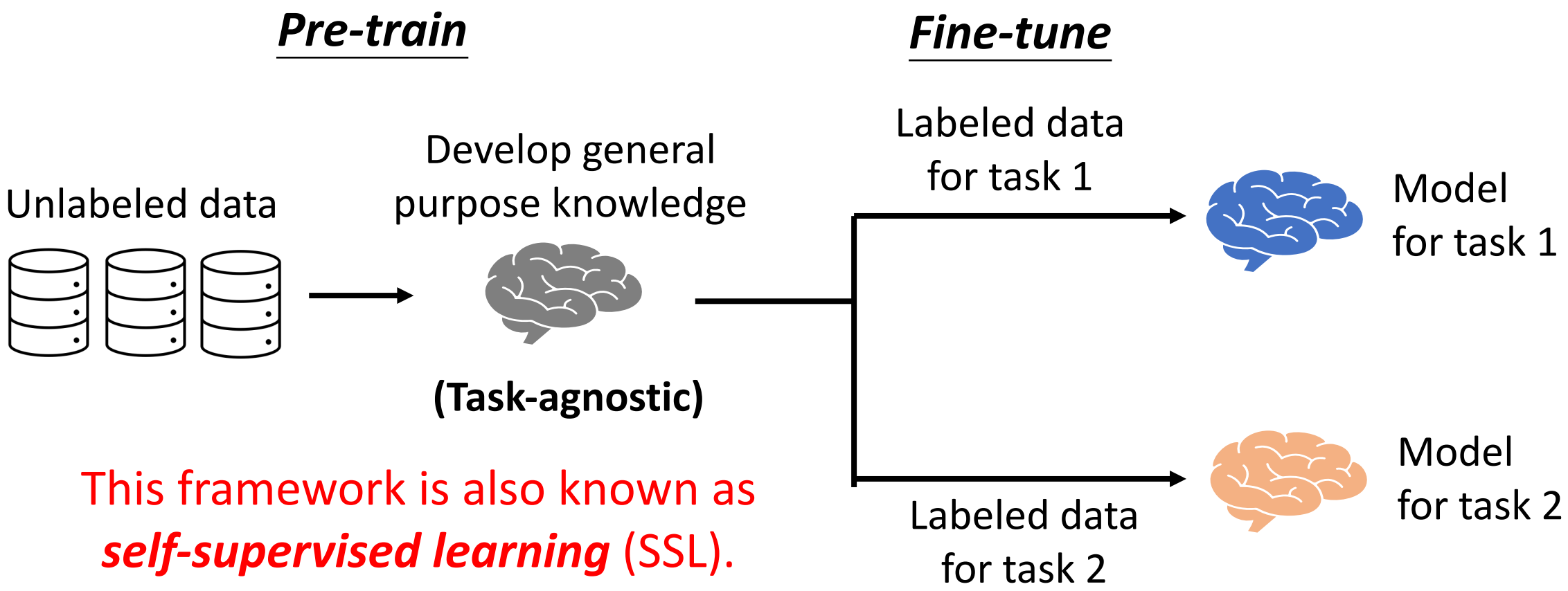
Leveraging Pre-training Models for Speech Processing

Research Group @ JSALT 2022

Speaker: Hung-yi Lee, National Taiwan University

Framework of Pre-training

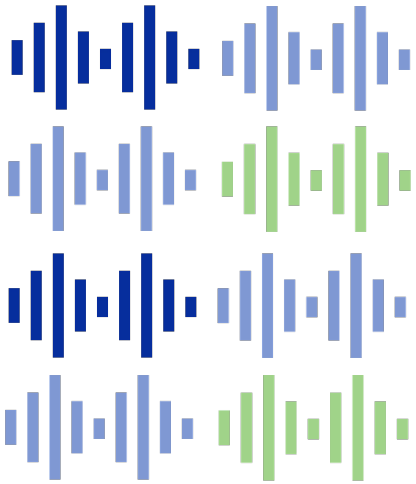


This framework is also known as ***self-supervised learning*** (SSL).

With pre-training, less labeled data for each task is required.

Self-supervised Learning for Speech

Pre-train



Unlabeled speech data

Pre-trained Speech Model

PASE+ | APC | NPC

Tera | DeCoAR | Wav2vec

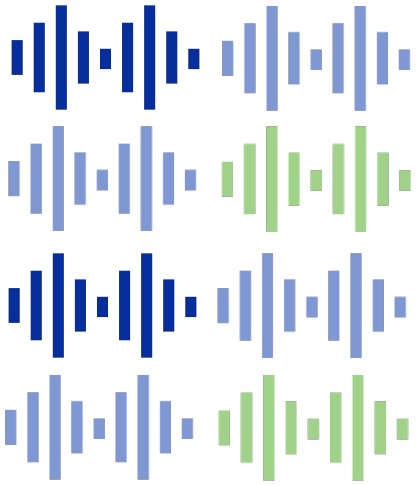
HuBERT | WavLM | Wav-BERT

Just name a few …

Self-supervised Learning for Speech

Pre-train

Fine-tune



Unlabeled speech data

Pre-trained Speech Model

"how are you"

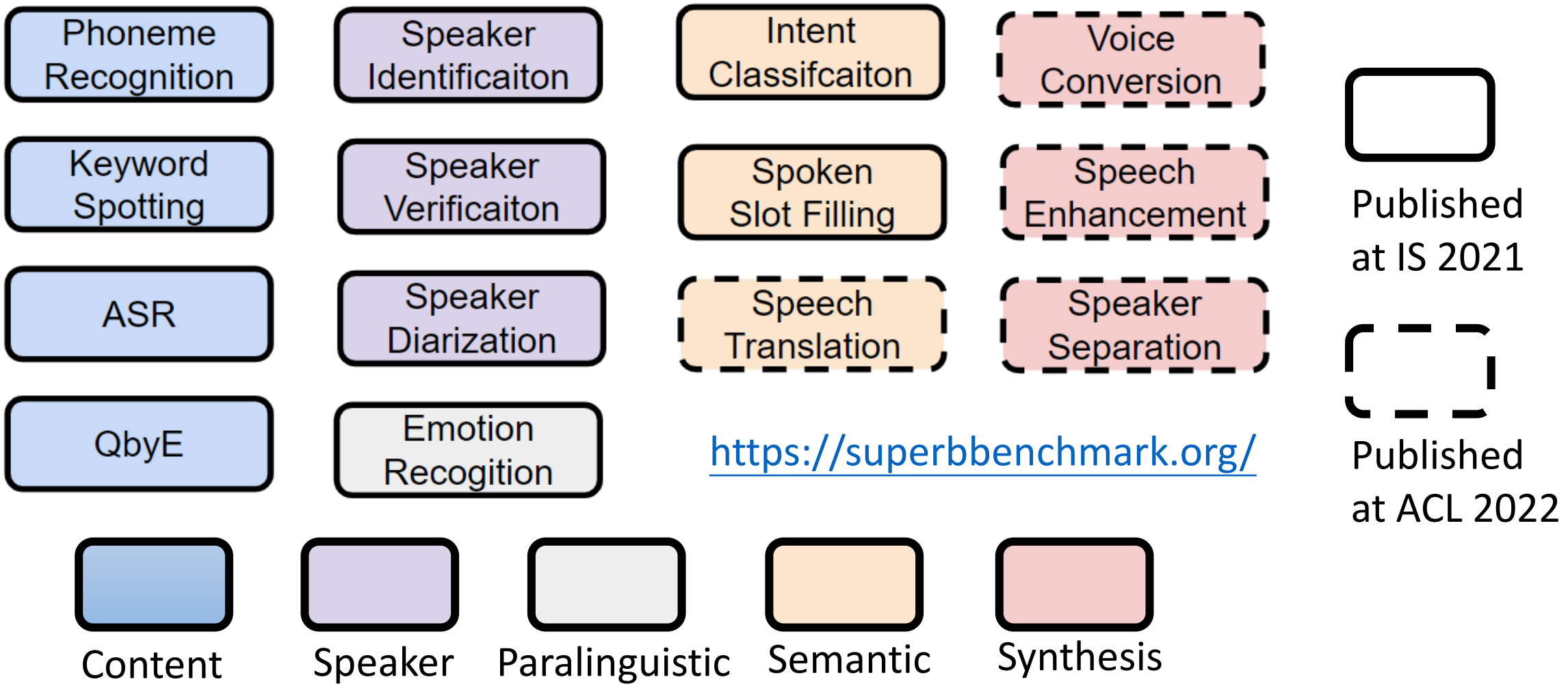
Speech Recognition

Speaker 42

Speaker Recognition

Self-supervised Learning for Speech

Speech processing Universal PERformance Benchmark (SUPERB)



Pre-training is good for Small Players

Pre-trained Model vs. **Downstream Tasks**

Operating Systems

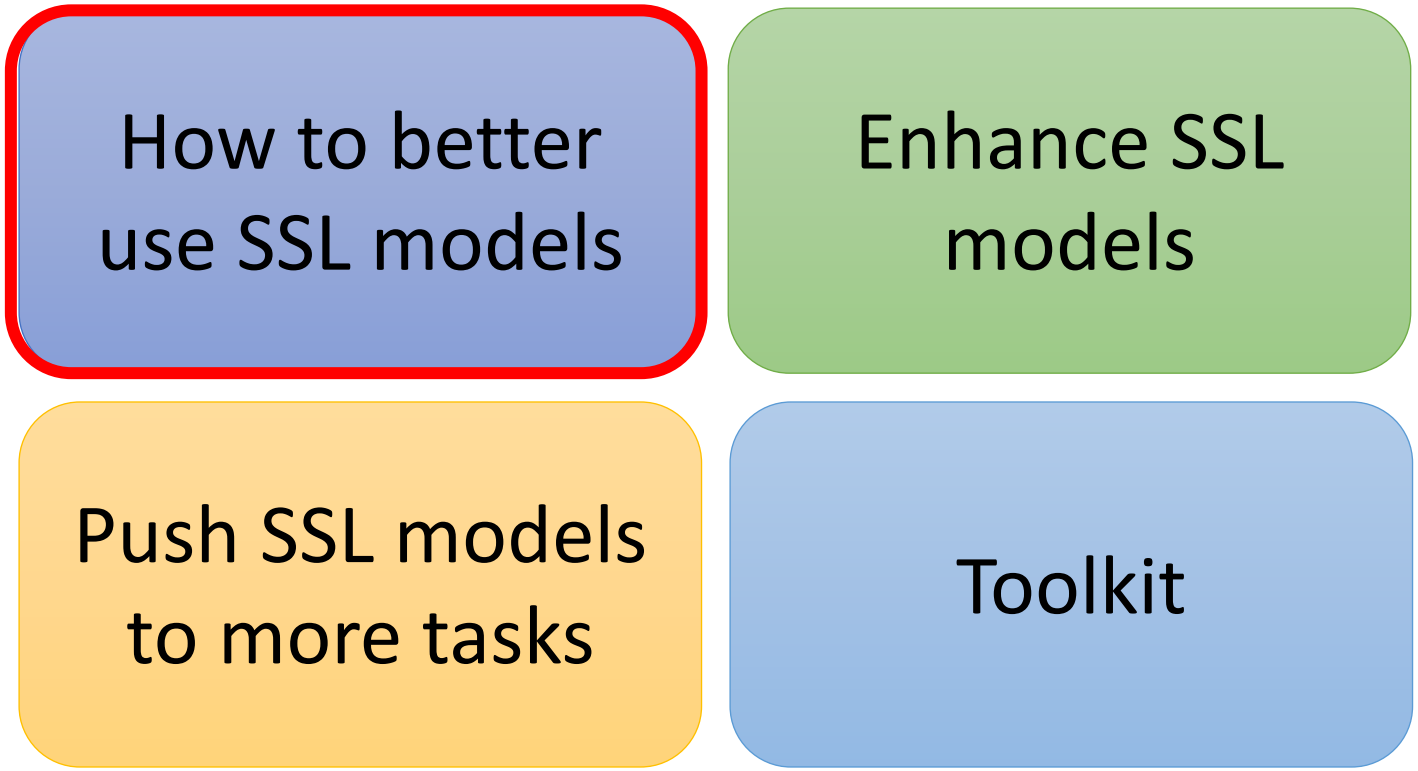
Applications

Less labeled data

Less training

Good for small players

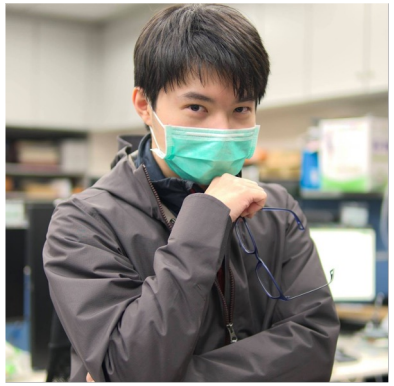
Goal


How to better use SSL models
Enhance SSL models
Push SSL models to more tasks
Toolkit

