Recent Progress of Self-supervised Learning for Speech Processing

Speaker: Hung-yi Lee

2022 Eighth Frederick Jelinek Memorial Summer Workshop

The application for the undergraduate research internships will be available on February 10th. Please read the **"AI Research Internships for Undergraduates"** for more information

The Workshop June 27 to August 5, 2022

About the Eighth Frederick Jelinek Memorial Summer Workshop

The JSALT 2022 Program

JHU Summer School on Human Language Technology (June 13 June 24)

Opening Day Presentations Schedule (June 27)

Plenary Lectures by Invited Speakers (June 29, July 6, 13, 20, 27)

Closing Day Presentations (August 4 and 5)

Research Groups

- Speech Translation for Under-Resourced Languages
- Multilingual and Code-Switching Speech Recognition
- Leveraging Pre-Training Models for Speech Processing

Team's Webpage

https://jsalt-2022-ssl.github.io/member

Goal

Better Pre-trained Model

- More Efficient
- Better Generalization
- Learn from Multimodality

+

Better Use of Pre-trained Model

- Efficient Usage
- New Applications
- Toolkit

2 Interspeech papers & 5 SLT papers

Speech processing Universal PERformance Benchmark

https://superbbenchmark.org/

SUPERB: Speech processing Universal PERformance Benchmark

Shu-wen Yang[1], Po-Han Chi[1], Yung-Sung Chuang[1*], Cheng-I Jeff Lai[2*], Kushal Lakhotia[3*], Yist Y. Lin[1*], Andy T. Liu[1*], Jiatong Shi[4*], Xuankai Chang[6], Guan-Ting Lin[1], Tzu-Hsien Huang[1], Wei-Cheng Tseng[1], Ko-tik Lee[1], Da-Rong Liu[1], Zili Huang[4], Shuyan Dong[5†], Shang-Wen Li[5†], Shinji Watanabe[6], Abdelrahman Mohamed[3], Hung-yi Lee[1]*

Presented at INTERSPEECH'21

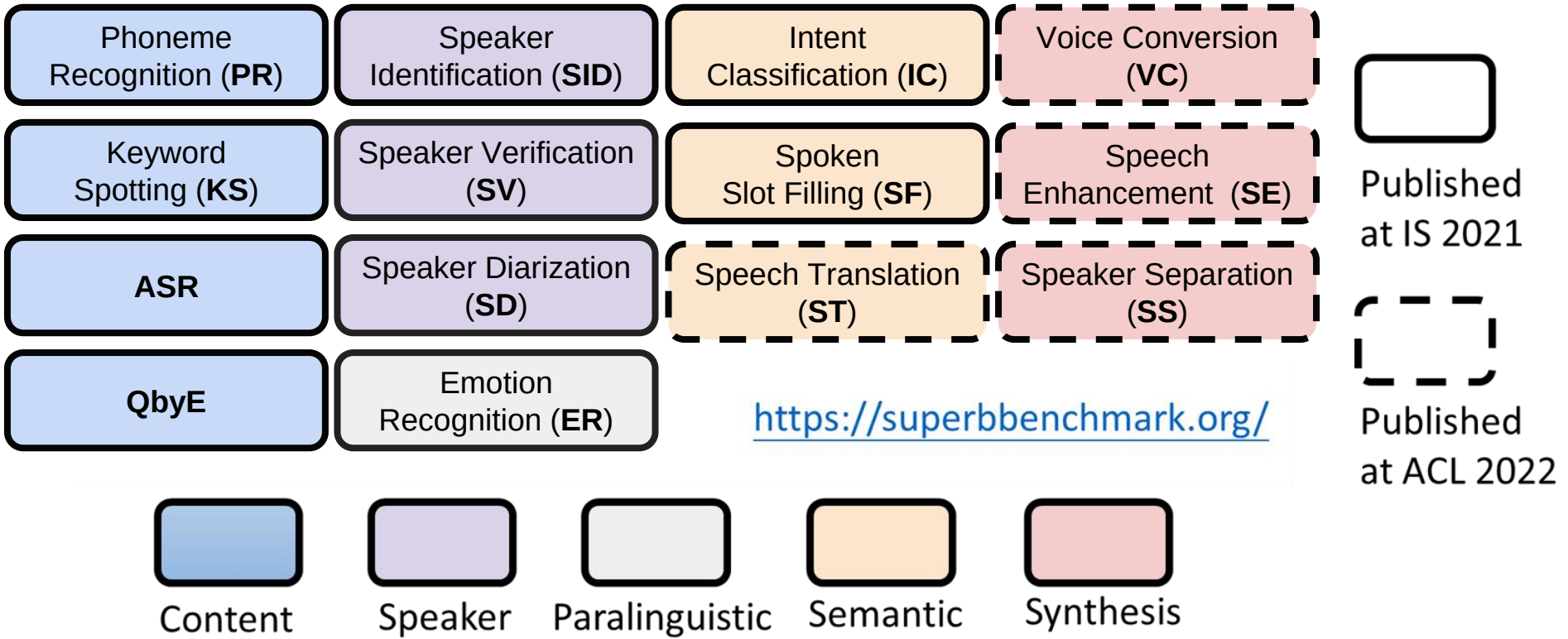
SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities

Hsiang-Sheng Tsai[1*], Heng-Jui Chang[1*], Wen-Chin Huang[2*], Zili Huang[3*], Kushal Lakhotia[4*], Shu-wen Yang[1], Shuyan Dong[5], Andy T. Liu[1], Cheng-I Lai[6], Jiatong Shi[7], Xuankai Chang[7], Phil Hall[8], Hsuan-Jui Chen[1], Shang-Wen Li[5], Shinji Watanabe[7], Abdelrahman Mohamed[5], Hung-yi Lee[1]

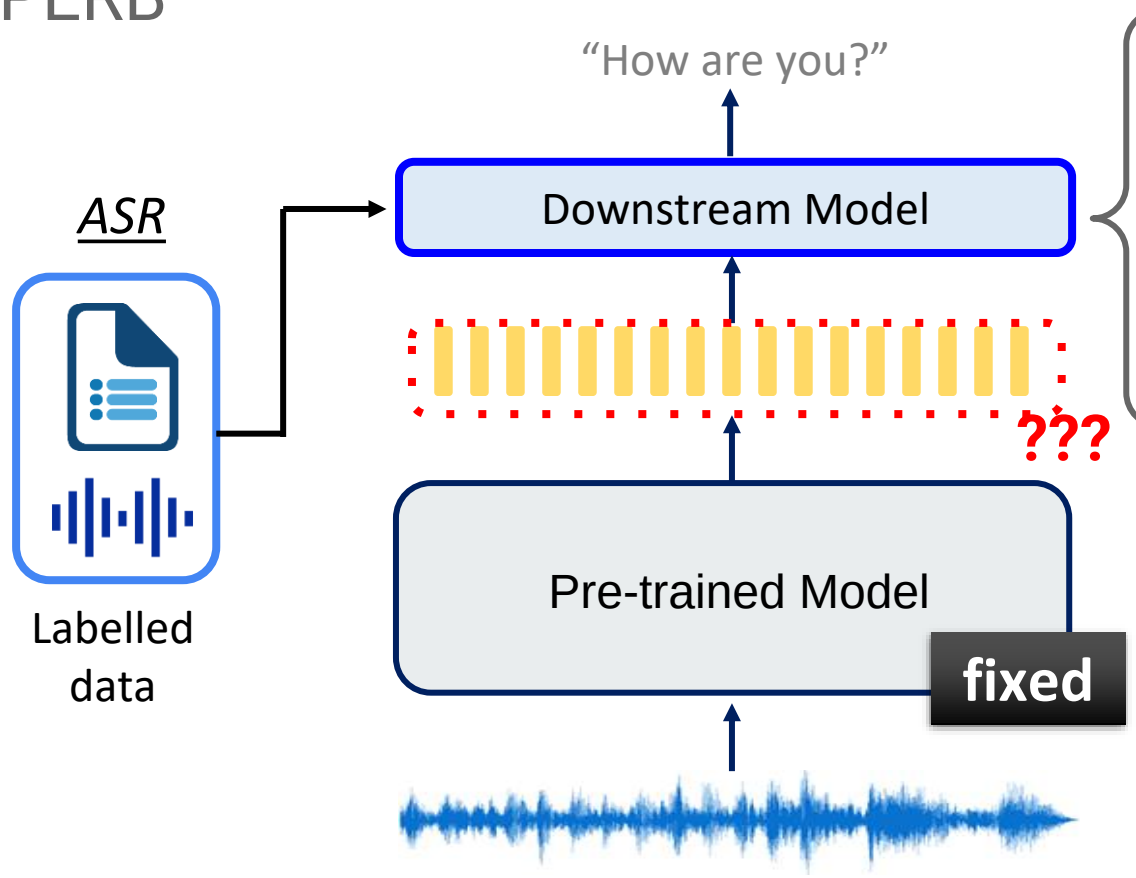
Presented at ACL'22

SUPERB

Benchmark pre-trained models on **a wide range of speech processing tasks**

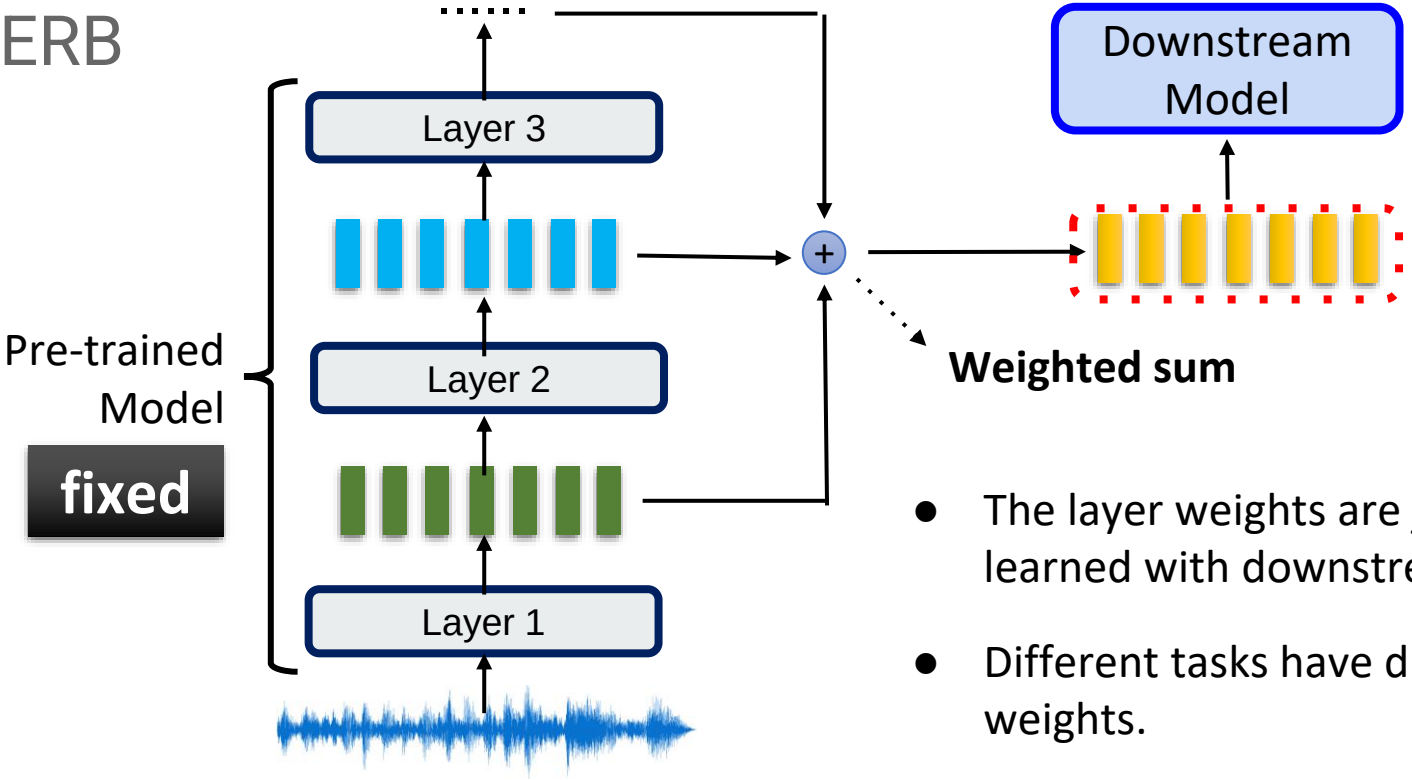


SUPERB



- Architecture is fixed for each task.
- E.g. For ASR, we always use 2-layer LSTM.
- The same set of parameters for all tasks.

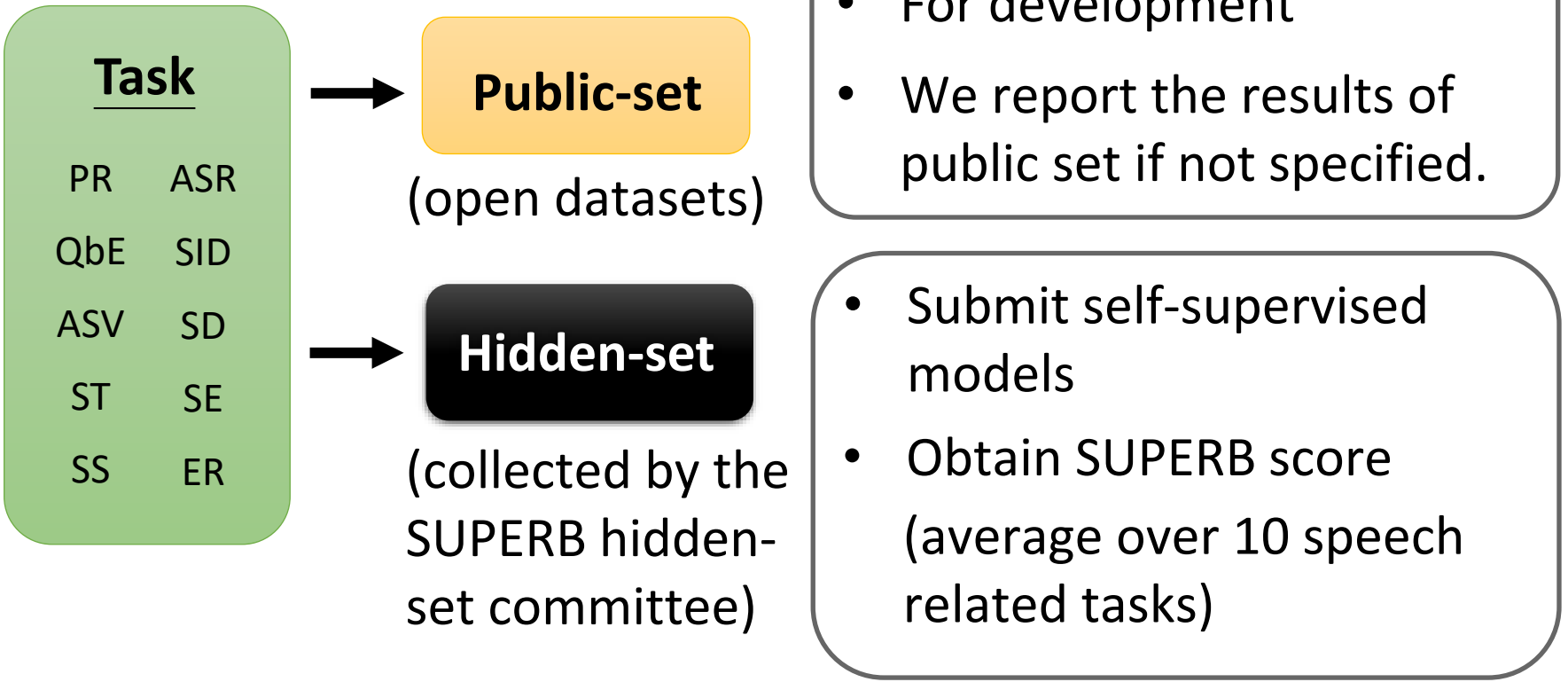
SUPERB



- The layer weights are jointly learned with downstream models.
- Different tasks have different weights.

I will show you this approach is a very efficient way to use pre-trained speech model.

SUPERB


Task
PR
ASR
QbE
SID
ASV
SD
ST
SE
SS
ER
Public-set
(open datasets)
Hidden-set
(collected by the SUPERB hidden-set committee)
• For development
• We report the results of public set if not specified.
• Submit self-supervised models
• Obtain SUPERB score (average over 10 speech related tasks)

Outline

Compression

Robust

Adapter/Prompt

Unsupervised ASR

Visual-enhanced

Prosody

Model Compression & Sequence Compression

Tzu-Quan Lin

Tsu-Hsun Feng

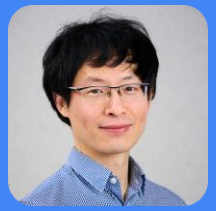
Tsung-Huan Yang

Chun-Yao Chang

Guang-Ming Chen

Yen Meng

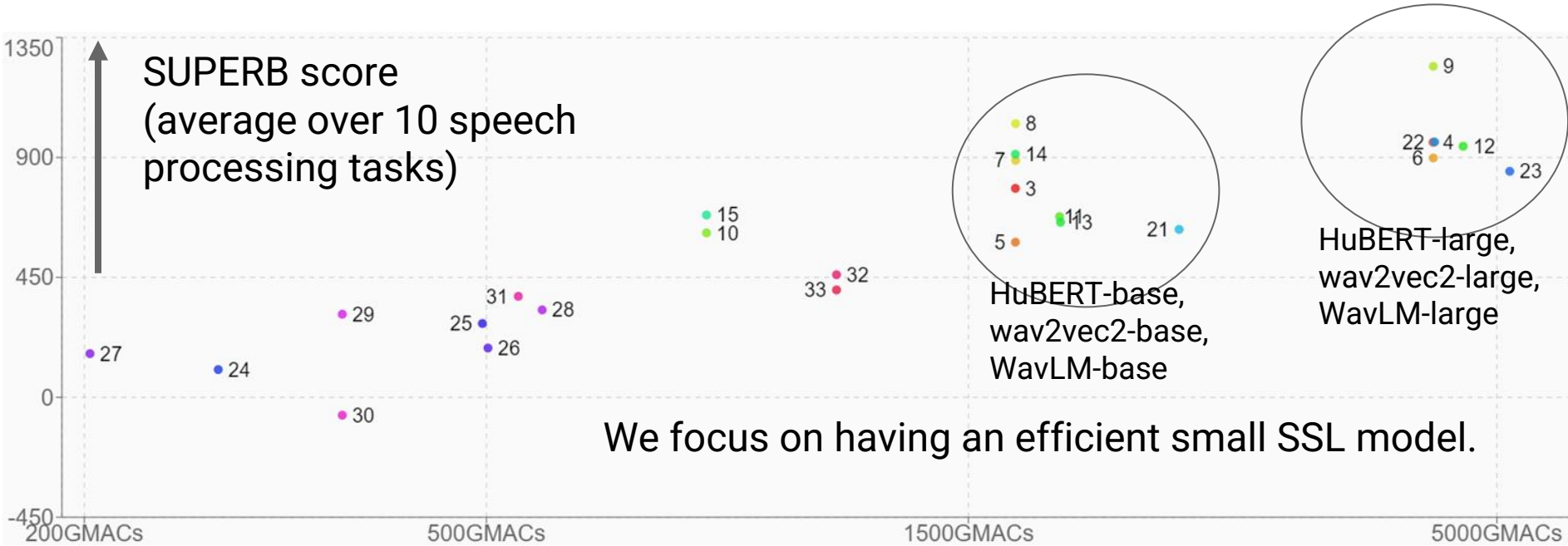
Hsuan-Jui Chen



Hao Tang

Hung-yi Lee

SUPERB Leaderboard - Hidden-set Track



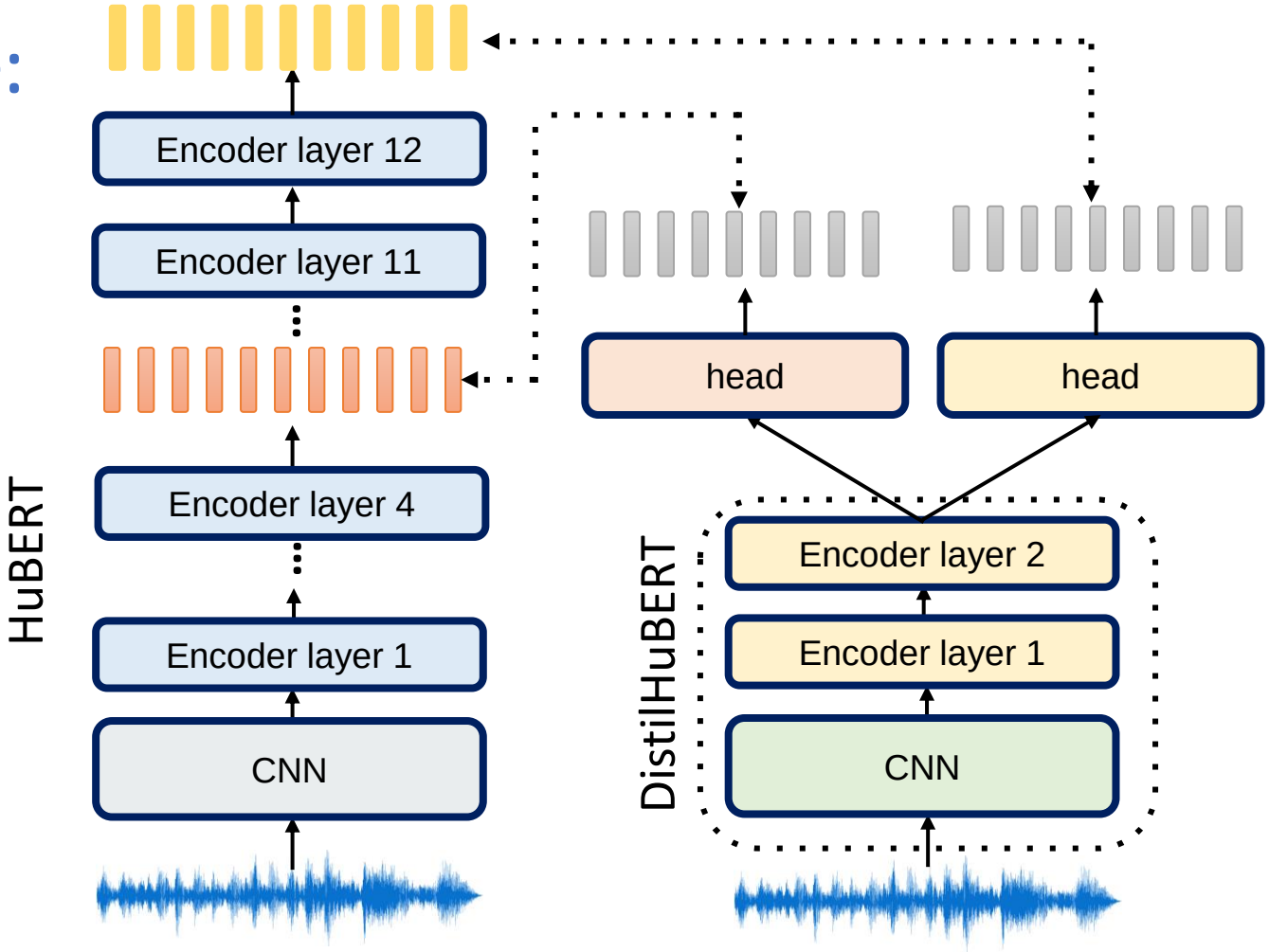
MACs = Multiply-accumulate operation

Common Network Compression Approach

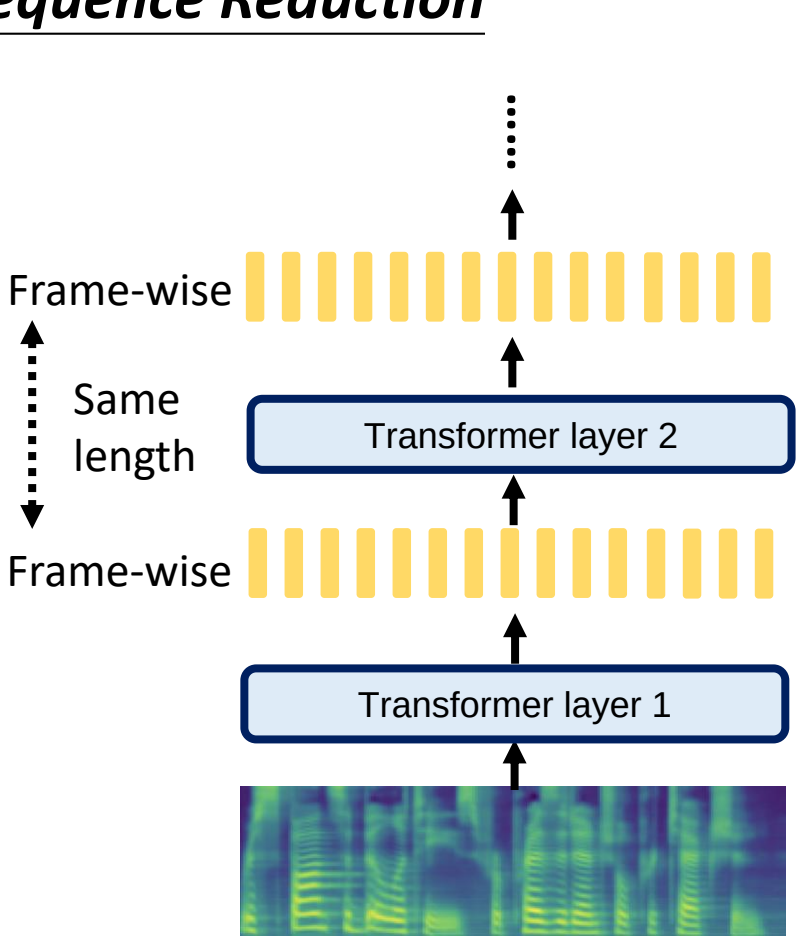
- We explored many network compression approaches.
  - Weight pruning
  - Head pruning
  - Low-rank approximation
  - Knowledge Distillation
- Focus on the *sequence reduction* here.

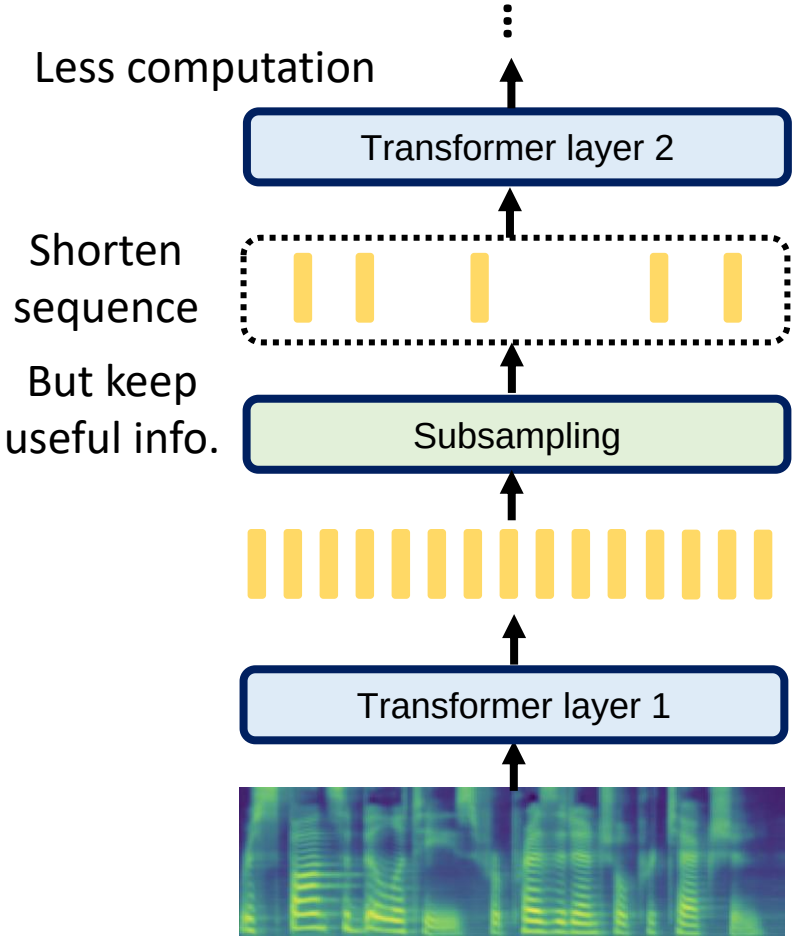
Prerequisite: Knowledge Distillation

https://arxiv.org/abs/2110.01900

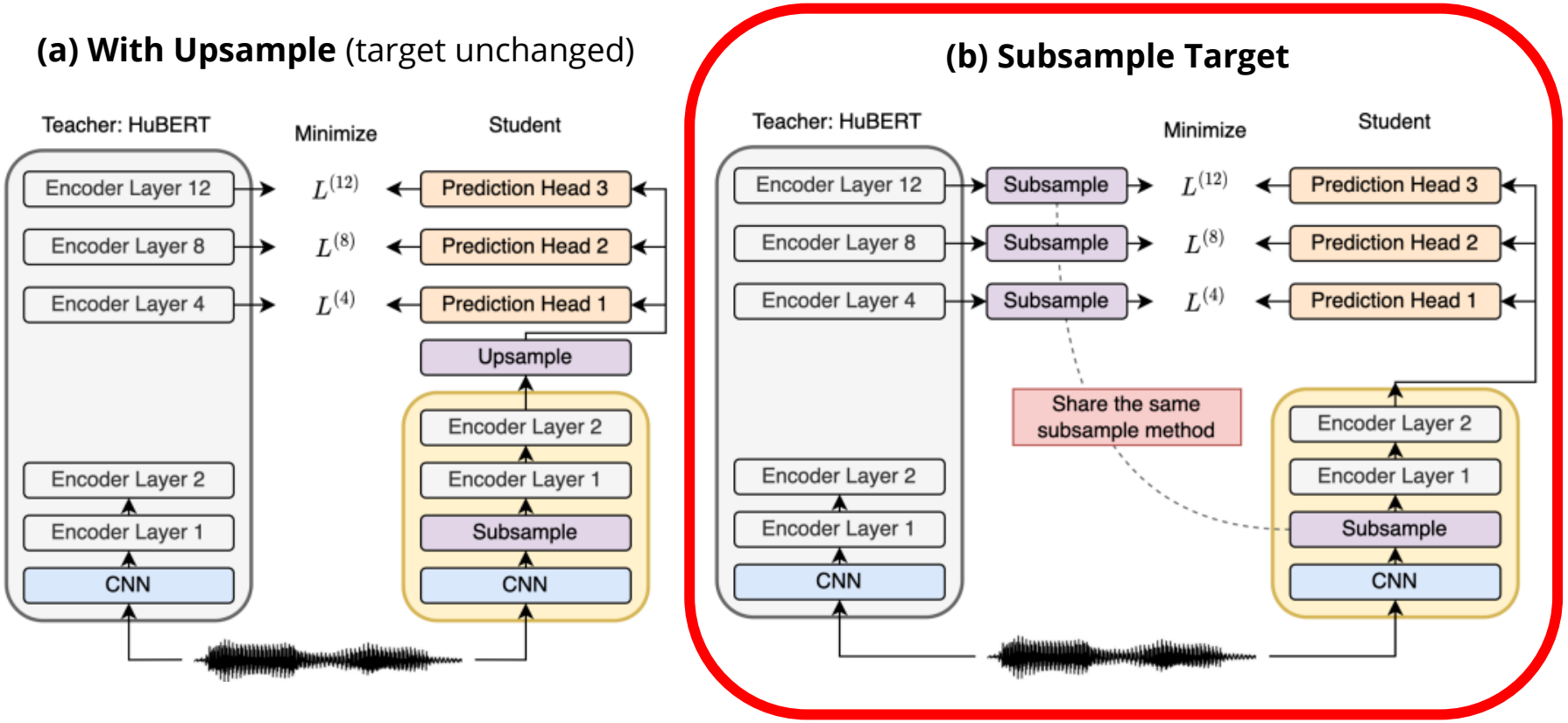


Sequence Reduction

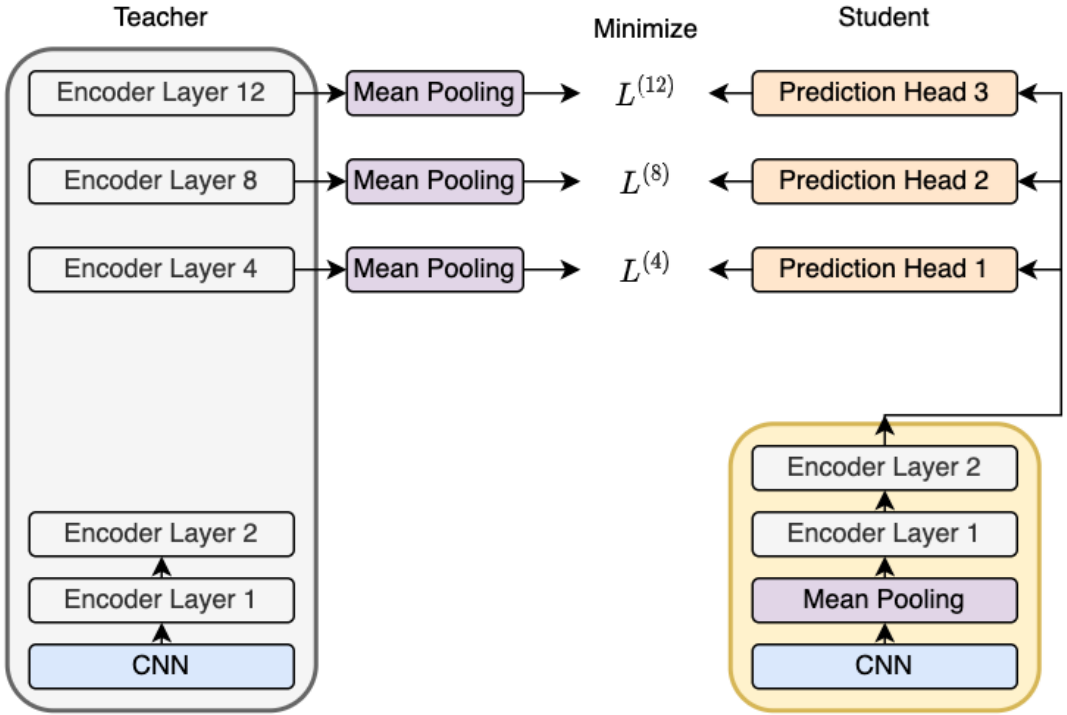


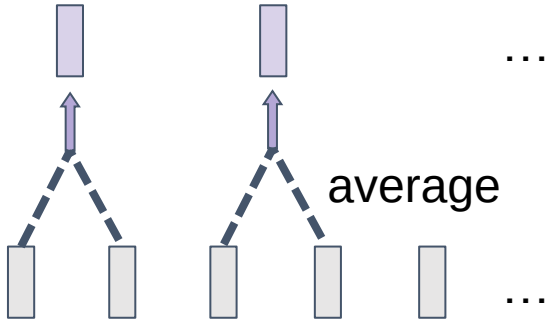


Framework for learning subsampling module



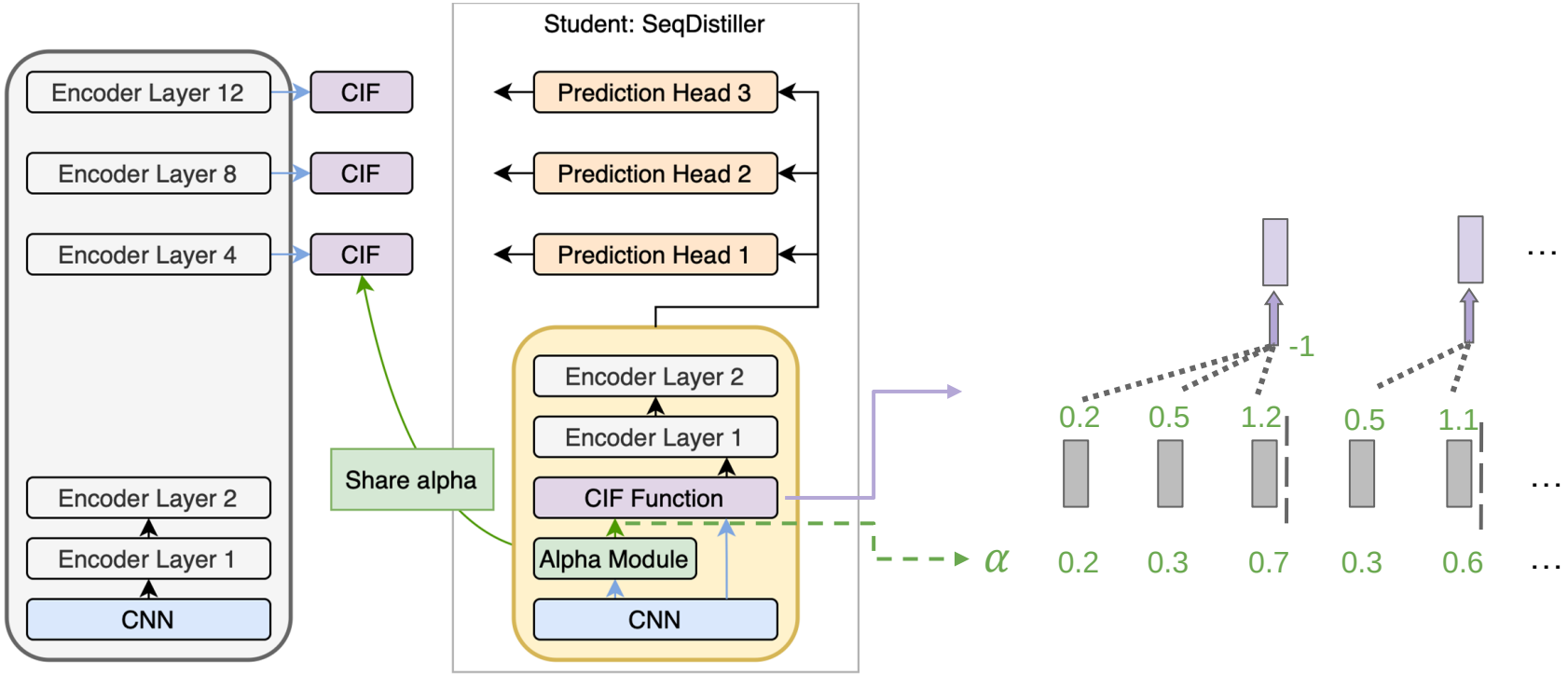
Subsampling – Fixed-length





50% subsampling

Subsampling – variable-length



[1] CIF: Continuous Integrate-and-Fire for End-to-End Speech Recognition

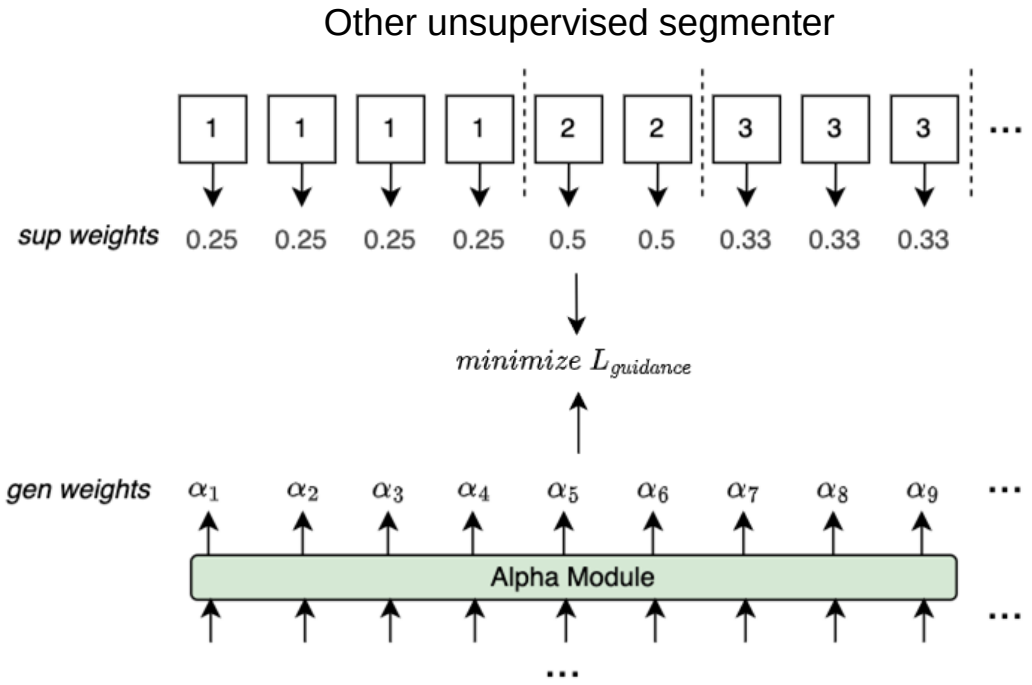
Segmentation guidance

Unsupervised

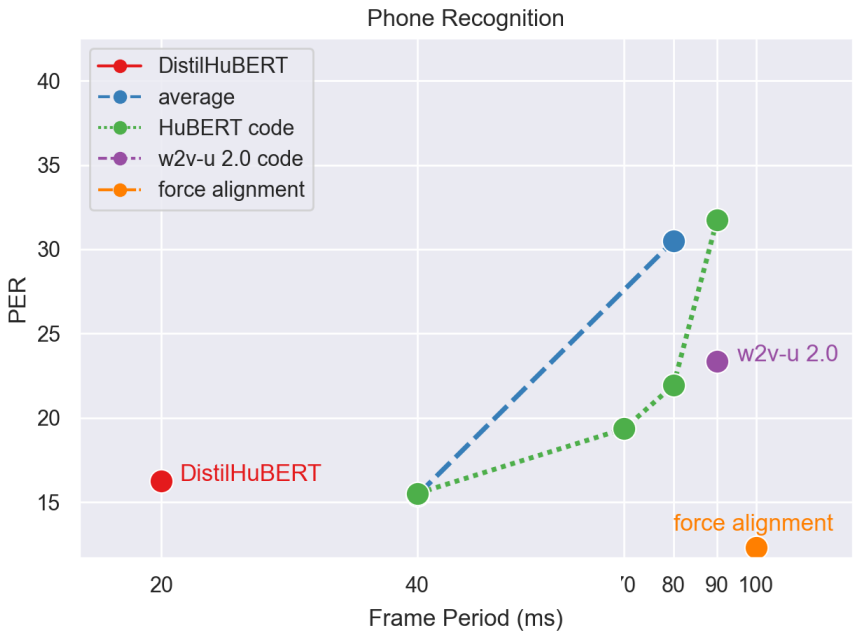
- Repetition in HuBERT codes
- Repetition in wav2vec-U 2.0 codes

Supervised

- Forced alignments

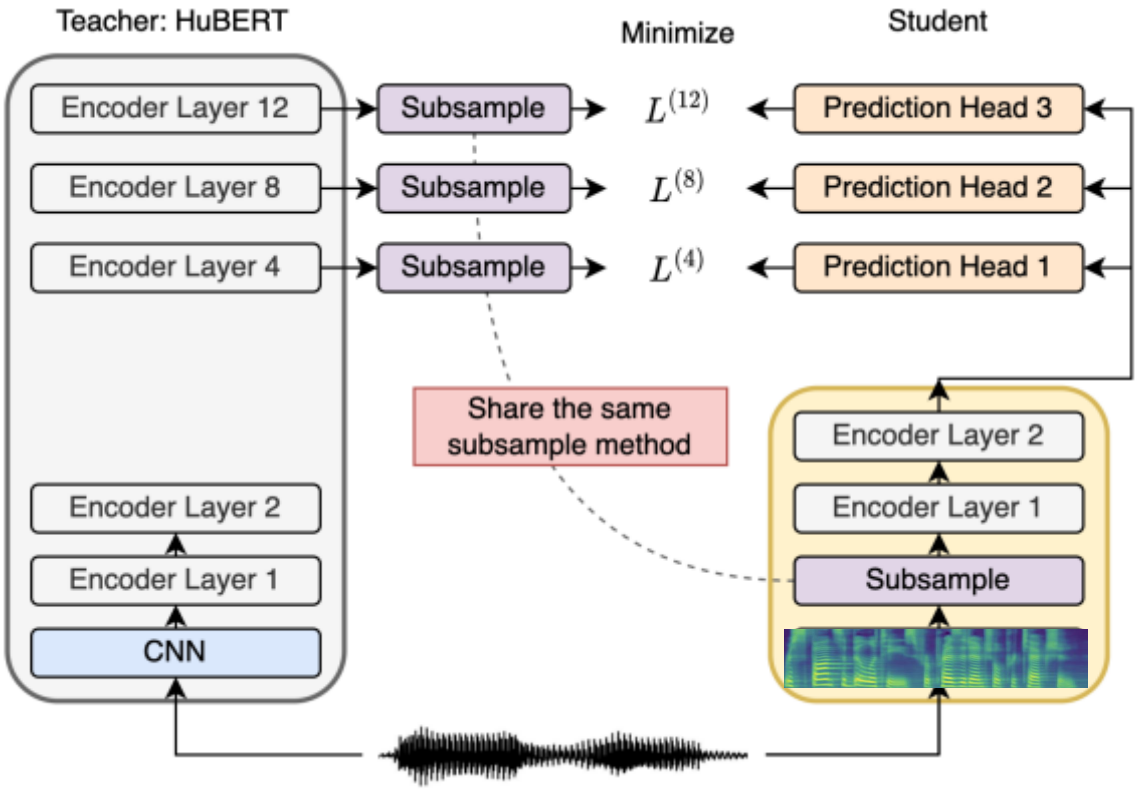


Results



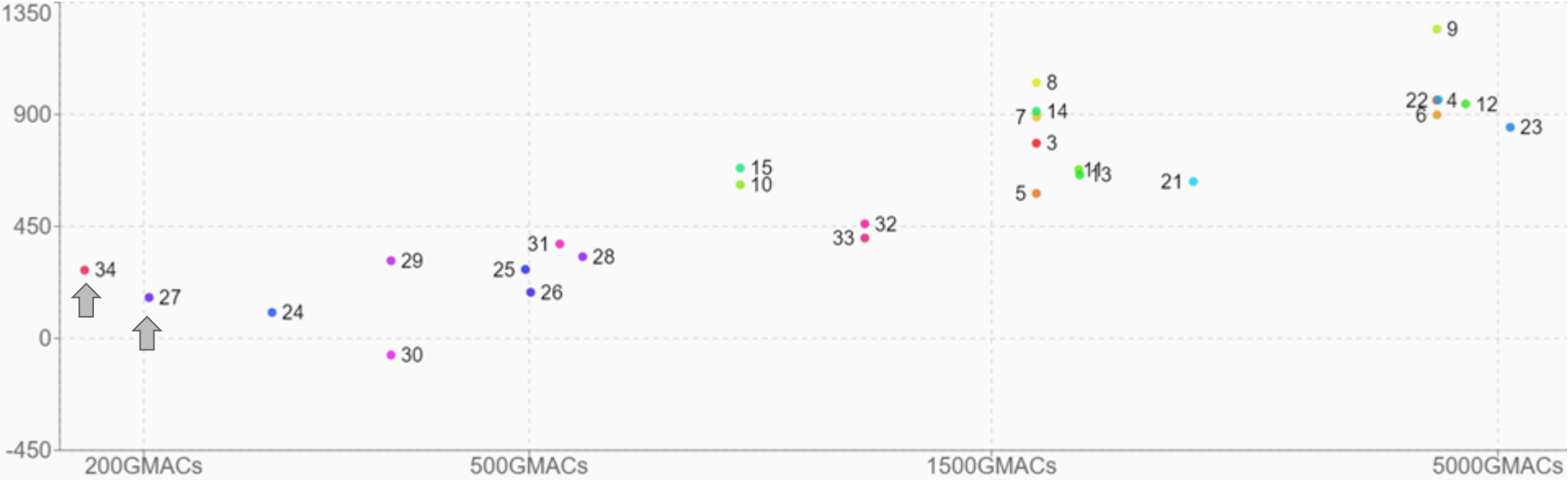
- Fixed-length subsampling can perform well with smaller compress rate.
- Variable-length subsampling works particularly well under low frame rates.
- Different tasks have their preferred frame rates for better performance.
  - E.g., Emotion recognition has better performance after subsampling.

Spectrogram as input, remove CNN layers



- Only report average down sampling
- Variable-length approach is under investigation.

SUPERB Leaderboard - Hidden-set Track



34: Seqeunce reduction + MelHuBERT (1k hours pre-trained data)

27: Modified CPC (60k hours pre-trained data)

Outline

Compression

Robust

Adapter/Prompt

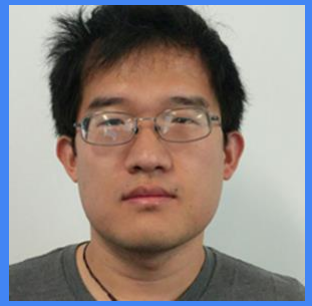
Unsupervised ASR

Visual-enhanced

Prosody

Generalization of SSL

Hung-yi Lee (NTU)

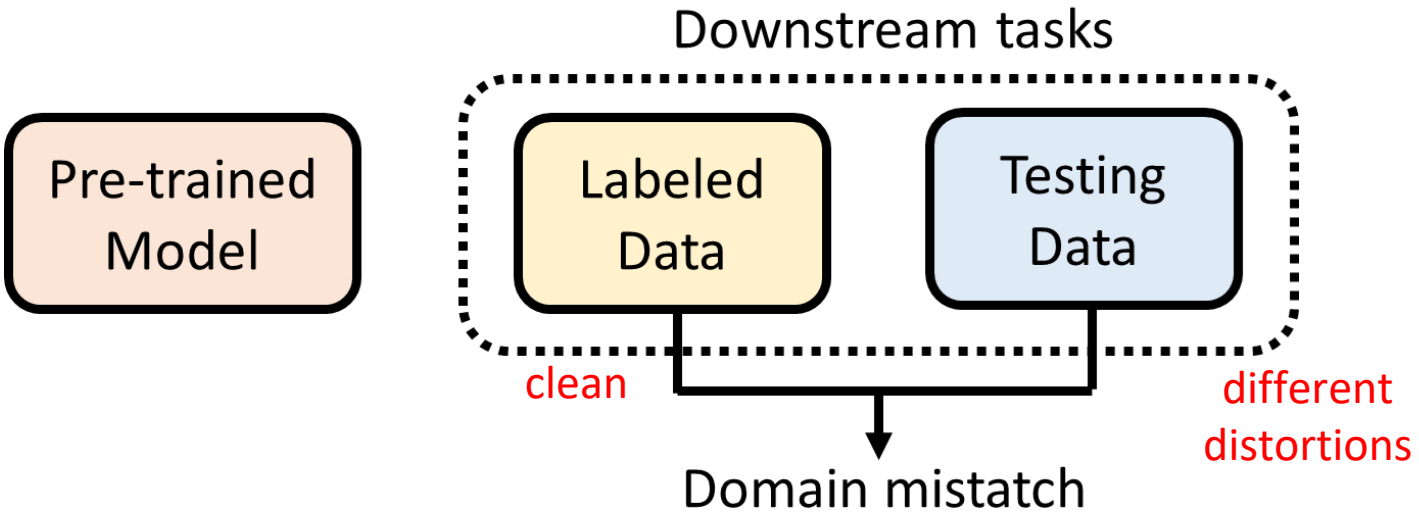


Yu Zhang (Google)

Kuan-Po Huang (NTU)

Fabian Ritter (NUS)

Generalization of SSL

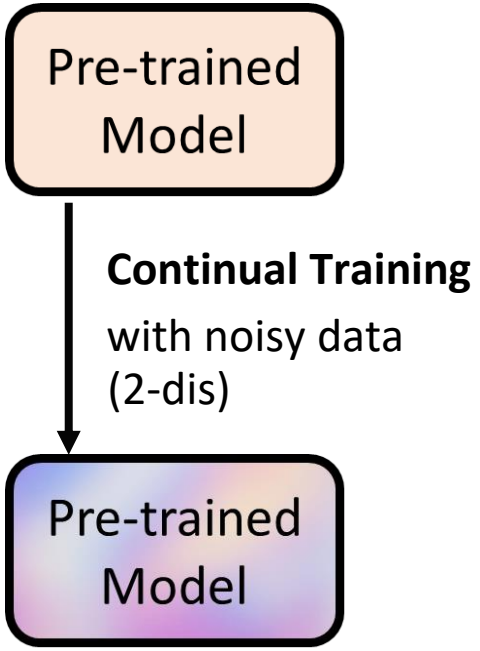


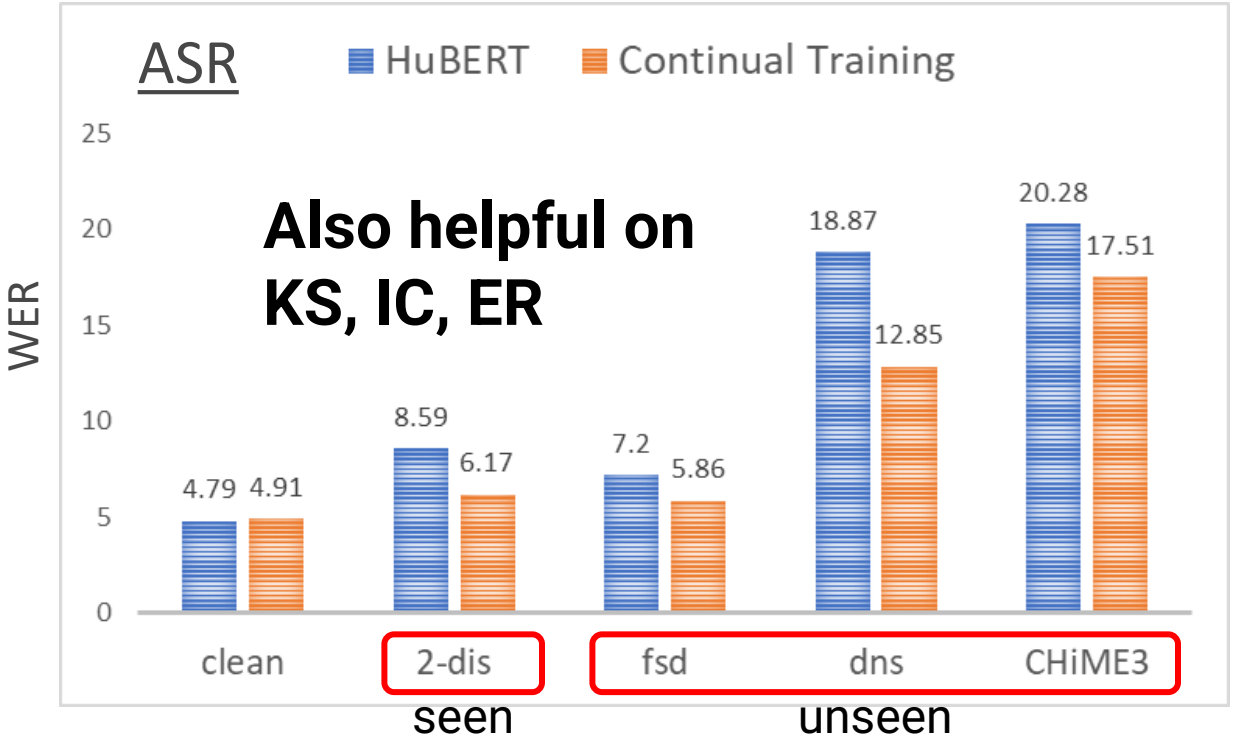
Different domains: speech distortions, speaking styles (read vs. spontaneous), accents/dialects, languages

Can self-supervised models maintain good performance? **NO**

Generalization of SSL

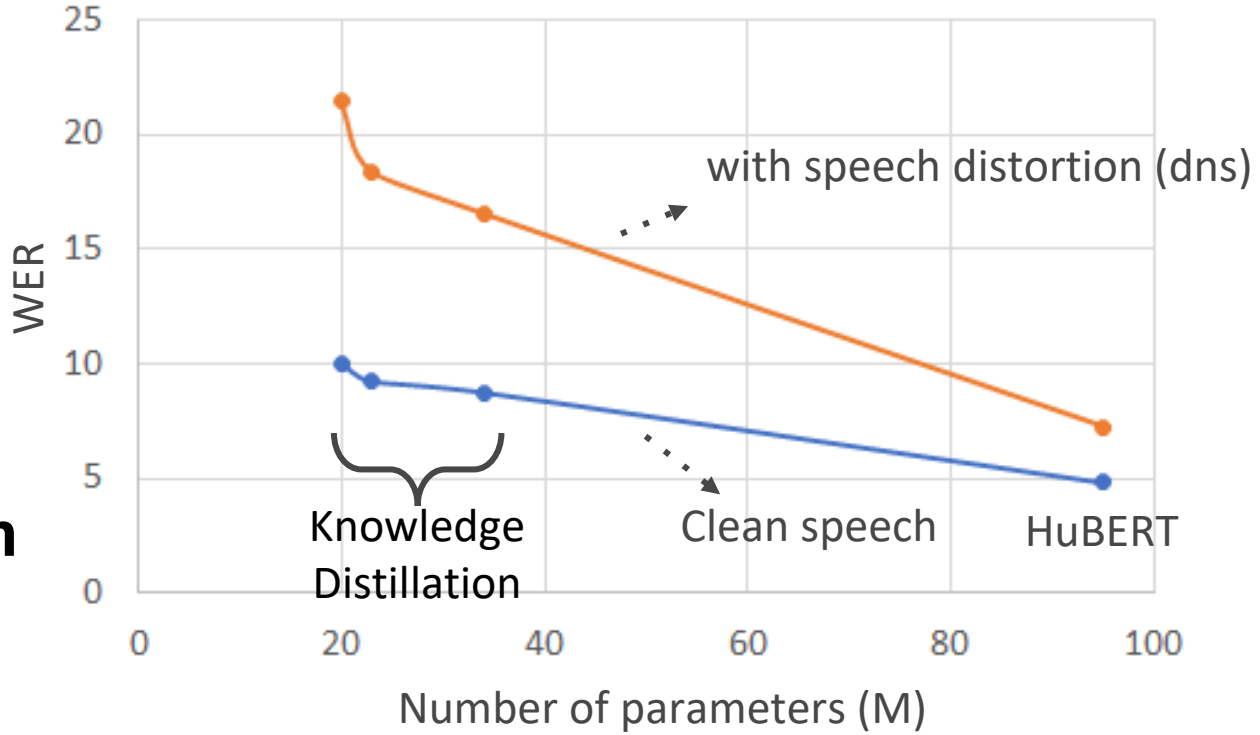
Kuan Po Huang, Yu-Kuan Fu, Yu Zhang, Hung-yi Lee, Improving Distortion Robustness of Self-supervised Speech Processing Tasks with Domain Adaptation, Interspeech, 2022





Compressed SSL Models are less robust.

ASR



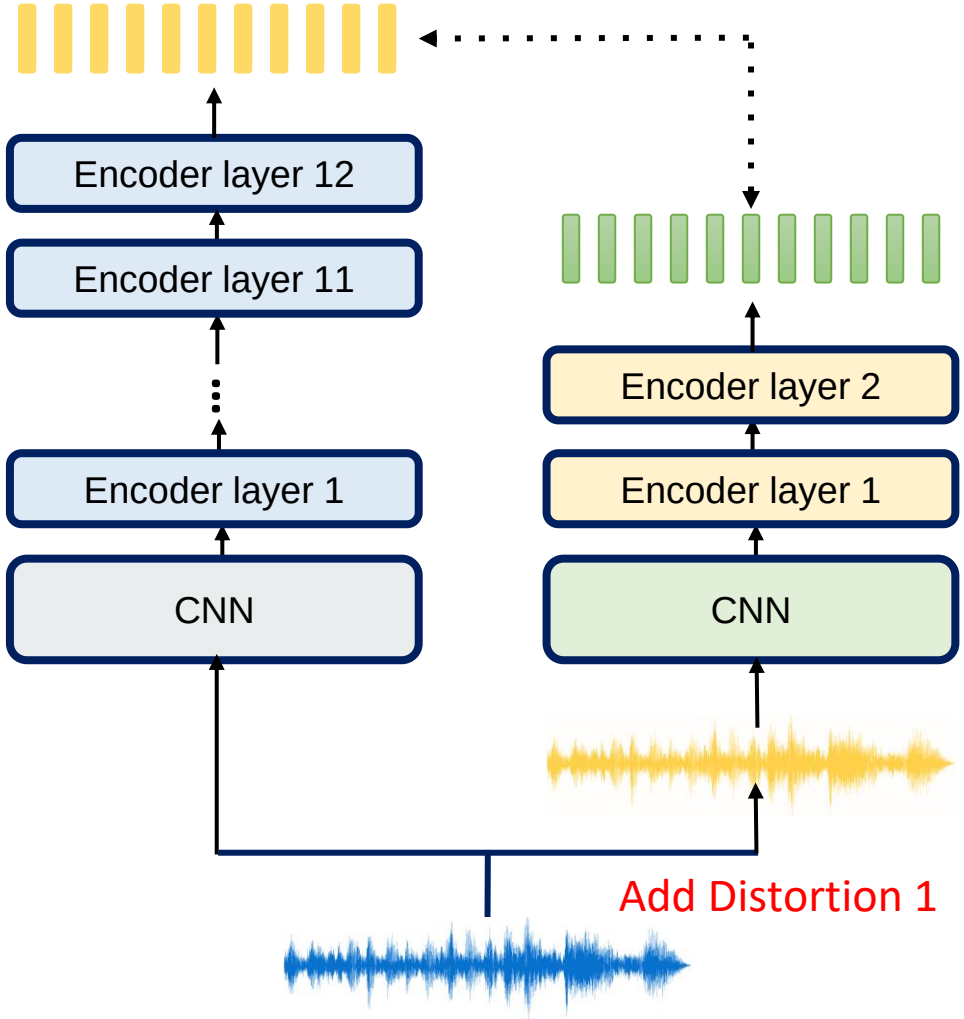
The same observation on KS, IC, ER

DistilHuBERT

+ Cross-Distortion Mapping

Setup 1:

- Student input: distortion
- Teacher input: clean



DistilHuBERT

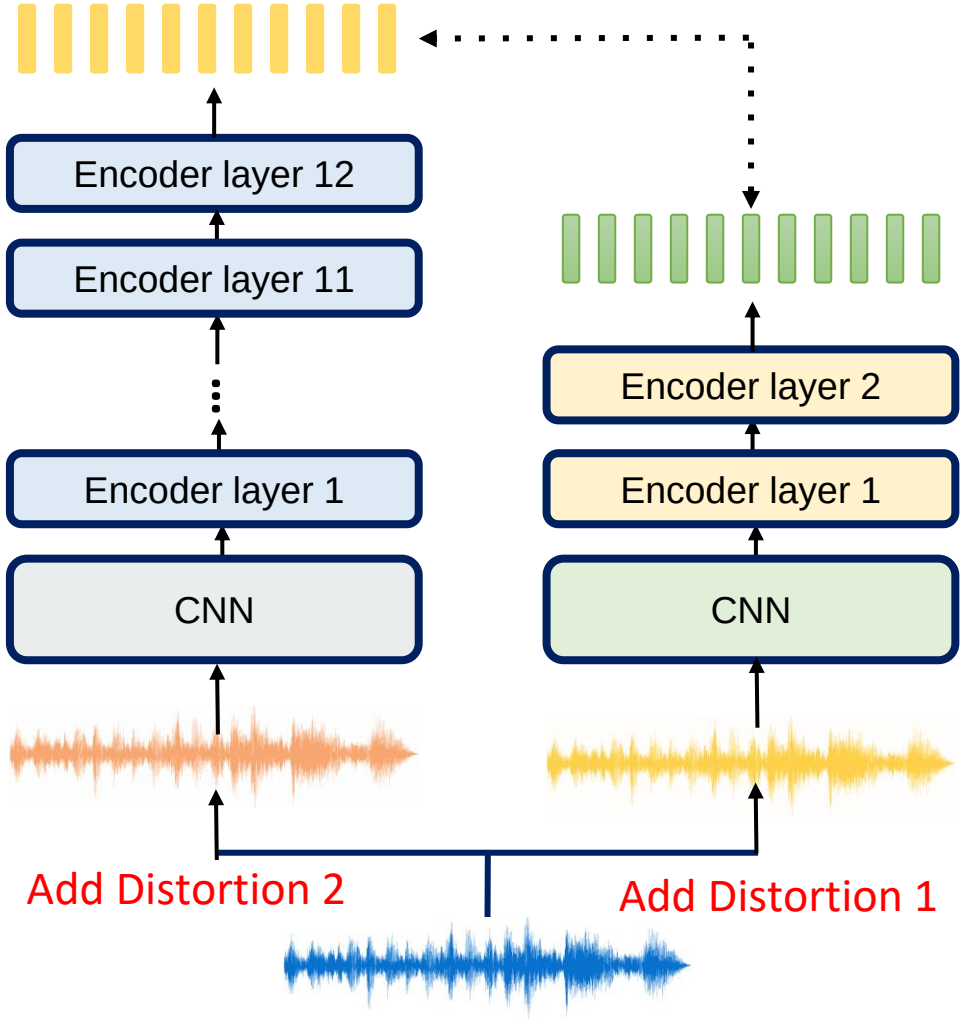
+ Cross-Distortion Mapping

Setup 1:

- Student input: distortion
- Teacher input: clean

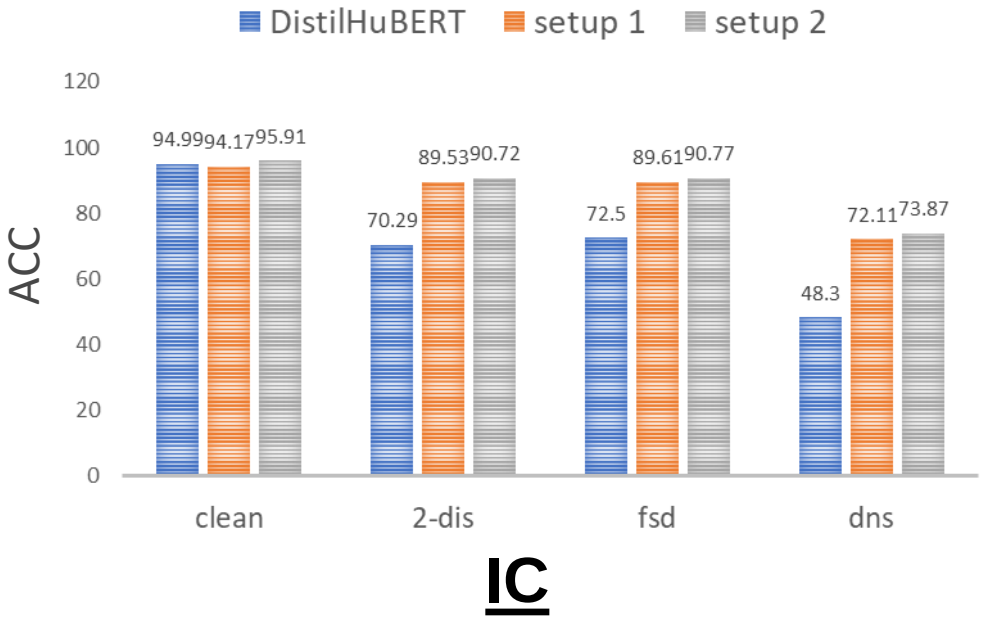
Setup 2:

- Student input: distortion 1
- Teacher input: distortion 2

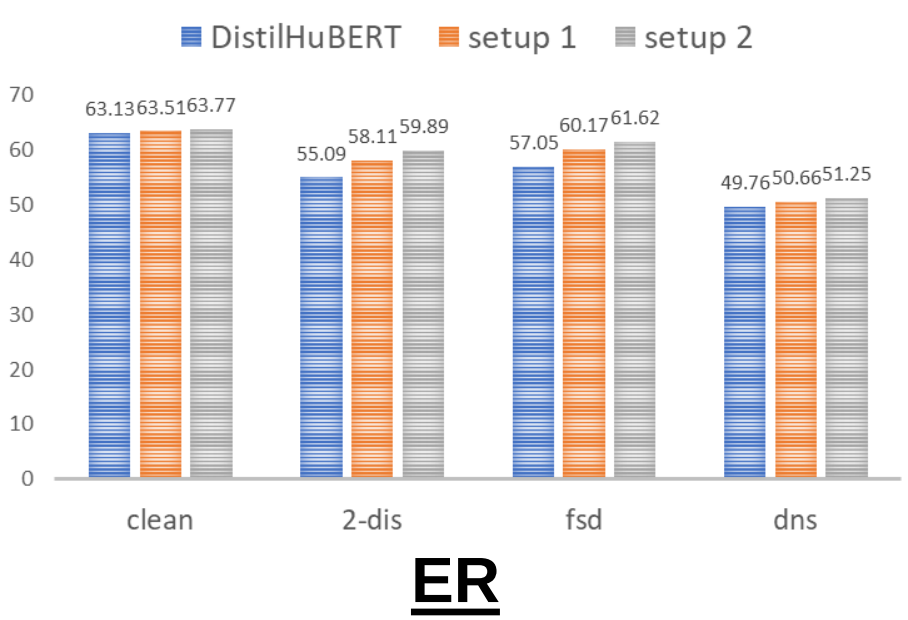


Experimental Results

Both setup 1 & 2 are helpful on all tasks, setup 2 is better than setup 1 on KS, IC, ER.

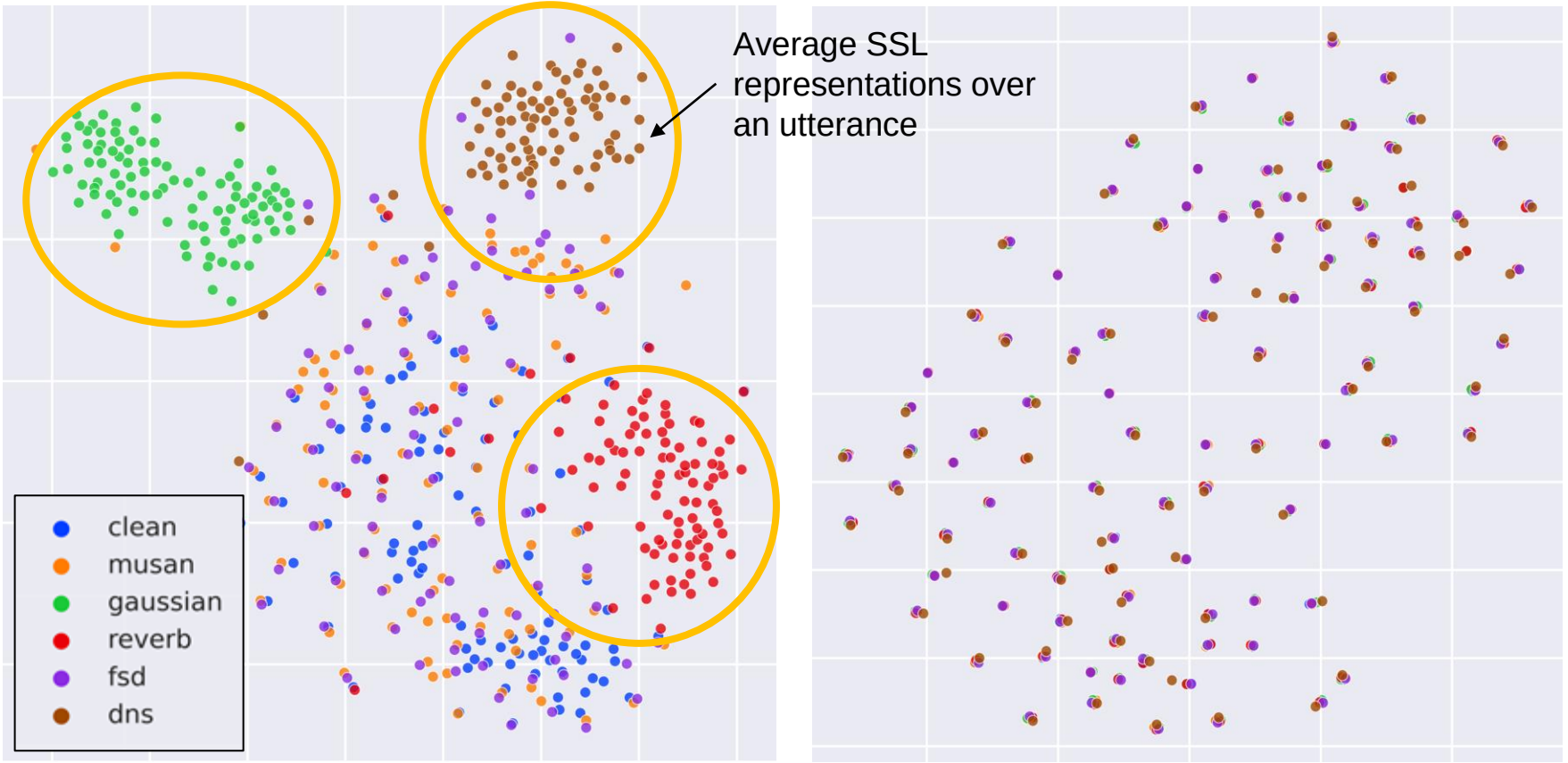


IC



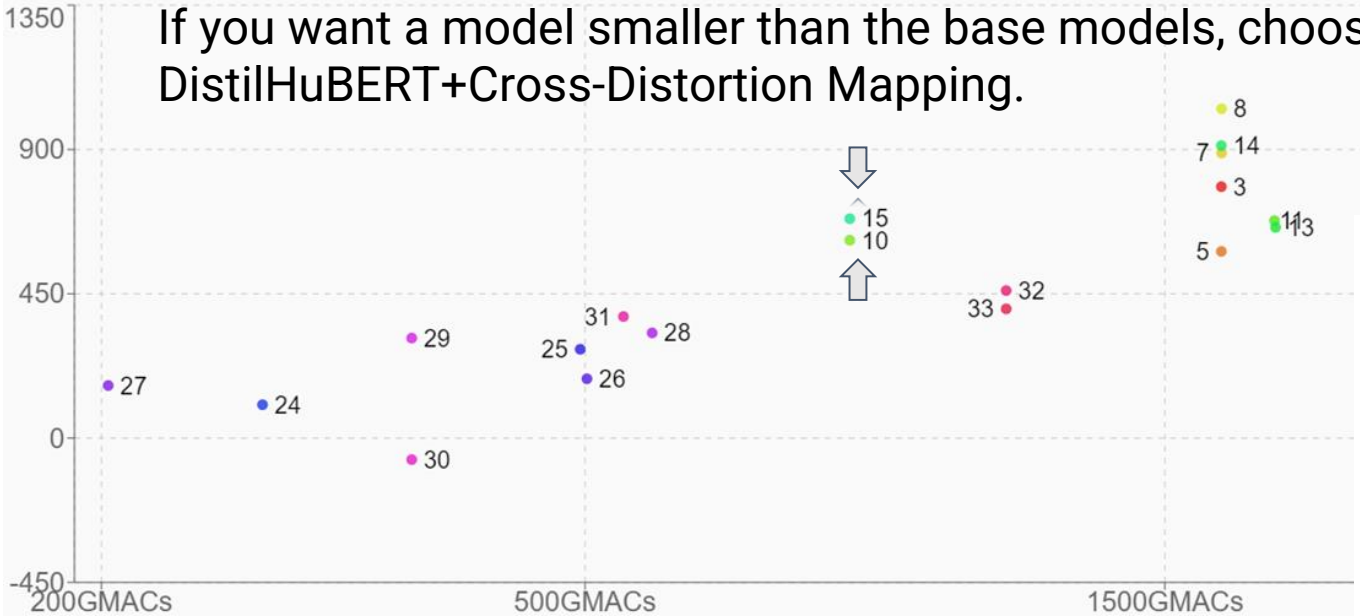
ER

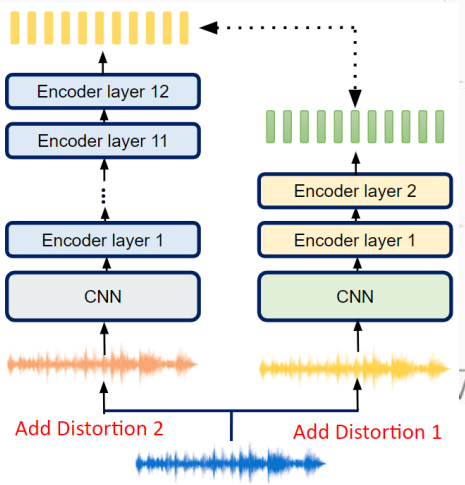
DistilHuBERT + Cross-Distortion Mapping (Setup 2)



SUPERB Leaderboard - Hidden-set Track

If you want a model smaller than the base models, choose DistilHuBERT+Cross-Distortion Mapping.





10: DistilHuBERT (CNN + 2-layer transformer)

15: DistilHuBERT (CNN + 2-layer transformer) + Cross-Distortion Mapping

Outline

Compression

Robust

Adapter/Prompt

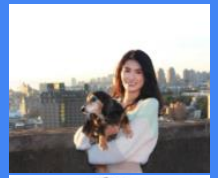
Unsupervised ASR

Visual-enhanced

Prosody

Adapter & Prompt

Kai-Wei Chang (NTU)



Zih-Ching Chen (NTU)

Allen Fu (NTU)

Chih-Ying Liu (NTU)

Hua Shen (Penn State)

Fabian Ritter (NUS)

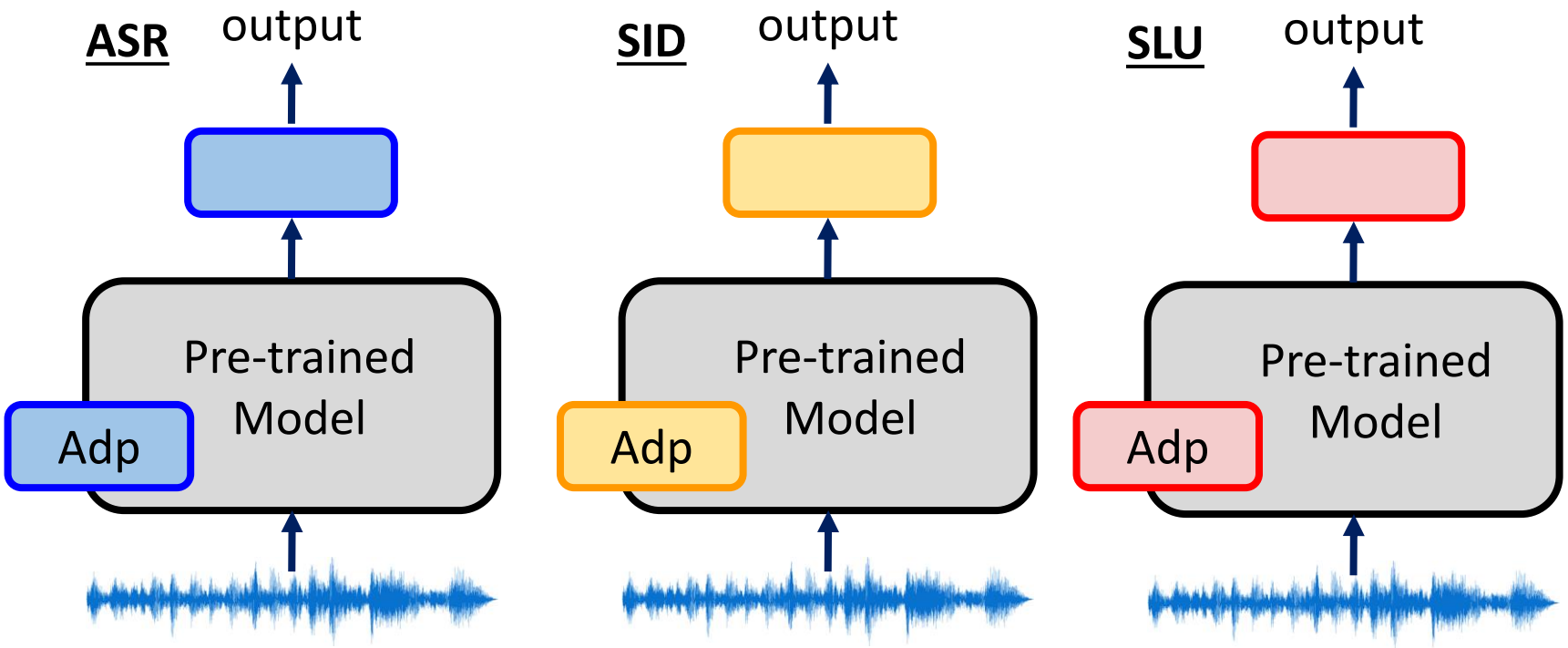
Yu-Kai Wang (NTU)

Shih-Ju Hsu (NTU)

Daniel Li (Meta)

Hung-yi Lee (NTU)

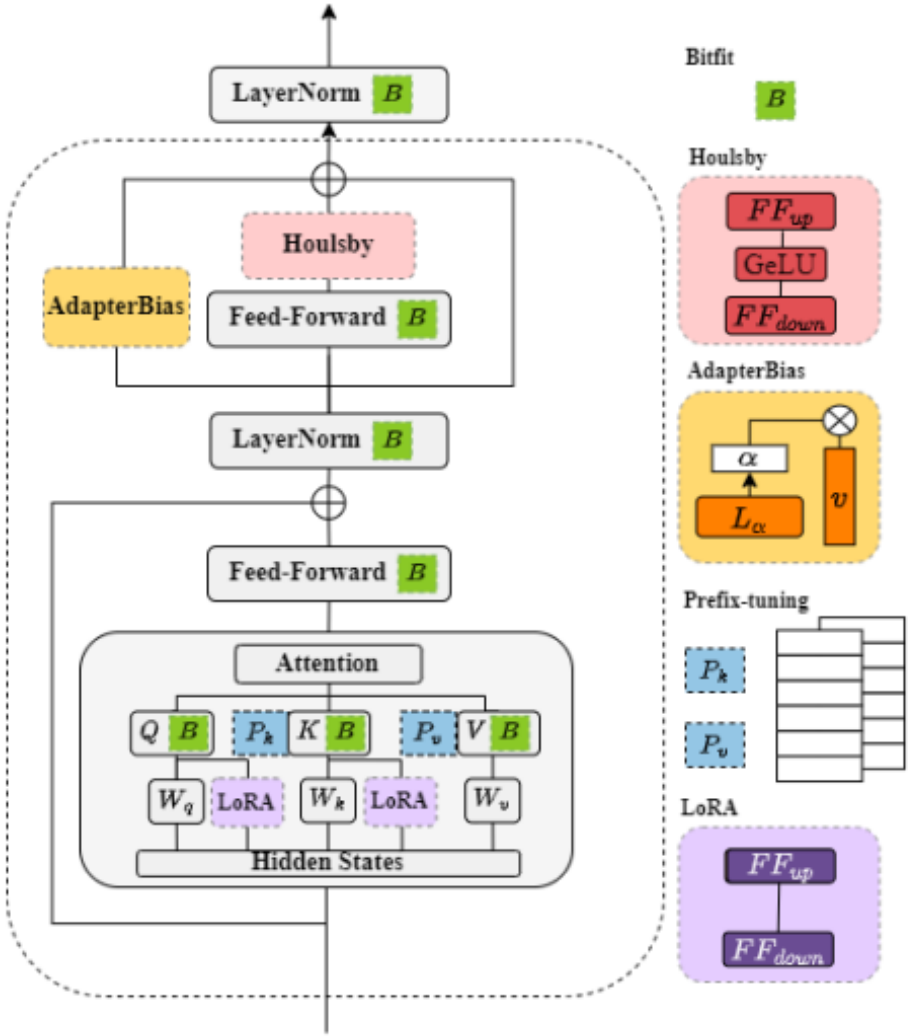
How to use pre-trained Models? **Adapter**



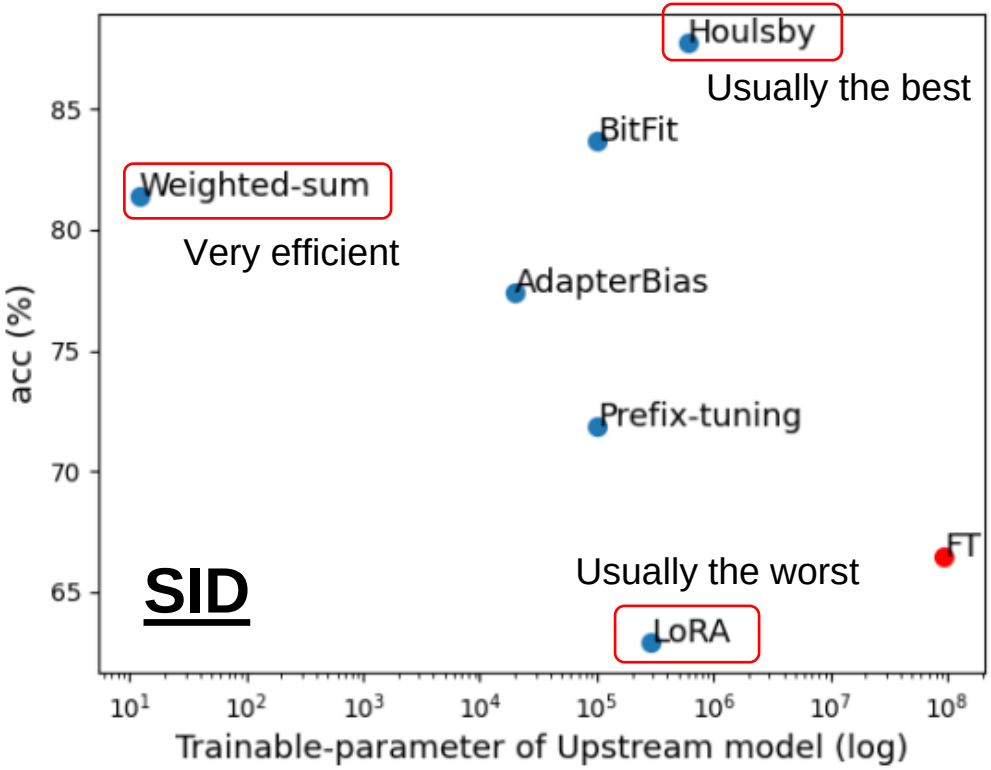
Instead the whole SSL model, only store an Adapter for each task

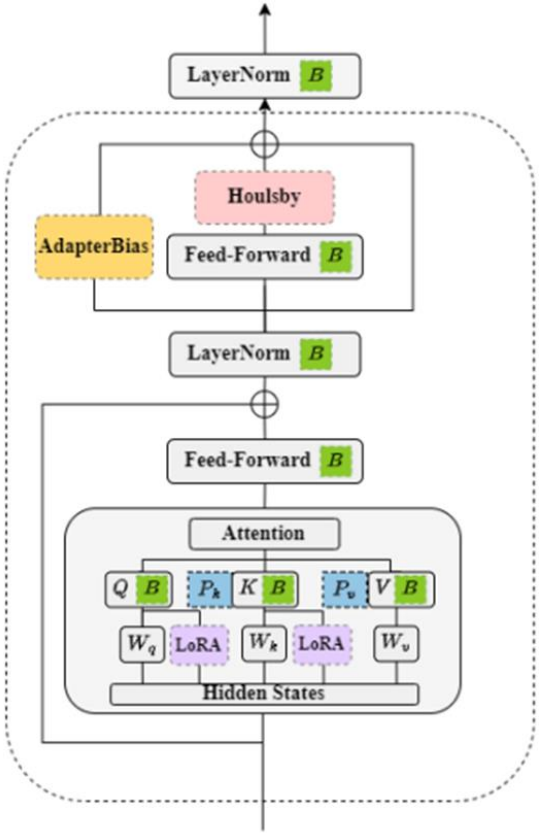
Adapter has been widely studied in NLP.

https://adapterhub.ml/

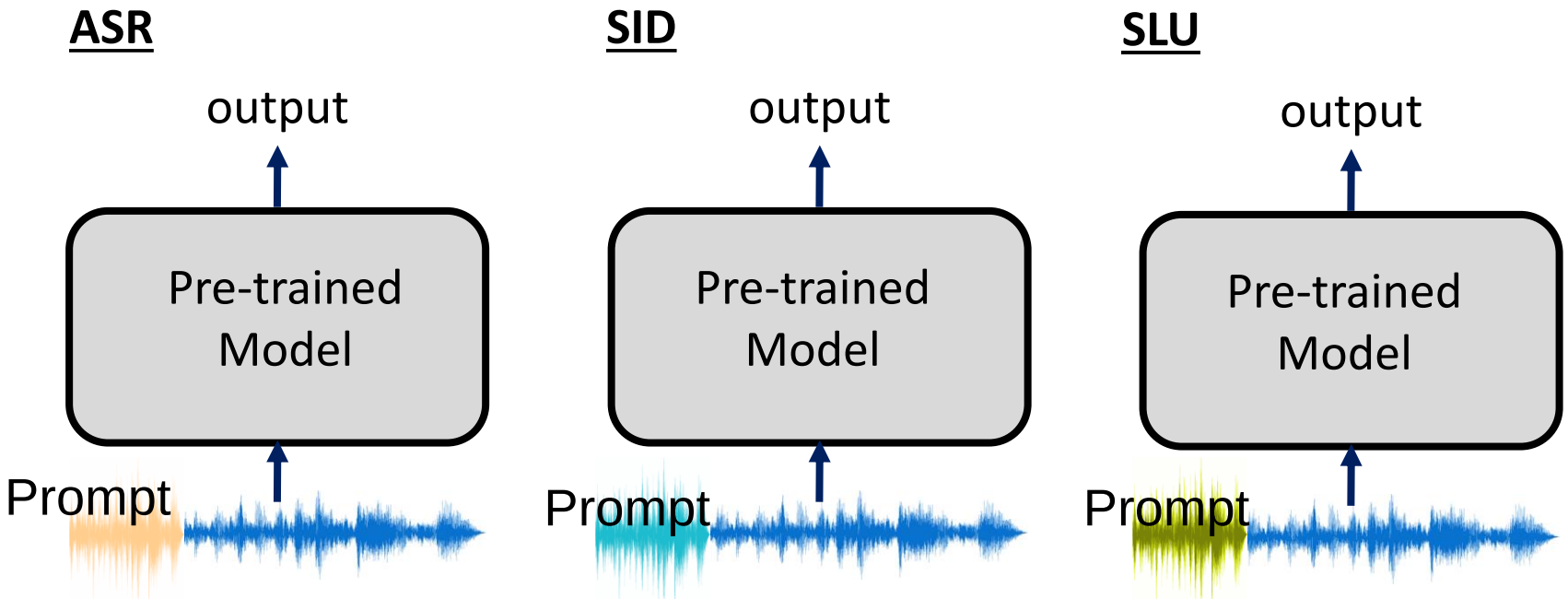


How to use pre-trained Models? **Adapter**



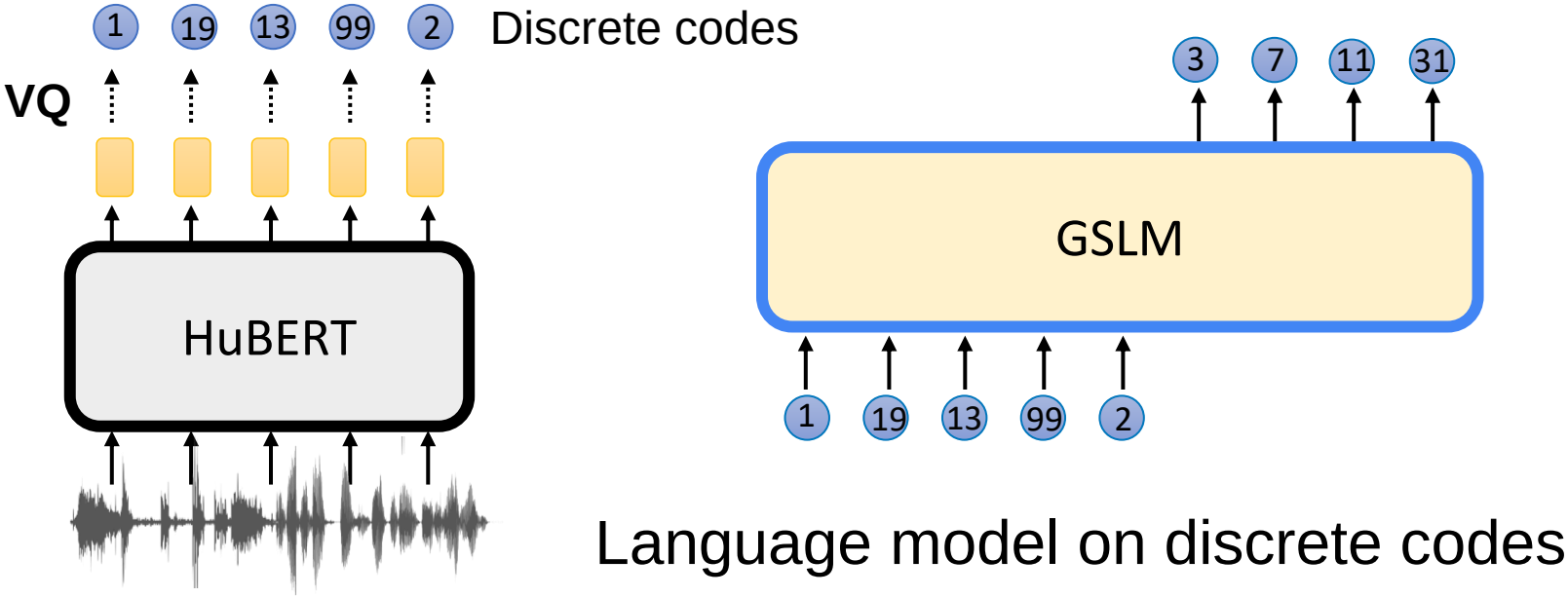


How to use pre-trained Models? **Prompting**



Here we demonstrate how to prompt *Generative Spoken Language Model* (GSLM).

Generative Spoken Language Model (GSLM)



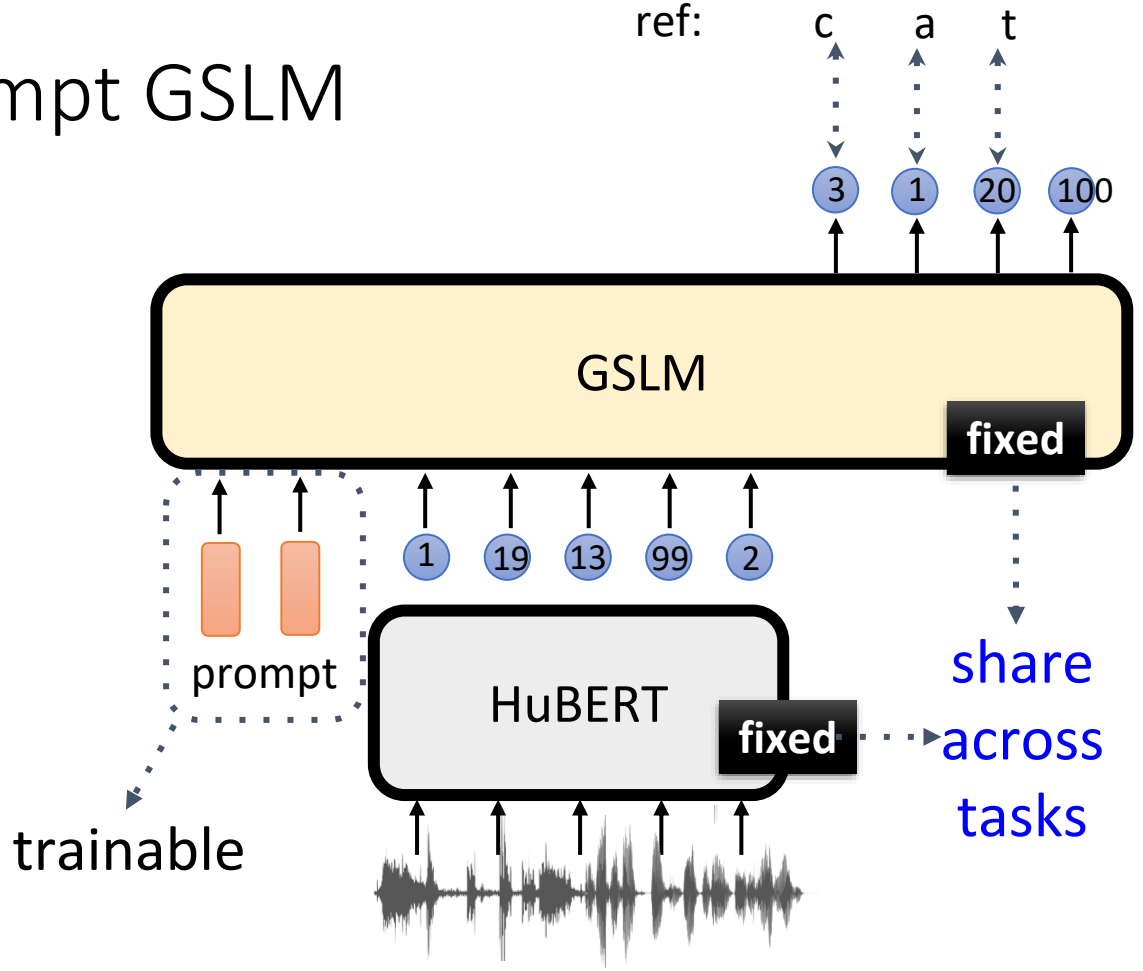
Language model on discrete codes

Attempt to Prompt GSLM

E.g., ASR

Char	Unit ID
a	1
b	2
c	3
d	4
…	…
<EOS>	100

Verbalizer



Attempt to Prompt GSLM

- KS: Keyword Spotting - Single-label Classification
- IC: Intent Classification - Multi-label Classification

		KS		IC	
		ACC ↑	**# param.**	**ACC ↑**	**# param.**
Default SUPERB framework		96.30	0.2M	98.34	0.2M
Prompt	Fixed Verbalizer	94.32	0.15M	98.10	0.15M
	Learnable Verbalizer	94.68	0.16M	98.66	0.16M

Attempt to Prompt GSLM

- ASR: Character-based speech recognition
- SF: Slot Filling

	ASR		SF	
	WER↓	# param.	F1↑	# param.
Prompt	34.17	4.5M	66.90	4.5M
Fine-Tuning	**6.42**	43M	**88.53**	43M

Kai-Wei Chang, Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee, SpeechPrompt: An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks, Interspeech, 2022

Outline

Compression

Robust

Adapter/Prompt

Unsupervised ASR

Visual-enhanced

Prosody

Unsupervised ASR with SSL and its extension use

Ann Lee (Meta)

Paola Garcia (JHU)

David Harwath (UT)

Shinji Watanabe (CMU)

Hung-yi Lee (NTU)

Dongji Gao (JHU)

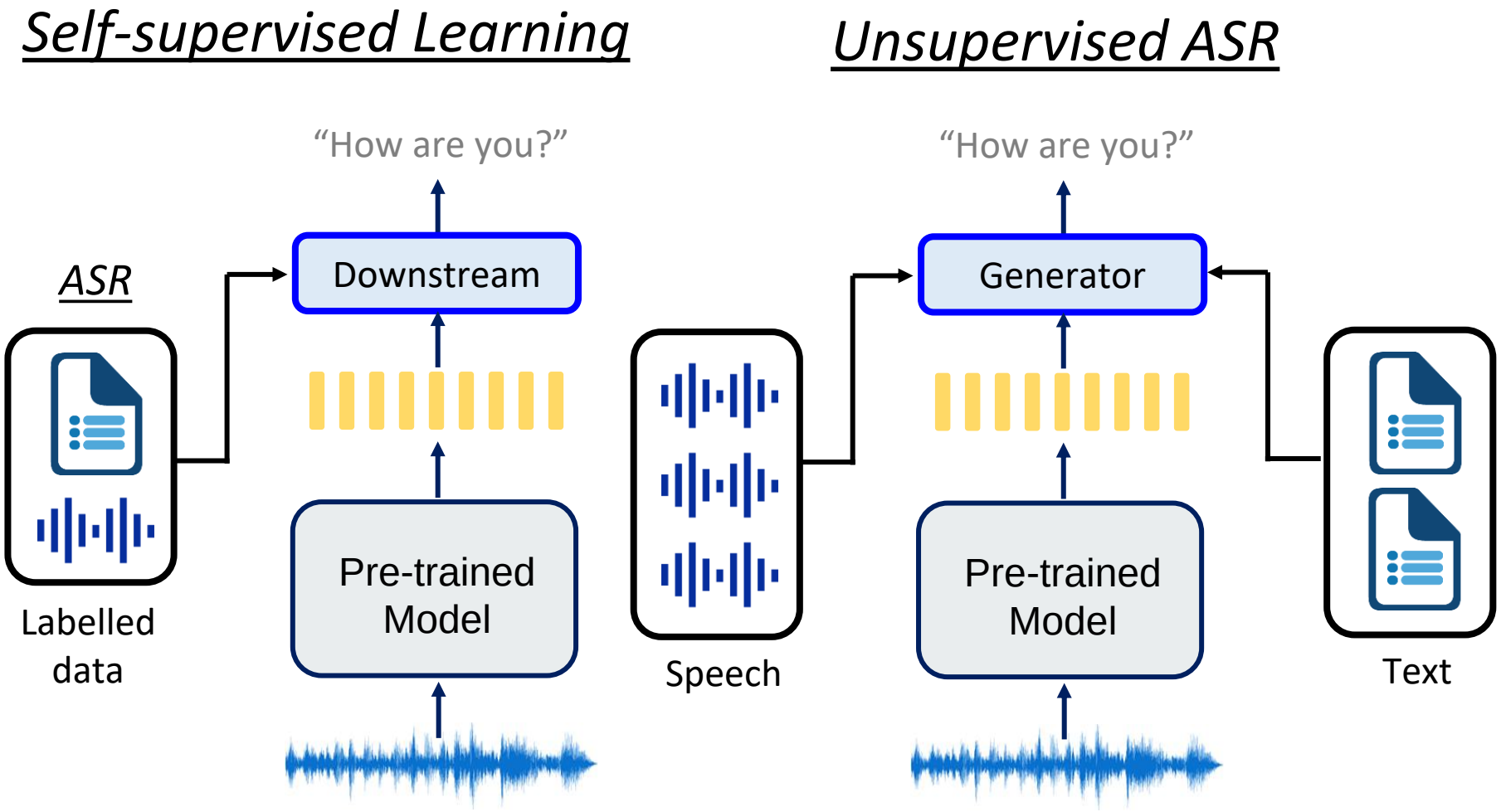
Virginia Layne Berry (UT)

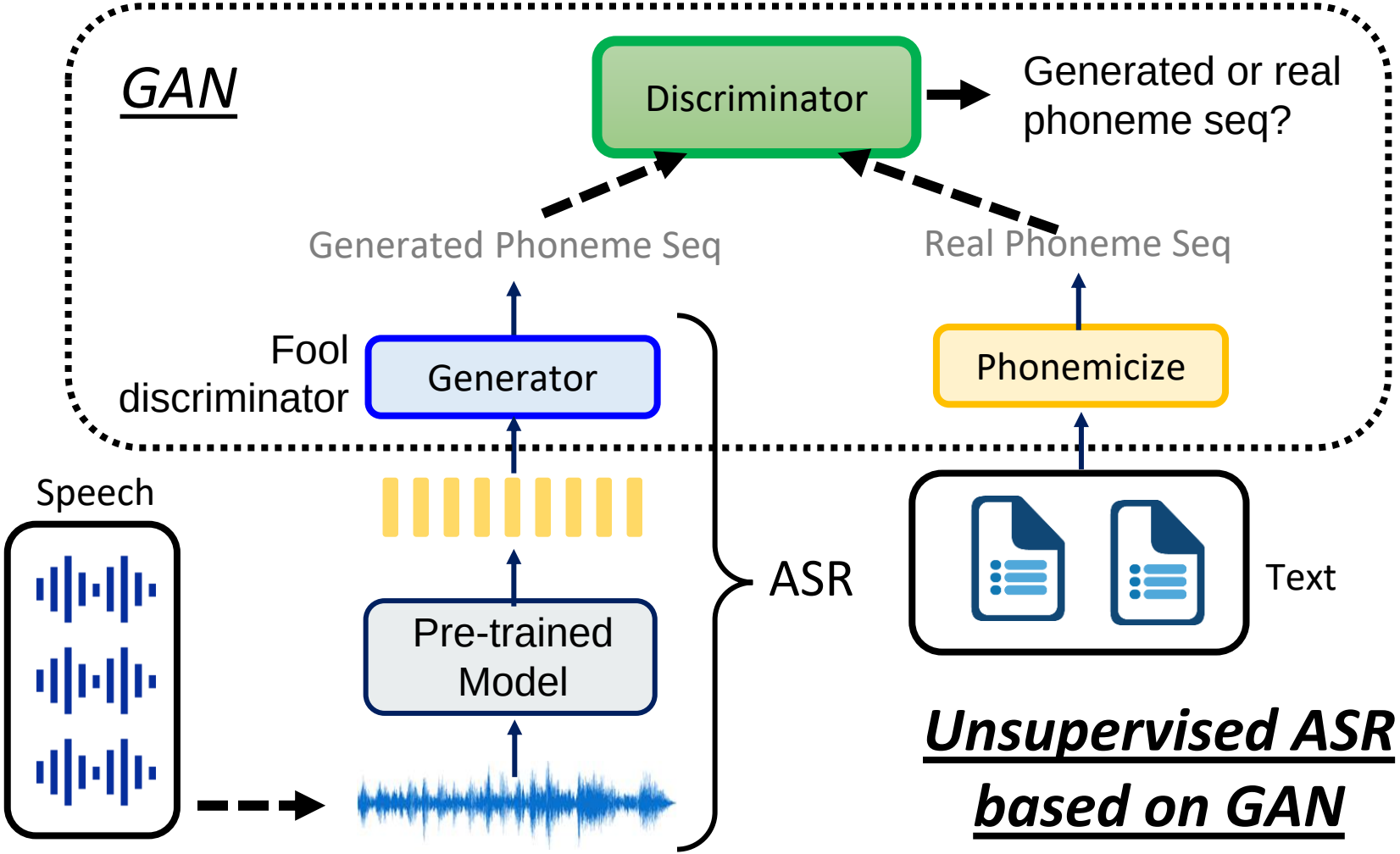
Jiatong Shi (CMU)

Yen Meng (NTU)

Hsuan-Jui Chen (NTU)

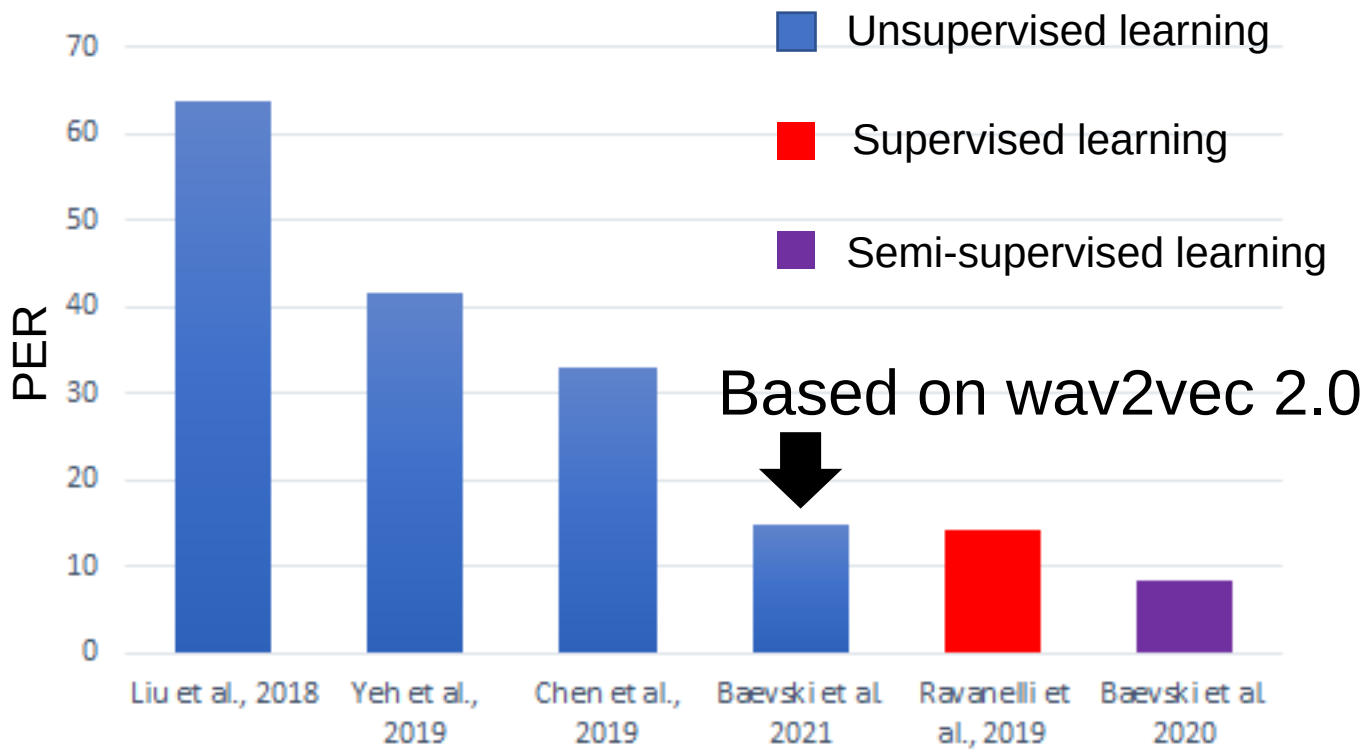
Andy. Liu (NTU)


Self-supervised Learning
Unsupervised ASR
"How are you?"
"How are you?"
ASR
Downstream
Generator
Pre-trained Model
Pre-trained Model
Labelled data
Speech
Text


GAN
Discriminator
Generated or real phoneme seq?
Generated Phoneme Seq
Real Phoneme Seq
Fool discriminator
Generator
Phonemicize
Speech
ASR
Pre-trained Model
Text
Unsupervised ASR based on GAN

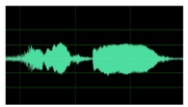
Unsupervised Speech Recognition

- TIMIT



Extention usage of Unsupervised ASR

Of course, use for ASR!



$S = \{s_n \in \mathbb{Z} | n = 1, ..., N\}$

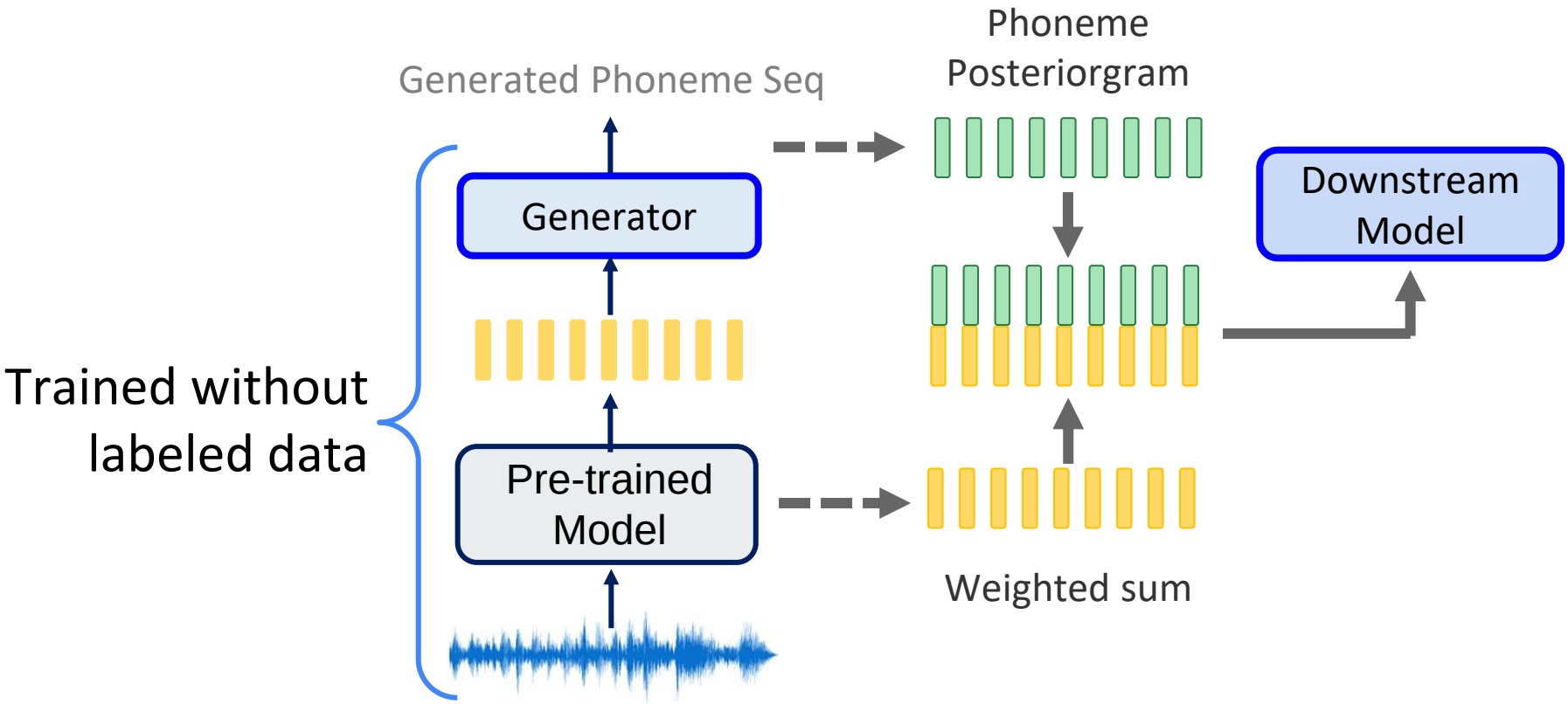
I'm not you

$W = \{w_l \in \mathcal{V} | l = 1, ..., L\}$

Next?

- Use as **a segmenter**
  - Provide phoneme boundaries
  - Used in sequence reduction
- Use as **a self-supervised model**
  - No supervised data needed
- Use as **a connector**
  - Connecting Speech SSL with Text SSL

Unsupervised ASR as an SSL Model



Unsupervised ASR as an SSL Model (SUPERB Hidden-set Leaderboard)

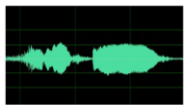
Models	Phone Recognition (↓)	Speech Recognition (↓)	Emotion Recognition (↑)	Query by Example (↓)	SUPERB Score (↑)
Wav2vec2	22.55	23.58	60.99	22.48	902
HuBERT	18.22	**22.03**	64.84	33.05	959
U ASR	**17.22**	23.75	**65.11**	**21.99**	**962**

- **Better** performances in **PR**
- **Similar** performances in **ASR**
- **Outperforms Hubert** on several tasks

- SUPERB Score is a scaled score over 10 superb hidden-set tasks (from 0 - 1000). Calculation is based on https://superbbenchmark.org/challenge-slt2022/metrics
- All numbers are evaluated by SUPERB **hidden sets** (training & evaluation)

Extention usage of Unsupervised ASR

Of course, use for ASR!



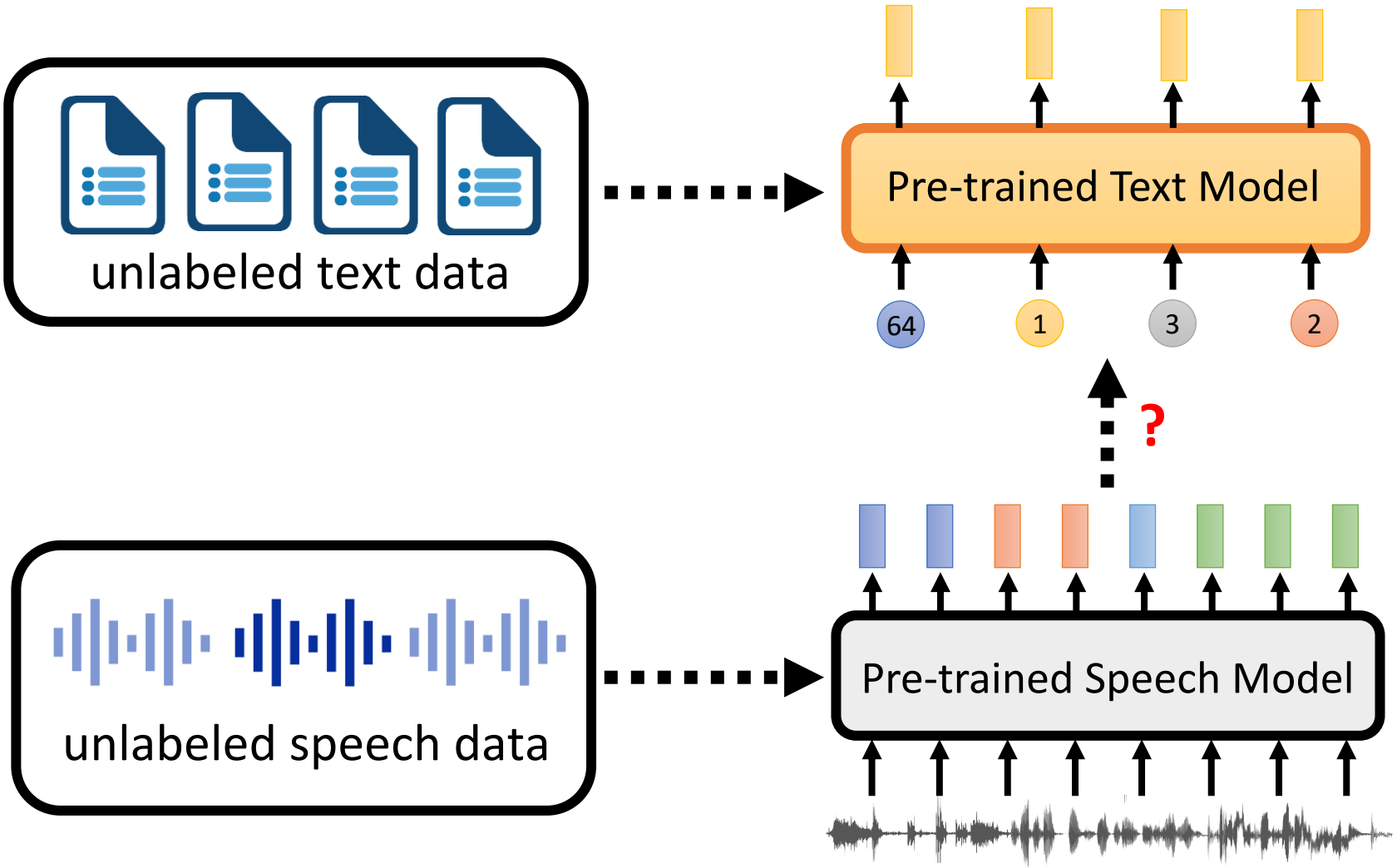
$S = \{s_n \in \mathbb{Z} | n = 1, ..., N\}$

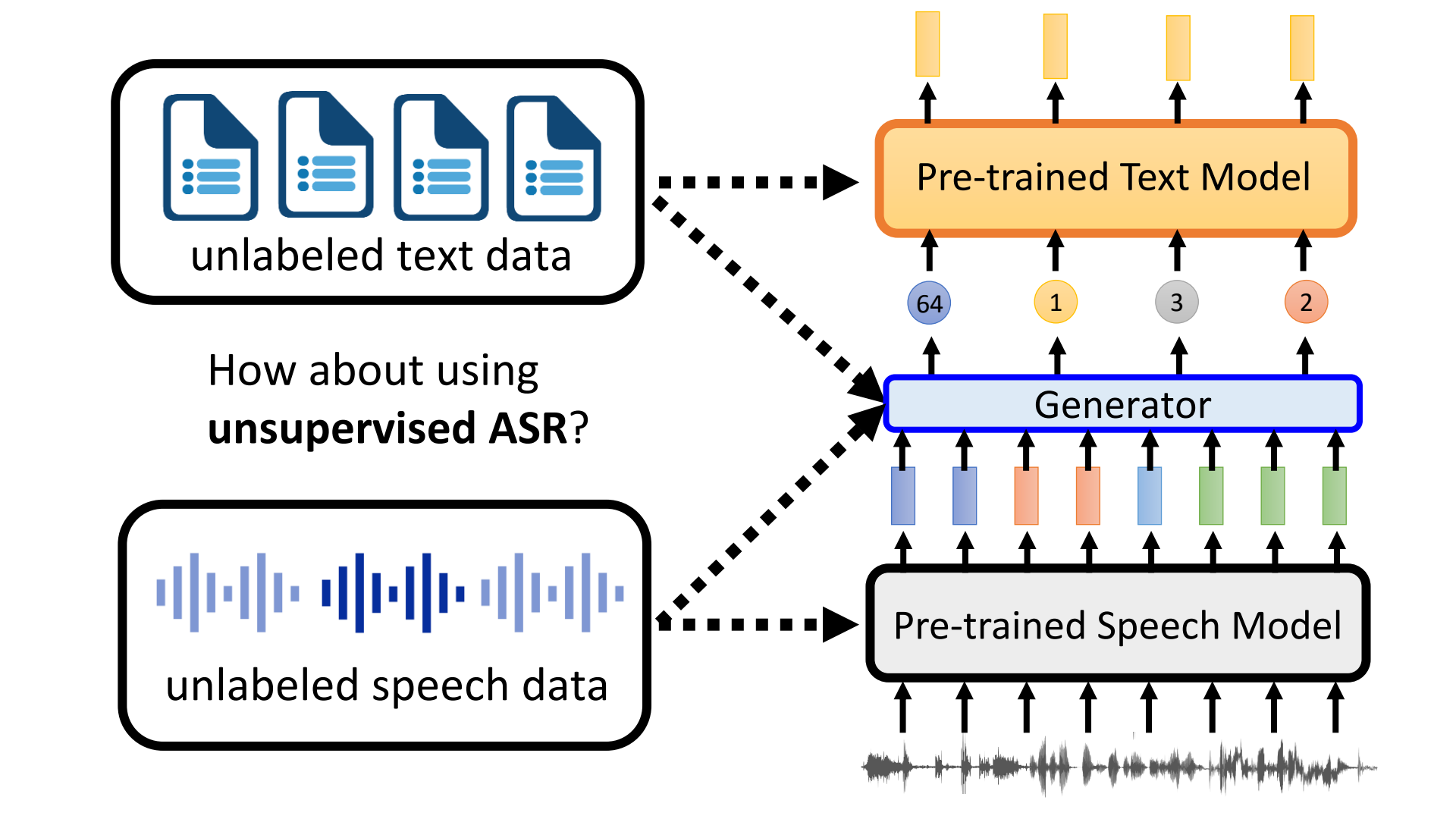
I'm not you

$W = \{w_l \in \mathcal{V} | l = 1, ..., L\}$

Next?

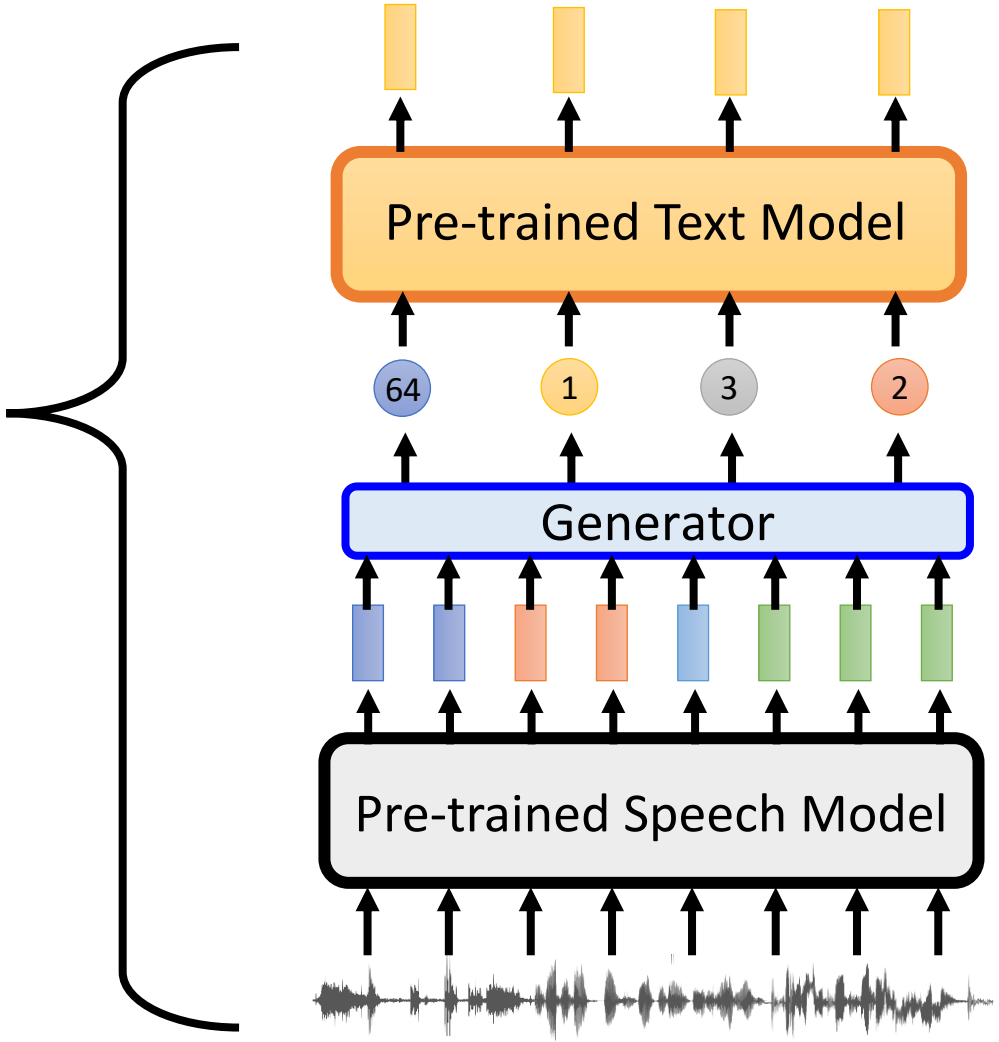
- Use as **a segmenter**
  - Provide phoneme boundaries
  - Used in sequence reduction
- Use as **a self-supervised model**
  - No supervised data needed
- Use as **a connector**
  - Connecting Speech SSL with Text SSL


unlabeled text data
Pre-trained Text Model
64
1
3
2
?
unlabeled speech data
Pre-trained Speech Model


unlabeled text data
How about using
unsupervised ASR?
unlabeled speech data
Pre-trained Text Model
64
1
3
2
Generator
Pre-trained Speech Model

New self-supervised Model

For spoken language understanding?



Unsupervised ASR as a Connector (Connector Options)

Tasks	Fixed - FSC (↑)	Fine-tuning - SLURP (↑)
Baseline (wav2vec2)	94.38	82.82
KM	93.69	85.31
U ASR + Phoneme T5	**94.88**	**86.14**

- KM methods **cannot** function well **without fine-tuning**
- **UASR** as a connector **outperforms KM** methods in both **fixed** and **fine-tuning** cases

Outline

Compression

Robust

Adapter/Prompt

Unsupervised ASR

Visual-enhanced

Prosody

Visually-Enhanced SSL Models

PIs:

David Harwath
(UT Austin)

Hung-yi Lee
(NTU)

Students:

Layne Berry
(UT Austin)

Heng-Jui Chang
(MIT)

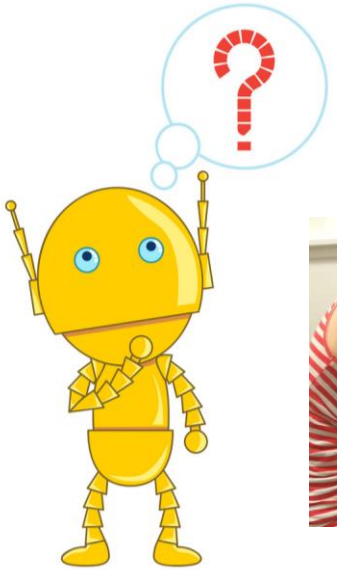
Ian Shih
(NTU)

Jeff Wang
(NTU)

Visually Enhanced Pre-trained Speech Model

(There is a little girl wearing sunglasses.)

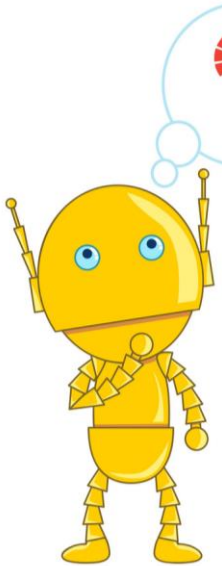
Associated Image



Pre-trained Model

Improving Visually Enhanced Model

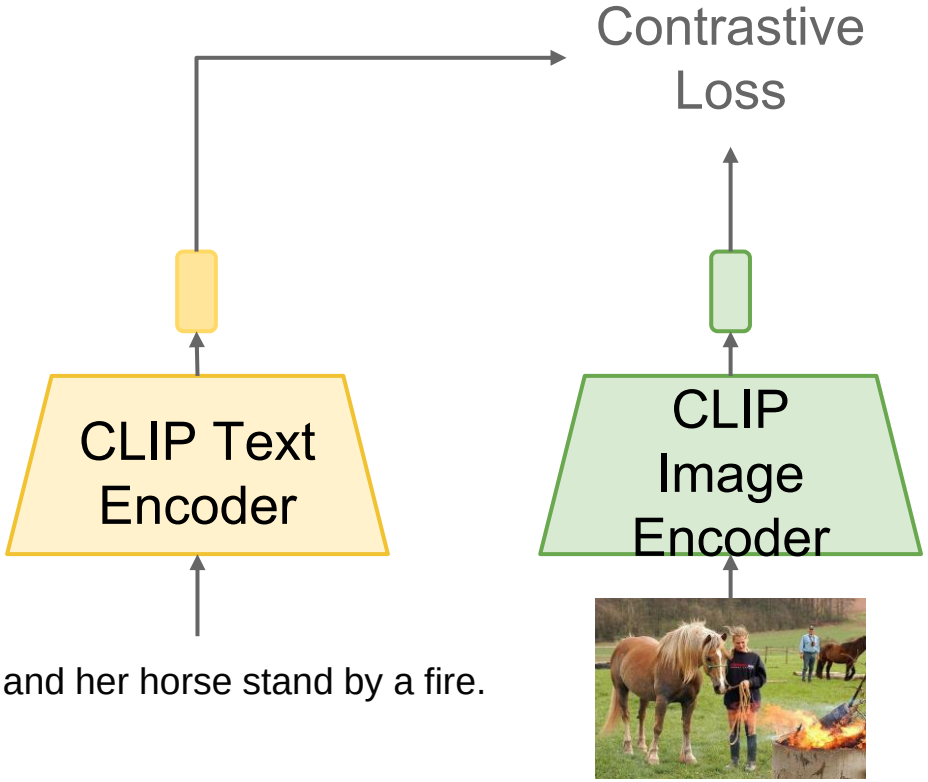
(There is a little girl ……)



Pre-trained Model

sheep on the grass

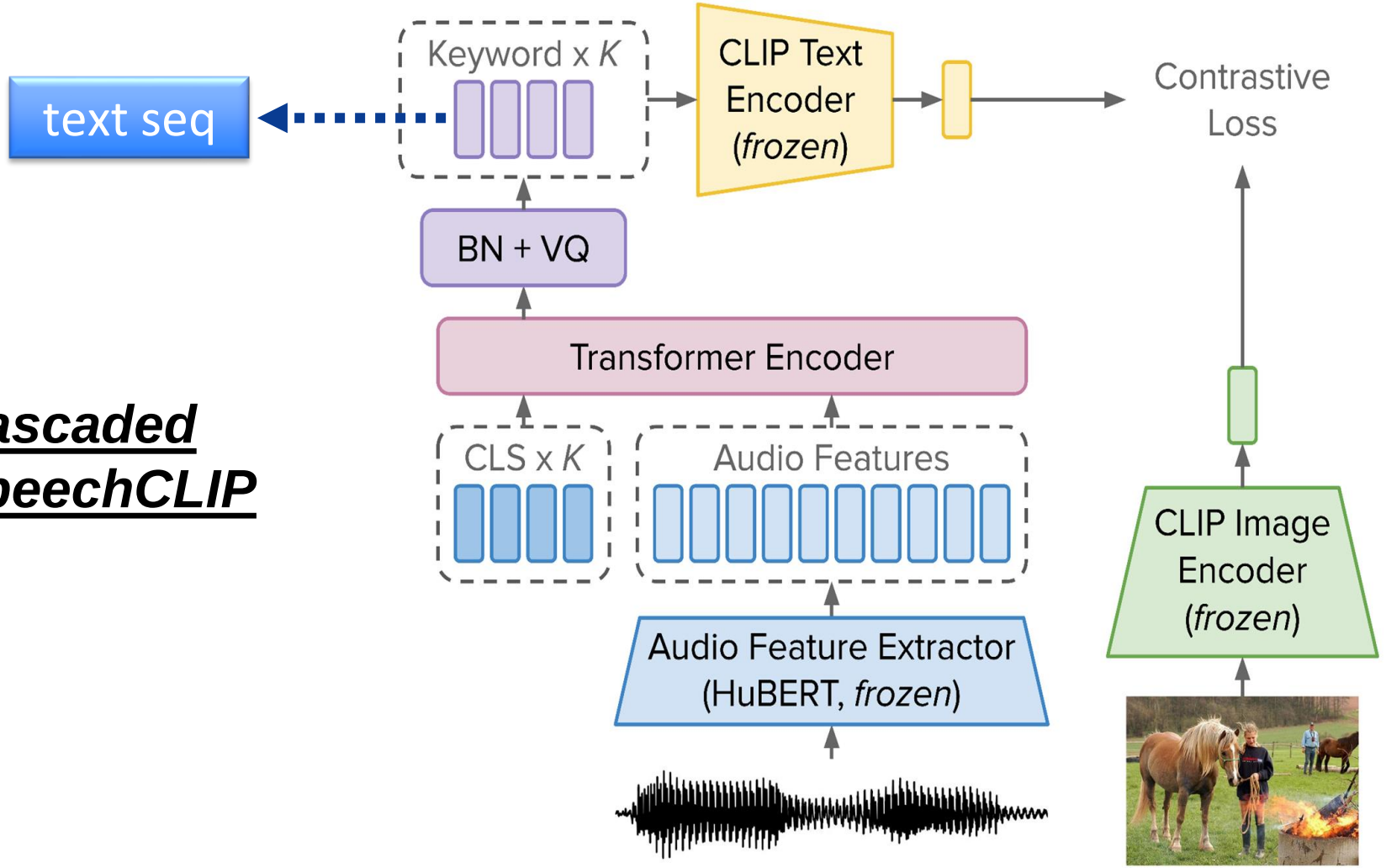
Contrastive **L**anguage-**I**mage **P**re-training **CLIP** (Radford et al.)



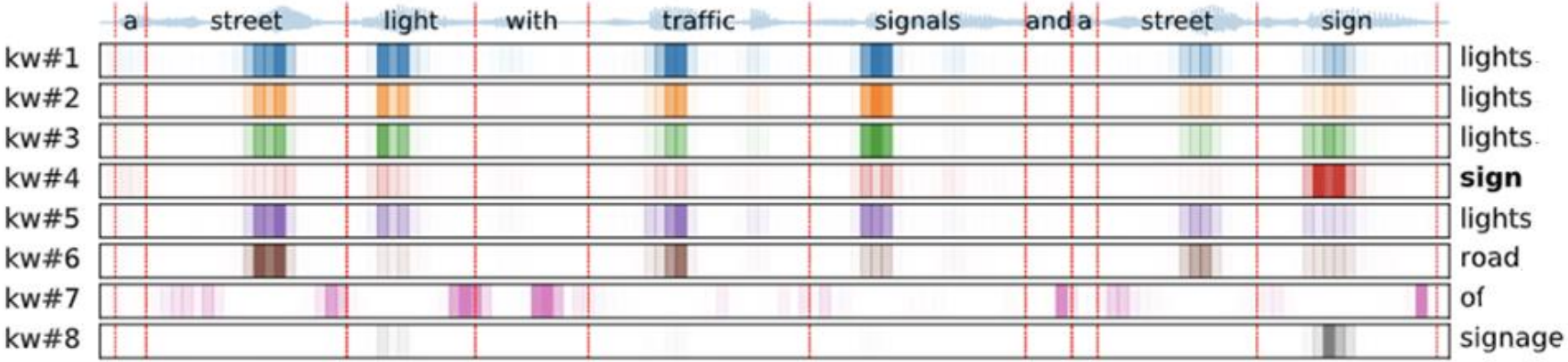
Some info about CLIP

1. Made by OpenAI
   Trained on **400M** Image-Text pairs
2. Training last for **12** days on **256** v100 GPUs
3. Trained by contrastive loss to learn a **shared embedding space for image and text**

Cascaded SpeechCLIP



Reference



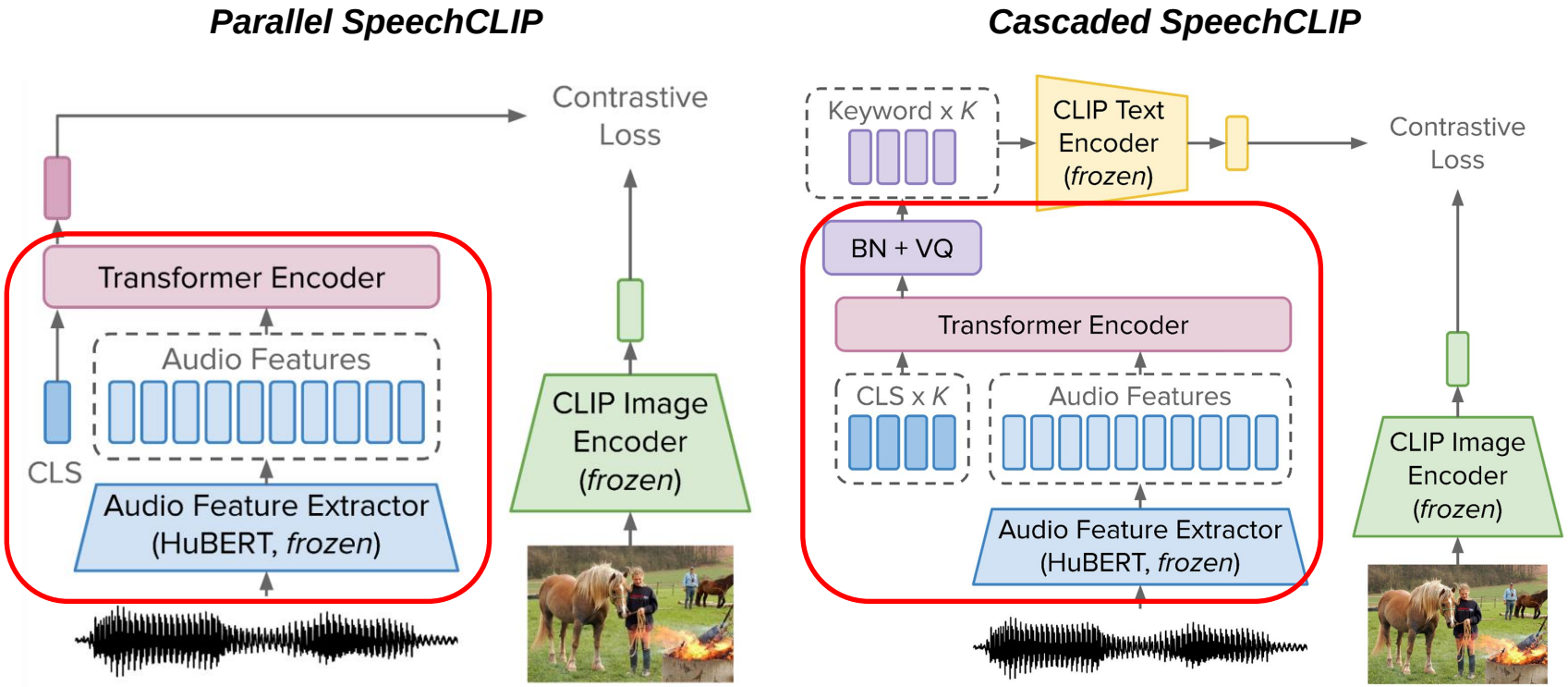
Model output

Keyword hit rates for cascaded SpeechCLIP

† : trained on Flickr8K, ‡ : trained on SpokenCOCO

Model	kw1	kw2	kw3	kw4	kw5	kw6	kw7	kw8	Avg
Base†	57.0	25.6	20.2	5.0	20.0	26.5	10.5	16.6	22.7
Large†	56.5	19.6	20.5	37.5	21.7	34.6	26.4	44.7	32.7
Large‡	27.5	22.4	35.8	61.0	21.6	54.2	60.1	22.9	38.2

Unsupervised ASR without paired speech-text data?



Future work: Evaluate SpeechCLIP on SUPERB tasks

Outline

Compression

Robust

Adapter/Prompt

Unsupervised ASR

Visual-enhanced

Prosody

SSL for Prosody

Guan-Ting Lin

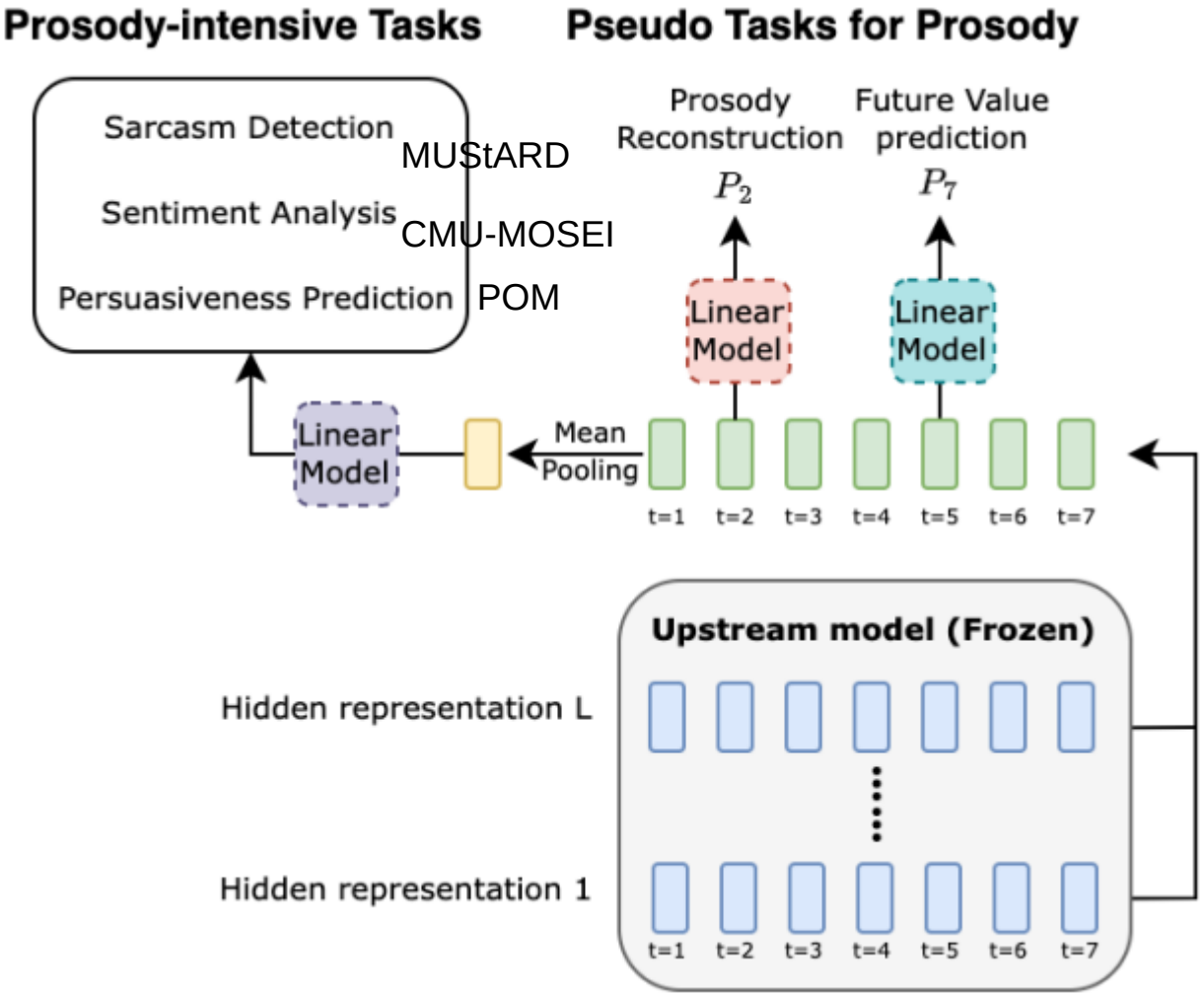
Chi-Luen Feng

Samuel Miller

Nigel Ward

Hung-yi Lee

Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li



- New audio-only SOTA on the prosody-intensive tasks.
- The representations contain prosodic features.

  (conclusion is the same if the input is not English)

Outline

Compression

Robust

Adapter/Prompt

Unsupervised ASR

Visual-enhanced

Prosody

s3prl

s3prl

Self-Supervised Speech Pre-training and Representation Learning Toolkit.

youtu.be/PkMFnS6cjAc

1.4k stars 315 forks

https://github.com/s3prl/s3prl/

Used by 14

Contributors 38

Creators

Leo

Andy

+ 27 contributors

Prof. Hung-yi Lee, Advisor & Sponsor

Acknowledgement

An AIHPC & GPU Cloud Service Provider

Thanks for supporting computing resources!

Paper (2 INTERSPEECH, 5 SLT papers)

- Improving Distortion Robustness of Self-supervised Speech Processing Tasks with Domain Adaptation, Interspeech, 2022
- SpeechPrompt: An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks, Interspeech, 2022
- On Compressing Sequences for Self-Supervised Speech Models, SLT, 2022
- SpeechCLIP: Integrating Speech with Pre-Trained Vision and Language Model, SLT, 2022
- Improving generalizability of distilled self-supervised speech processing models under distorted settings, SLT, 2022
- On the Utility of Self-supervised Models for Prosody-related Tasks, SLT, 2022
- Exploring Efficient-tuning Methods in Self-supervised Speech Models, SLT, 2022