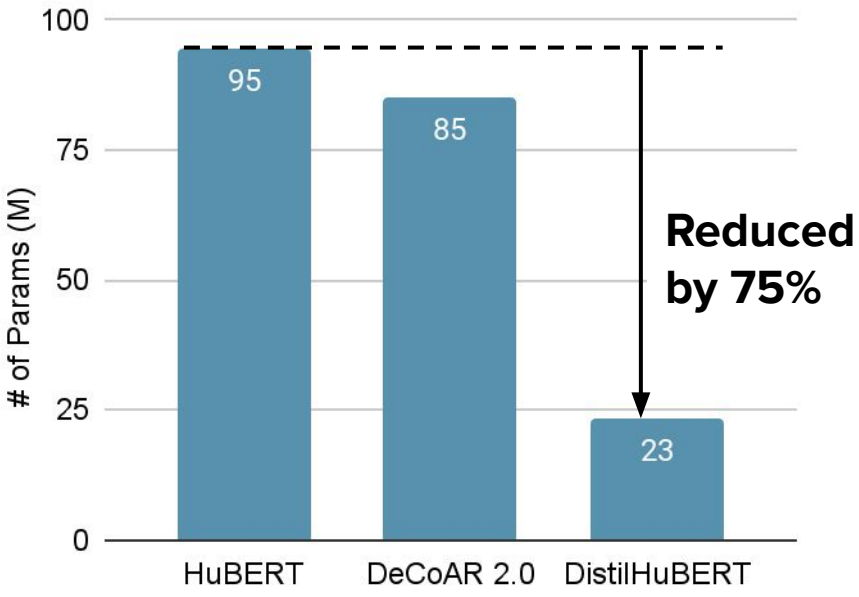
conventional knowledge distillation fails in some tasks

Speaker Identification (speaker)

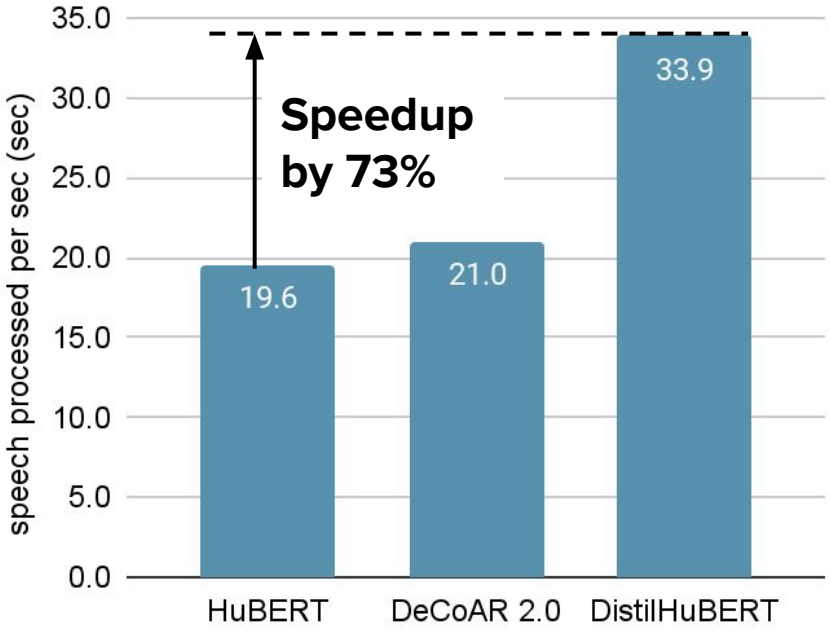


Automatic Speech Recognition (content)