Senior Member

Shang-Wen (Daniel) Li
(Meta)

Member



Kai-Wei Chang
(NTU)

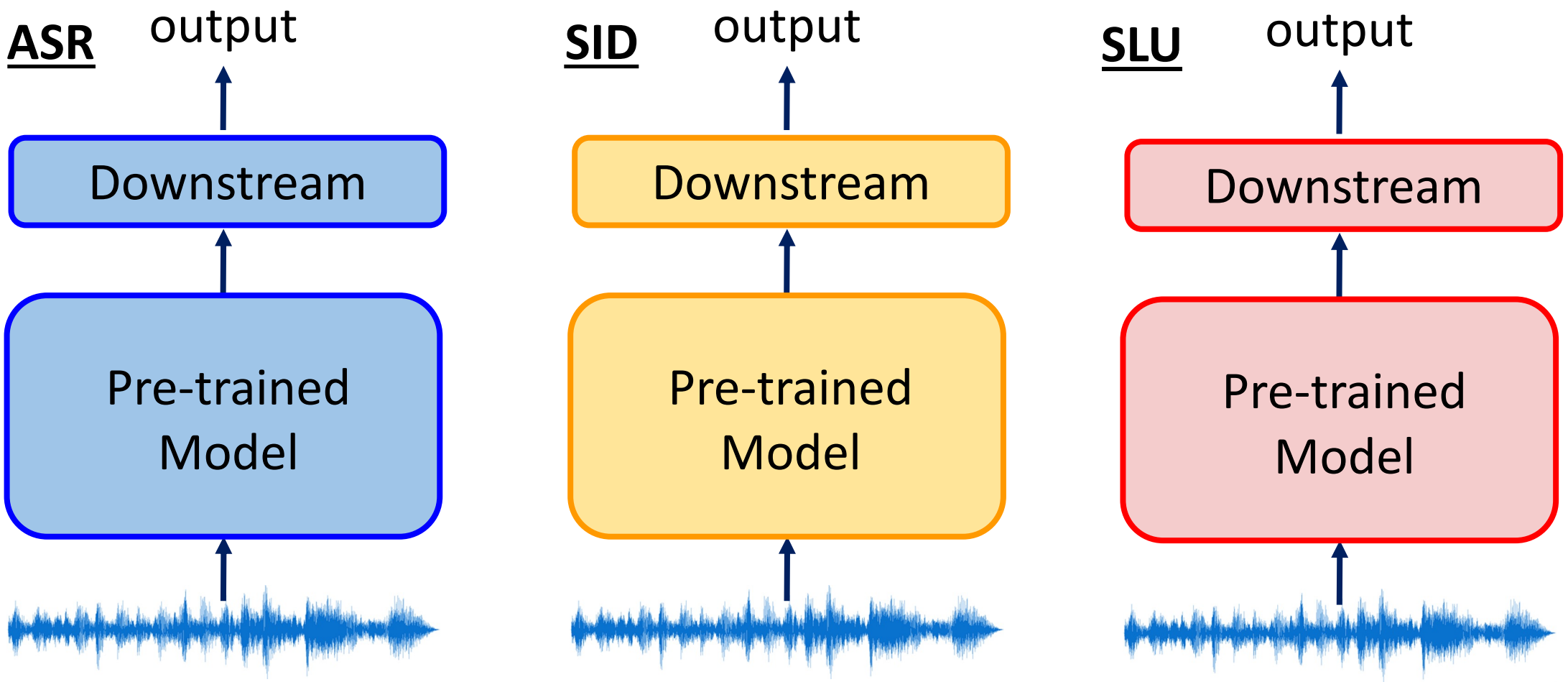
Fabian Ritter
(NUS)

Zih-Ching Chen
(NTU)

Chin-Lun Fu
(NTU)

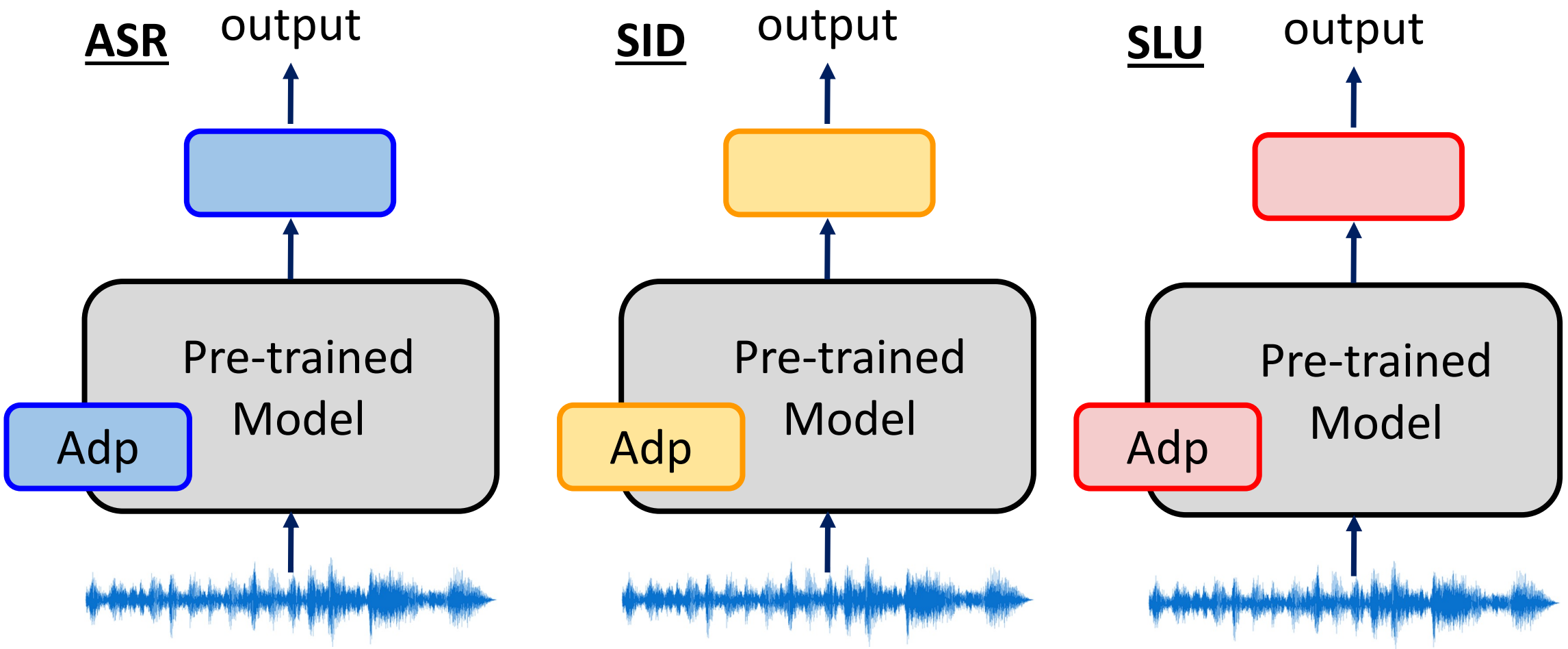
Hua Shen
(PennState)

How to use pre-trained Models? **Fine-tuning**



Weakness: Have to store a gigantic pre-trained model for each task

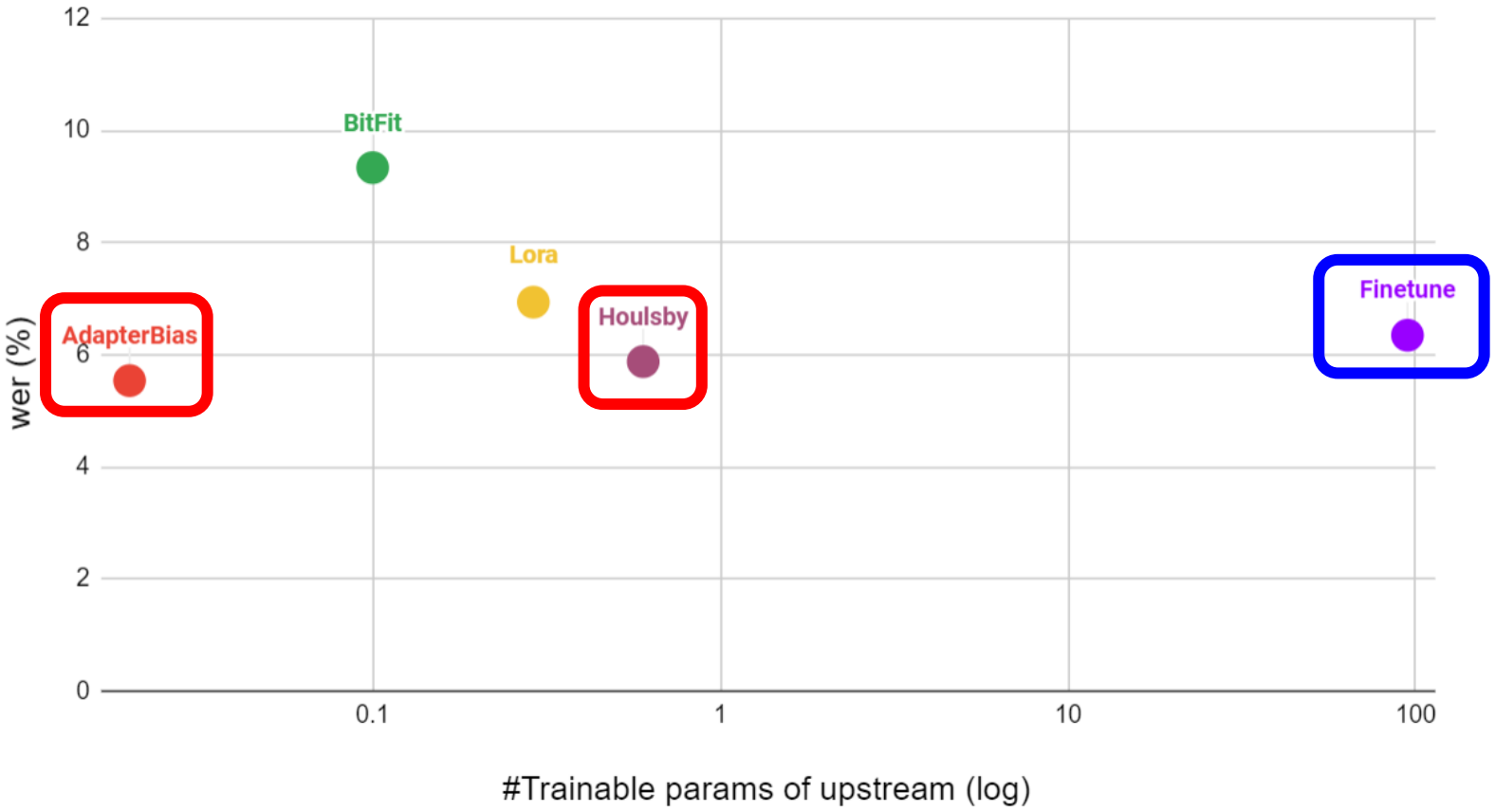
How to use pre-trained Models? **Adapter**



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

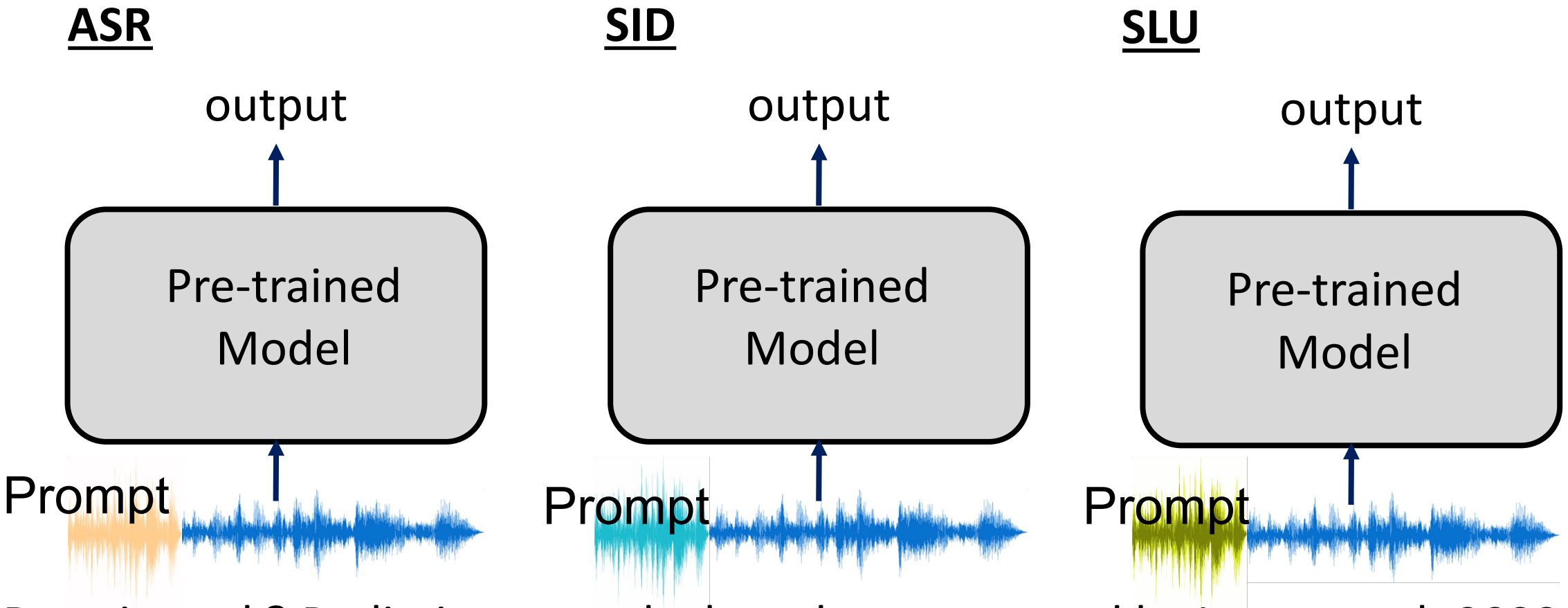
How to use pre-trained Models? **Adapter**

Adapter on HuBERT for ASR



Plan to study all kinds of adapters on various SSL models for wide range of tasks.

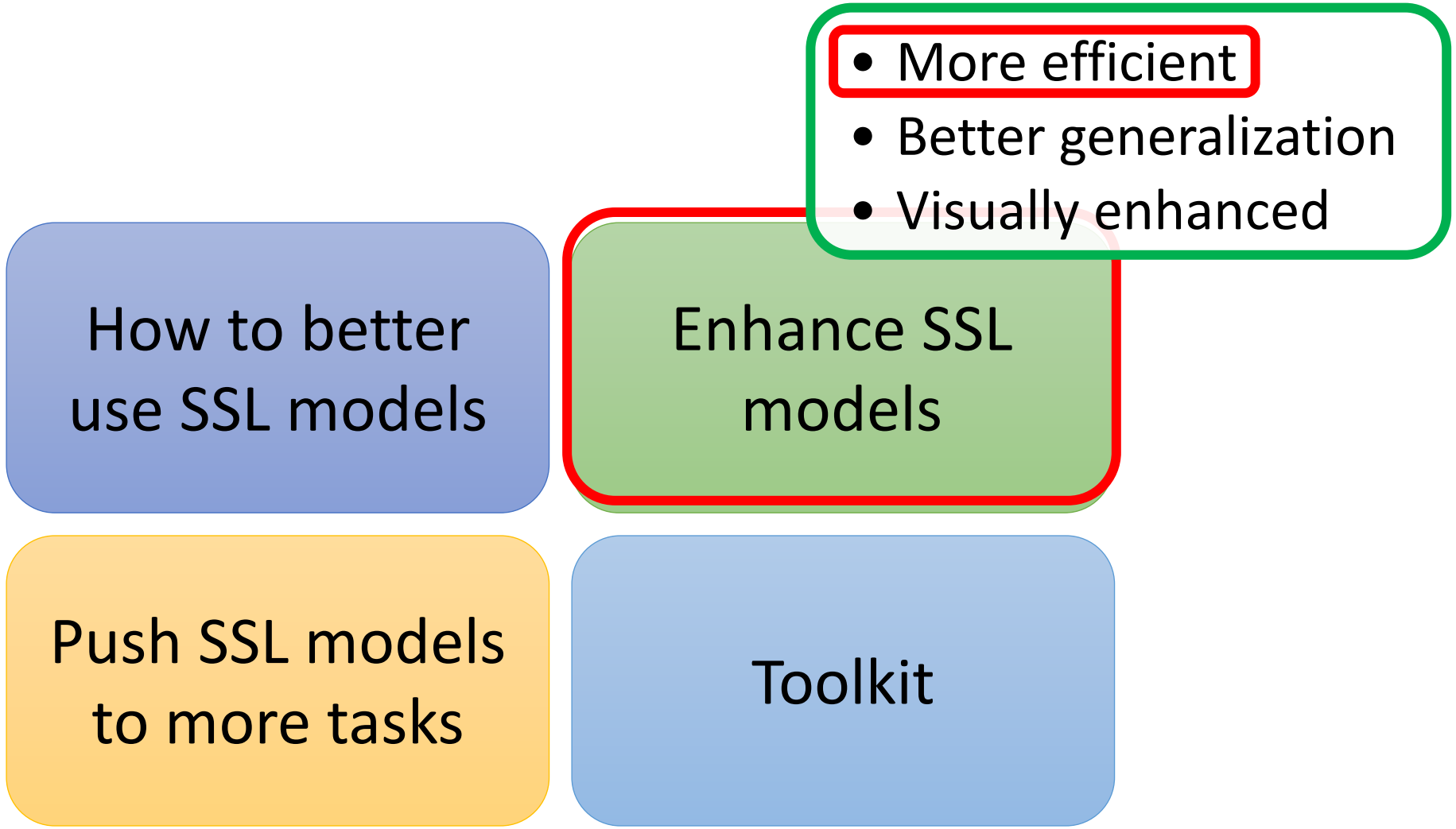
How to use pre-trained Models? **Prompting**



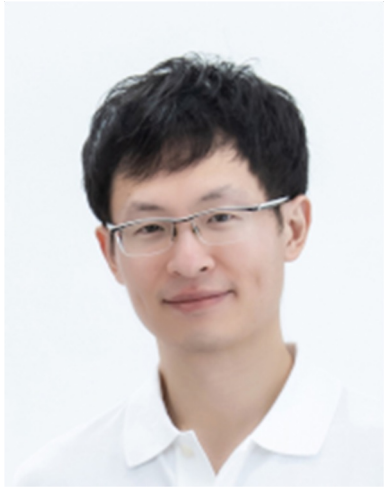
Does it work? Preliminary results have been accepted by Interspeech 2022.

https://arxiv.org/abs/2203.16773

Goal


How to better use SSL models
Enhance SSL models
Push SSL models to more tasks
Toolkit
More efficient
Better generalization
Visually enhanced

Senior Member



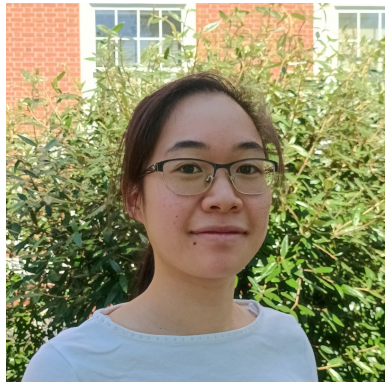
Hao Tang (University of Edinburgh)

Lucas Ondel (SONOS)

Diego Aguirre (UTEP)

Member

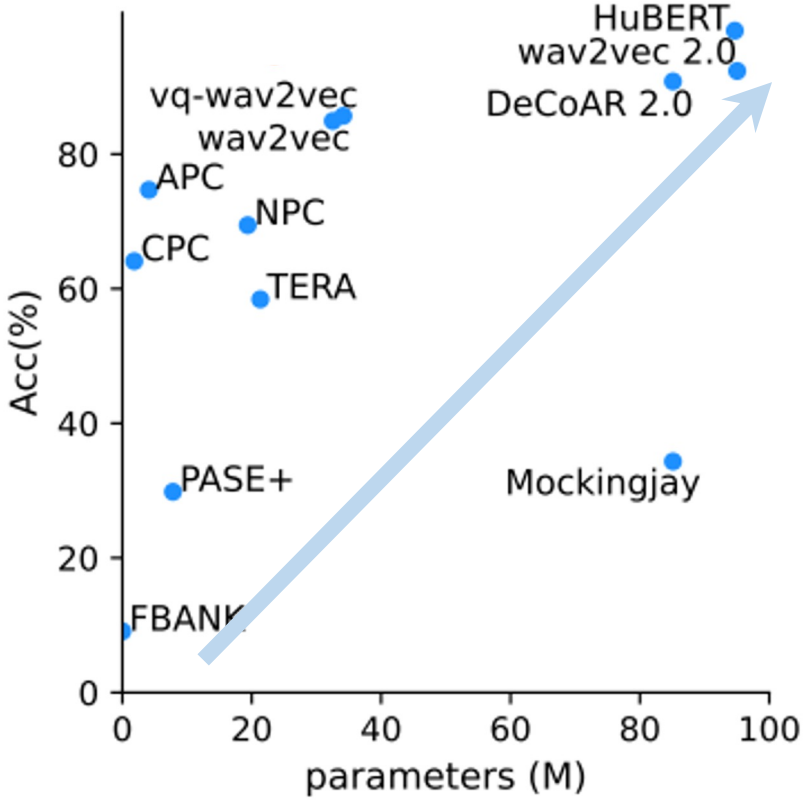
Tzu-Quan (Jack) Lin (NTU)



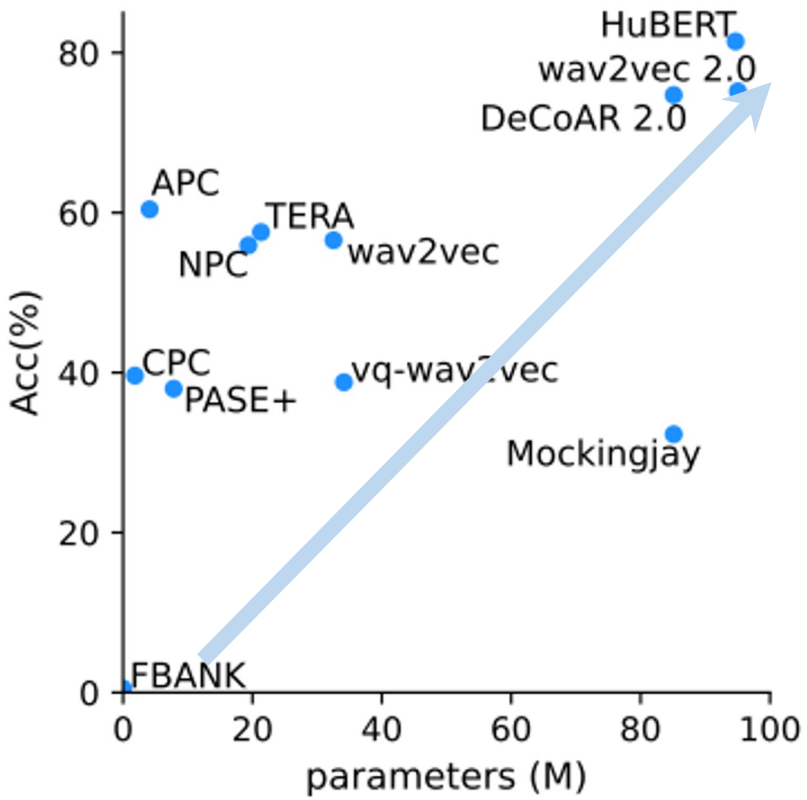
Léa-Marie Lam-Yee-Mui (LISN)

Larger Models lead to better Results …

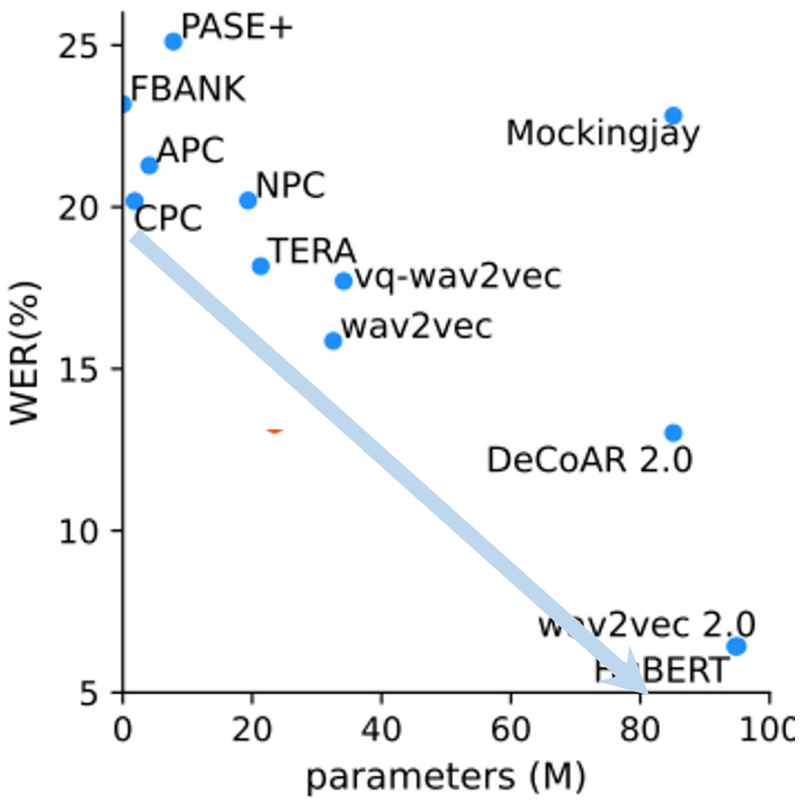
(From SUPERB benchmark)



Intent Classification

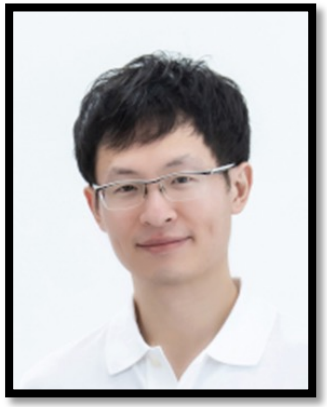


Speaker Identification



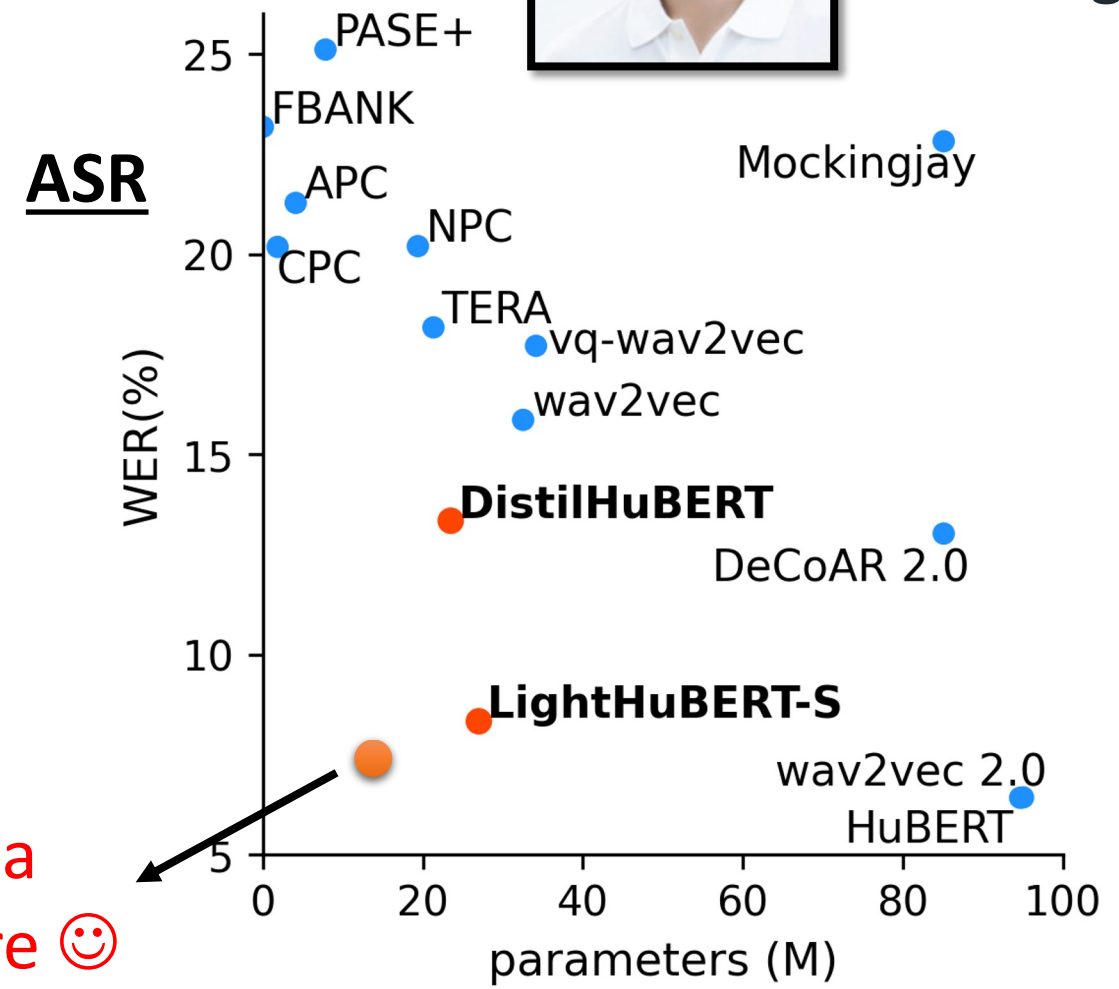
ASR (without LM)

Efficient Pre-trained Models



Hao Tang (University of Edinburgh)

- Reducing No. of Parameters & Computation
  - Knowledge Distillation
  - Low-rank Approximation
  - Head Pruning / Weight Pruning
  - MelHuBERT: Removing CNN encoder
  - Sequence reduction



Let's put a point here ☺

MelHuBERT: Removing CNN encoder

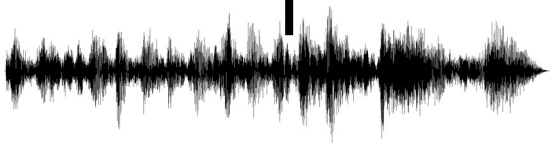
HuBERT

Transformer layer 12

Transformer layer 11

Transformer layer 1

CNN



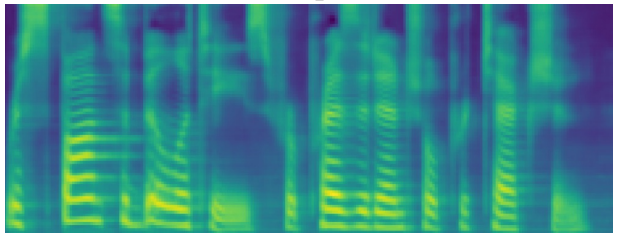
33% computational overhead

MelHuBERT

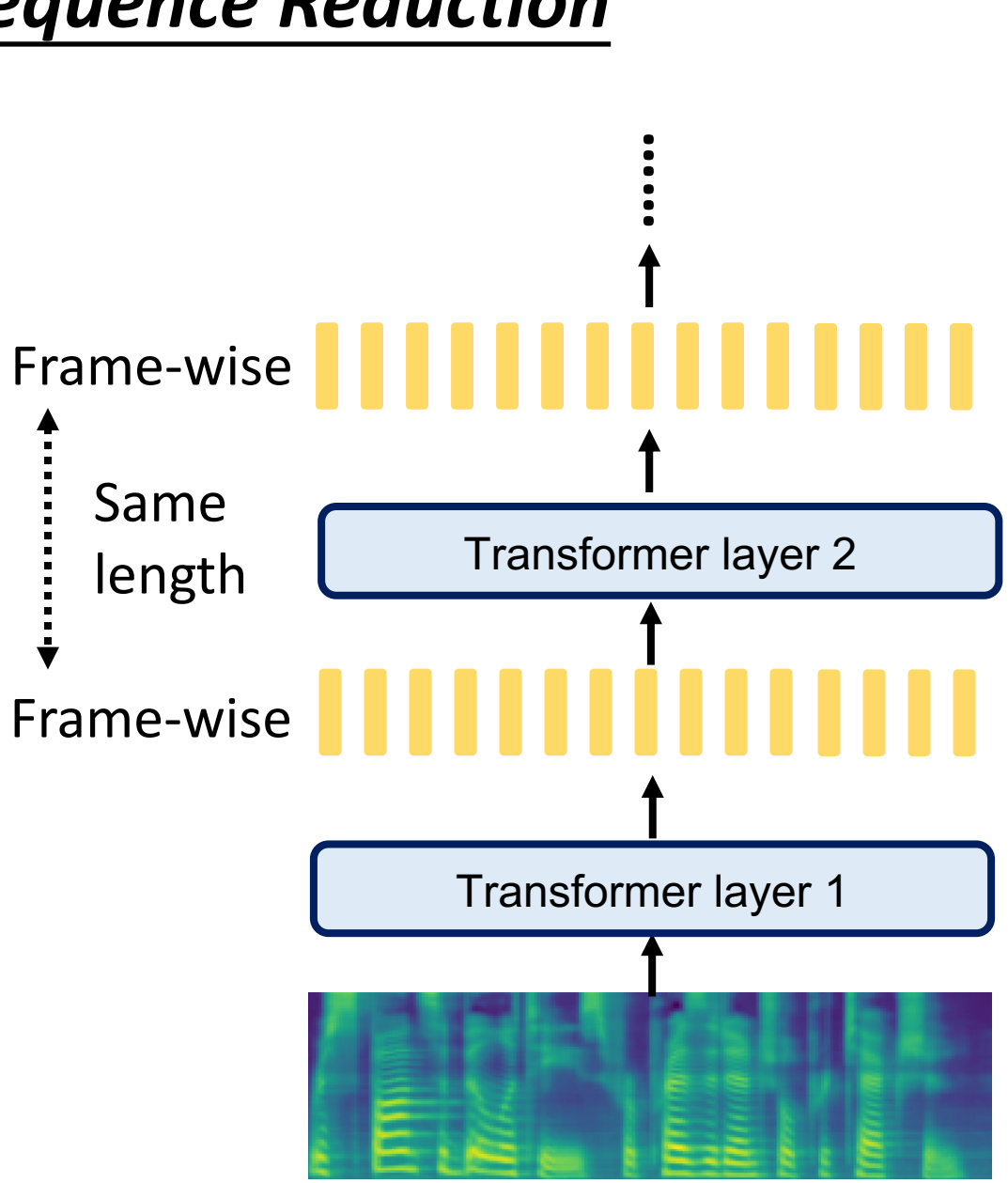
Transformer layer 12

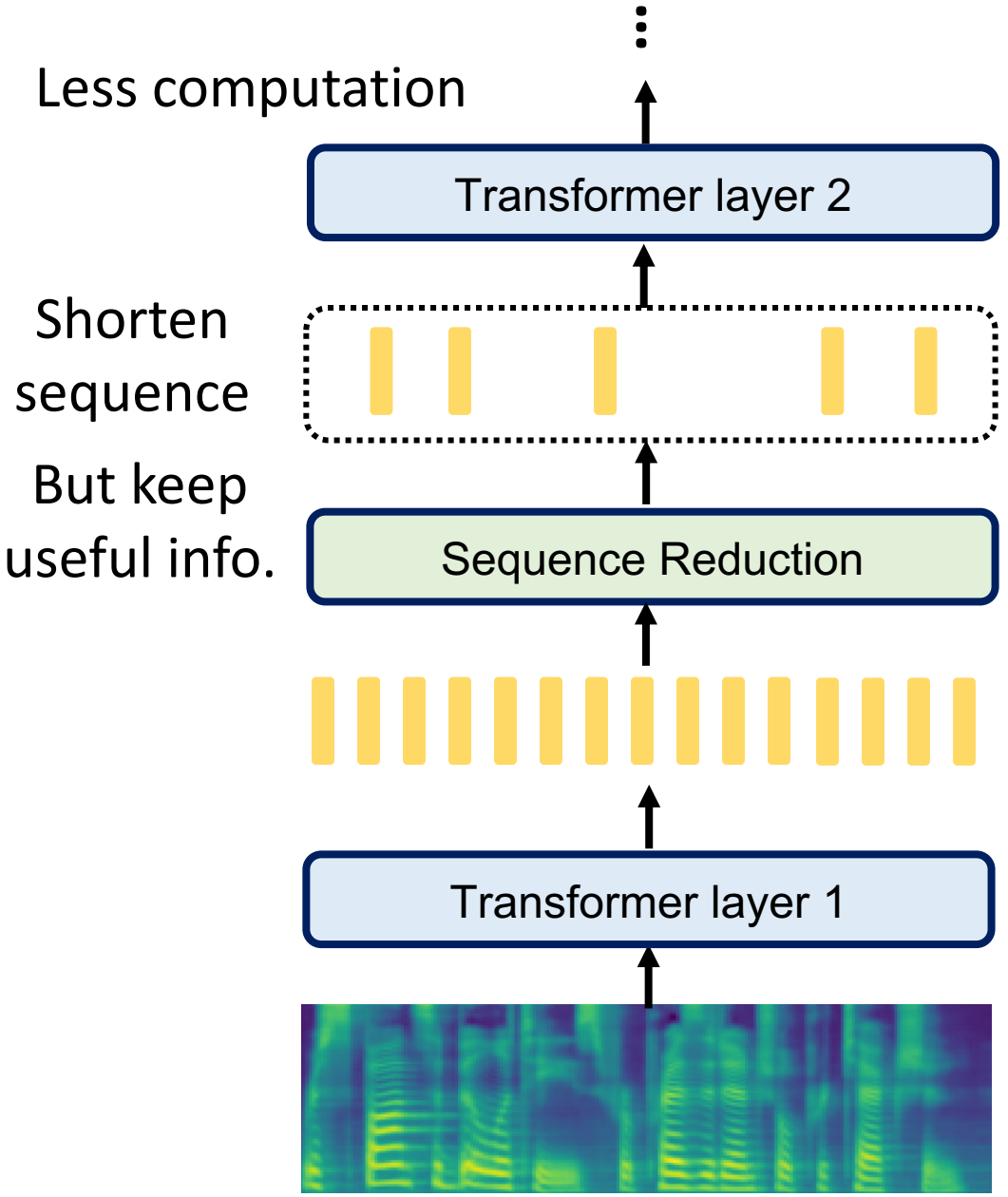
Transformer layer 11

Transformer layer 1



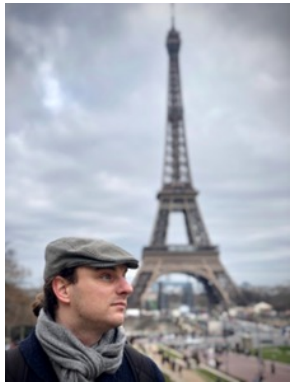
Sequence Reduction





Efficient Pre-trained Models

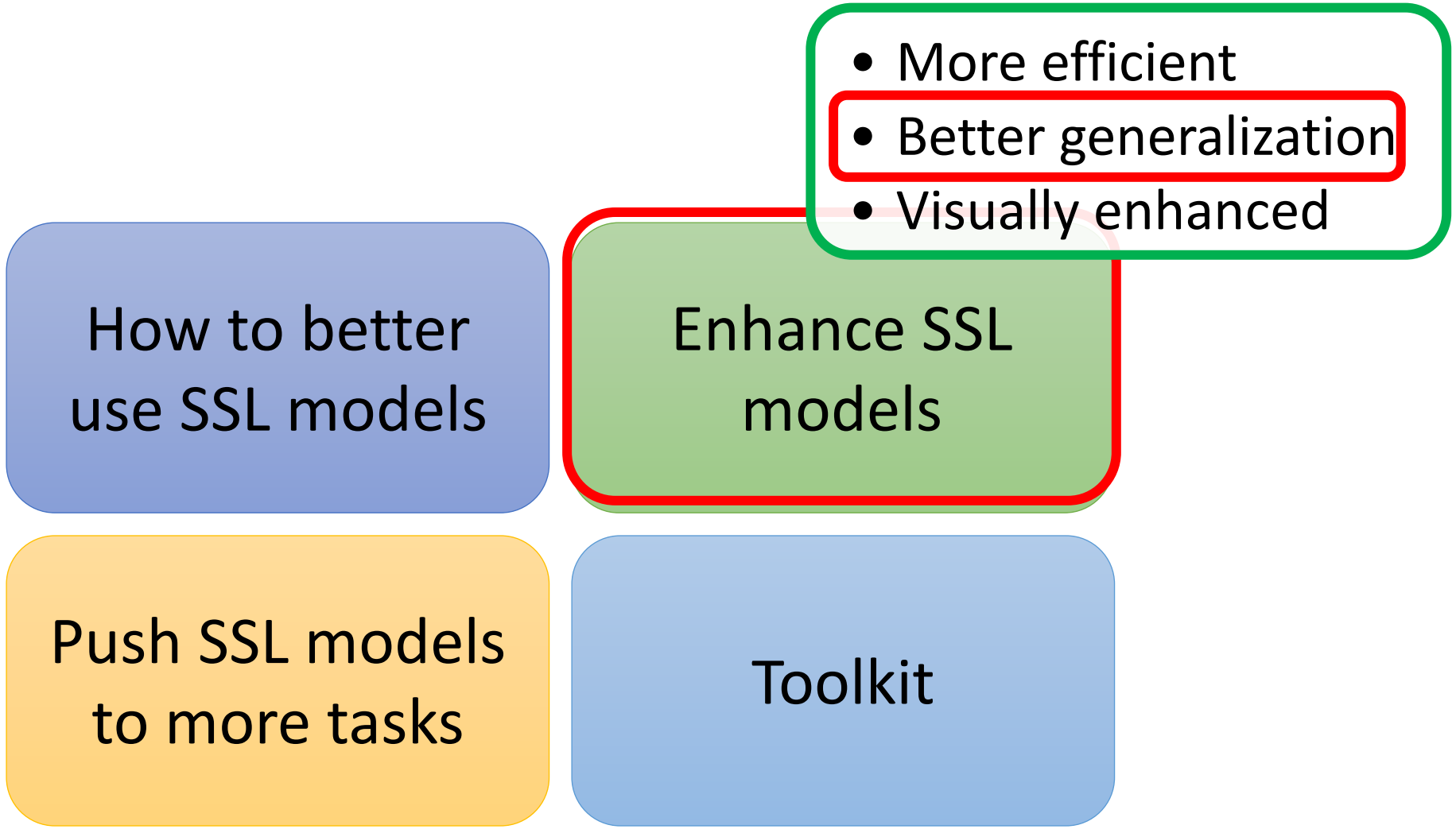
- **Efficient network architecture**
  - Typical pre-trained models using self-attention
  - Explore more efficient approaches (e.g., FNet)
- **Efficient weight initialization**
  - Pre-training a model also needs weight initialization
  - Towards faster training and better performance via weight initialization



Lucas Ondel (SONOS)

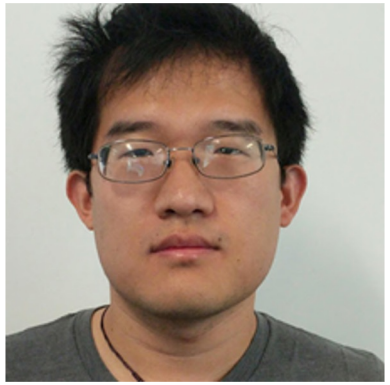
Diego Aguirre (UTEP)

Goal


How to better use SSL models
Enhance SSL models
Push SSL models to more tasks
Toolkit
More efficient
Better generalization
Visually enhanced

Senior Member

Hung-yi Lee
(NTU)



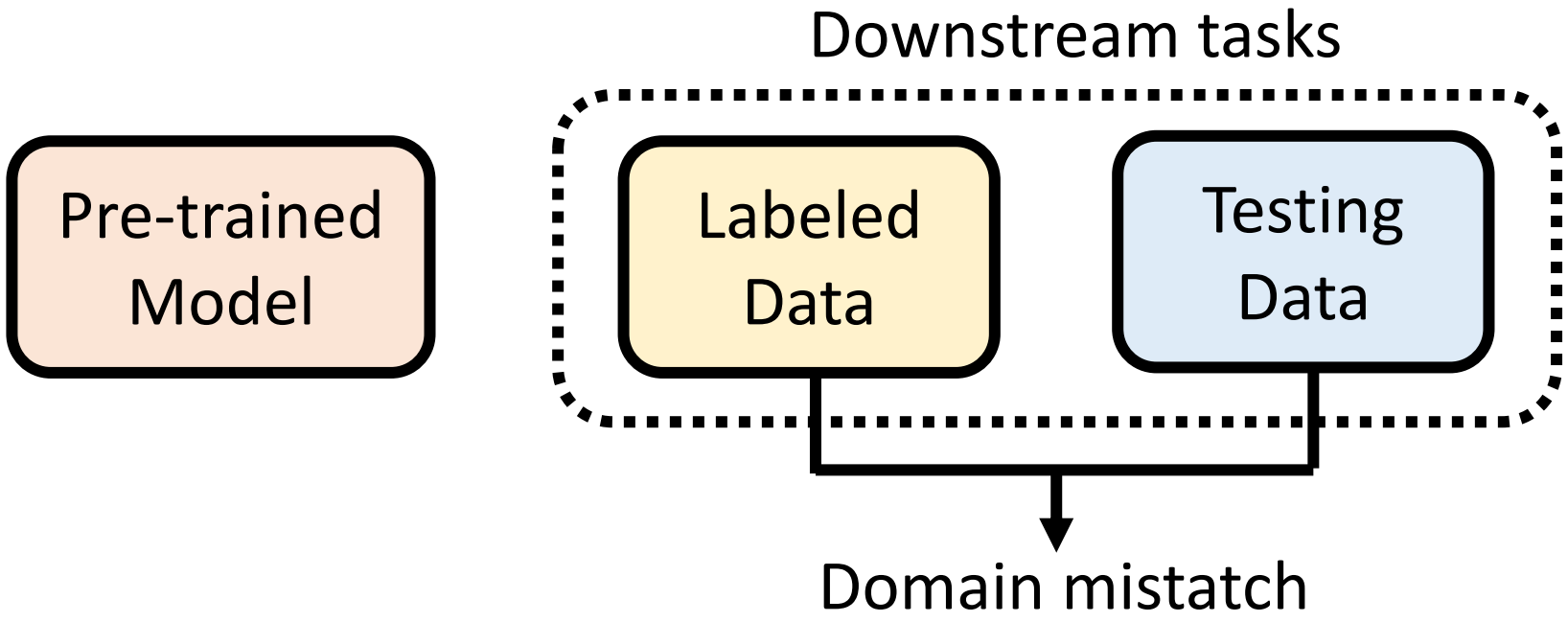
Yu Zhang
(Google)

Member

Kuan Po Huang
(NTU)

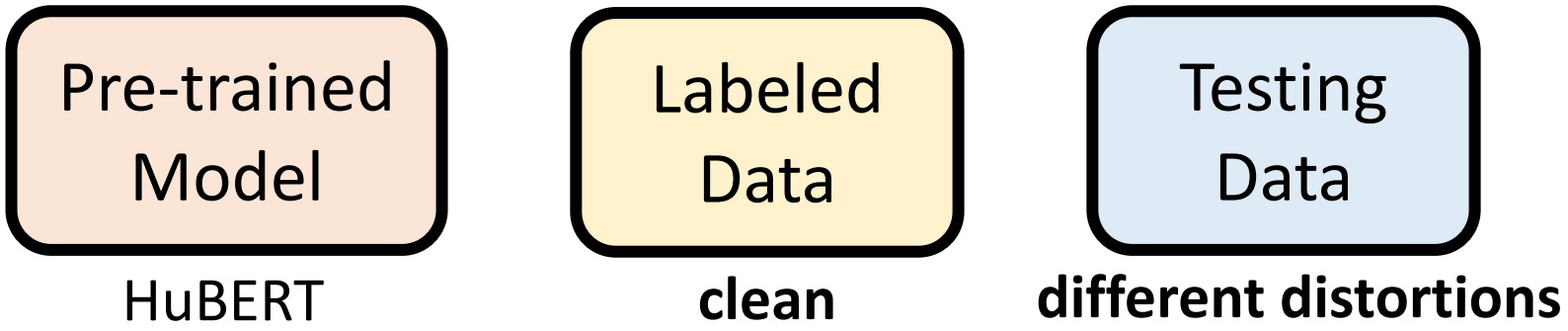
Fabian Ritter
(NUS)

Generalization Capability of Pre-trained Model



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

Can self-supervised models maintain good performance?



	Intent Classification ↑			Emotion Recognition ↑			Keyword Spotting ↑		
Testing Data	clean	m+g+r	fsd50k	clean	m+g+r	fsd50k	clean	m+g+r	Fsd50k
HuBERT	99.47	96.94	97.47	63.96	57.33	60.55	97.14	93.87	93.80

	Speaker Identification ↑			ASR (WER) ↓			
Testing Data	clean	m+g+r	fsd50k	clean	m+g+r	fsd50k	CHiME3
HuBERT	84.97	65.51	77.61	4.88	7.94	7.57	29.26

m+g+r = Musan + Gaussian + reverberation

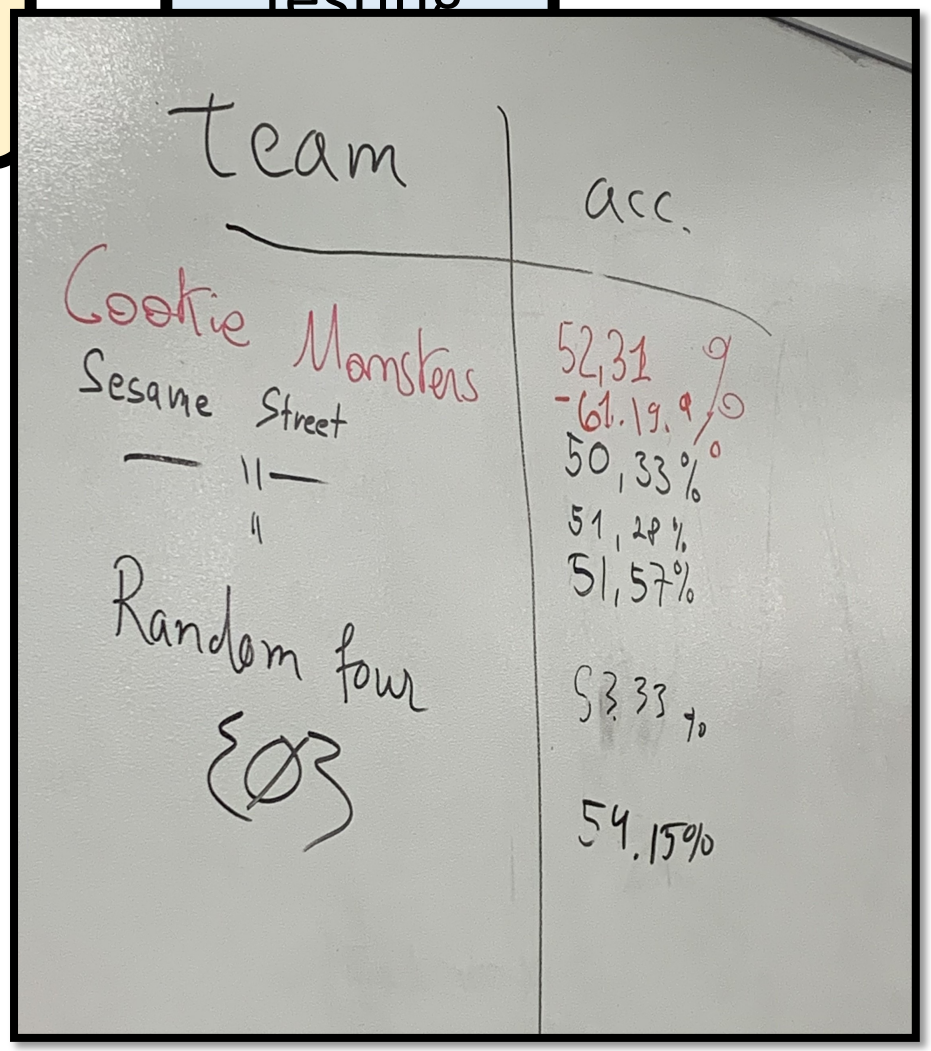
Continuously train with noisy data (m+g+r)

Pre-trained Model

Labeled Data

clean

Testing



Remember this? ☺

Preliminary results have been accepted by INTERSPEECH 2022.

https://arxiv.org/abs/2203.16104

The continuously trained models have been public.

This is the model you got the best performance in the mini-challenge last Friday.

Next Step

- Explore other types of domain mismatch, more pre-trained models
- Especially focus on improving the generalizability of the **compressed** pre-trained model.

	Intent Classification ↑		Emotion Recognition ↑		Keyword Spotting ↑	
Testing Data	clean	noisy	clean	noisy	clean	noisy
HuBERT	99.47	96.94	63.96	57.33	97.14	93.87
DistilHuBERT	94.78	66.41	63.87	53.92	96.04	89.84

	Speaker Identification ↑		ASR (WER) ↓	
Testing Data	clean	noisy	clean	m+g+r
HuBERT	84.97	65.51	4.88	7.94
DistilHuBERT	73.02	40.42	13.77	37.59

Goal

How to better use SSL models

Enhance SSL models

- More efficient
- Better generalization
- Visually enhanced

Push SSL models to more tasks

- Prosody-related Tasks
- Spoken Language Understanding

Toolkit

Senior Member

David Harwath (UT)

Member

Virginia Layne Berry (UT)

Heng-Jui (Harry) Chang (MIT)

Yi-Jen (Ian) Shih (NTU)

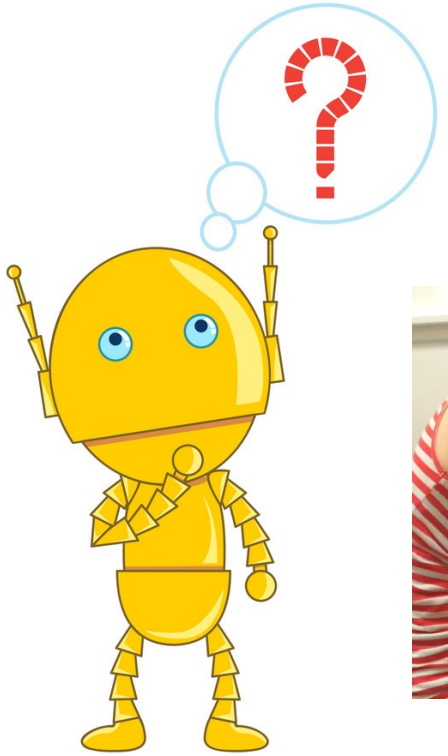
Hsuan-Fu (Jeff) Wang (NTU)

Elizabeth Boroda (JHU)

Visually Enhanced Pre-trained Speech Model

(There is a little girl wearing sunglasses.)

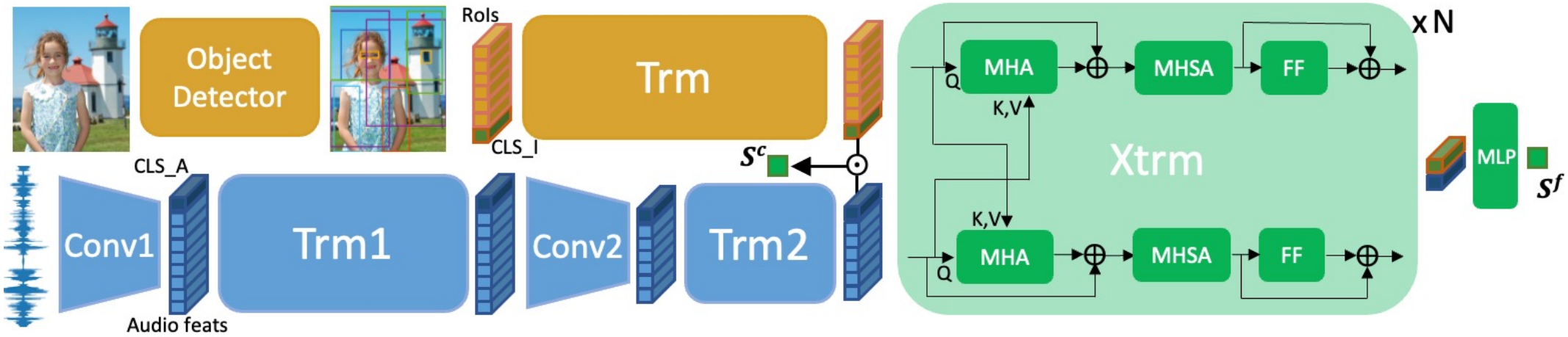
Associated Image



Pre-trained Model

Visually Enhanced Pre-trained Speech Model

FaST-VGS

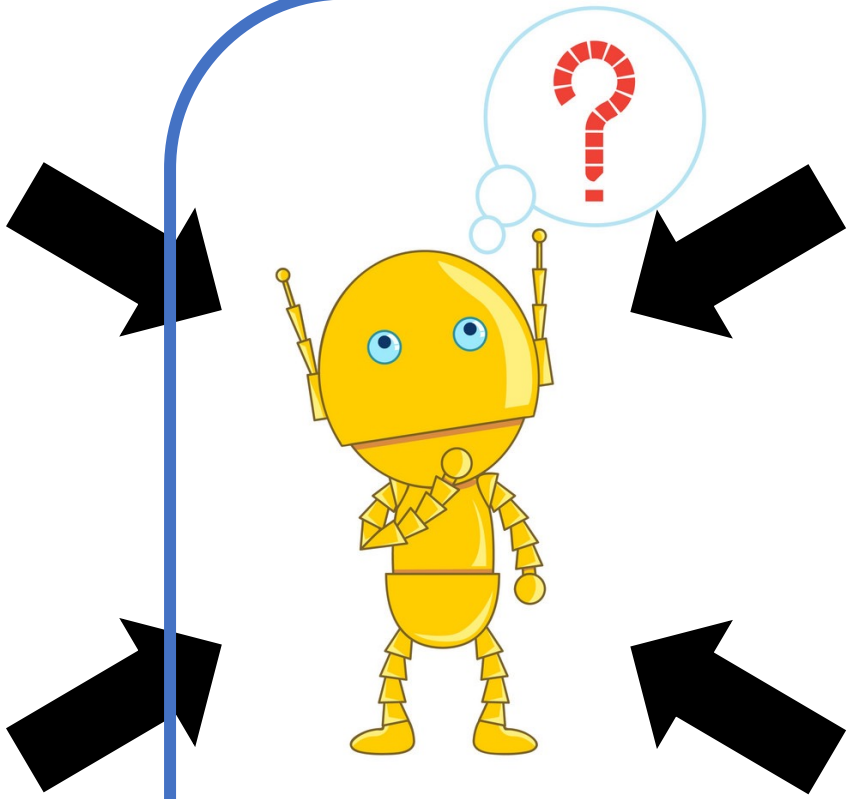


ICASSP 2022 (https://arxiv.org/abs/2109.08186)

Show good performance on some SUPERB tasks, but still struggle on some tasks (e.g., ASR).

How to further improve visually grounded pre-trained models?

(There is a little girl ……)



Pre-trained Model

sheep on the grass

CLIP

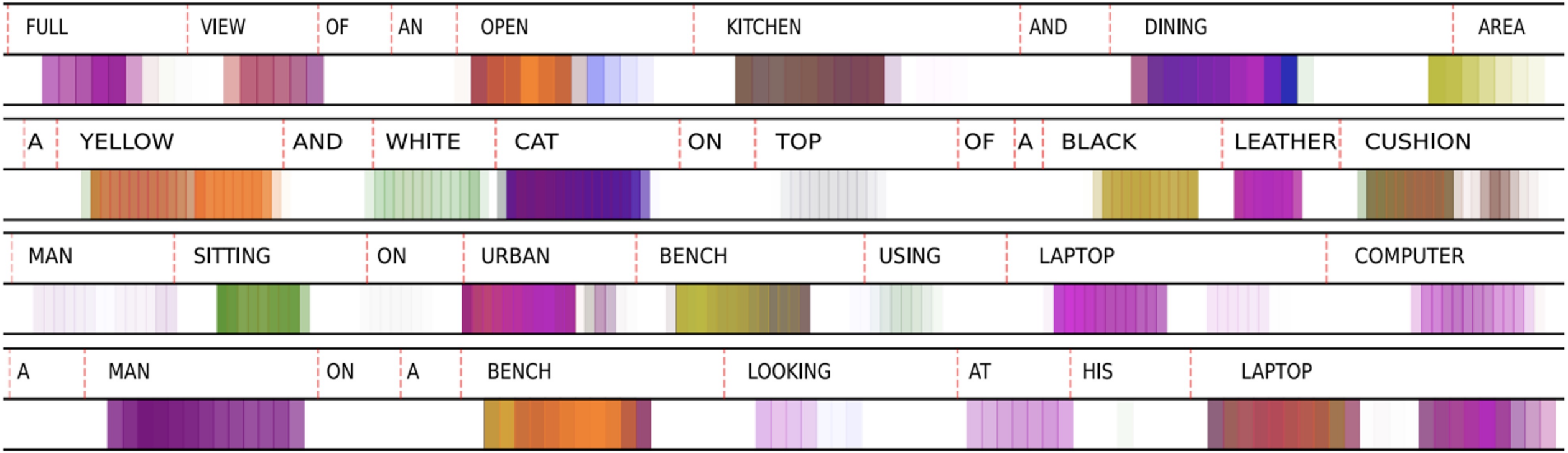
https://openai.com/blog/clip/

- Can we use the CLIP image and text encoders as a teacher model for a speech encoder like HuBERT?
- Promising results on SUPERB tasks, unsupervised ASR, unsupervised speech translation

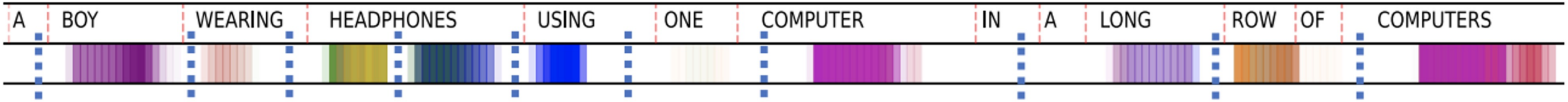
VG-HuBERT

Outstanding at detecting and segmenting Words

https://arxiv.org/abs/2203.15081



Generate word boundaries: midpoints of adjacent boundaries of attention segments:

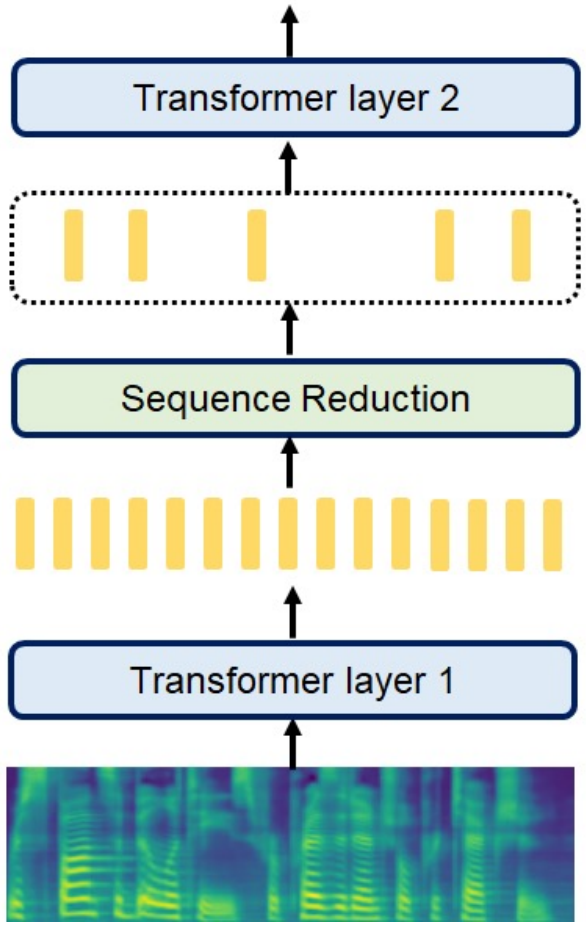


VG-HuBERT

Outstanding at detecting and segmenting Words

https://arxiv.org/abs/2203.15081

- Use automatic word segmentation for non-uniform downsampling of speech signal
- Use discovered word segments/clusters as masks/targets for SSL model training
- Use discovered words to improve unsupervised ASR (will talk about this later)



Goal

- More efficient
- Better generalization
- Visual enhanced

How to better use SSL models

Enhance SSL models

Push SSL models to more tasks

Toolkit

- Prosody-related Tasks
- Spoken Language Understanding

Senior Member

Nigel Ward (UTEP)

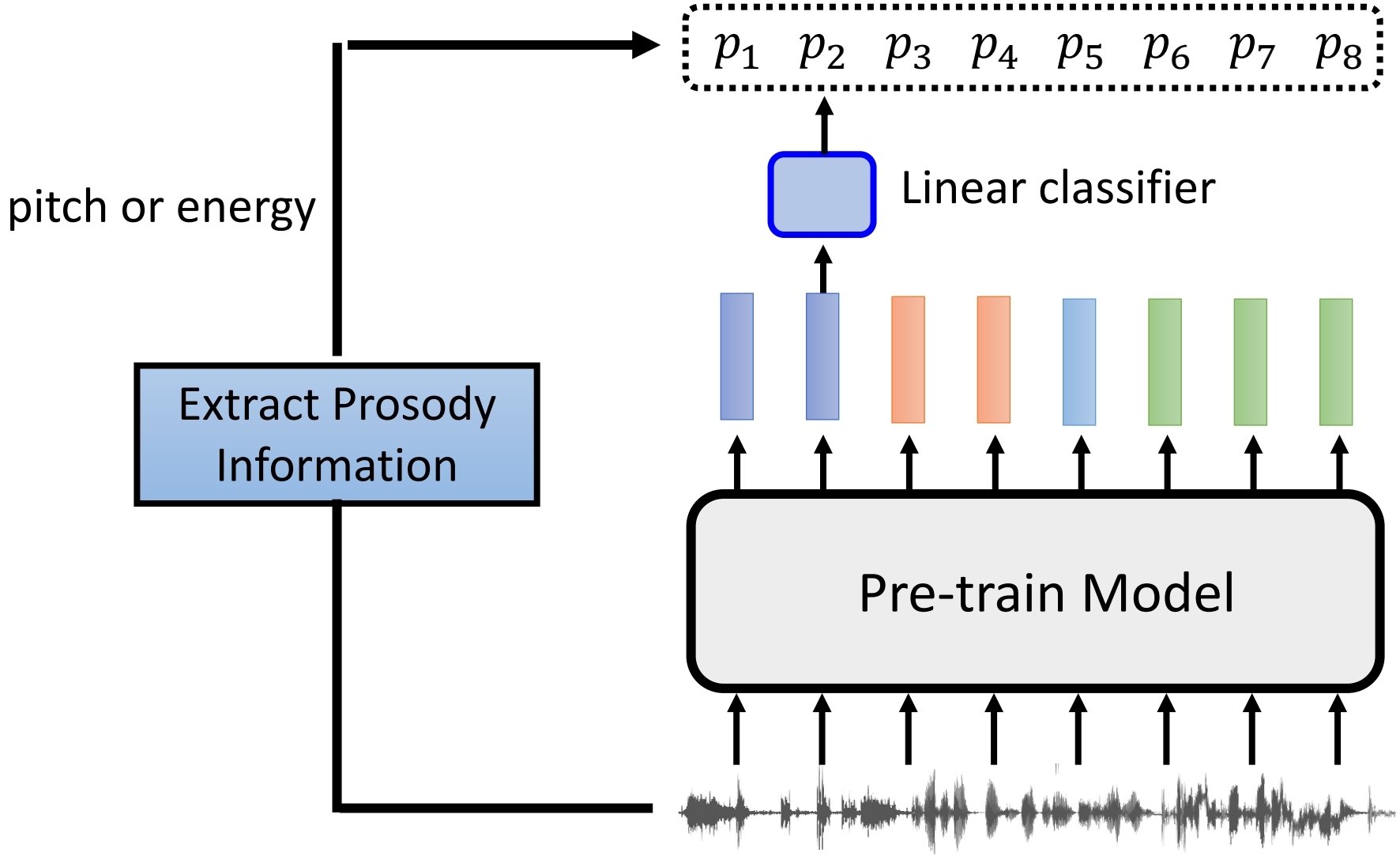
Member

Chi-Luen (Leo) Feng
(NTU)

Guan-Ting (Daniel) Lin
(NTU)

Samuel Miller
(University of Maryland)

Do SSL models extract prosody information?



Do SSL models extract prosody information? Yes

- Smaller values represent better reconstructions, i.e., contain more prosodic information.

	fbank	HuBERT	wav2vec	wav2vec 2.0
Pitch	0.089	0.018	0.035	0.022
Energy	0.517	0.243	0.318	0.240

- Analysis is in English only, and we will do more analysis during the workshop.
- Working on prosody-related downstream tasks
  - Pitch / Energy Prediction, Sentiment Analysis (CMU-MOSEI), Sarcasm detection (MUStARD dataset), Persuasiveness prediction (CMU-POM dataset), etc.

Goal

- More efficient
- Better generalization
- Visual enhanced

How to better use SSL models

Enhance SSL models

Push SSL models to more tasks

Toolkit

- Prosody-related Tasks
- Spoken Language Understanding

Senior Member

Ann Lee
(Meta)

Paola Garcia
(JHU)

Member

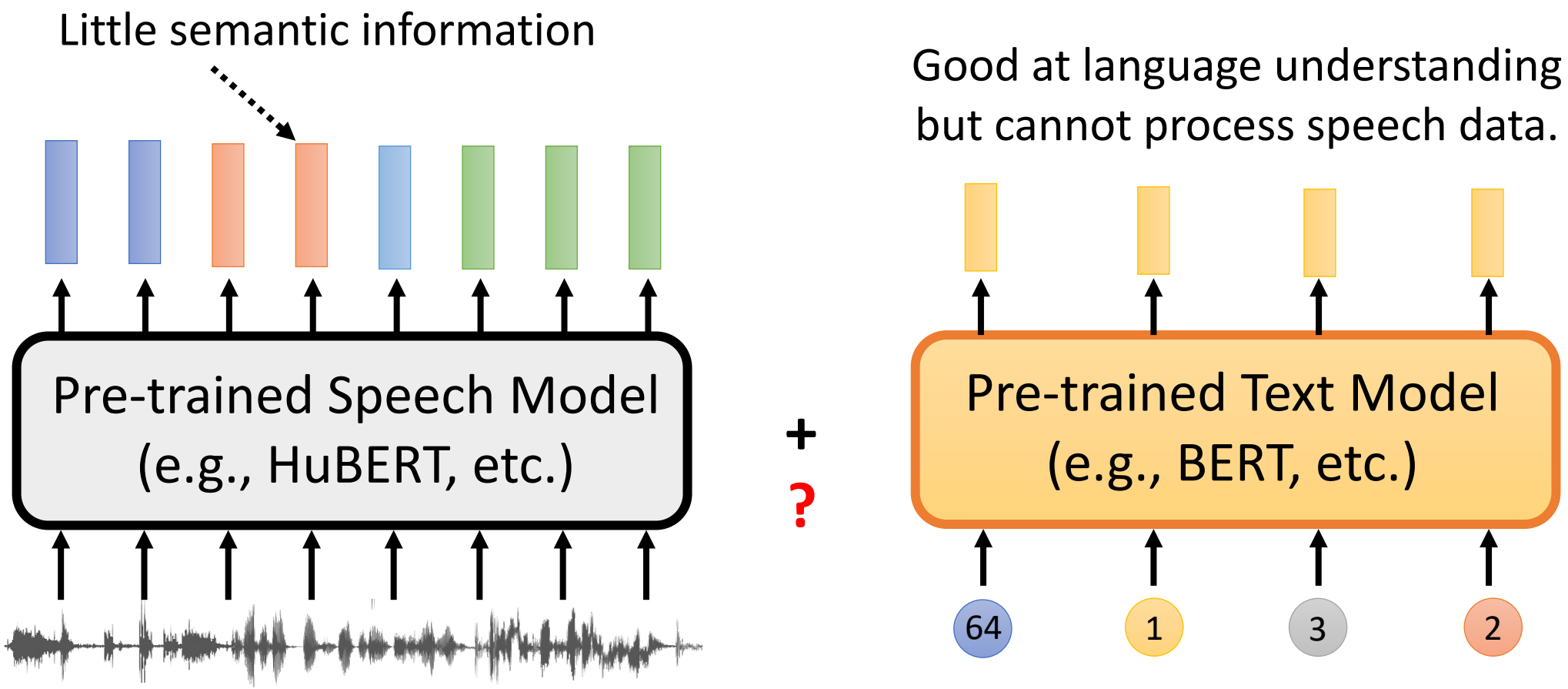
Jiatong Shi
(CMU)

Dongji Gao
(JHU)

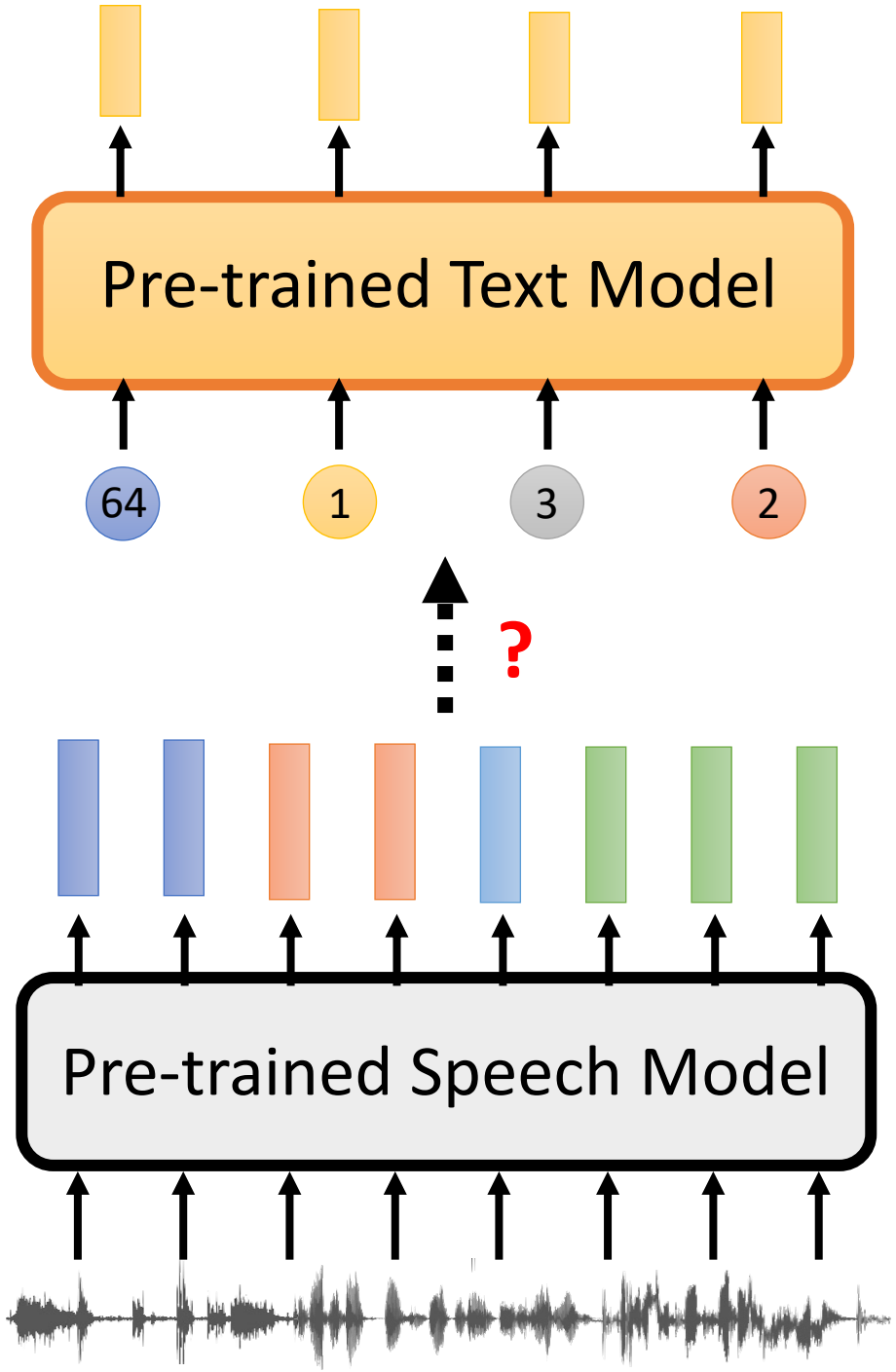
Yen Meng
(NTU)

Hsuan-Jui (Ray)
Chen (NTU)

Spoken Language Understanding Tasks



New self-supervised Model



Related work: SLAM (https://arxiv.org/abs/2110.10329)

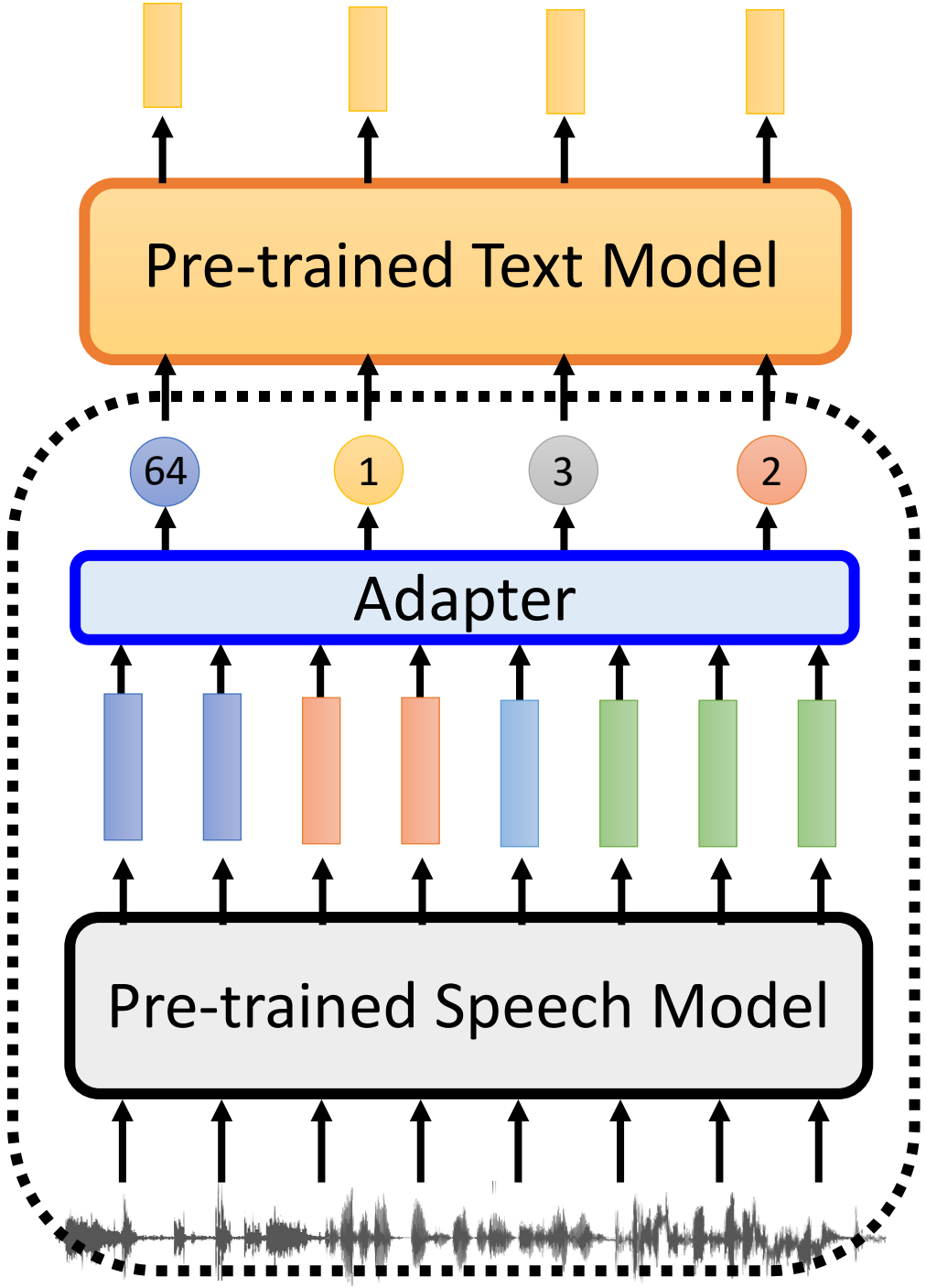
Difference: leveraged pre-trained text model, and not to use speech-text paired data

Train it as a GAN-based unsupervised ASR model

Progress: Have reproduce wav2vec-u 2.0

More:

- Trying more SSL speech model
- Using more robust, visual-enhanced model
- Using sequence reduction approach



Goal

How to better use SSL models

Enhance SSL models

- More efficient
- Better generalization
- Visual enhanced

Push SSL models to more tasks

- Prosody-related Tasks
- Spoken Language Understanding

Toolkit

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

S3PRL

Self-**S**upervised **S**peech **P**re-training and **R**epresentation **L**earning

Pre-training

Pre-trained model collection

Downstream Benchmarking & SUPERB

2019

Andy T. Liu

Shu-wen Yang

Po-Han Chi

2020

Shu-wen Yang

Andy T. Liu

2021

Shu-wen Yang

Andy T. Liu

Po-Han Chi

Heng-Jui Chang

Xuankai Chang

Yung-Sung Chuang

Zili Huang

Wen-Chin Huang

Tzu-Hsien Huang

Kushal Lakhotia

Yist Lin Y.

Guan-Ting Lin

Jiatong Shi

Hsiang-Sheng Tsai

Wei-Cheng Tseng

Shu-wen (Leo) Yang
(NTU)

s3prl

s3prl

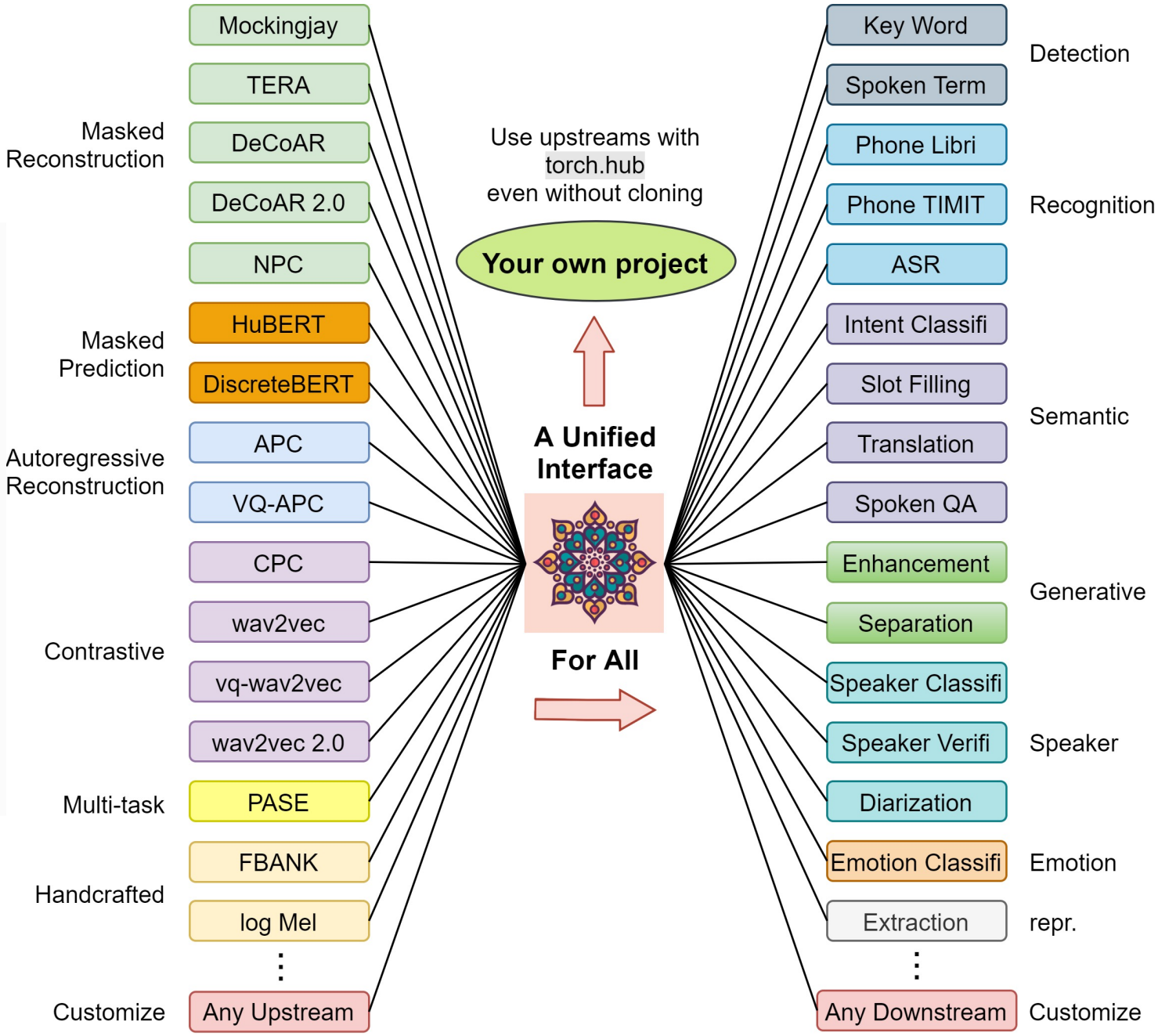
Self-Supervised Speech Pre-training and Representation Learning Toolkit.

youtu.be/PkMFnS6cjAc

1.3k stars 273 forks

+ Add to list

Restructure S3PRL to be more reusable, and more flexible for different pre-training & benchmarking experiment setup.



Goal

- More efficient
- Better generalization
- Visual enhanced

How to better use SSL models

Enhance SSL models

Push SSL models to more tasks

Toolkit

- Prosody-related Tasks
- Spoken Language Understanding

Thanks to Advisors of the Team

Shinji Watanabe
(CMU)

Bhuvana Ramabhadran
(Google)

Nancy Chen
(A*STAR)

Self-supervised Learning for Speech

Speech processing **U**niversal **PER**formance **B**enchmark (SUPERB)

SUPERB Challenge @ SLT 2022

Please visit the webpage to learn more: https://superbbenchmark.org/

Webpage and Twitter

Webpage

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

Twitter Account

@JSALT_pretrain