Model Size



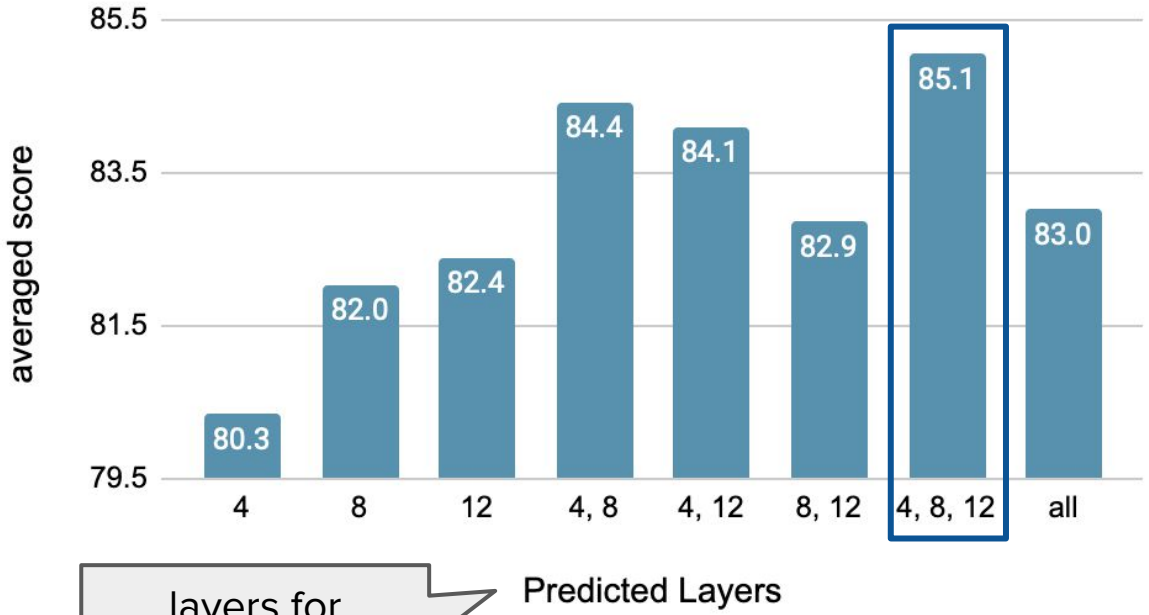
Inference Speed



Layer Selection

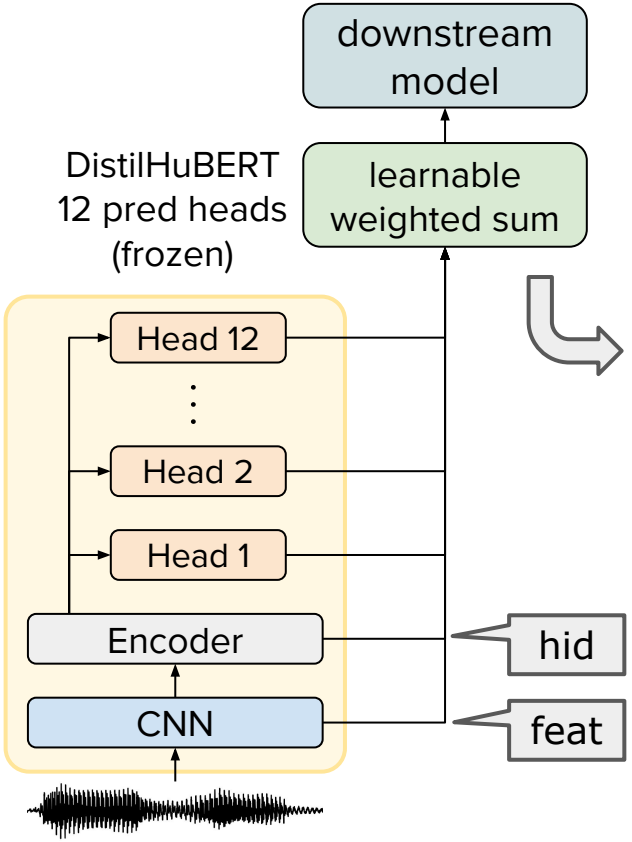
offered **balanced representations** for different types of tasks

averaged accuracy over IC, SID, ASR

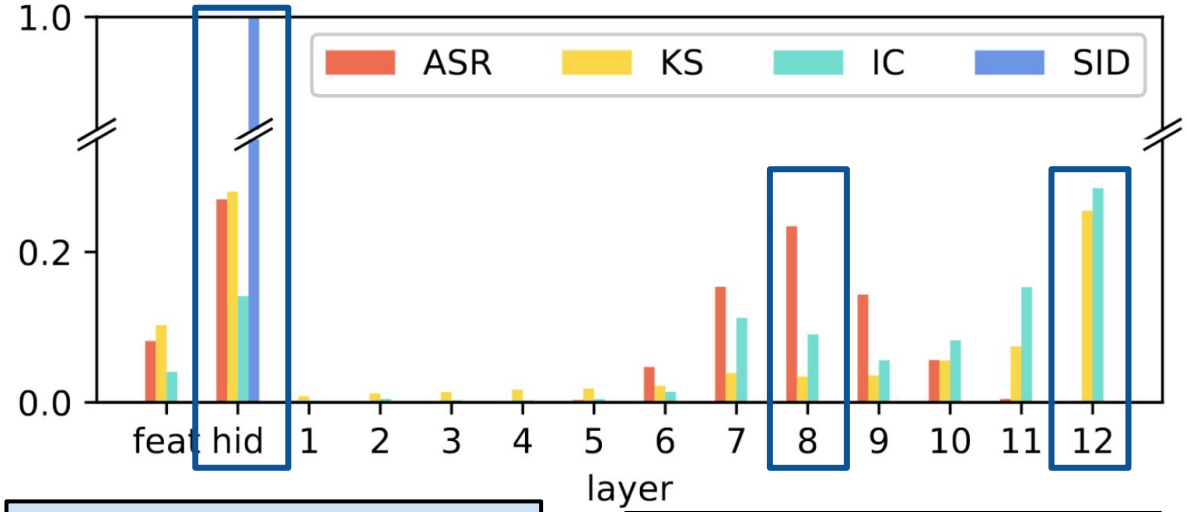


layers for DistilHuBERT to learn

Layer Selection



Normalized Weights for the Learnable Weighted Sum



shared representations -> rich information

content & semantics

multi-task learning distilled good representations

4th, 8th, 12th layers -> learn balanced representations

Conclusion

DistilHuBERT: a novel framework to layer-wise distill knowledge from HuBERT.

Retained most of HuBERT's performance with **significant speedup** and **model size reduction**.

Methods can be easily applied to other SSL models. (code is open-sourced on S3PRL)

More details can be found in the paper.
Thanks for listening!

National Taiwan University

AAAI-22

Self-supervised Learning for Audio and Speech Processing

Paper:

Contact: