Leveraging Pre-training Models for Speech Processing

Research Group @ JSALT 2022

Overview

Hung-yi Lee
(NTU)

Goal

Better Pre-trained Model

- More Efficient
- Better Generalization
- Learn from Multimodality

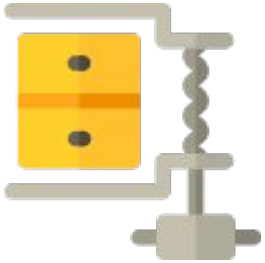
+

Better Use of Pre-trained Model

- Efficient Usage
- New Applications
- Toolkit

2 Interspeech papers (pre-workshop discussion) & **5 SLT papers** (submitted)

Outline of Part 1: Better Pre-trained Model



Compression

9:35 - 9:55

Robust

9:55 - 10:10

Visual-enhanced

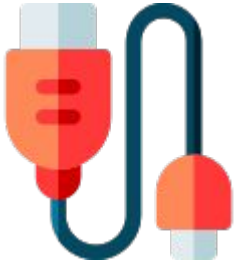
10:10 - 10:25

Integration

10:25 - 10:30

10 mins Q&A
+ 10 mins break

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

More usage

11:50 - 12:00

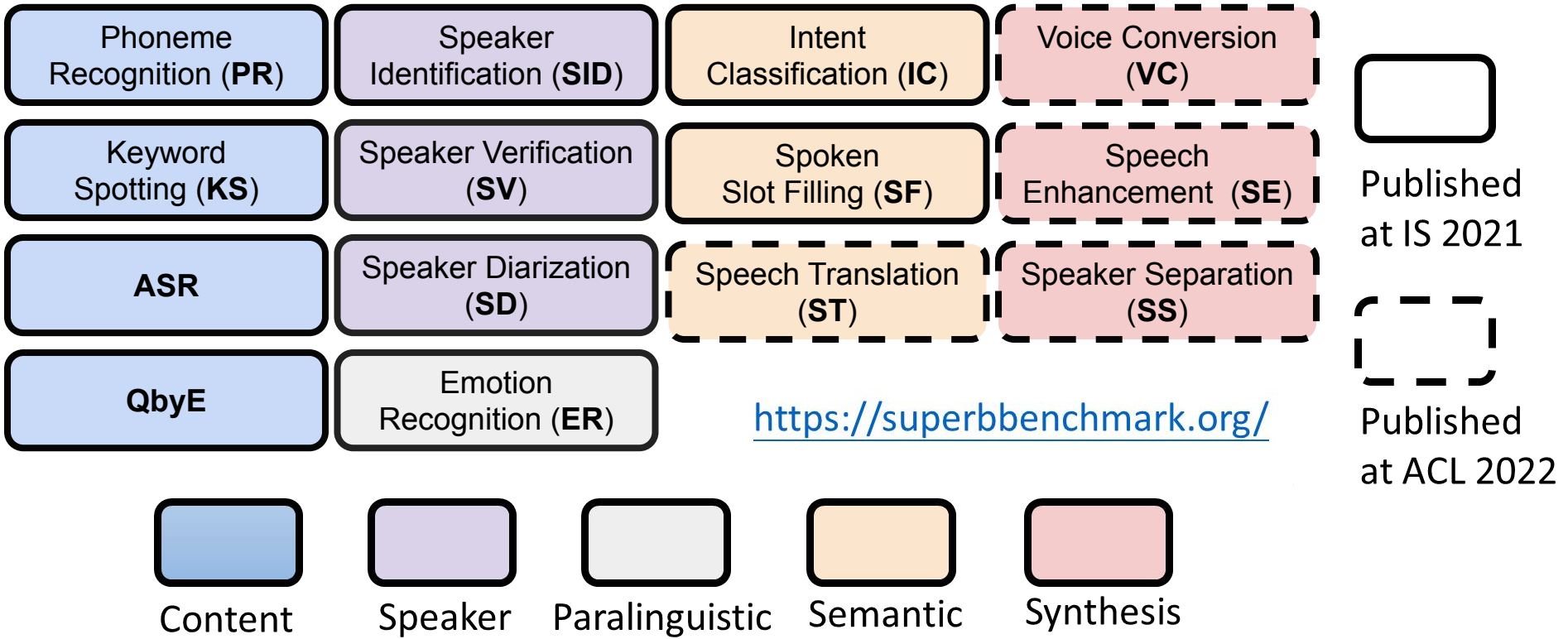
Toolkit for Speech Pre-training 12:00 - 12:10

Experimental Setup (if not specified)

https://superbbenchmark.org/

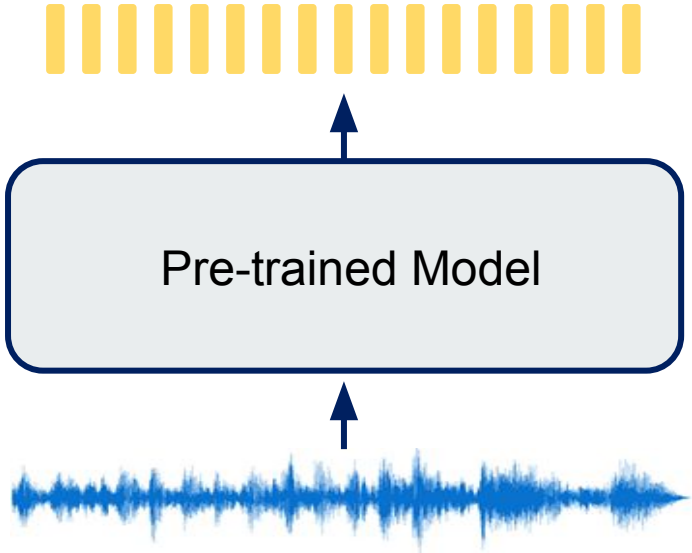
SUPERB

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

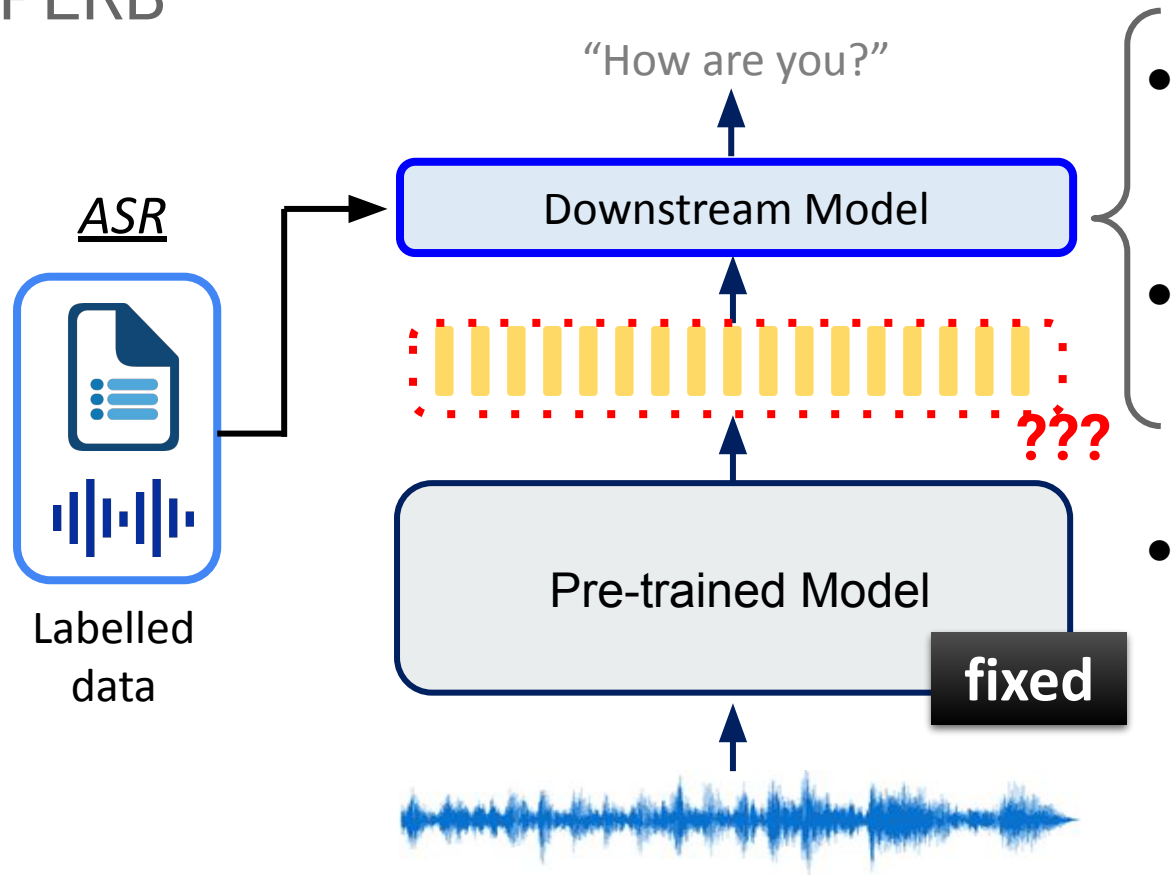


SUPERB

- The pre-trained models are learned from **unlabeled data**.
- Also known as **Self-supervised Learning (SSL)**

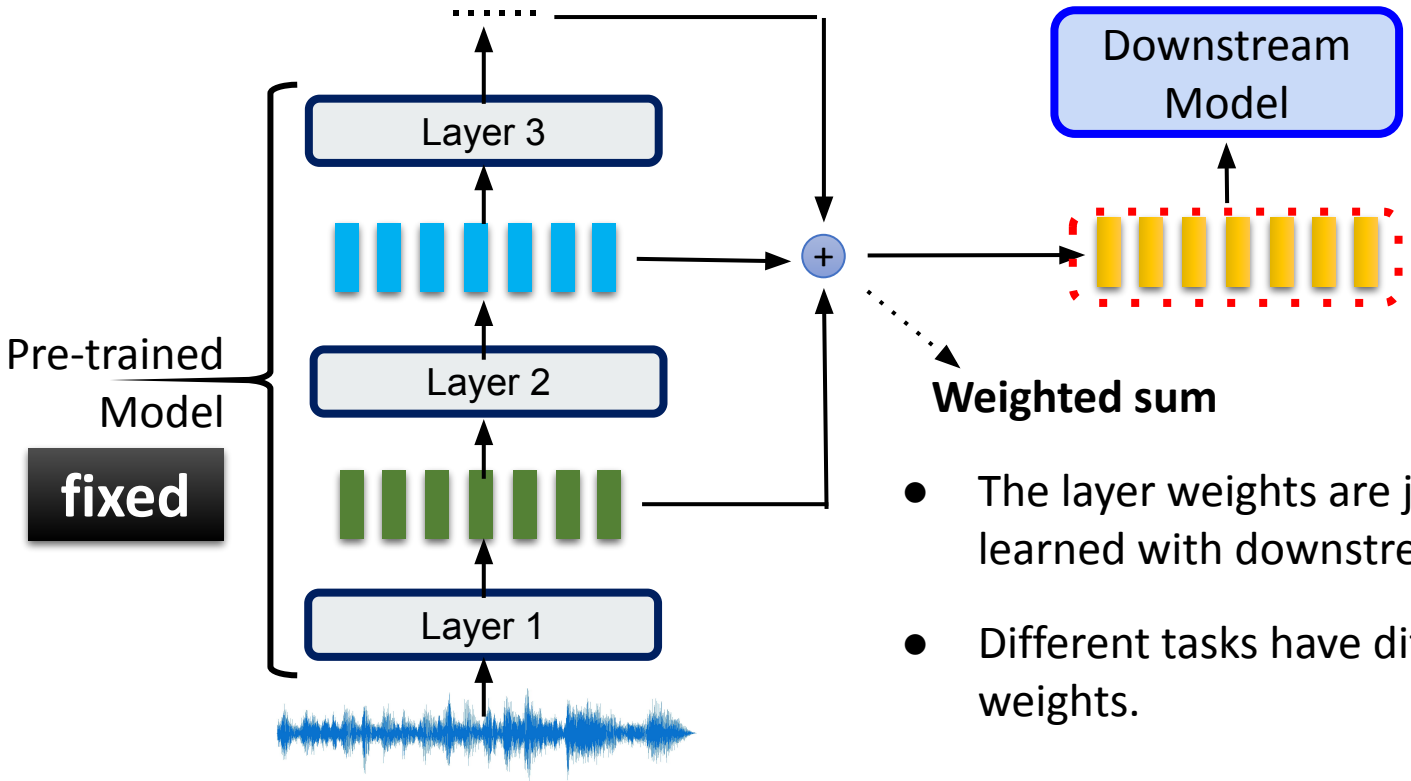


SUPERB



- 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

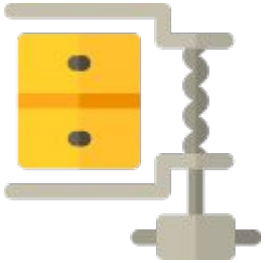


- The layer weights are jointly learned with downstream models.
- Different tasks have different weights.

Acknowledgement

Thanks for supporting computing resources!

Outline of Part 1: Better Pre-trained Model



Compression

9:35 - 9:55

Robust

9:55 - 10:10

Visual-enhanced

10:10 - 10:25

Integration

10:25 - 10:30

10 mins Q&A
+ 10 mins break

Model Compression & Sequence Compression

Tzu-Quan Lin

Tsu-Hsun Feng

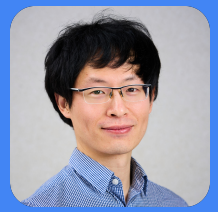
Tsung-Huan Yang

Chun-Yao Chang

Guang-Ming Chen

Yen Meng

Hsuan-Jui Chen



Hao Tang

Hung-yi Lee

Related work in compression

- A lot in supervised learning
  (LeCun et al., 1989; Han et al., 2015; Frankle et al., 2018; Pang et al., 2018; Lai et al., 2021)

- Several in self-supervised learning of vision and NLP
  (Sanh et al., 2019; Sajjad et al., 2020; Chen et al., 2021; Shi et al., 2021)

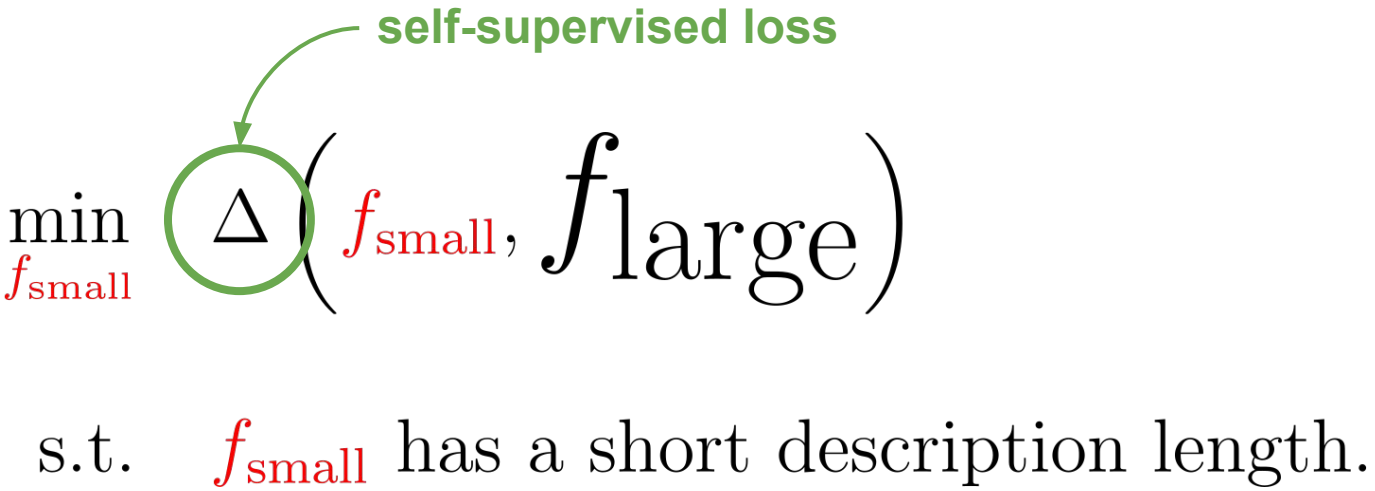
- Few in self-supervised speech models
  (Chang et al., 2022; Lee et al., 2022; Wang et al., 2022)

Goal

- Studying the landscape of compression techniques for self-supervised speech models

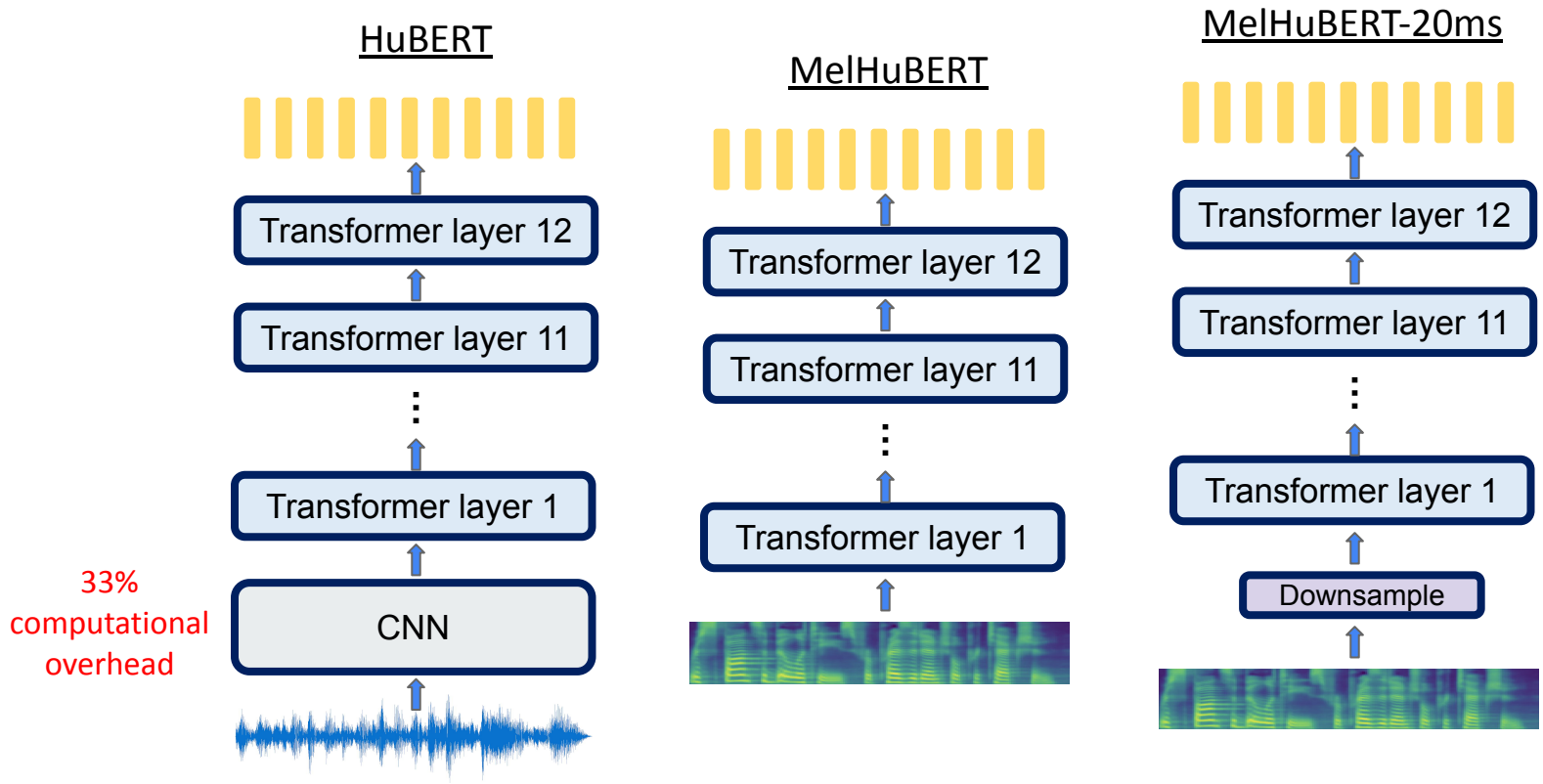
- Understanding the impact of compression on the learned representations

Compression

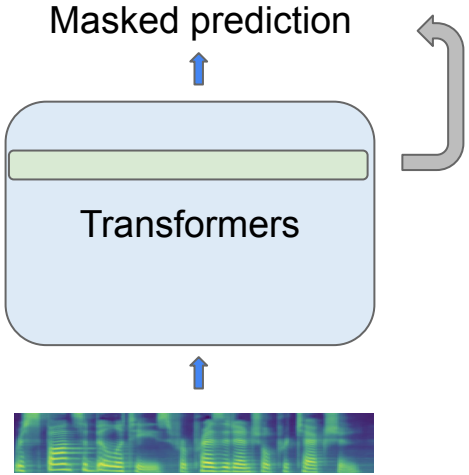


MelHuBERT f_{large}

Basic concept of MelHuBERT

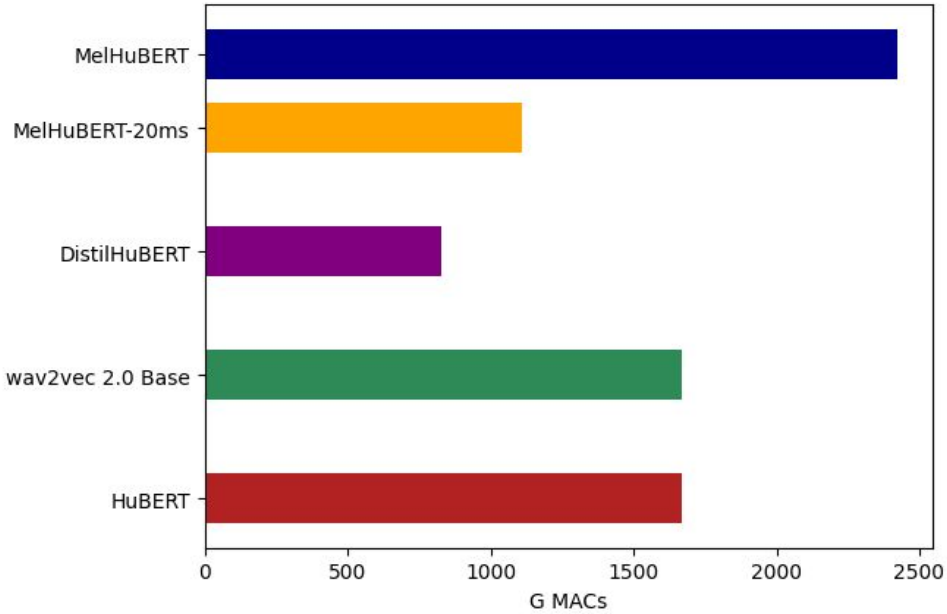


Two-stage pre-training



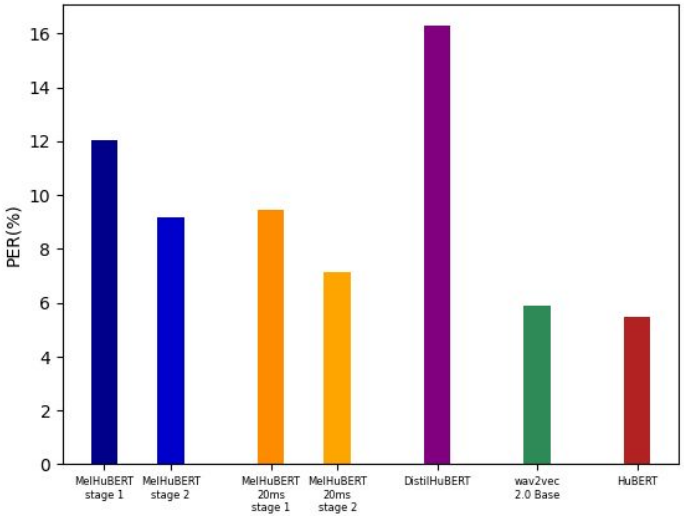
- Stage 1: predicting cluster IDs of log Mel features
- Stage 2: predicting cluster IDs of the 8th hidden layer from stage 1

MACs of MelHuBERT

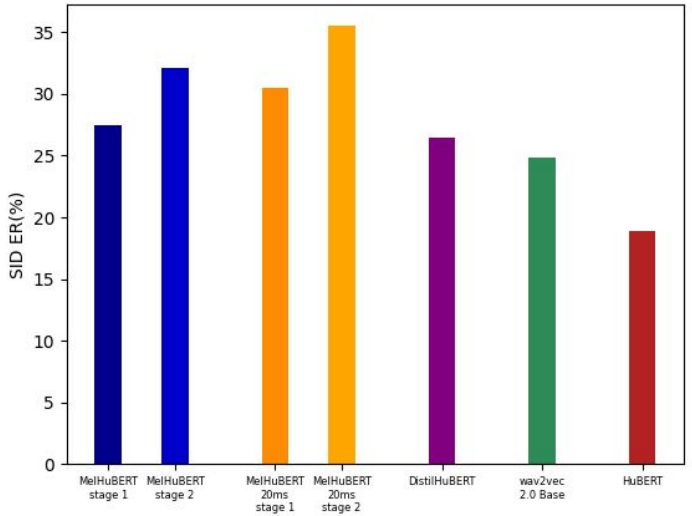


MACs = Multiply-accumulate operation

Downstream performance of MelHuBERT



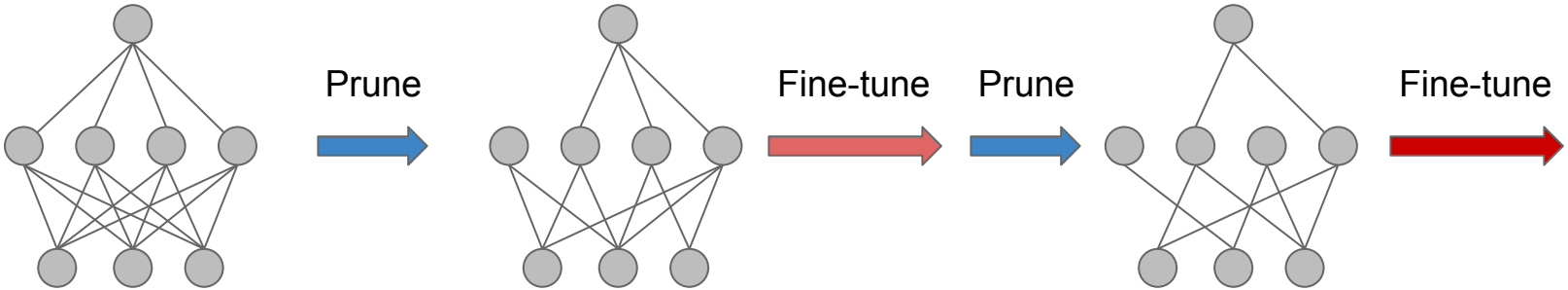
Phone Recognition



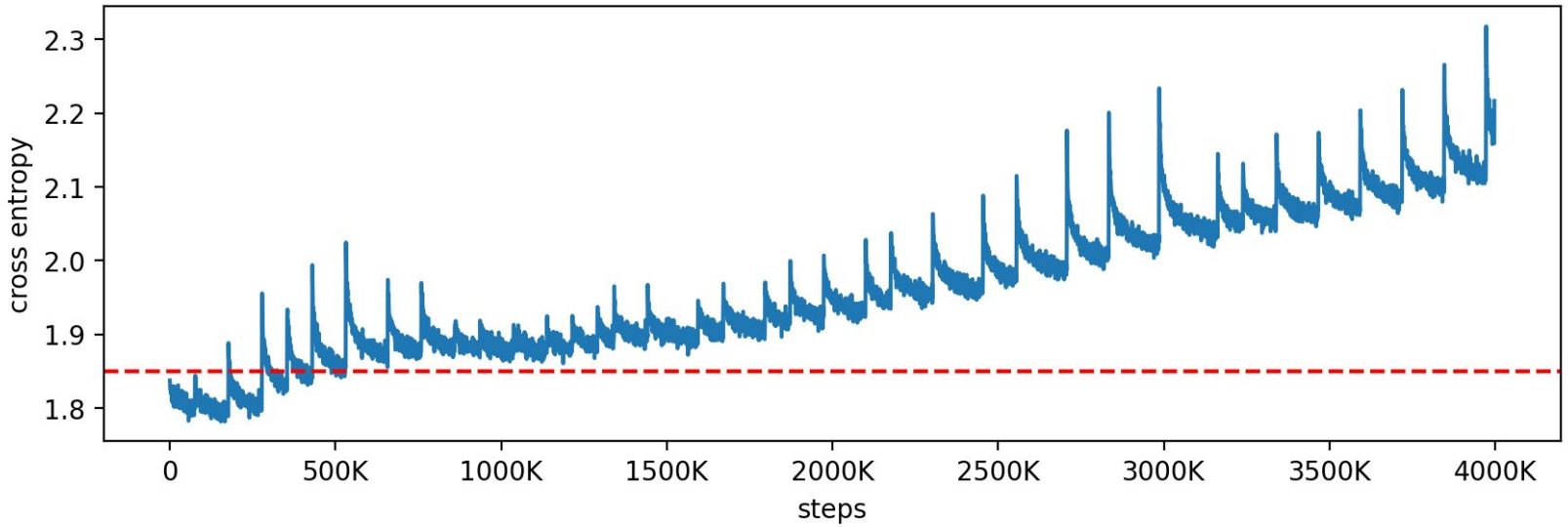
Speaker Identification

Weight Pruning

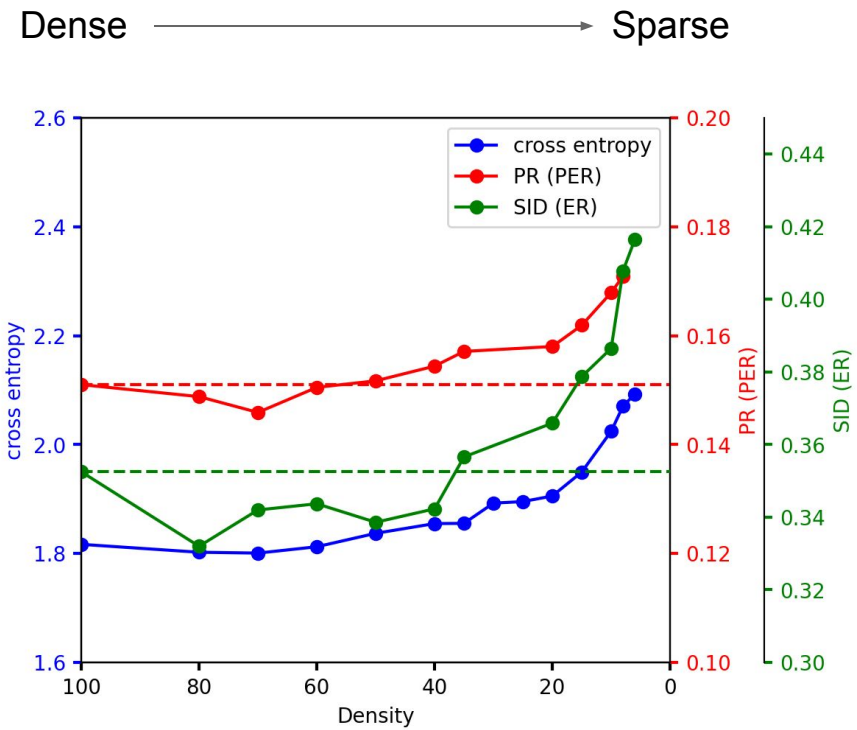
Iterative pruning



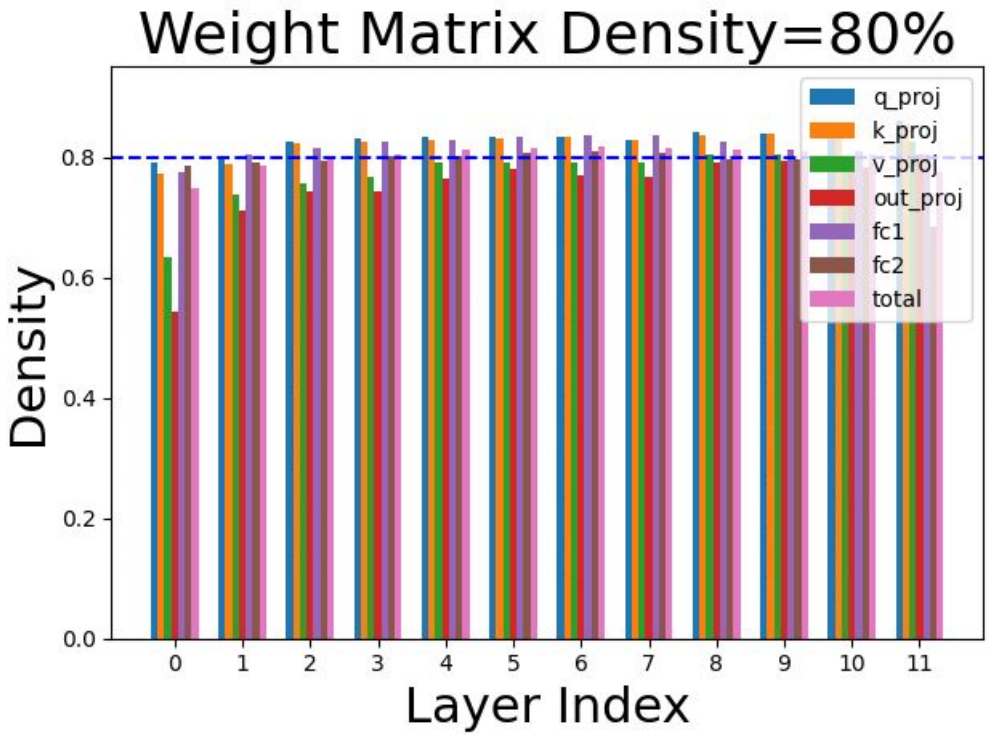
Training loss of weight pruning



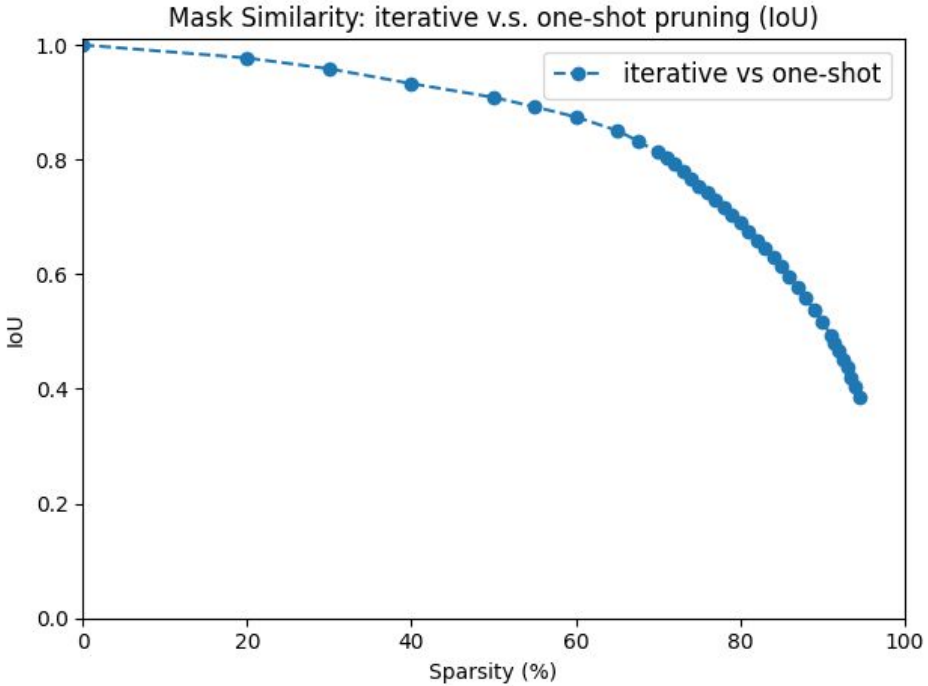
Downstream performance of weight pruning



Weight density



Mask similarity

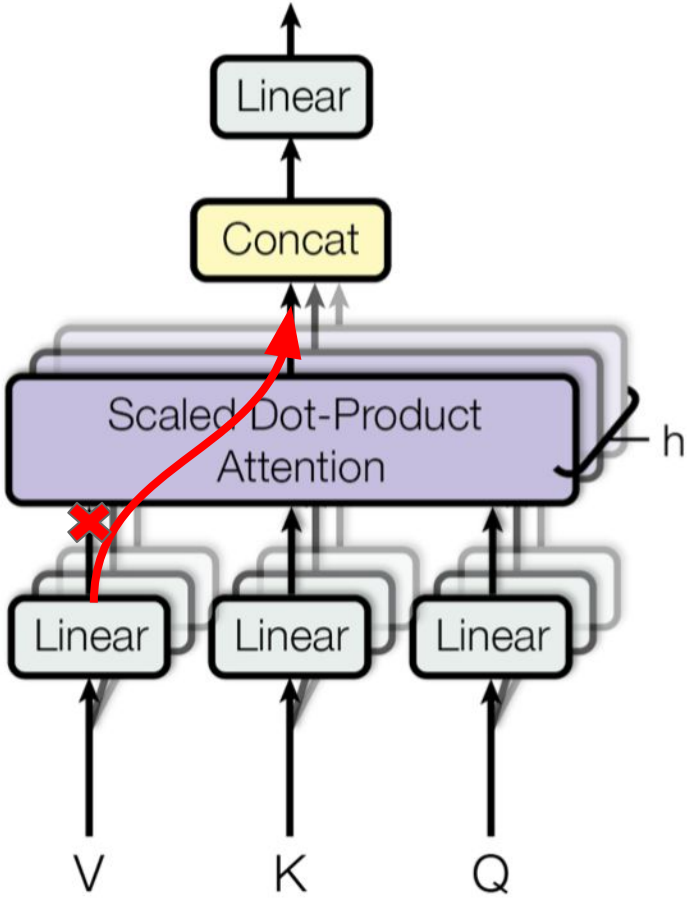


Head Pruning

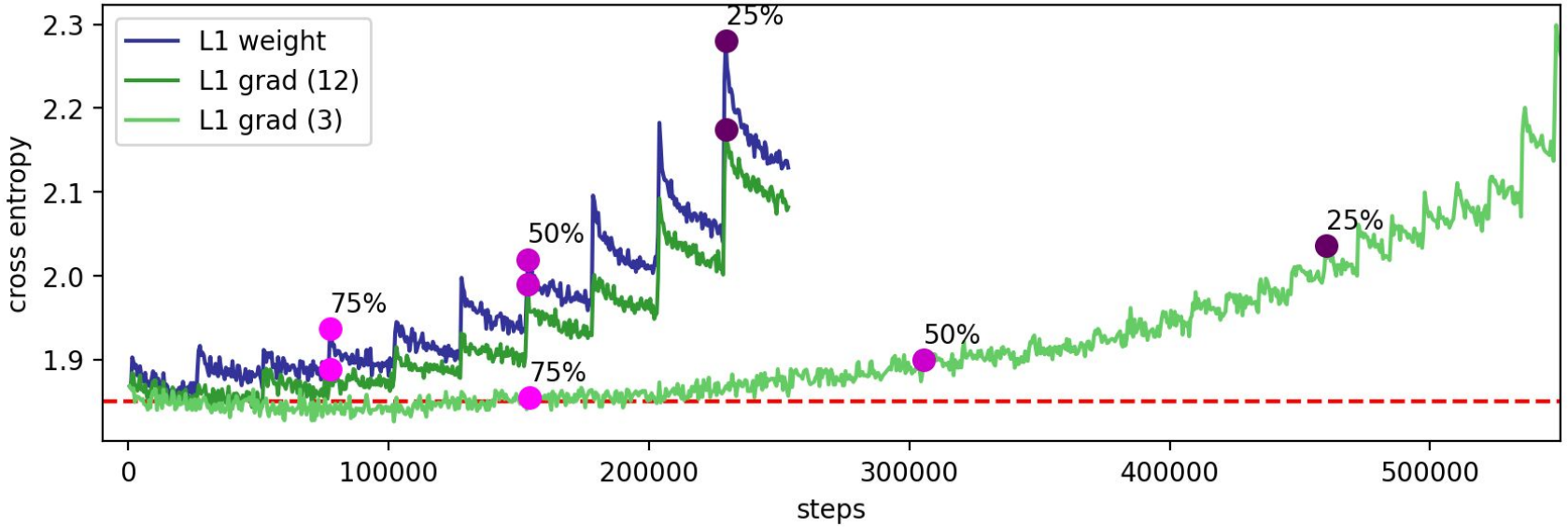
Head pruning approach

1. Prune based on L1 norm of the weights or L1 norm of the gradients

2. Skip ahead when attention map is sufficiently diagonal (replacing attention with identity matrix *I*)

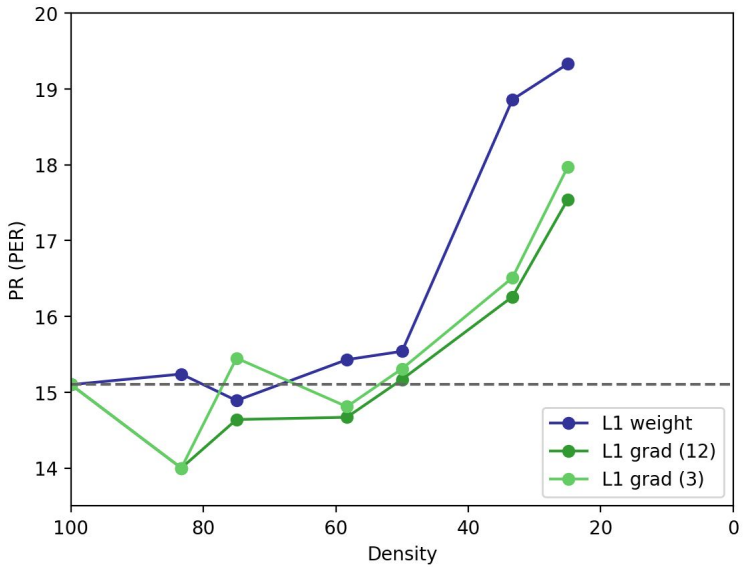


Training loss of head pruning



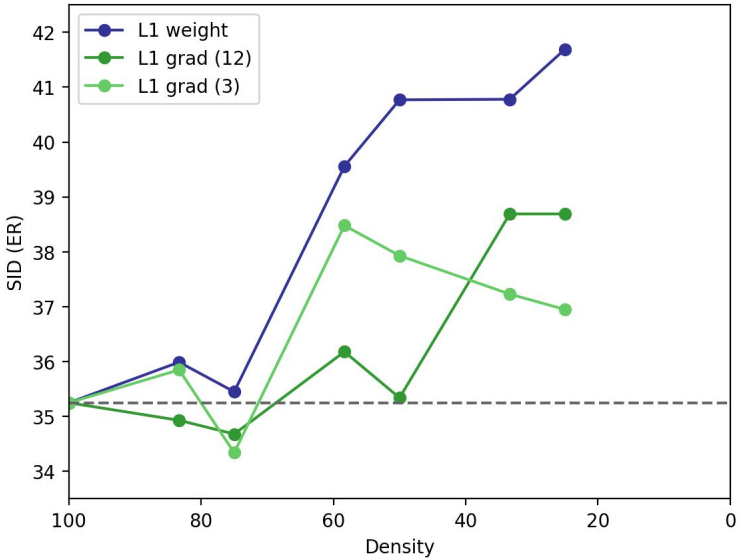
Downstream performance of head pruning

Dense → Sparse



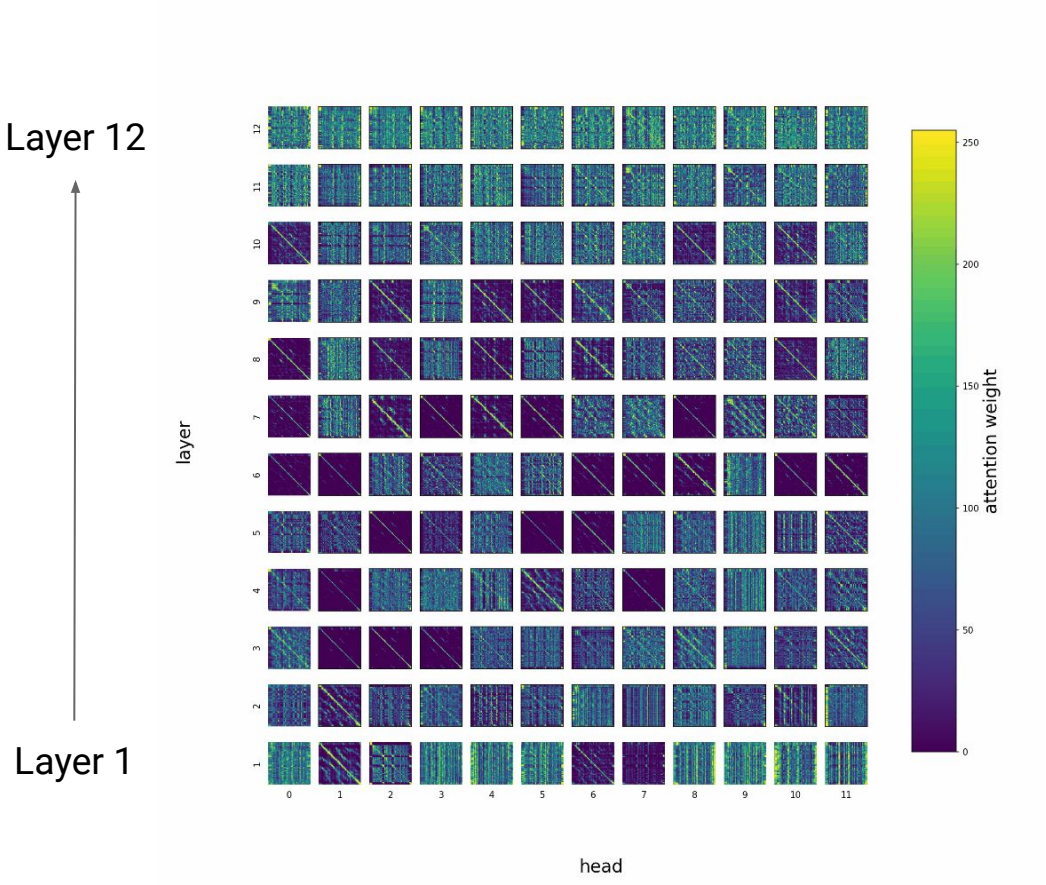
Phone recognition

Dense → Sparse

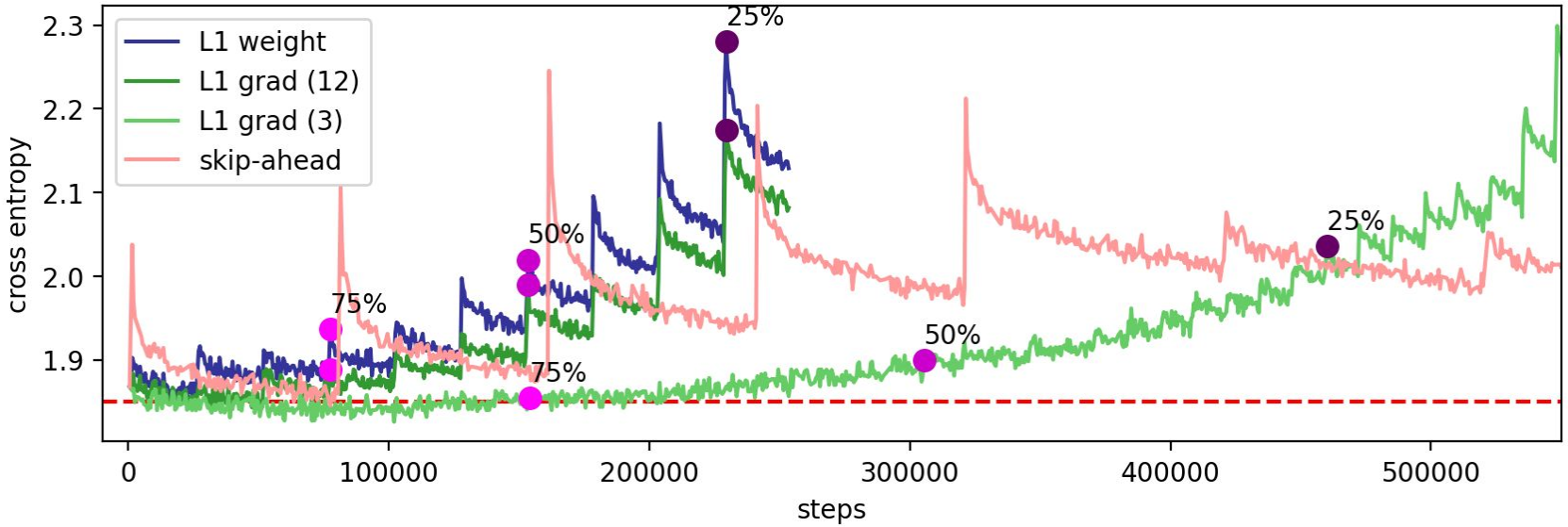


Speaker identification

Heads pruned over time

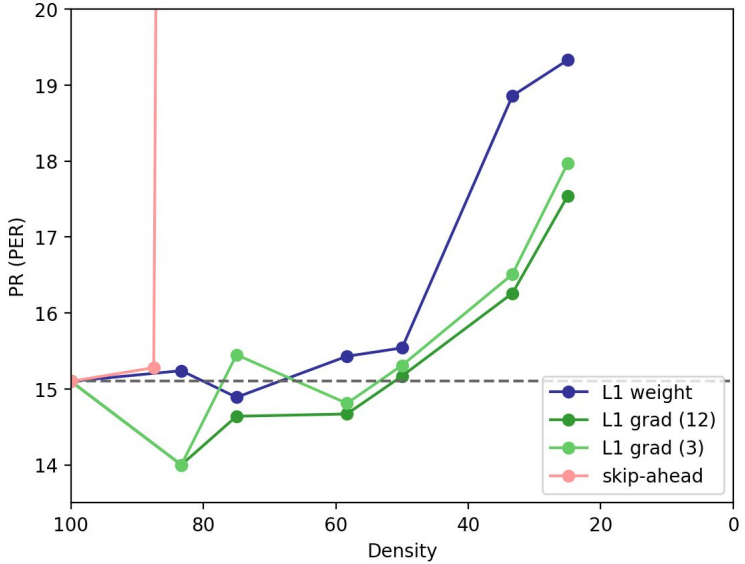


Training loss of the skip-ahead approach



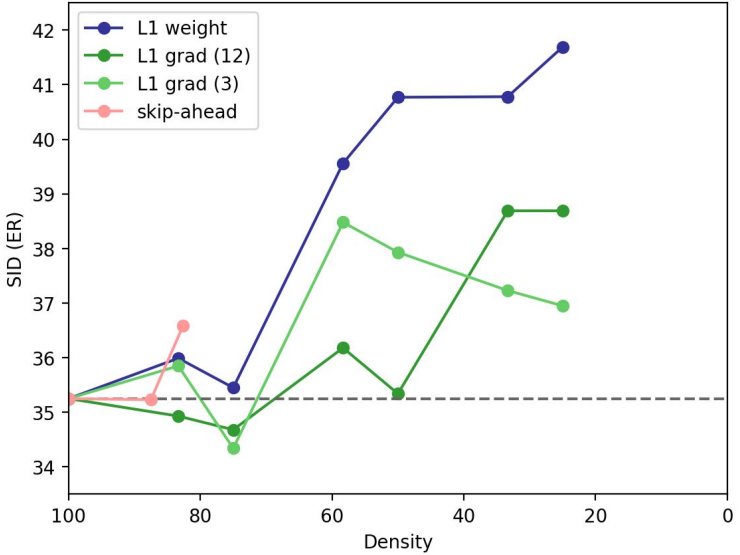
Downstream performance of the skip-ahead approach

Dense → Sparse



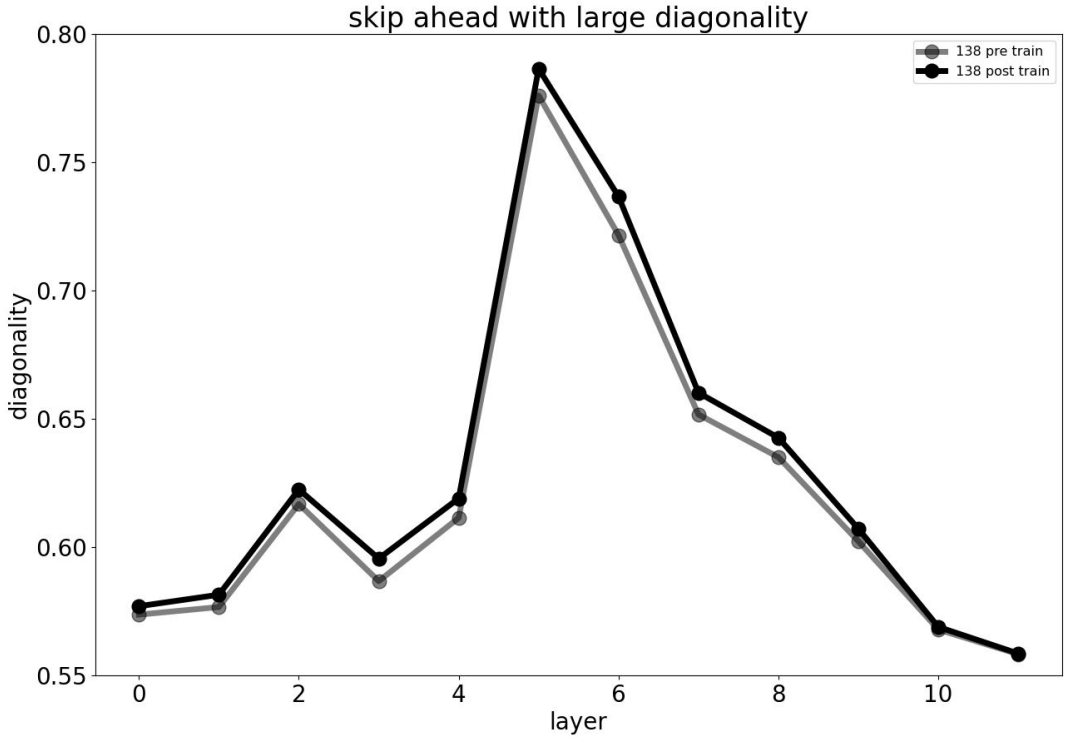
Phone recognition

Dense → Sparse

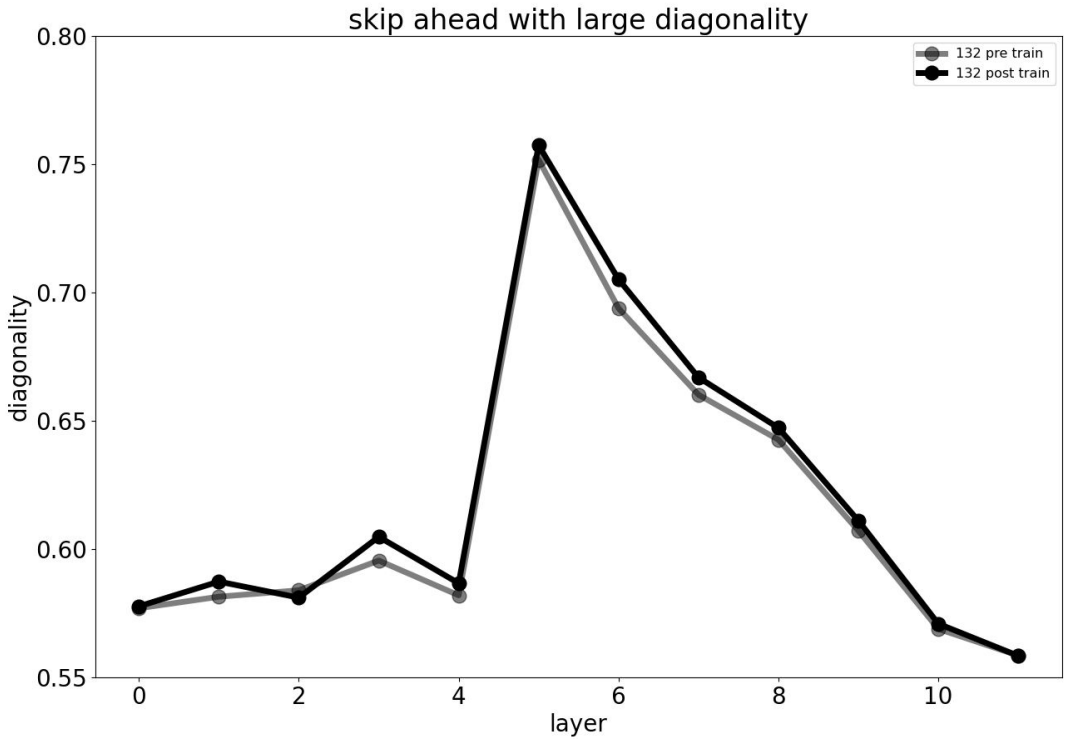


Speaker identification

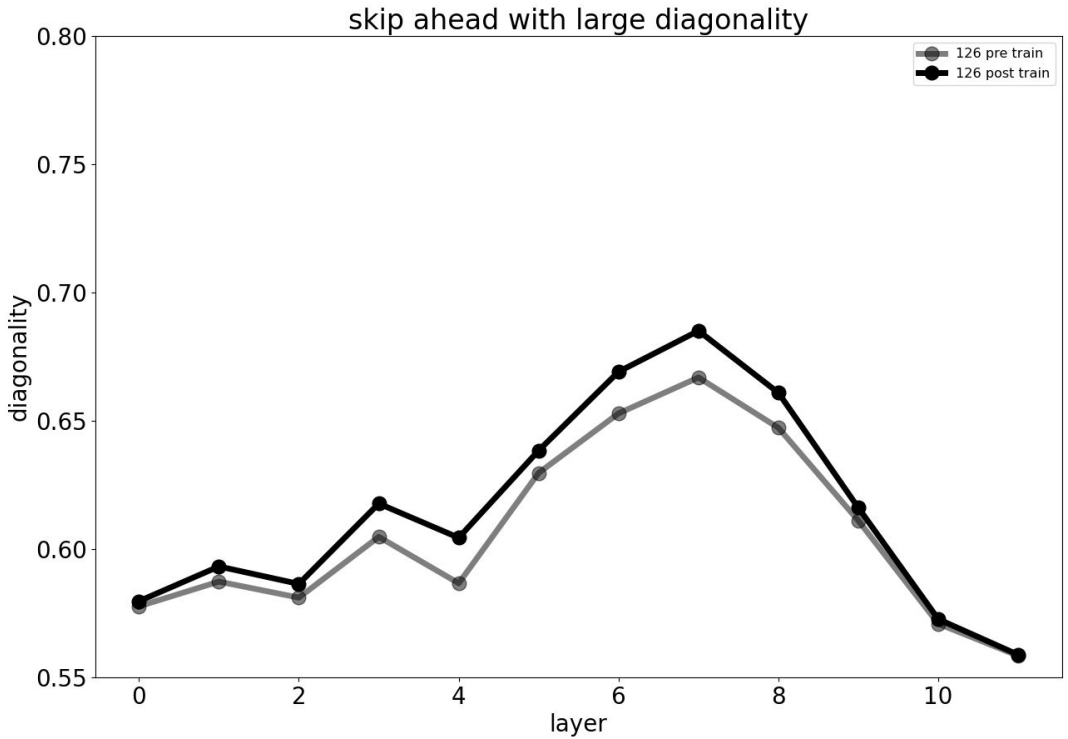
Diagonality of individual layers



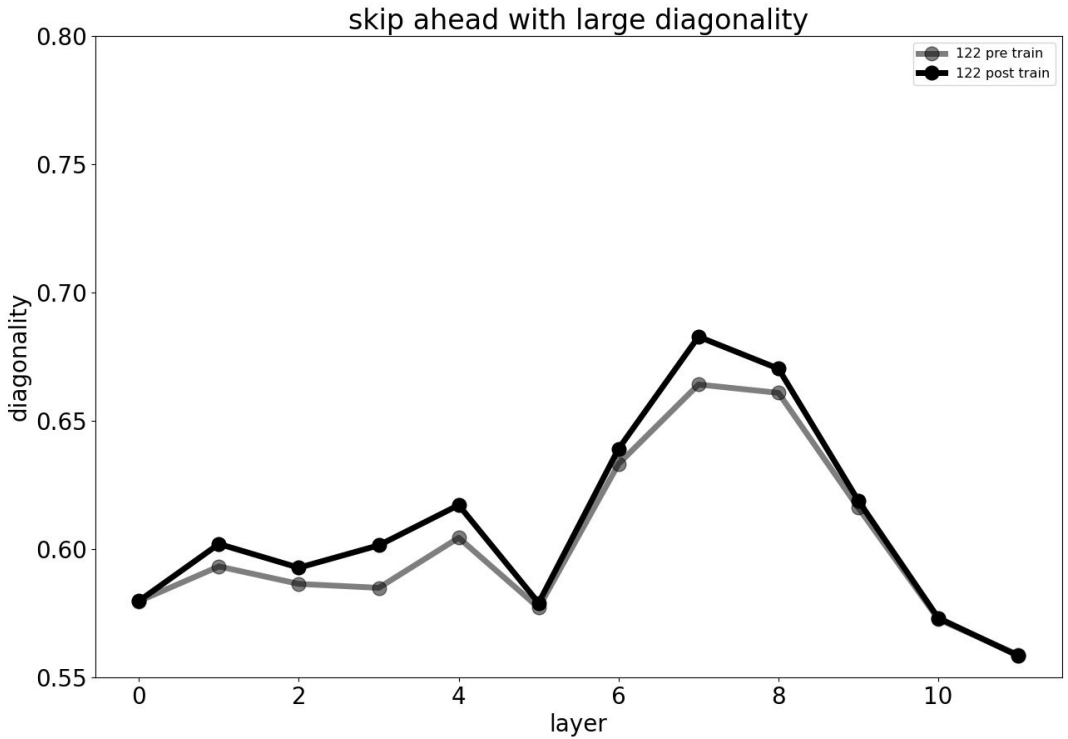
Diagonality of individual layers



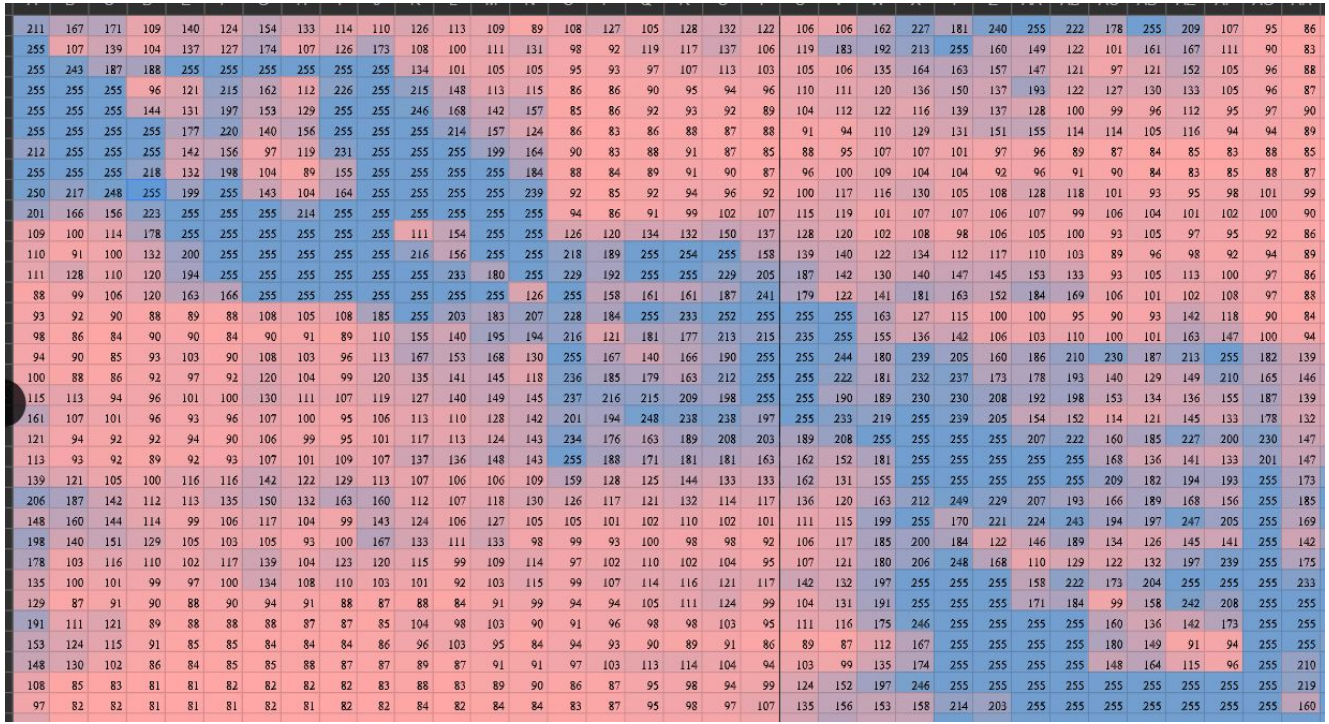
Diagonality of individual layers



Diagonality of individual layers

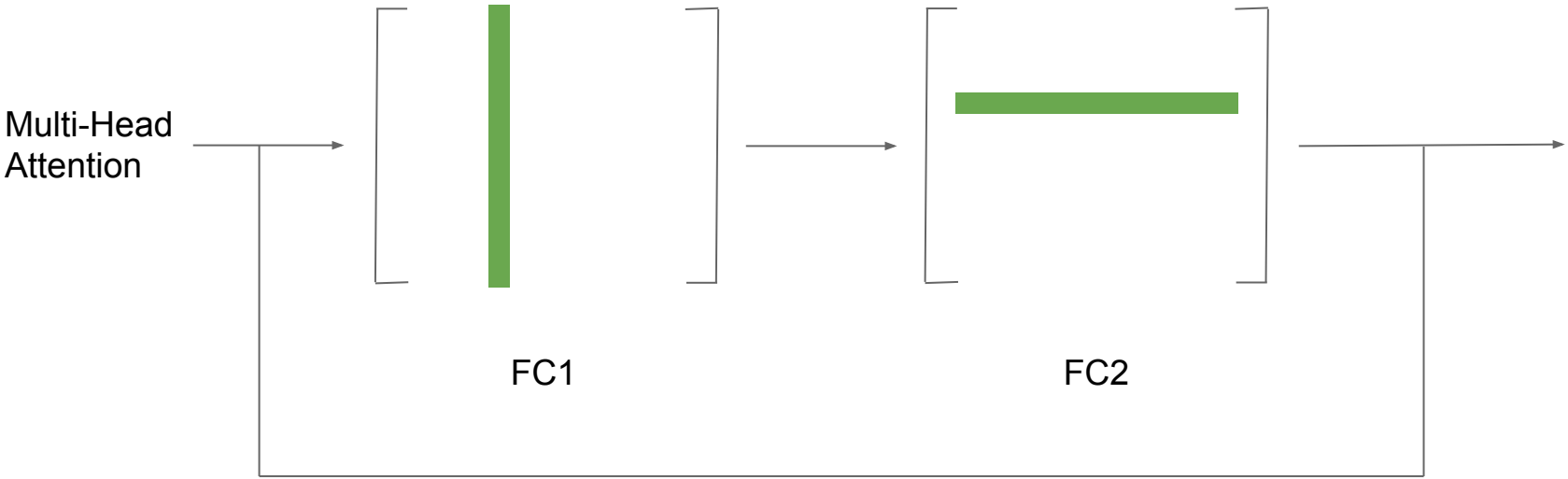


Attention map zoomed in



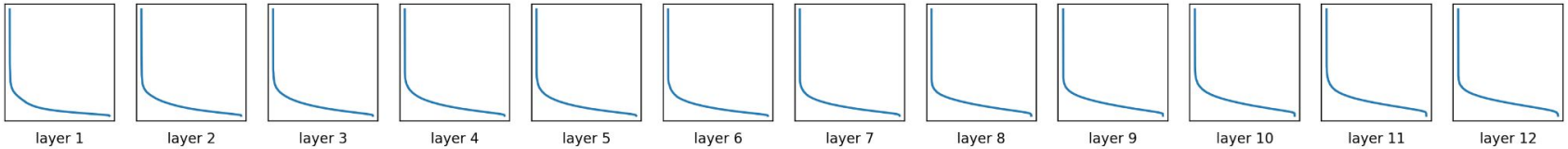
Low-Rank Approximation

Pruning rows and columns of FC1 and FC2

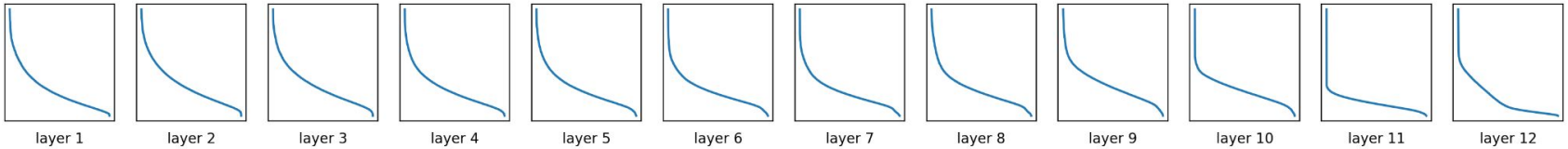


Singular values of FC1 and FC2

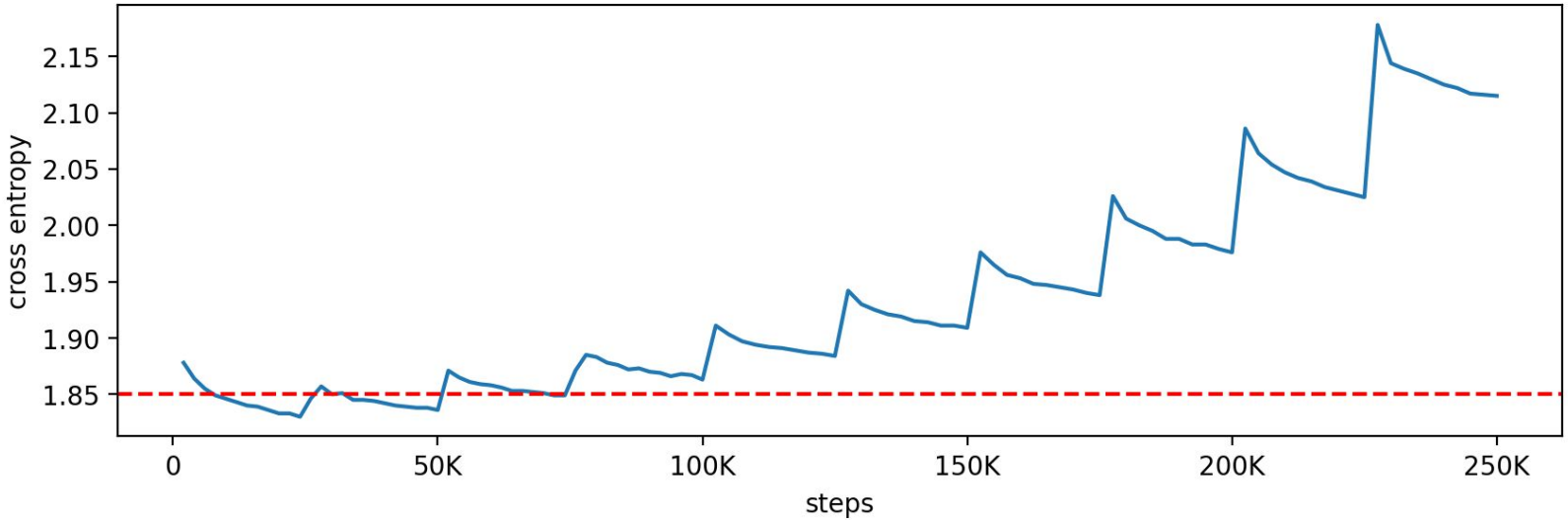
FC1



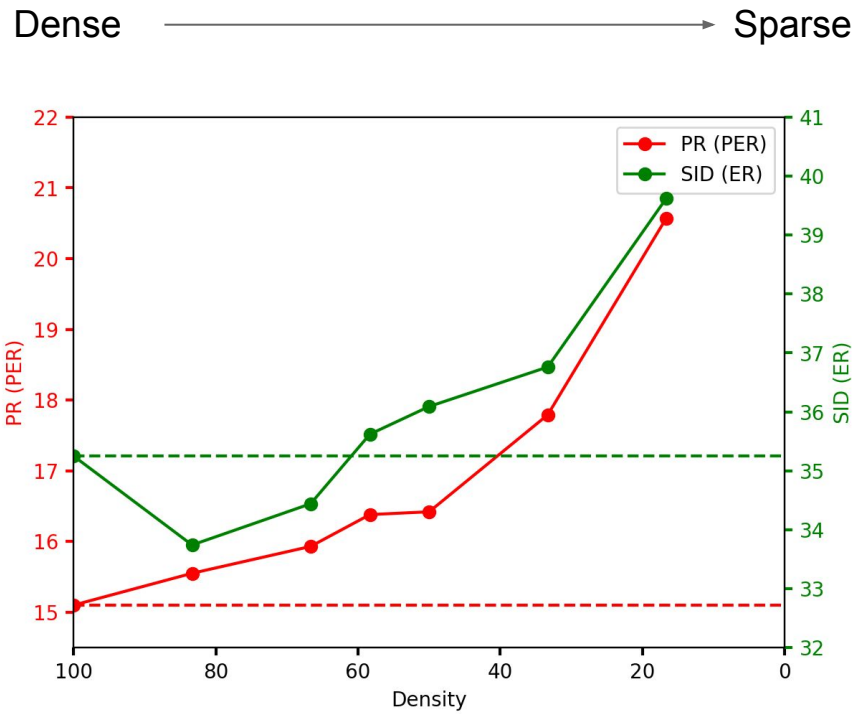
FC2



Training loss of pruning the FC layers

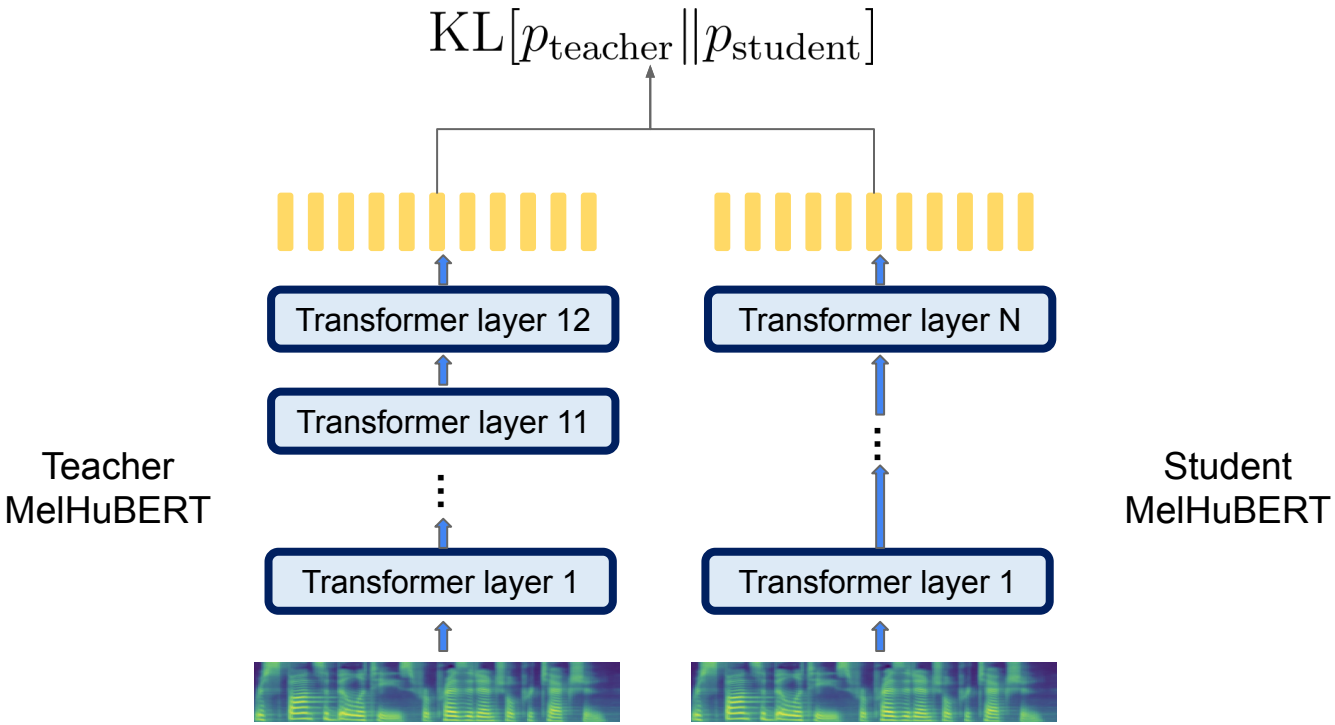


Downstream performance of pruning the FC layers

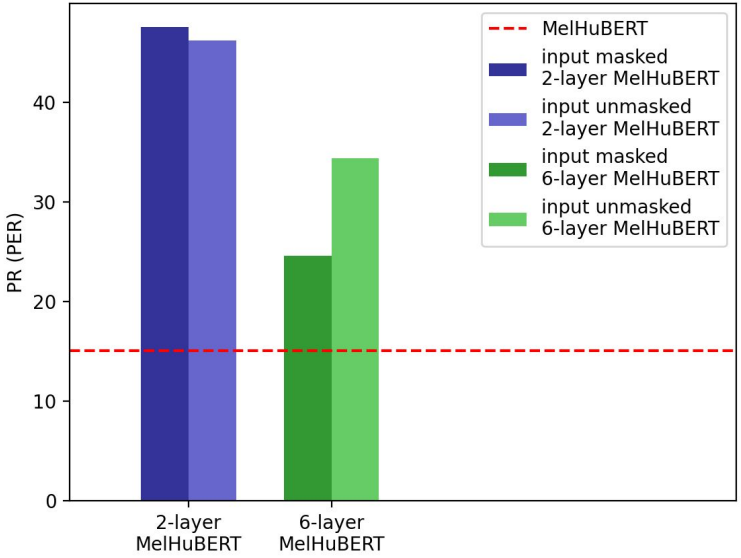


Knowledge Distillation

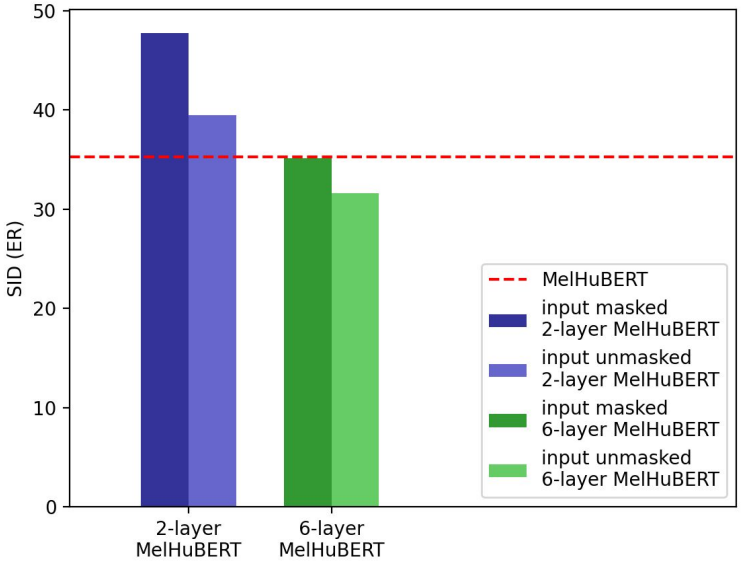
Knowledge distillation



Downstream performance of knowledge distillation



Phone recognition



Speaker identification

Sequence Compression

Introduction

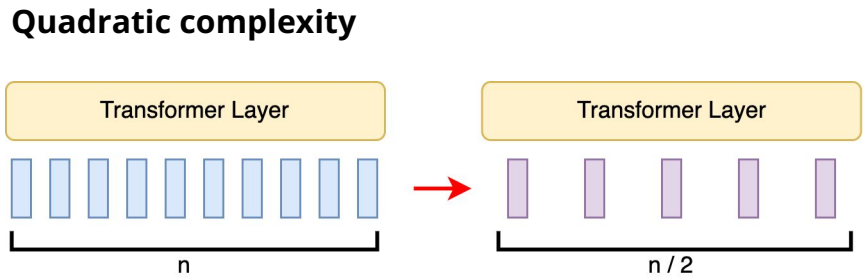
Why sequence compression?

Computational cost reduction

- Faster pre-training/inference speed
- Less operations & memory usage

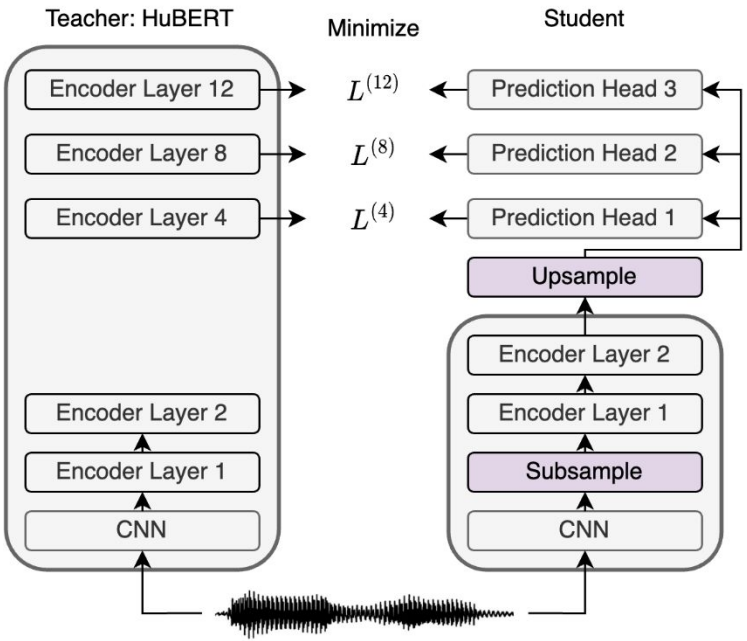
→ Impact of subsampling on different downstream tasks

→ How much can the sequence be compressed?

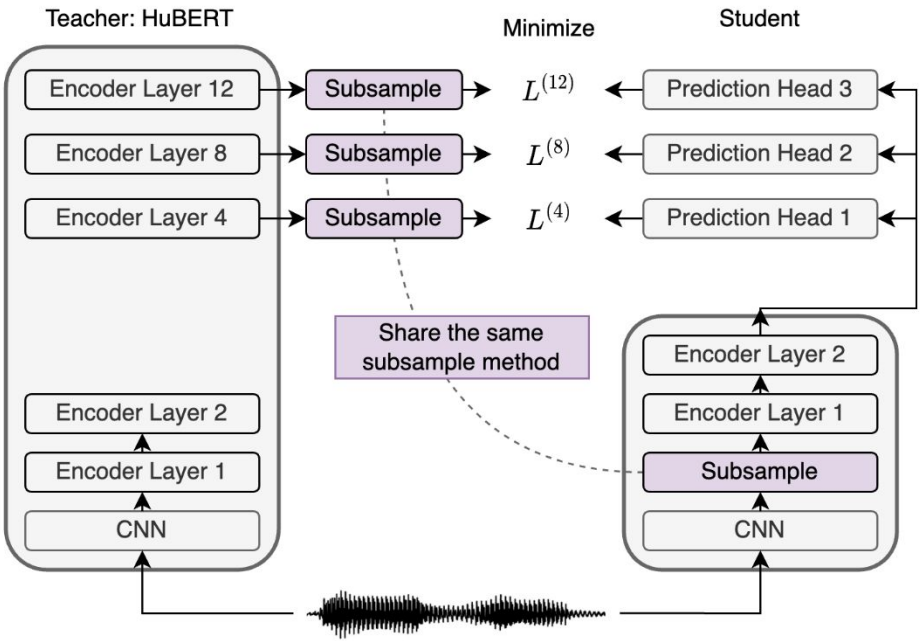


Framework

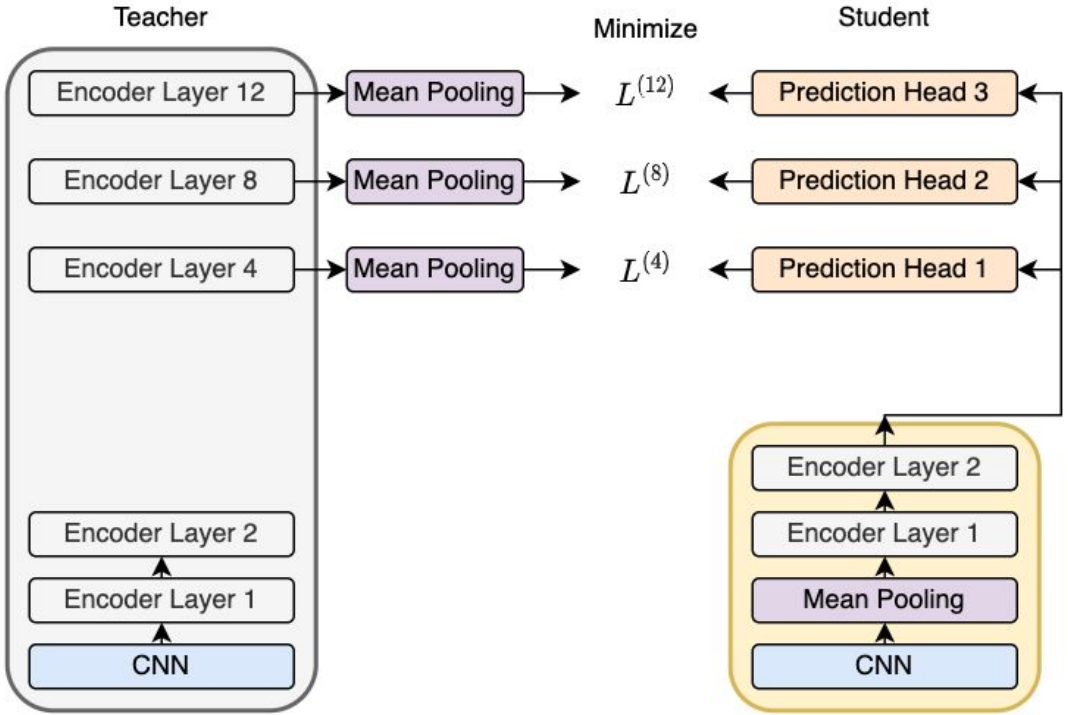
(a) With Upsample (target unchanged)



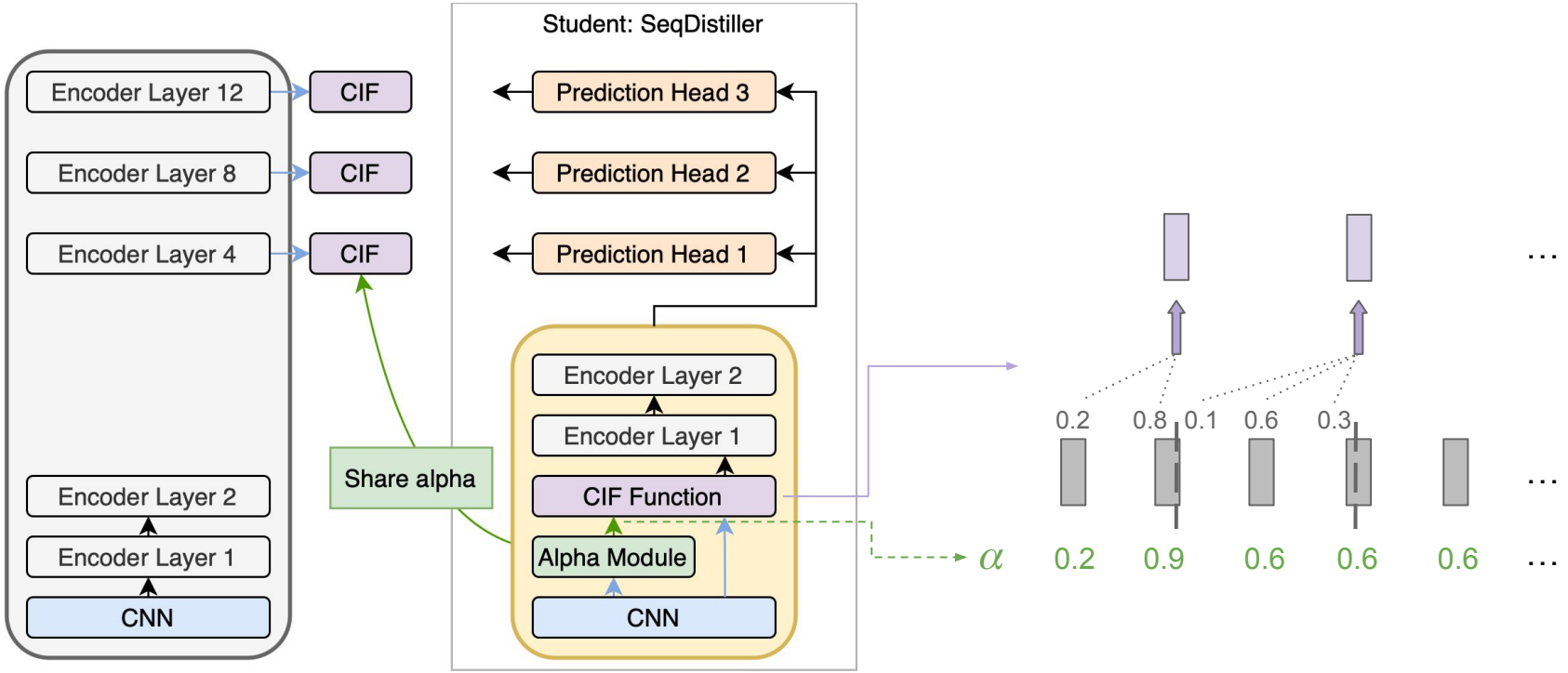
(b) Subsample Target



Subsampling – fixed-length



Subsampling – variable-length



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

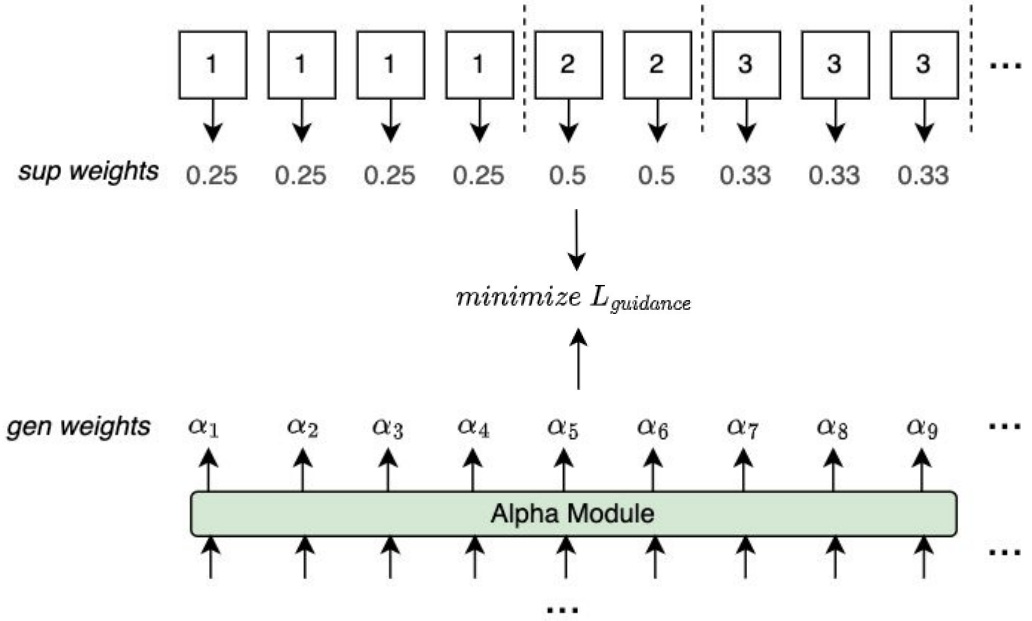
Segmentation guidance

Unsupervised

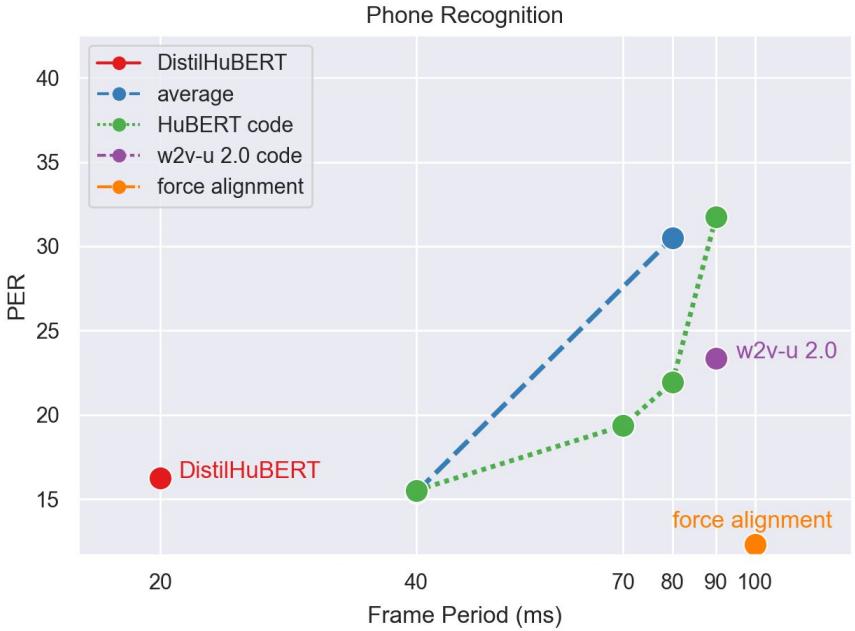
- Repetition in HuBERT codes
- Repetition in wav2vec-U 2.0 codes

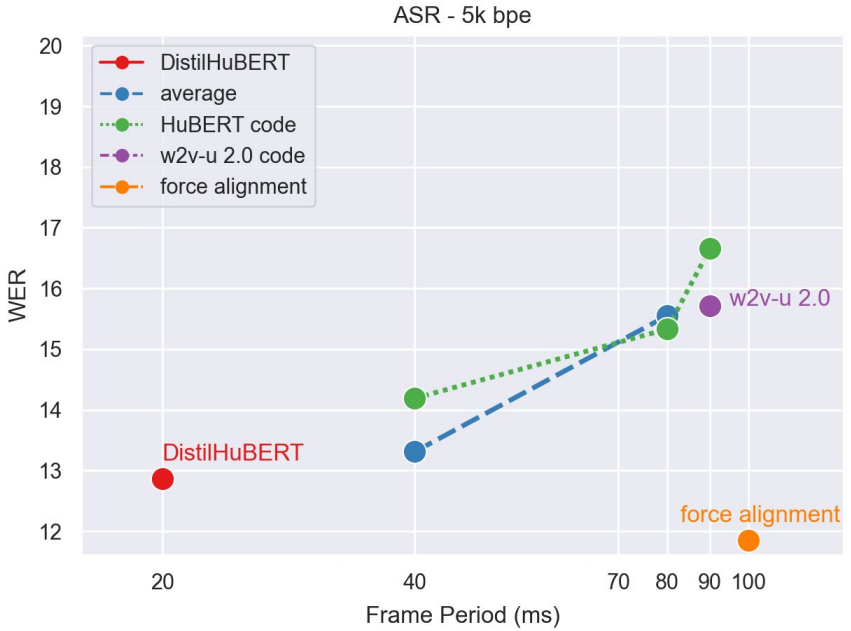
Supervised

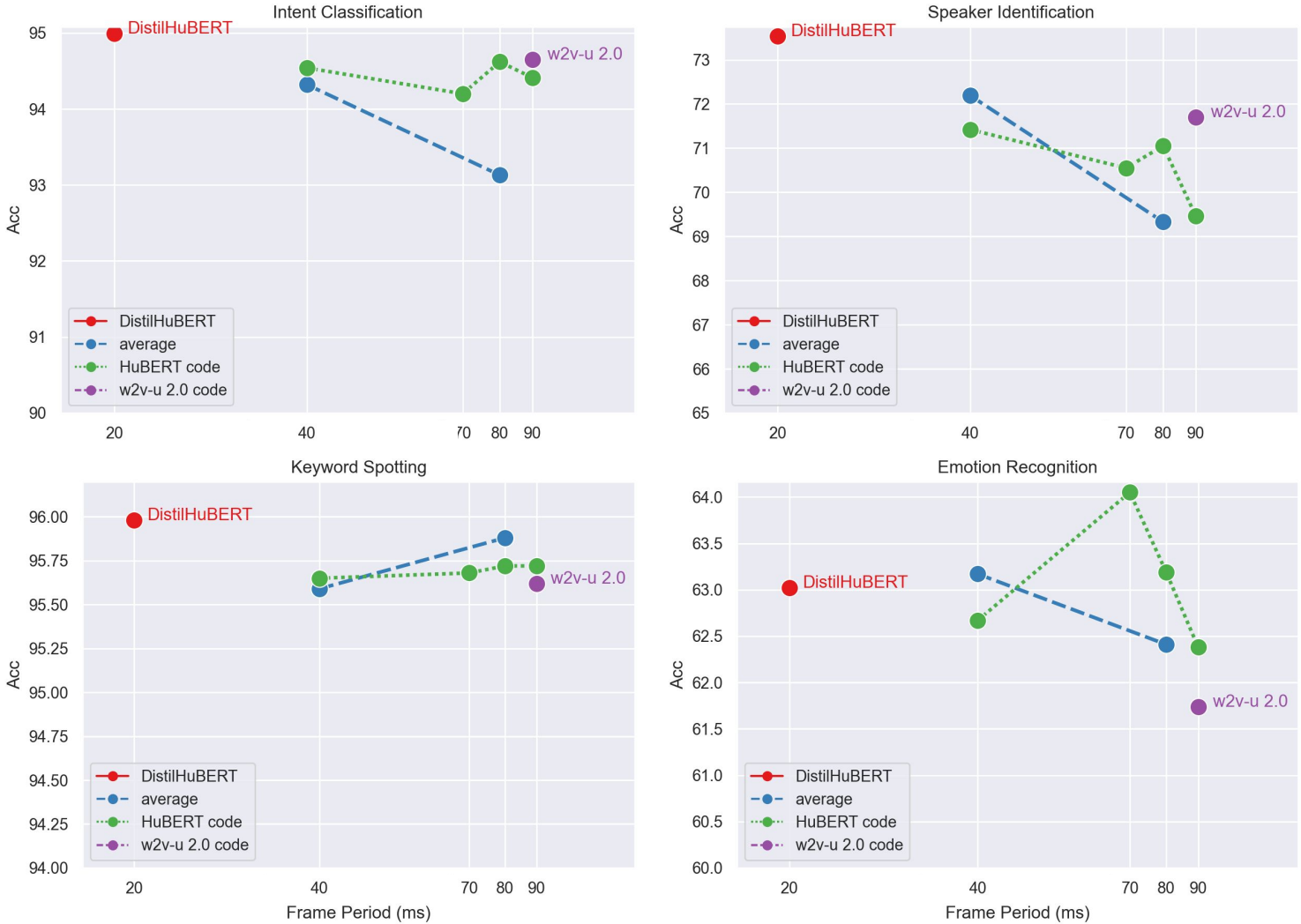
- Forced alignments



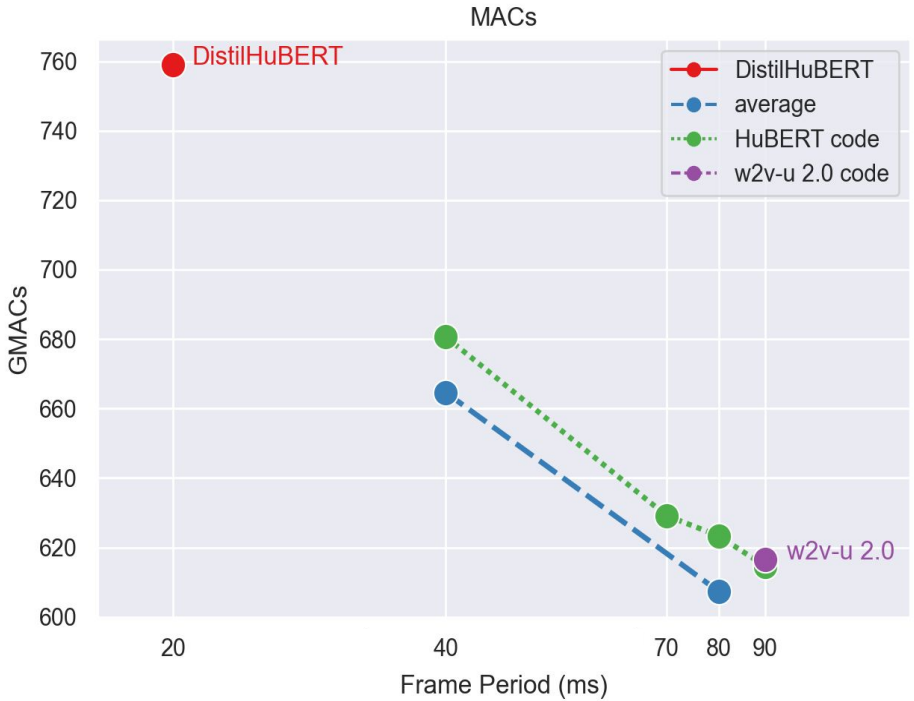
Results


Phone Recognition
DistilHuBERT
average
HuBERT code
w2v-u 2.0 code
force alignment
PER
40
35
30
25
20
15
w2v-u 2.0
DistilHuBERT
force alignment
20
40
70
80
90
100
Frame Period (ms)


ASR - 5k bpe
DistilHuBERT
average
HuBERT code
w2v-u 2.0 code
force alignment
WER
20
19
18
17
16
15
14
13
12
w2v-u 2.0
DistilHuBERT
force alignment
20
40
70
80
90
100
Frame Period (ms)


Intent Classification
DistilHuBERT
w2v-u 2.0
Acc
95
94
93
92
91
90
DistilHuBERT
average
HuBERT code
w2v-u 2.0 code
20
40
70
80
90
Speaker Identification
DistilHuBERT
w2v-u 2.0
Acc
73
72
71
70
69
68
67
66
65
DistilHuBERT
average
HuBERT code
w2v-u 2.0 code
20
40
70
80
90
Keyword Spotting
DistilHuBERT
w2v-u 2.0
Acc
96.00
95.75
95.50
95.25
95.00
94.75
94.50
94.25
94.00
DistilHuBERT
average
HuBERT code
w2v-u 2.0 code
20
40
70
80
90
Frame Period (ms)
Emotion Recognition
DistilHuBERT
w2v-u 2.0
Acc
64.0
63.5
63.0
62.5
62.0
61.5
61.0
60.5
60.0
DistilHuBERT
average
HuBERT code
w2v-u 2.0 code
20
40
70
80
90
Frame Period (ms)

Results – MACs



Sequence compression

- Different tasks have their preferred frame rates for better performance.
- While fixed-length subsampling can perform well with smaller compress rate, variable-length subsampling works particularly well under low frame rates.
- Forced alignment boundaries significantly improves the performance of phone recognition and ASR even with a frame rate as low as 10Hz.

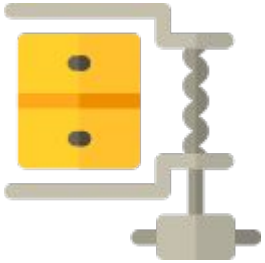
Final Remarks

Main findings

- There exists various small models that can achieve competitive performance.

- Different techniques have different strengths.
  - Weight pruning achieves the most compression, while maintaining performance.
  - Head pruning maintains phone recognition, while losing speaker identification.
  - Low-rank approximation reduces a significant amount of parameters.
  - Knowledge distillation has difficulty applying to shallow networks.
  - Fixed-length subsampling works well in high frame-rate settings, while variable-length subsampling works well in low frame-rate settings.

Outline of Part 1: Better Pre-trained Model



Compression

9:35 - 9:55

Robust

9:55 - 10:10

Visual-enhanced

10:10 - 10:25

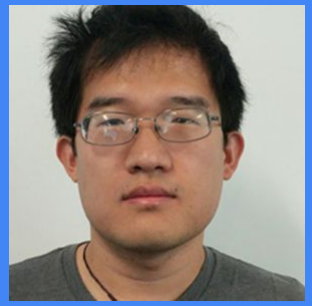
Integration

10:25 - 10:30

10 mins Q&A
+ 10 mins break

Generalization of SSL

Hung-yi Lee (NTU)

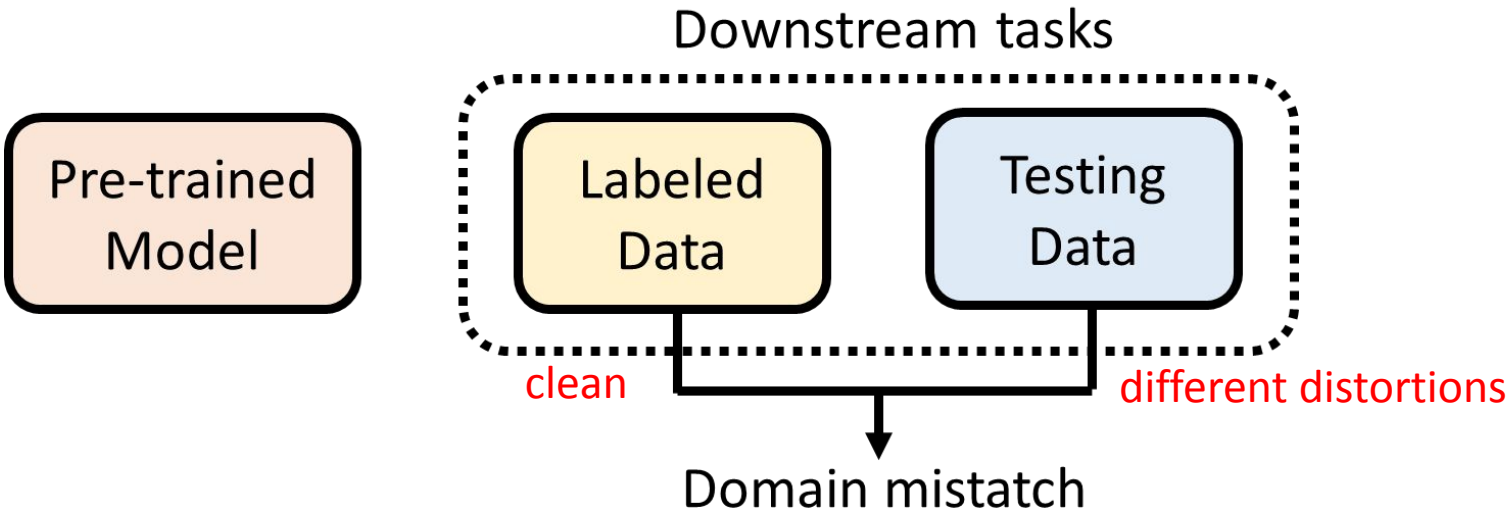


Yu Zhang (Google)

Kuan-Po Huang (NTU)

Fabian Ritter (NUS)

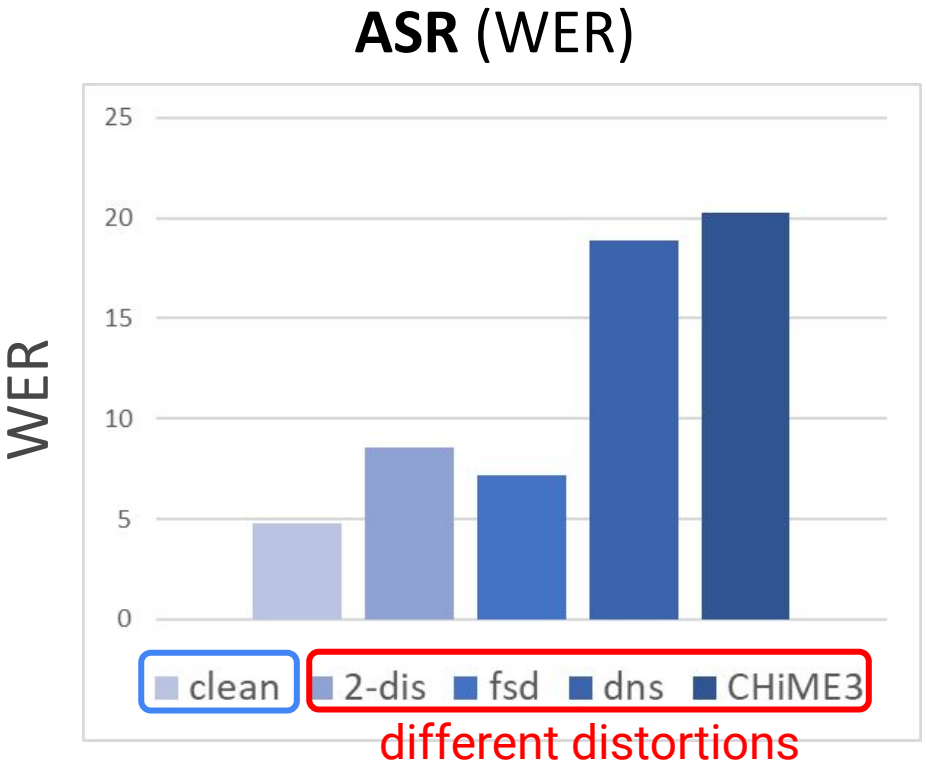
Generalization of SSL



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

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

Generalization of SSL

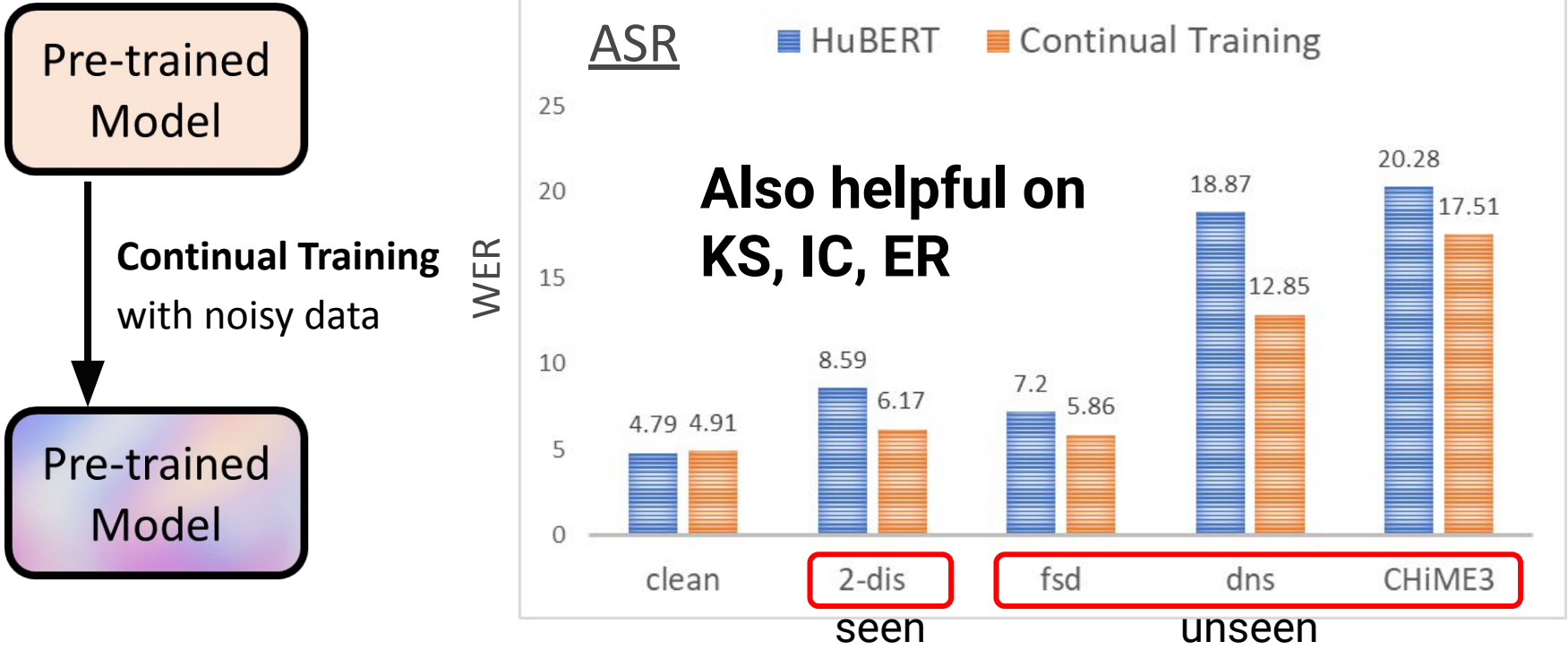


The same observation on

- Keyword Spotting (KS)
- Intent Classification (IC)
- Emotion Recognition (ER)

Generalization of SSL

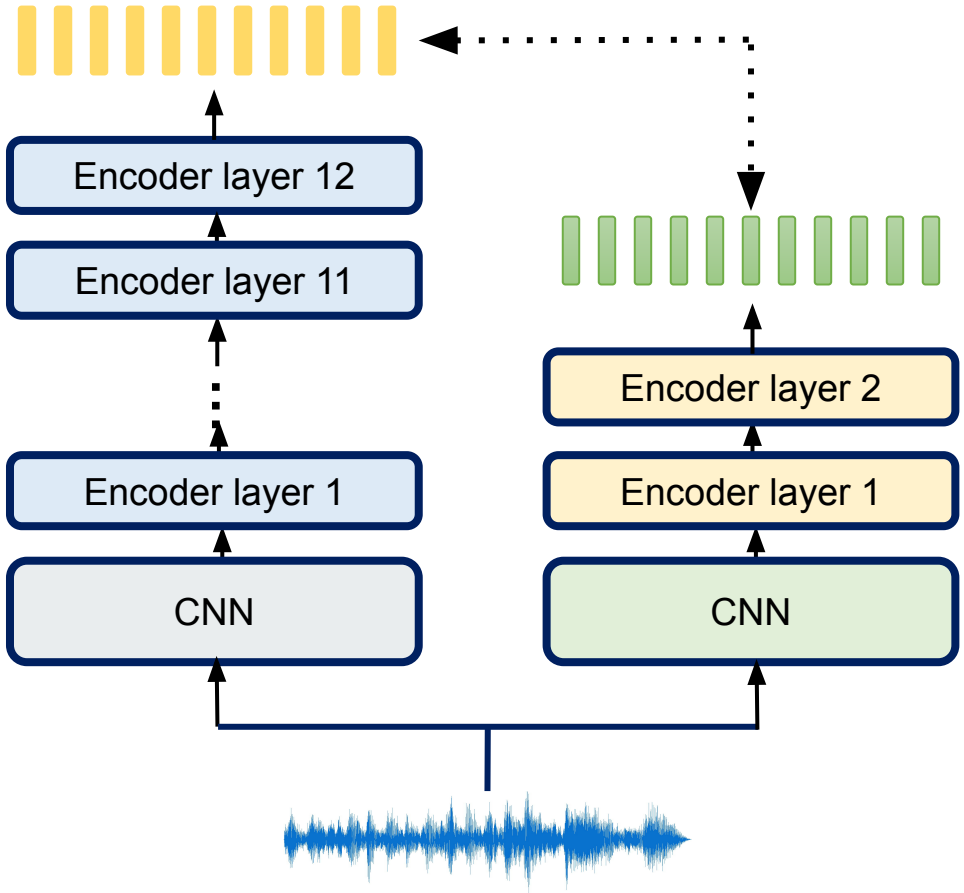
Kuan Po Huang, Yu-Kuan Fu, Yu Zhang, Hung-yi Lee, Improving Distortion Robustness of Self-supervised Speech Processing Tasks with Domain Adaptation, Interspeech, 2022



Compressed networks from Knowledge Distillation are less robust.

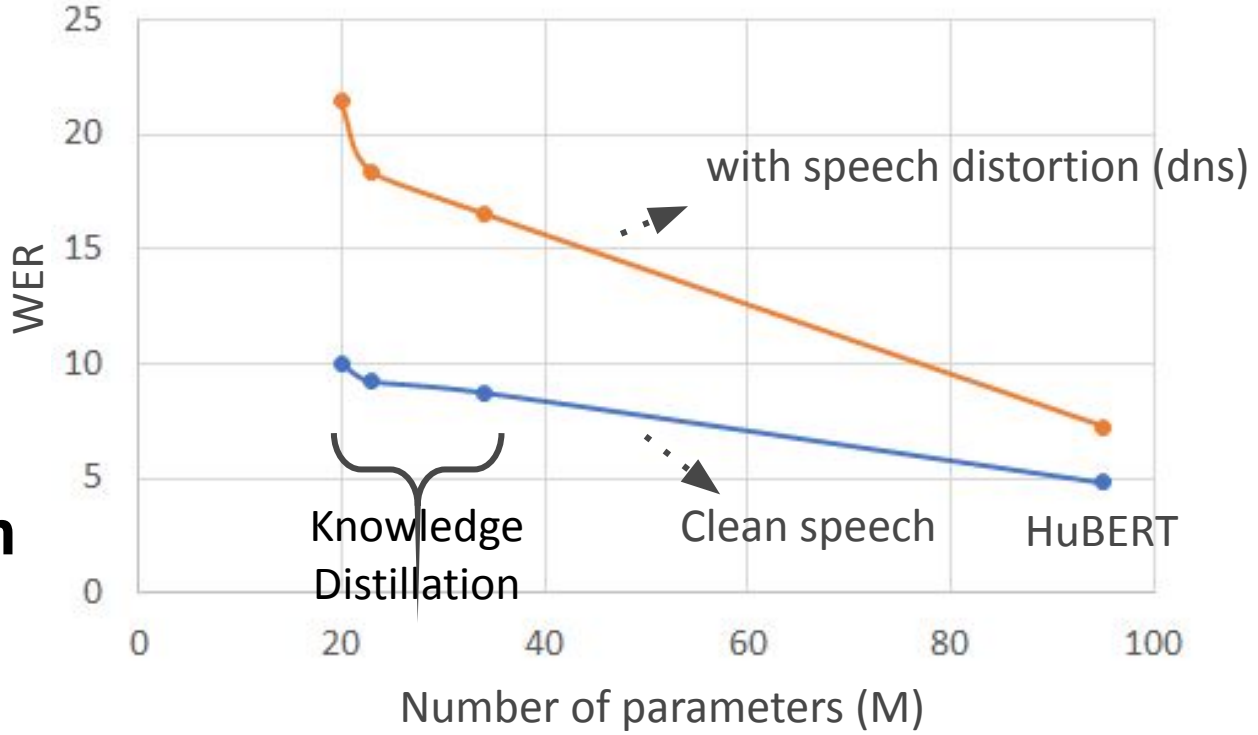
We use the framework of *DistilHuBERT*.

Heng-Jui Chang, Shu-wen Yang, Hung-yi Lee, Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert, ICASSP, 2022



Compressed SSL Models are less robust.

ASR



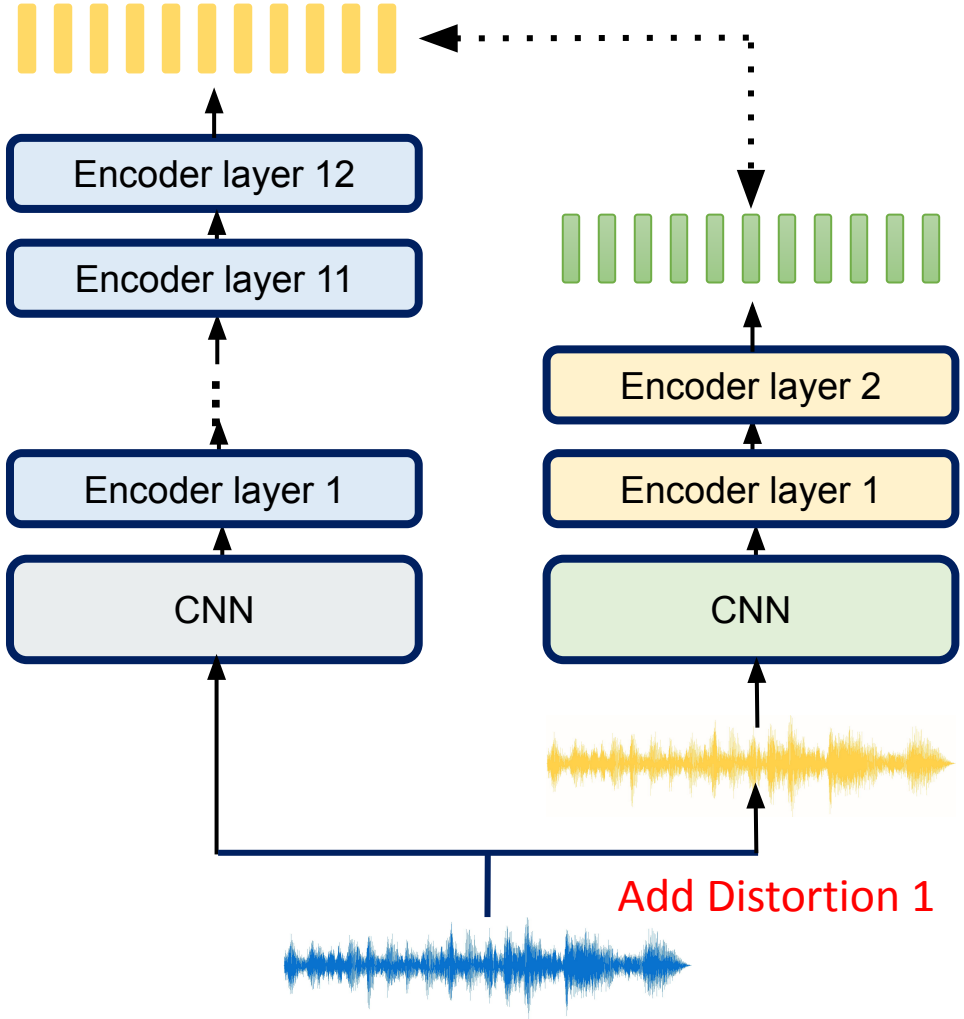
The same observation on KS, IC, ER

DistilHuBERT

+ Cross-Distortion Mapping

Setup 1:

- Student input: distortion
- Teacher input: clean



DistilHuBERT

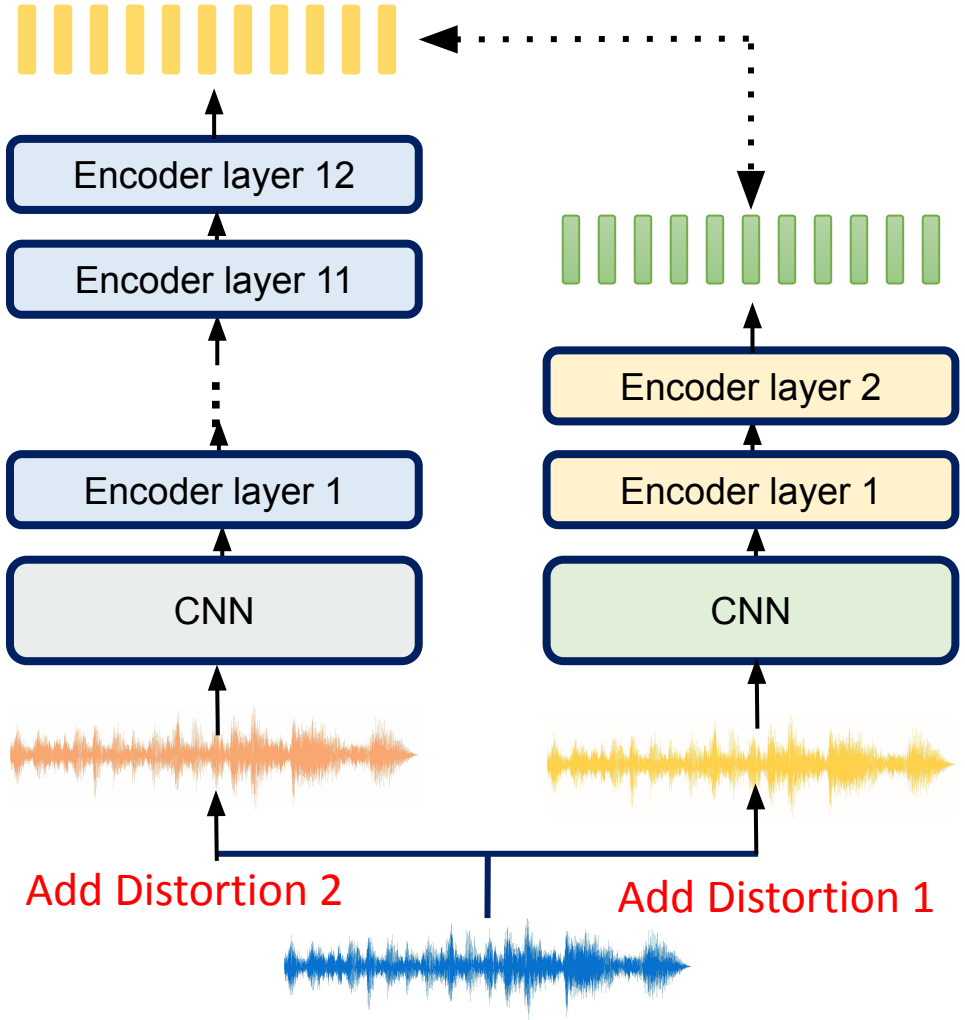
+ Cross-Distortion Mapping

Setup 1:

- Student input: distortion
- Teacher input: clean

Setup 2:

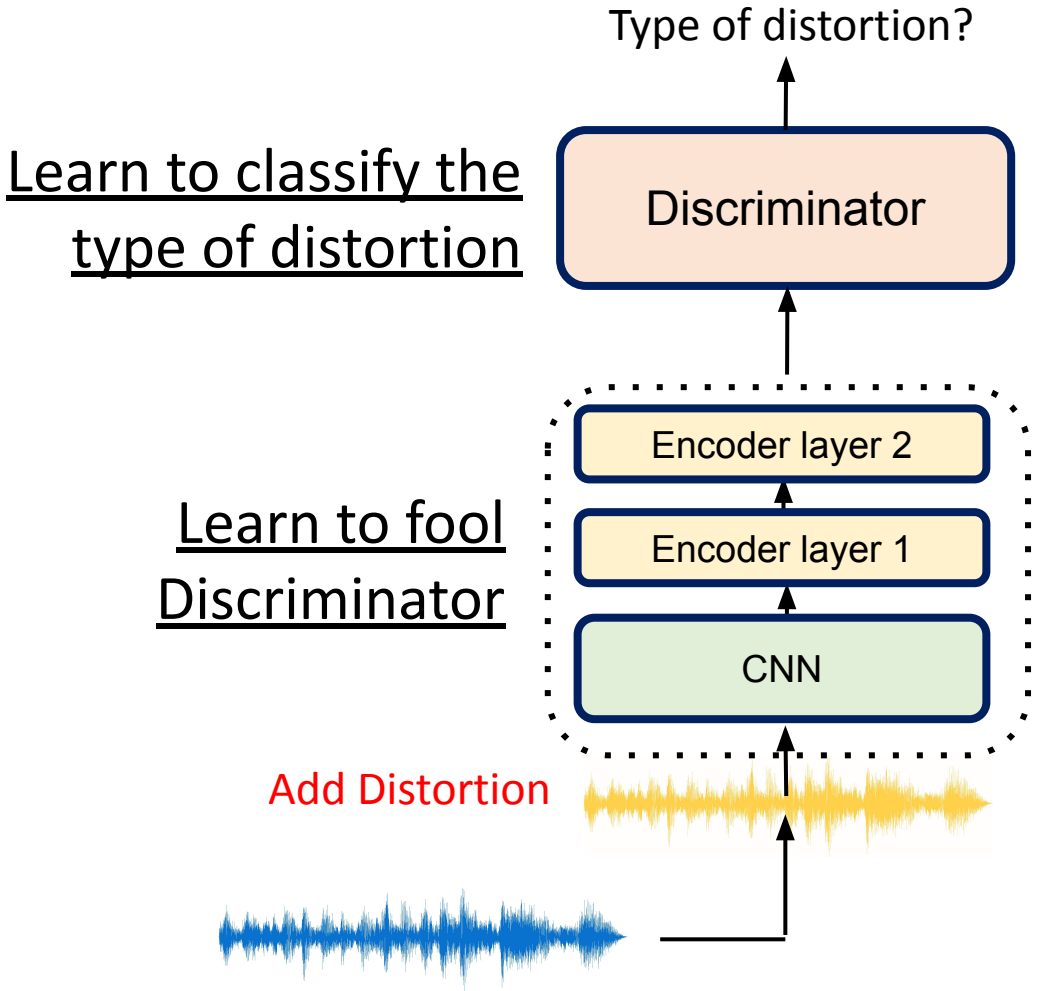
- Student input: distortion 1
- Teacher input: distortion 2



DistilHuBERT

+ Cross-Distortion Mapping

+ Domain Adversarial Training



Experimental Results

Kuan-Po, Huang

Questions to answer

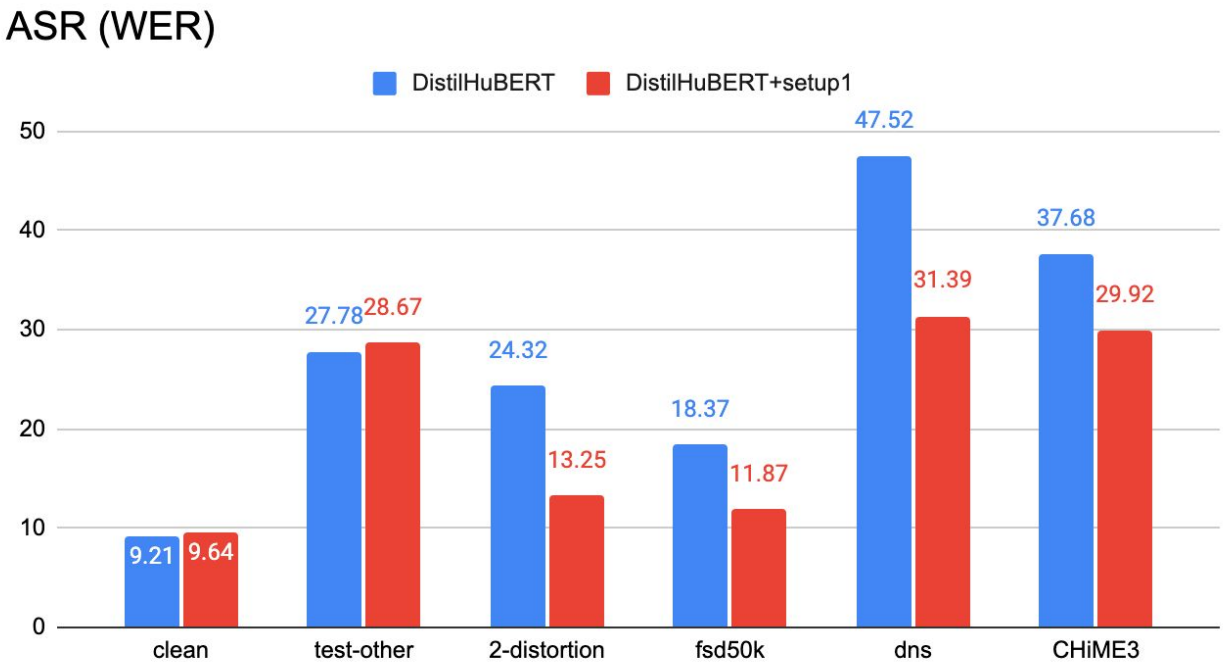
Does Cross-Distortion Mapping (CDM) enhance robustness ?

Does CDM have consistent results for models with different sizes ?

Does Domain Adversarial Training enhance robustness ?

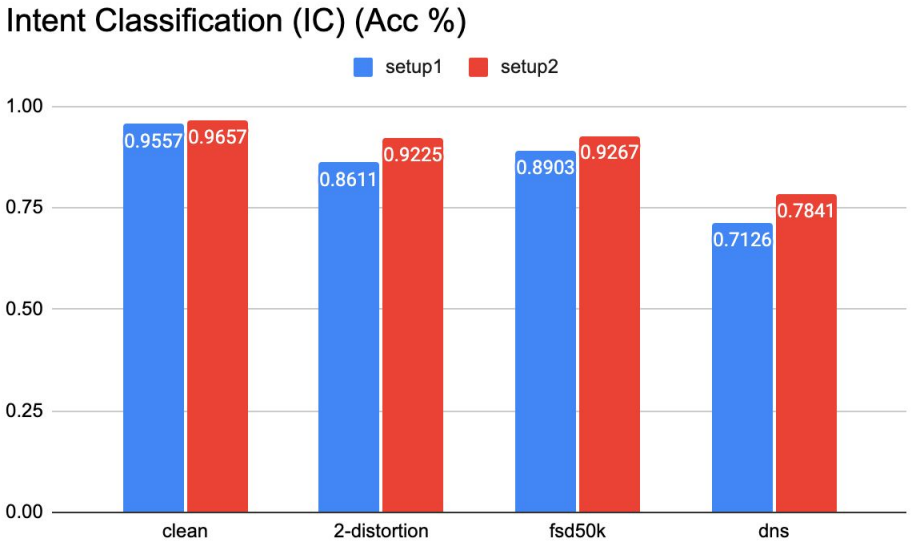
Cross-Distortion Mapping (setup 1)

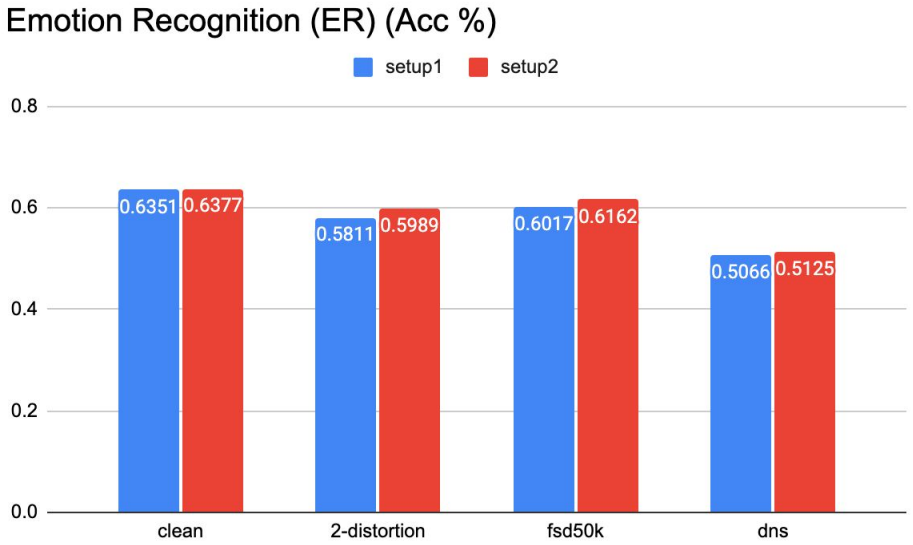
- Setup 1 (student: distorted, teacher:clean) improves robustness for KS, IC, ER, ASR.



Cross-Distortion Mapping (setup 1 vs setup 2)

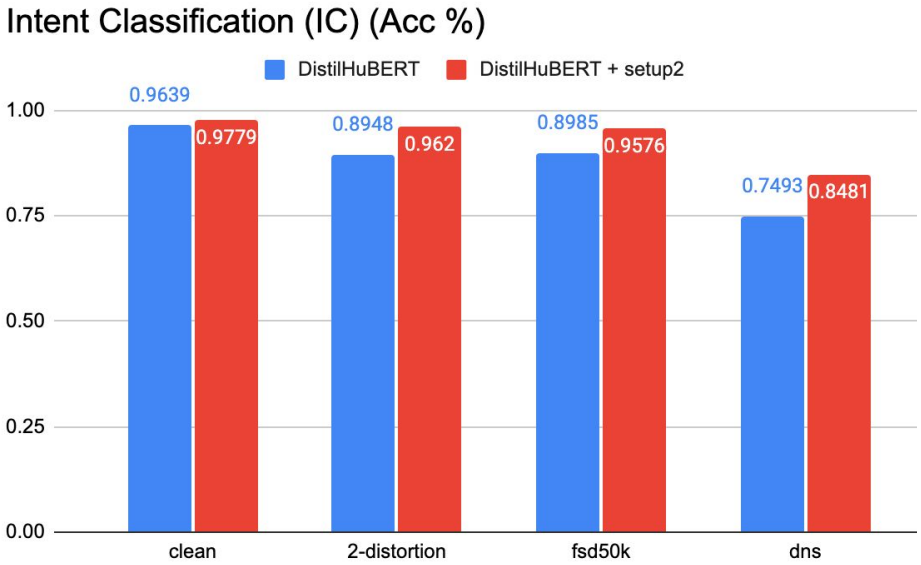
- Setup 2 (student: distorted, teacher:distorted) is better than setup 1 for KS, IC, ER.

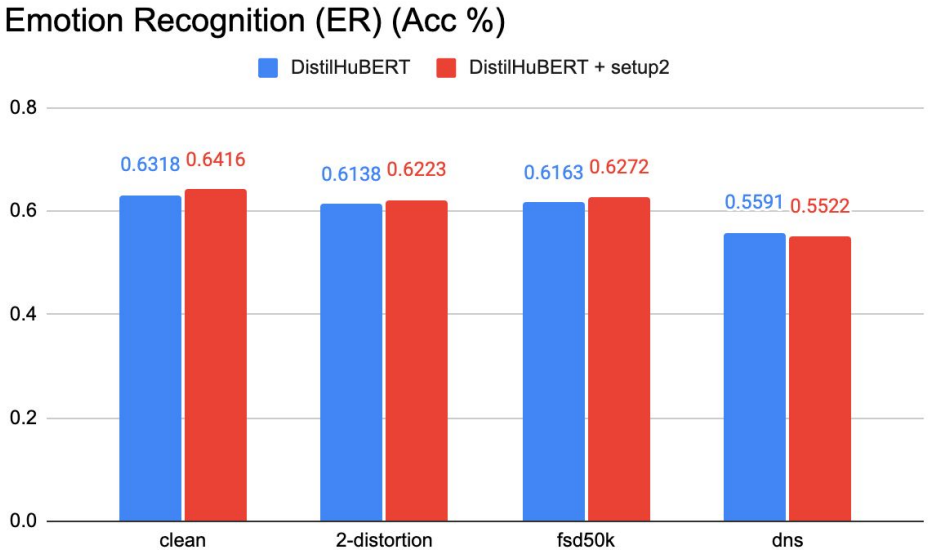




Distorted downstream training

- CDM still has better performance when distortions are added during training downstream tasks.

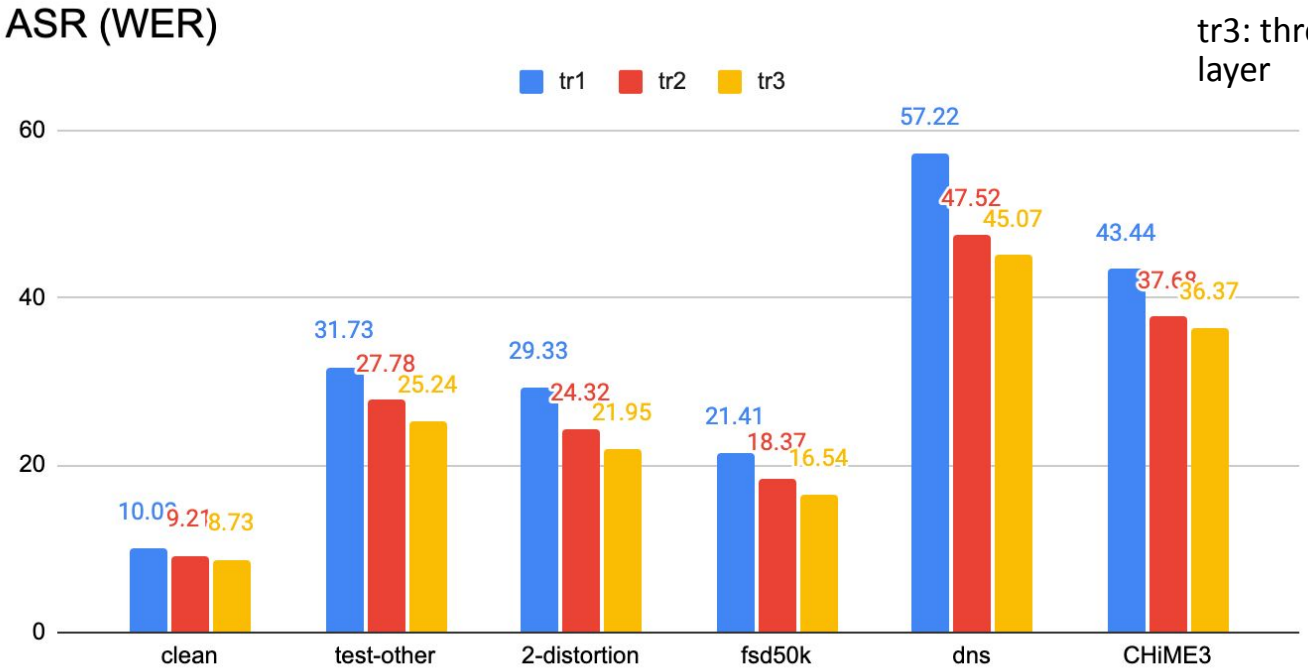




Different model sizes

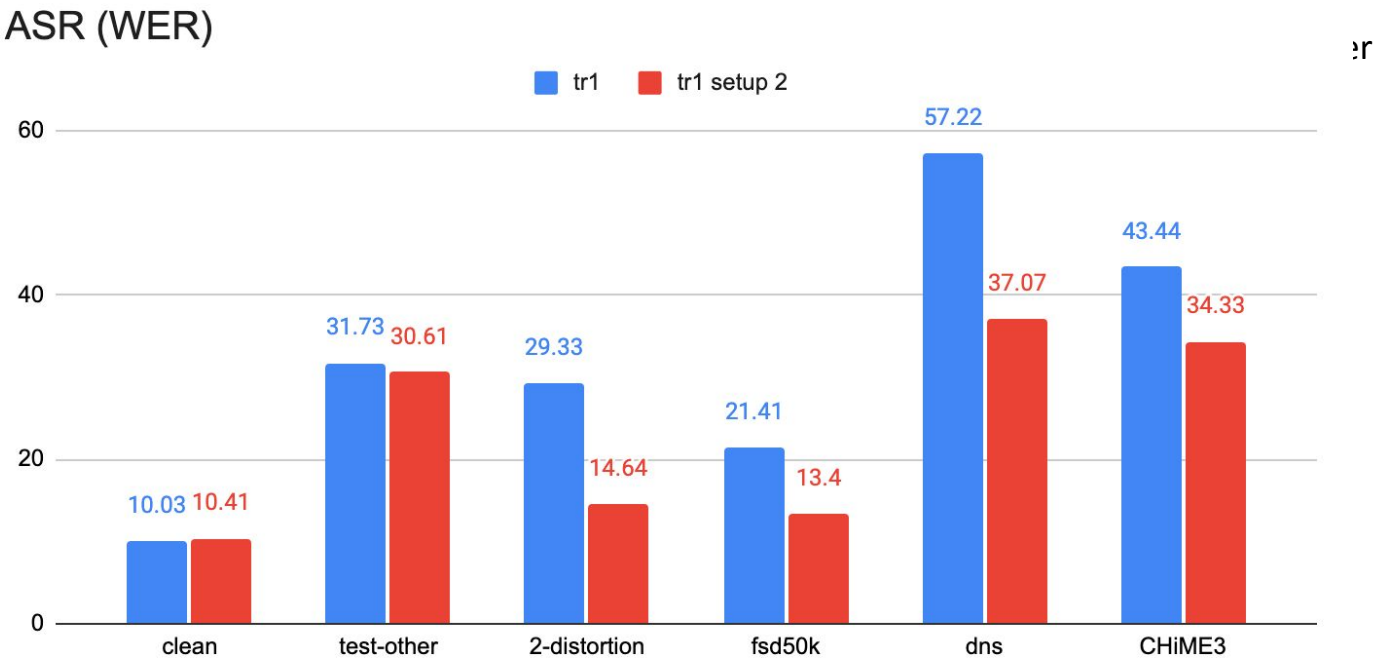
- Robustness of models with different sizes.

tr1: one transformer encoder layer
tr2: two transformer encoder layer (original DistilHuBERT)
tr3: three transformer encoder layer



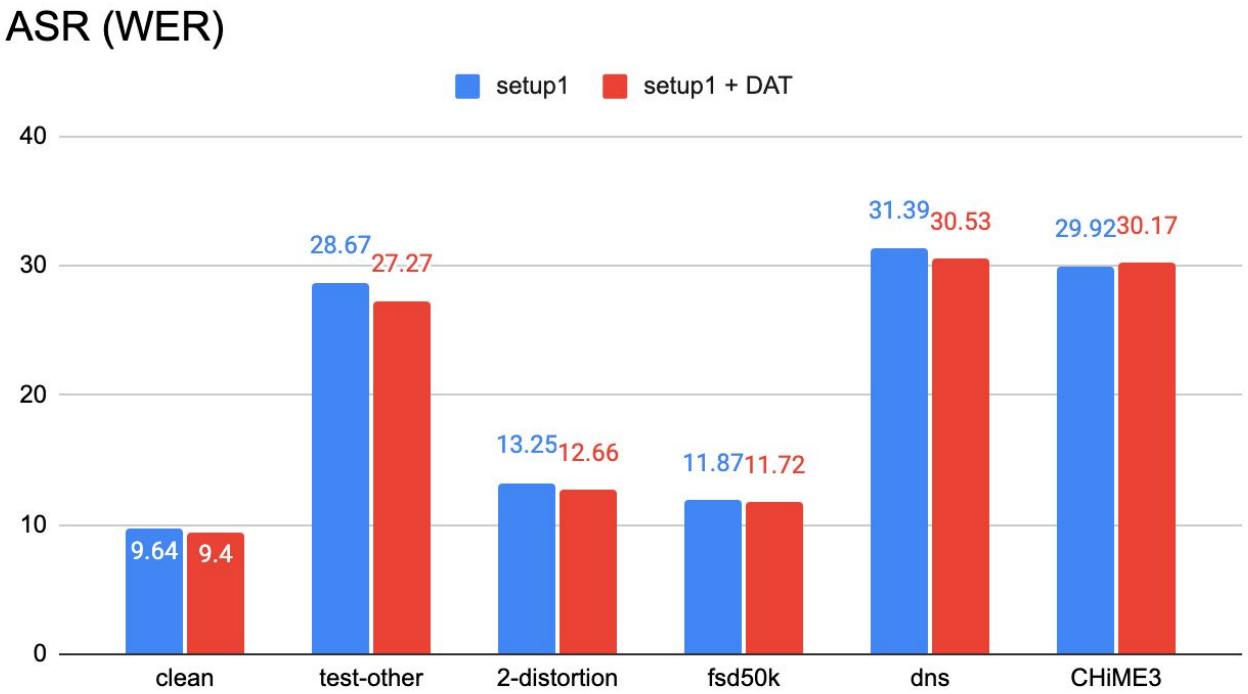
Different model sizes (CDM setup2)

- CDM setup2 has consistent results for smaller (or bigger) student models.



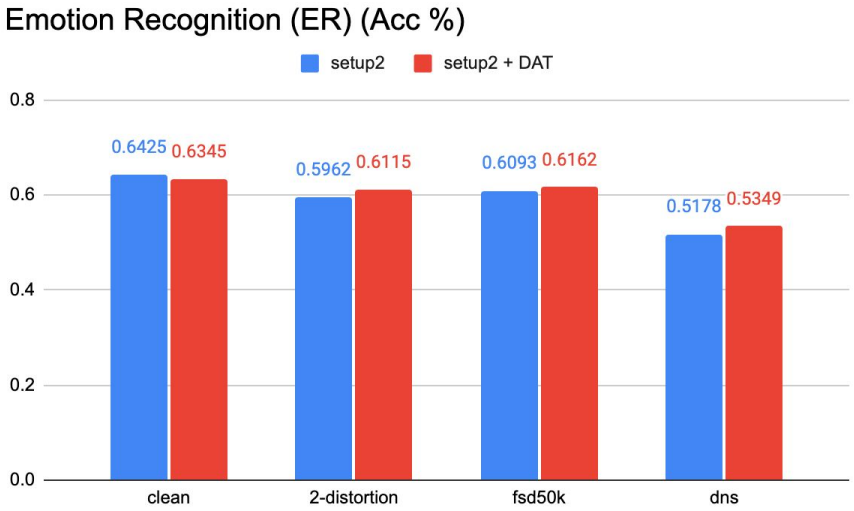
Domain Adversarial Training (DAT)

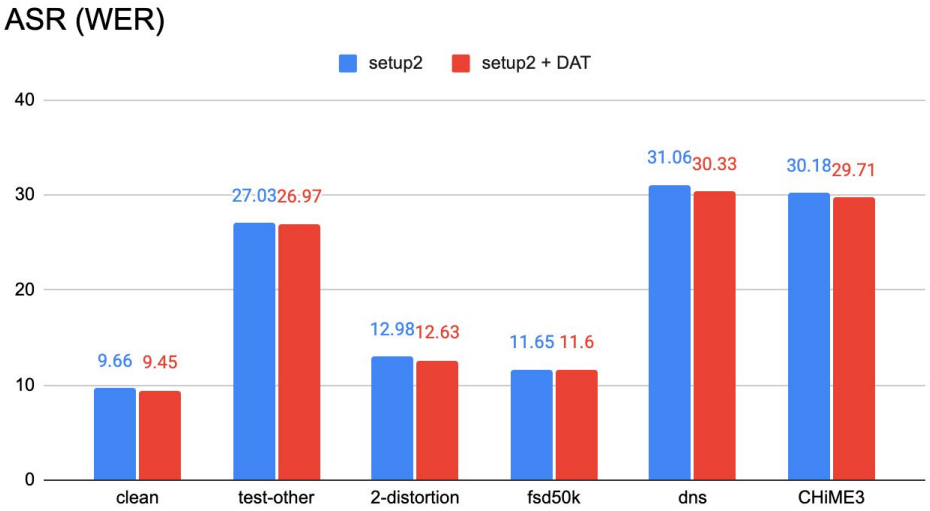
- DAT improves KS, IC, ER, ASR under setup1 setting.



Domain Adversarial Training (DAT)

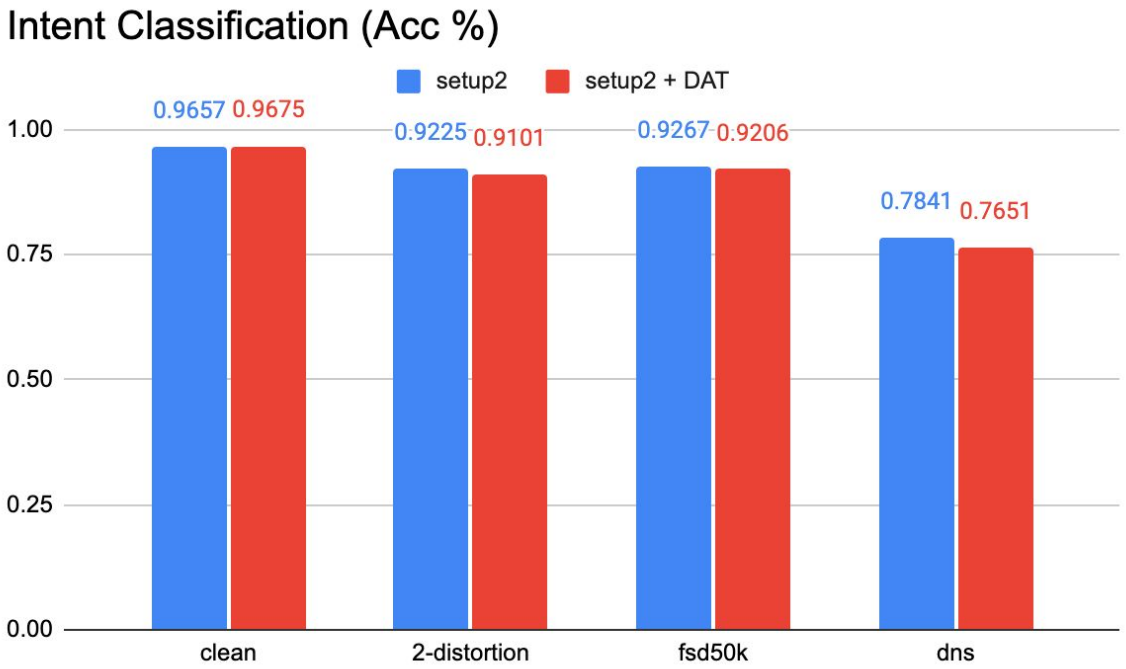
- DAT improves ER and ASR under setup2 setting.





Domain Adversarial Training (DAT)

- DAT does not improve IC under setup2 setting.



Questions to answer

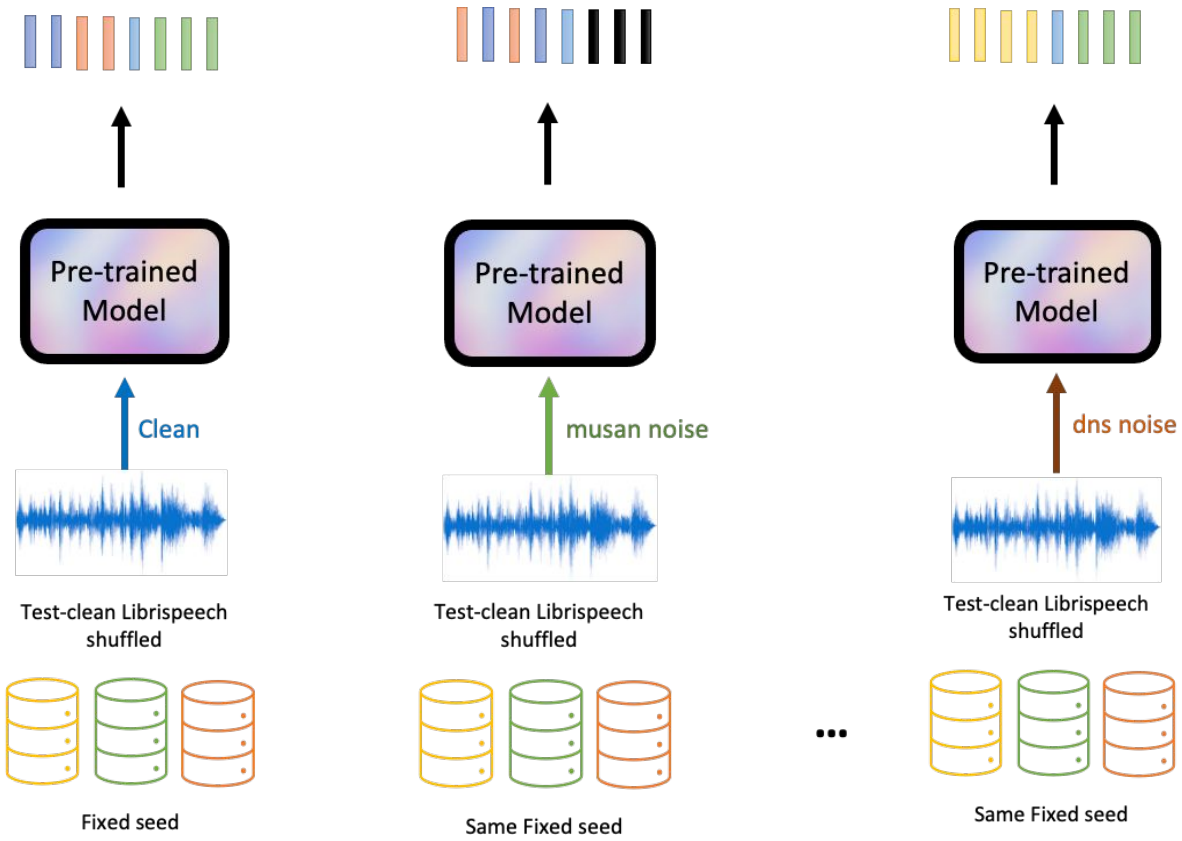
Does Cross-Distortion Mapping (CDM) enhance robustness ? YES

Does CDM have consistent results for models with different sizes ? YES

Does Domain Adversarial Training enhance robustness ? Sometimes

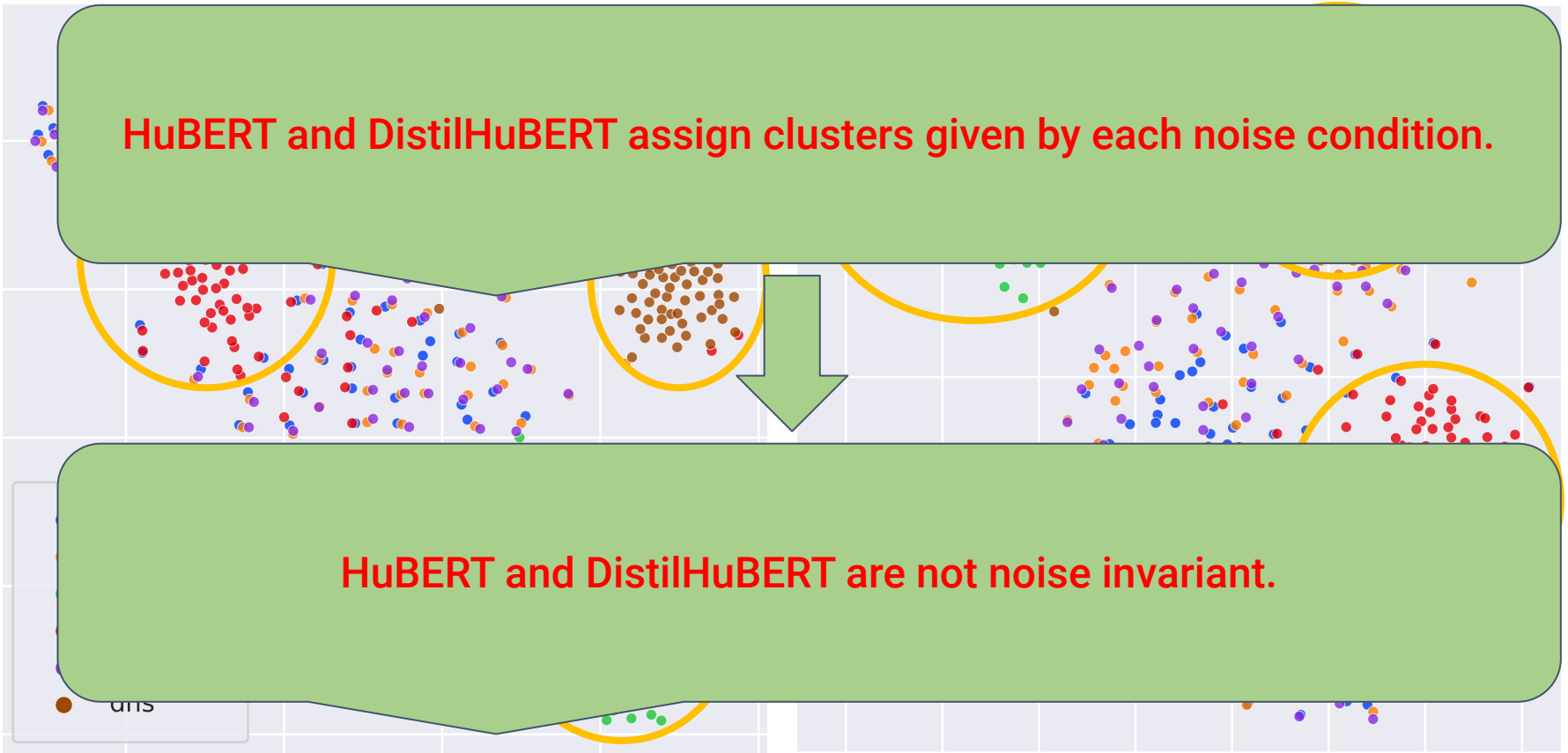
Visualization Experiments

Fabian Ritter (NUS)


Pre-trained Model
Clean
Test-clean Librispeech shuffled
Fixed seed
Pre-trained Model
musan noise
Test-clean Librispeech shuffled
Same Fixed seed
Pre-trained Model
dns noise
Test-clean Librispeech shuffled
Same Fixed seed

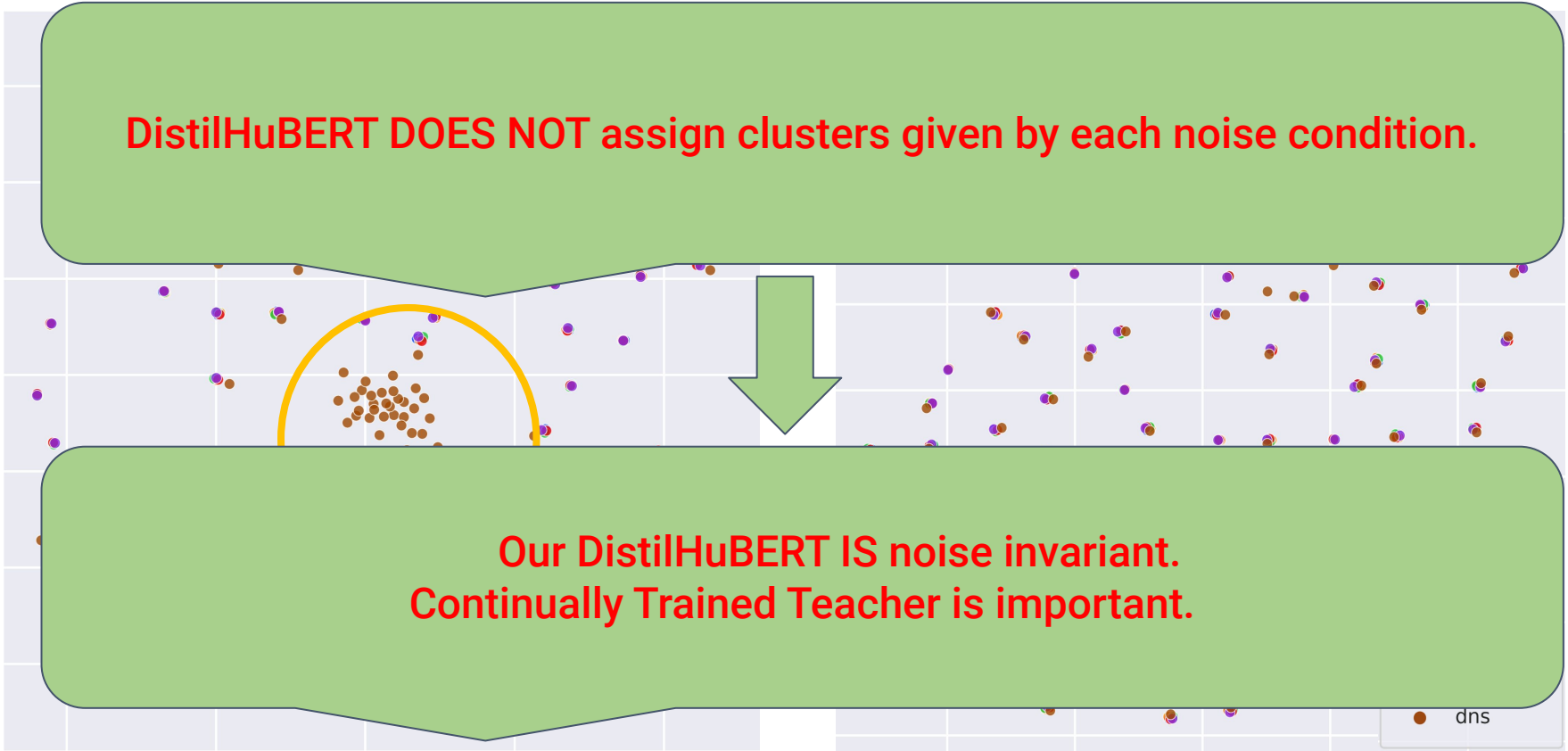
HuBERT

DistilHuBERT



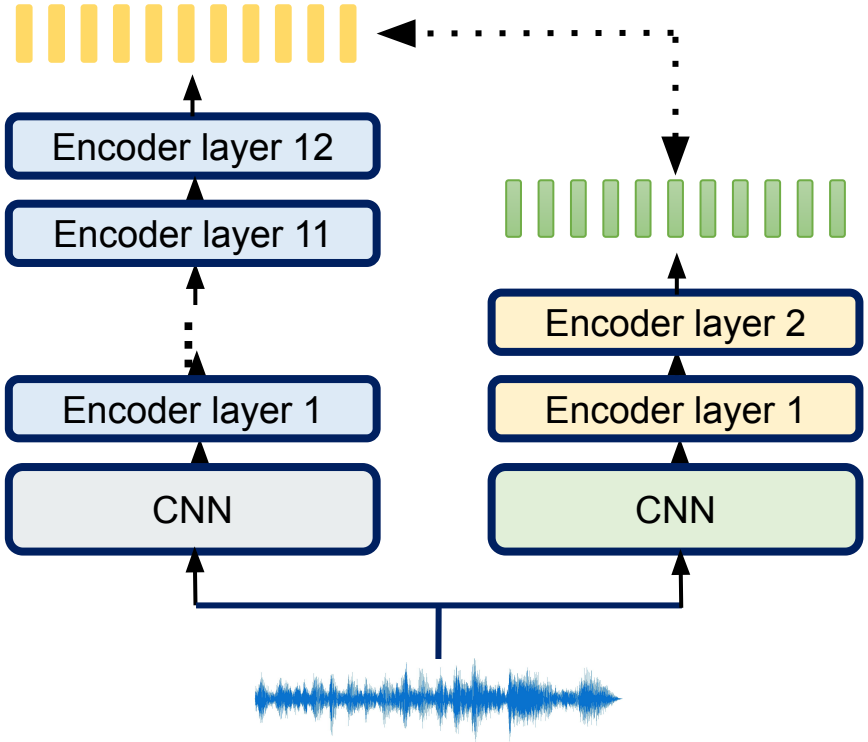
DistilHuBert CDM Setup 2

Same but Cont Trained Teacher

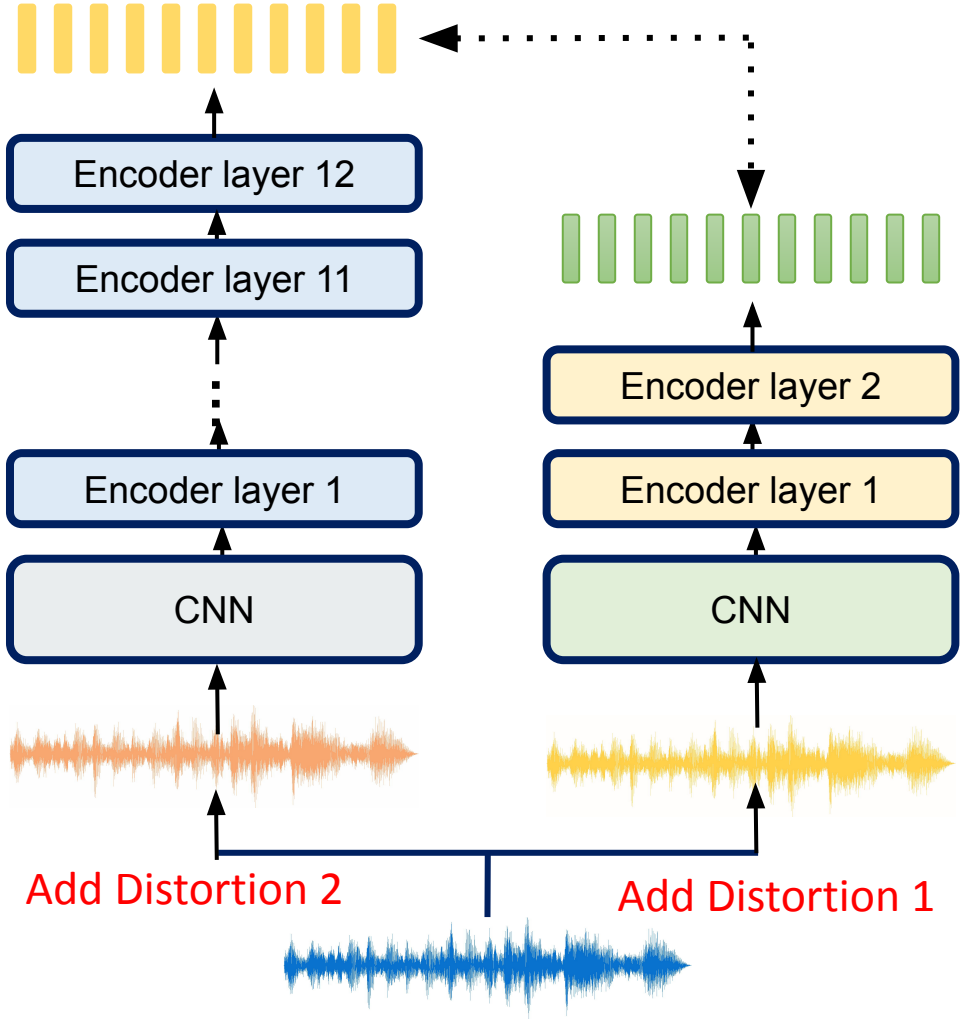


Takeaway Message & Future Work

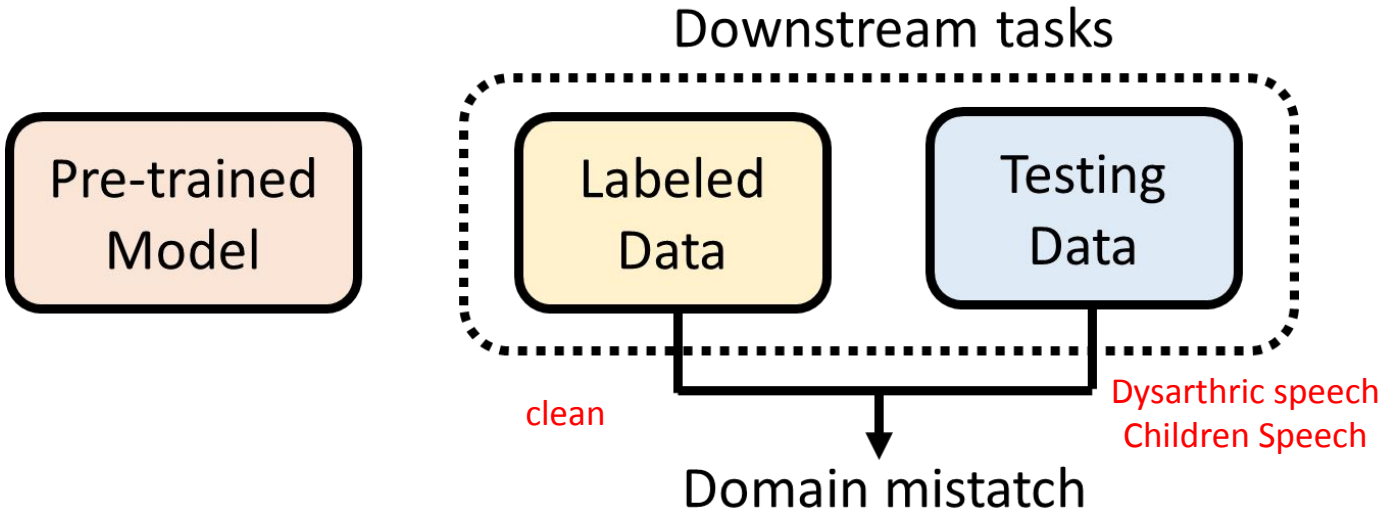
- We started JSALT with a big gap in performance for distilled models when subdued to distortions.



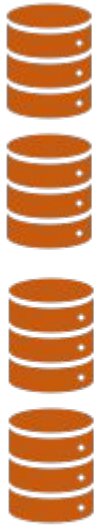
- We have reduced this gap by using Cross Domain Mapping and Domain Adversarial Training for pre-training



Future Work


Pre-trained Model
Downstream tasks
Labeled Data
Testing Data
clean
Dysarthric speech
Children Speech
Domain mistatch

Challenges

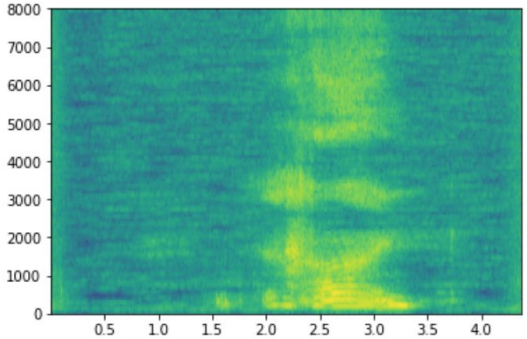


Adult Typical Speech

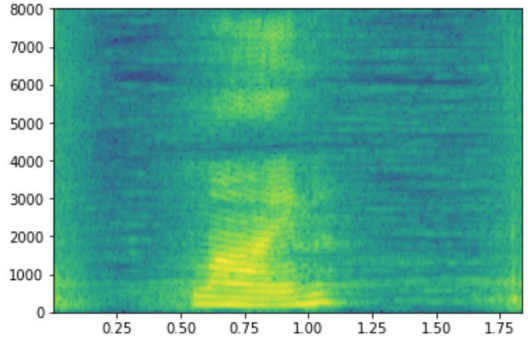
Data Scarcity

Dysarthric Speech

Spectrograms for the word “LINE”



Dysarthric Speech



Adult Typical Speech

Strong Mismatch !

How does a Self-Supervised Model generalizes without any training?

	ASR (WER) ↓ w/o LM		
Testing Data	test-clean-100	test-control	test-dysarthric
WavLM+ base	5.59	47.12	88.7
HuBERT	6.42	55.48	90.6
Robust Teacher (T1')	6.75	54.07	90.86

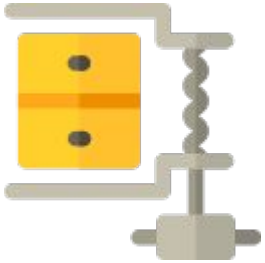
Systems finetuned on healthy speakers from UASPEECH helps but still behind

	ASR (WER) ↓	
Testing Data	test-control	test-dysarthric
GMM-HMM (fMLLR)	13.48 (w/LM)	67.16 (w/LM)
HuBERT	46	79.92
HuBERT with LM	41	78

All system were trained using the healthy training portion of UASPEECH dataset.

Test is healthy portion of test set and dysarthric test set

Outline of Part 1: Better Pre-trained Model



Compression

9:35 - 9:55

Robust

9:55 - 10:10

Visual-enhanced

10:10 - 10:25

Integration

10:25 - 10:30

10 mins Q&A
+ 10 mins break

Visually-Enhanced SSL Models

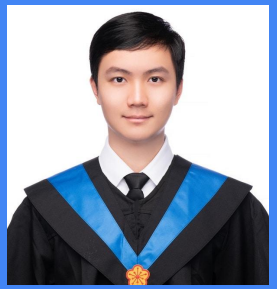
PIs:

David Harwath
(UT Austin)

Hung-yi Lee
(NTU)

Students:

Layne Berry
(UT Austin)



Heng-Jui Chang
(MIT)

Ian Shih
(NTU)

Jeff Wang
(NTU)

Visually-grounded speech (VGS)

two lambs standing in the grass

vs.

Why train models with visually-grounded speech?

- Spoken image descriptions are easy and cheap to collect ($0.18-$0.24 per minute vs. $1.50-$3.00/minute for text-transcribed speech)
- Provide a semantic training signal for languages that do not have a standard orthographic representation (Swiss German, Egyptian Arabic, many more…)
- We know that humans acquire language **without text** and sensory grounding (e.g. visual) is an important part of this process

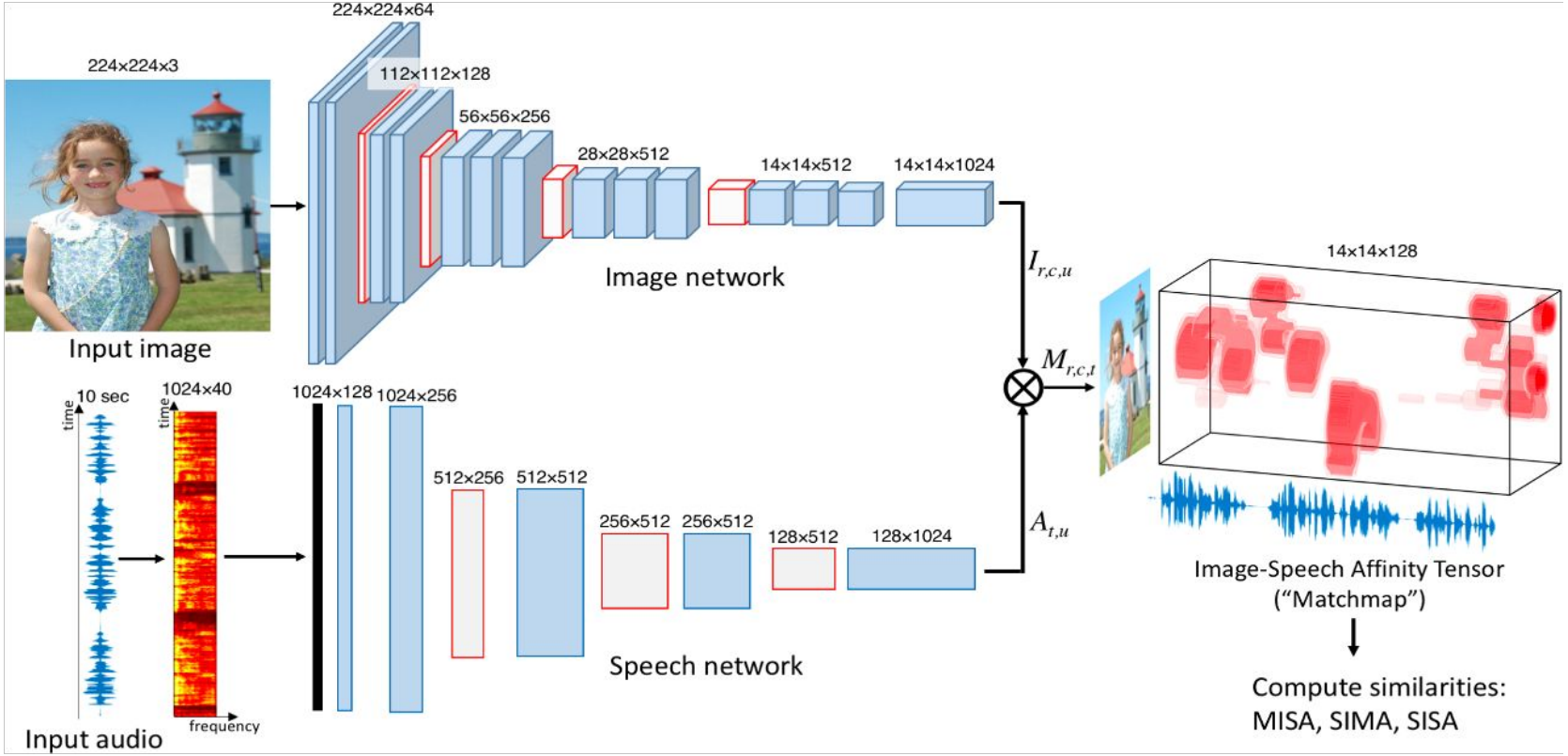
Image Caption Datasets

- Flickr 8k
  - Train: 6K Images 30K Captions (text/audio)
  - Dev: 1K Images 5K Captions (text/audio)
  - Test: 1K Images 5K Captions (text/audio)
- SpokenCOCO
  - Train: 113K Images 565k Captions (text/audio)
  - Dev: 5k images 25k Captions (text/audio)
  - Test: 5k images 25k Captions (text/audio)

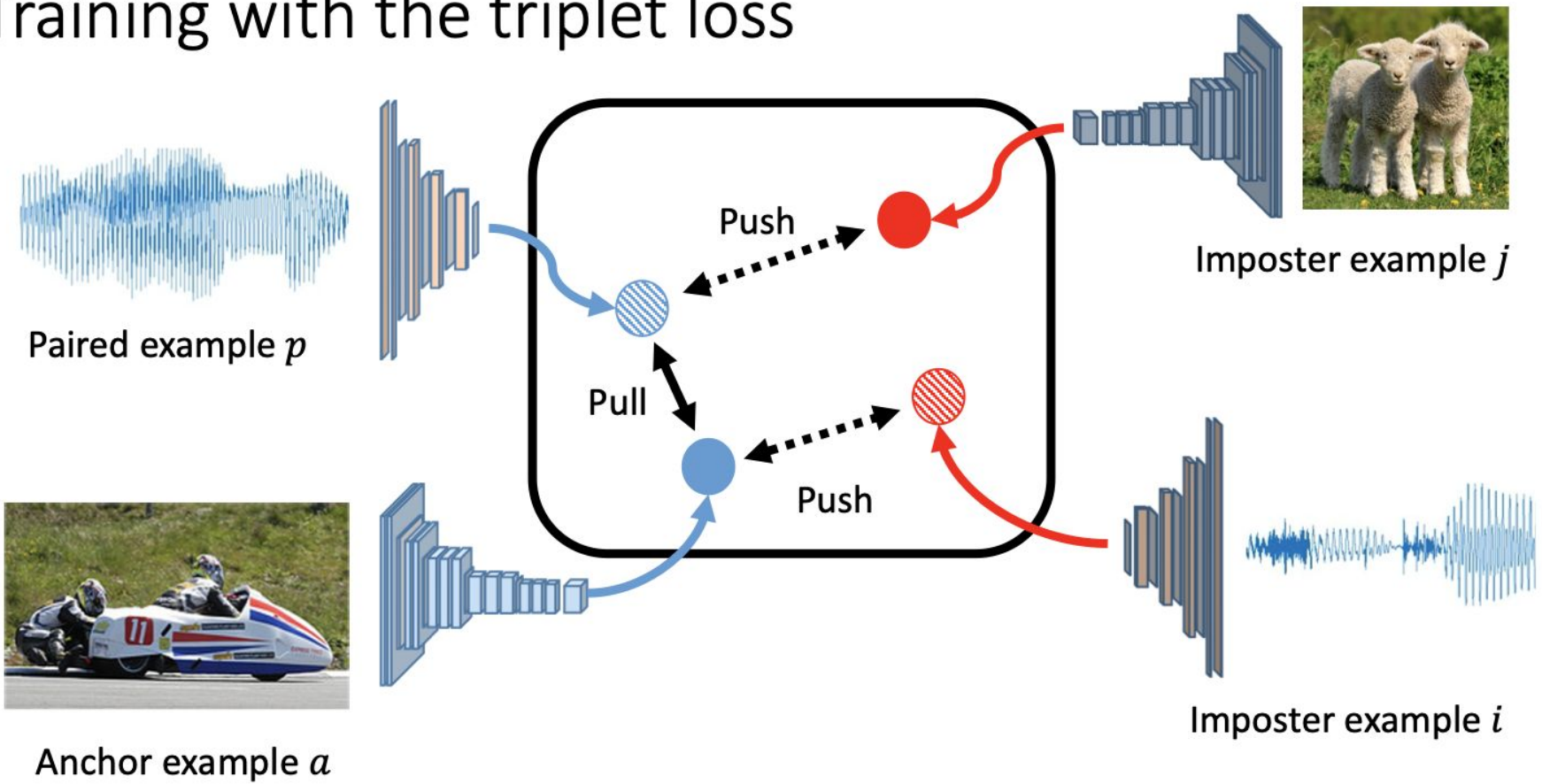
Example:

A blonde horse and a blonde girl in a black sweatshirt are staring at a fire in a barrel .
A girl and her horse stand by a fire .
A girl holding a horse 's lead behind a fire .
A man , and girl and two horses are near a contained fire .
Two people and two horses watching a fire .

Prior work on VGS (Harwath et al. 2018)



Training with the triplet loss

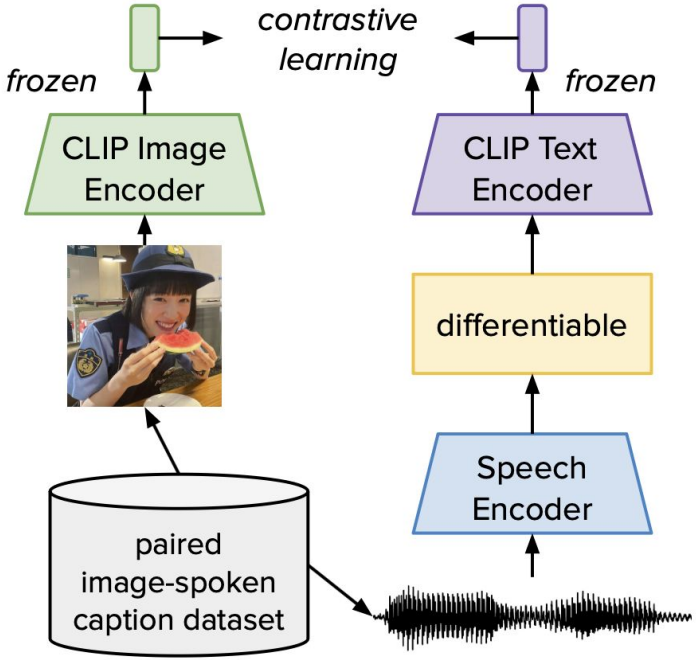


Questions addressed in this workshop

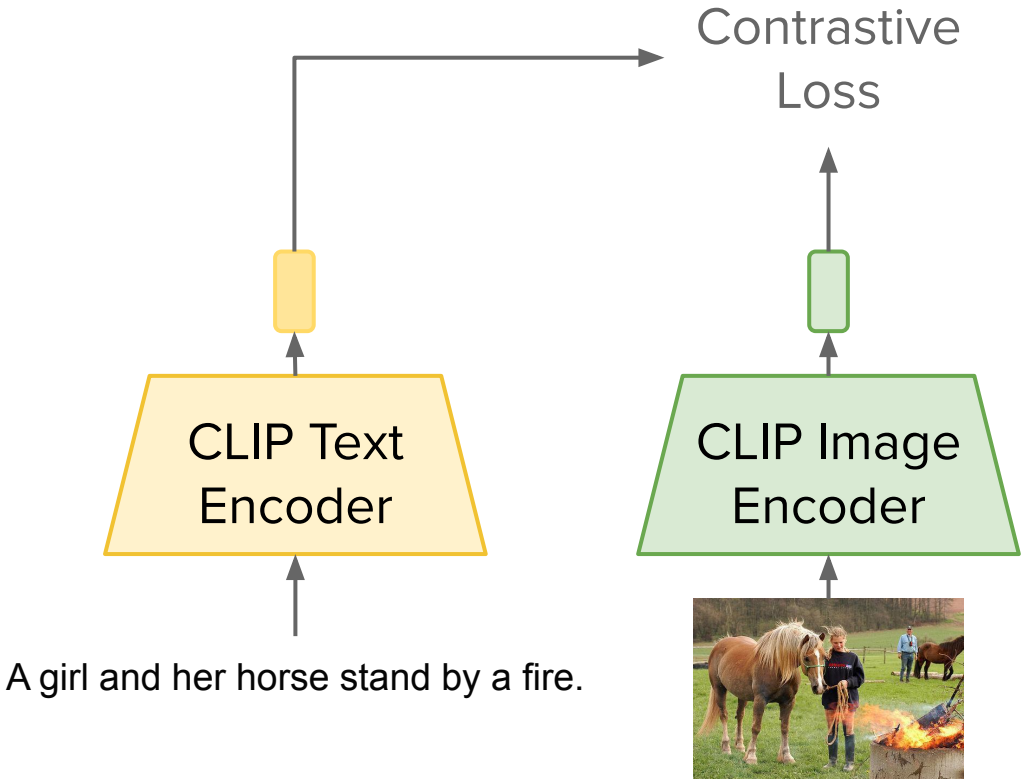
1) Can we leverage **non-parallel text and speech** via pre-trained vision+text models (**e.g. CLIP**) to improve visually-grounded, self-supervised speech models?

2) Can these models provide improved performance on different downstream speech tasks with limited labels? **Evaluate on SUPERB and image retrieval for this.** (Ian Shih will discuss this in a moment)

3) Can these models also be leveraged to **improve unsupervised ASR?** (Layne Berry will present this later today)



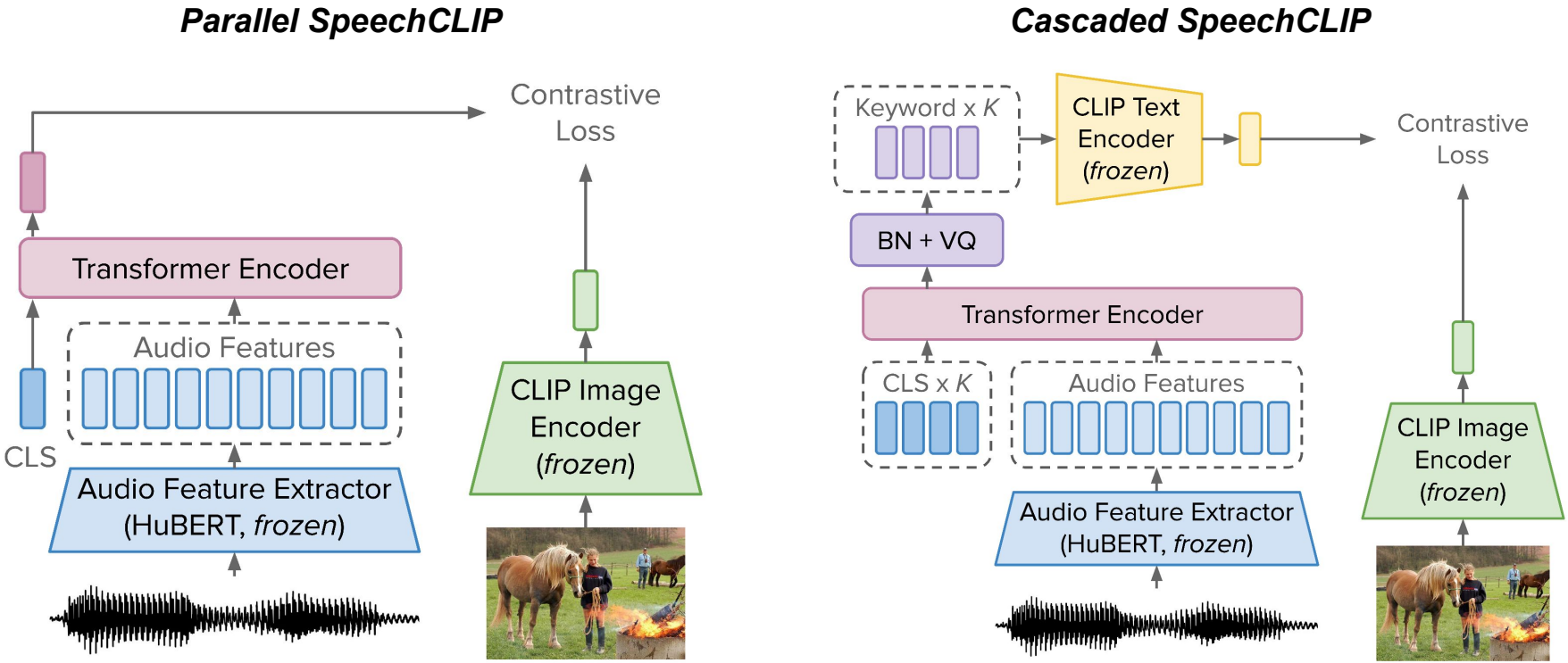
Contrastive **L**anguage-**I**mage **P**re-training **CLIP** (Radford et al.)



Some info about CLIP

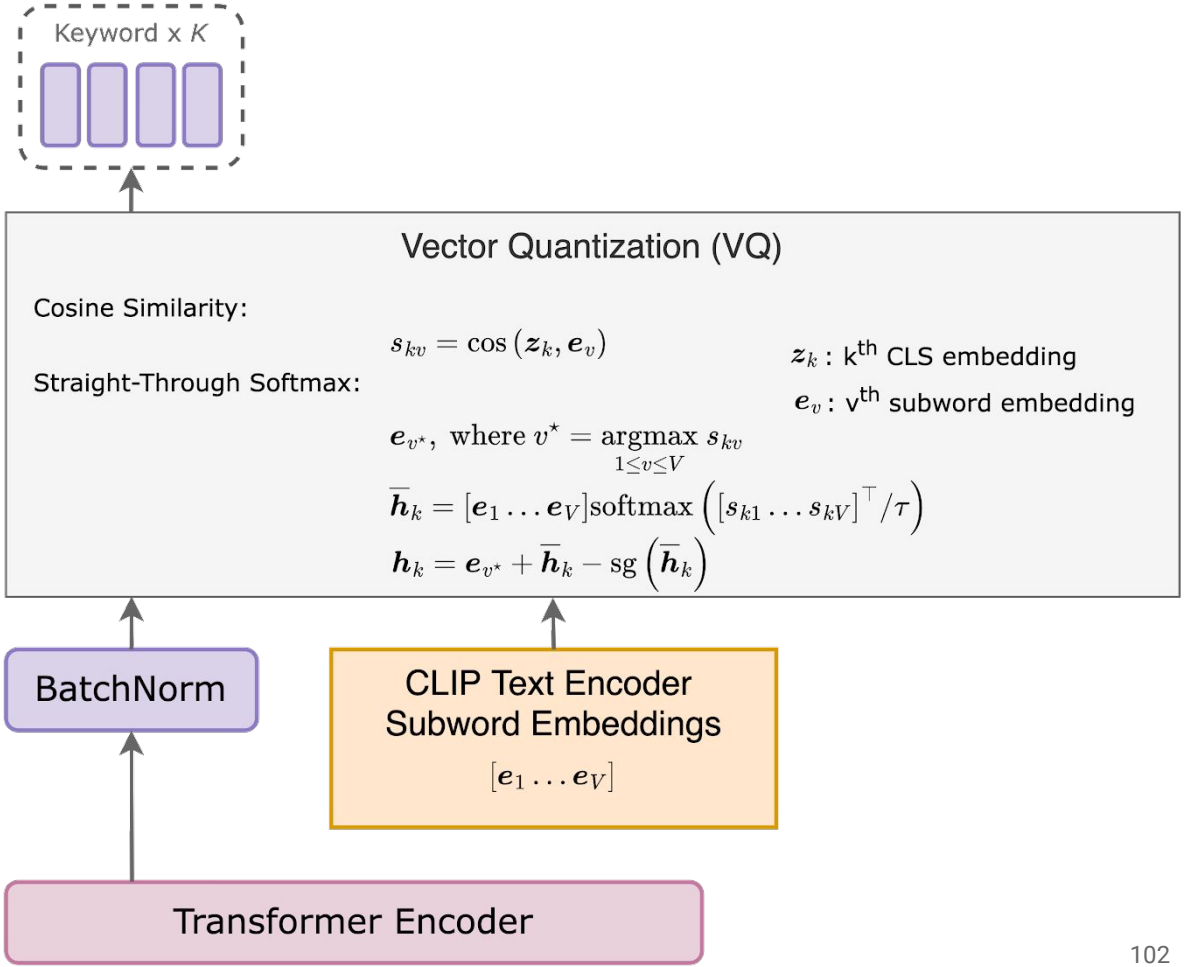
1. Made by OpenAI
   Trained on **400M** Image-Text pairs
2. Training last for **12** days on **256** v100 GPUs
3. Trained by contrastive loss to learn a **shared embedding space for image and text**

SpeechCLIP Architecture - Details



SpeechCLIP - VQ

- Select subword embedding with the **highest cosine similarity**
- The selected subword is called ***keyword***



Evaluation

1. **SUPERB** (KS, PR)
   a. **K**eyword **S**potting: detects preregistered keywords from speech
   b. **P**honeme **R**ecognition
2. **Speech-Image retrieval**
   a. Goal: given a library of ***N images*** and **a *spoken description*** of one of those images as input, search over the library and return the target image
   b. Results reported with **Recall @** top ***1***, ***5***, ***10*** returned results

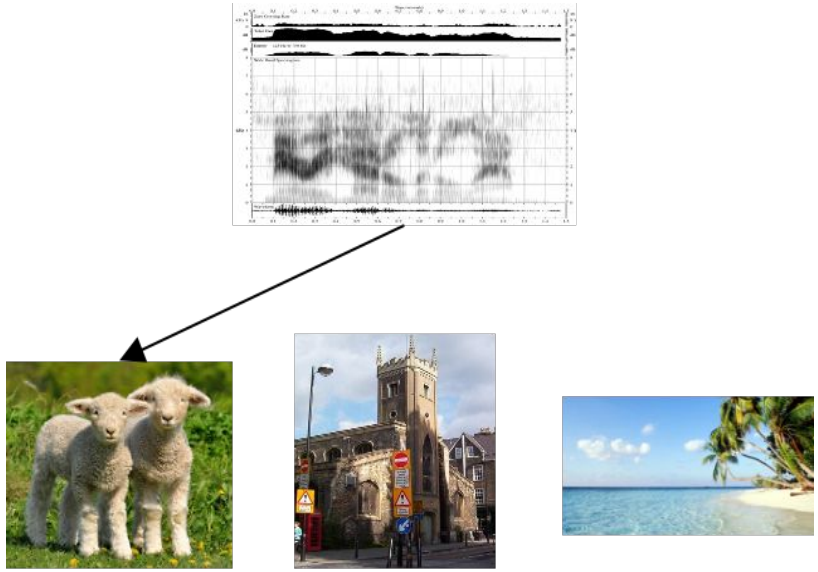


Image-Speech Retrieval

Model	Audio Encoder	CLIP Image Encoder	Trainable Params (M)	Total Params (M)
Base	HuBERT Base (95 M)	ViT-B/32 (250 M)	2.8 – 7.5	252 – 257
Large	HuBERT Large (316 M)	ViT-L/14 (422 M)	6.1 – 13.4	765 – 772

- **Outperform** previous VGS models on Flickr8k and SpokenCOCO
- Only **little fine-tuning** required
  - **2** days **2** GPUs for largest model
  - Only **10M+** trainable params
- Results show the benefits of **leveraging CLIP in VGS model**

Method	Speech → Image R@1	Speech → Image R@5	Speech → Image R@10	Image → Speech R@1	Image → Speech R@5	Image → Speech R@10
	Flickr8k					
FaST-VGS$_{CO}$ [36]	26.6	56.4	68.8	36.2	66.1	76.5
FaST-VGS$_{CTF}$ [36]	29.3	58.6	71.0	37.9	68.5	79.9
MILAN [35]	33.2	62.7	73.9	49.6	79.2	87.5
Parallel	26.7	57.1	70.0	41.3	73.9	84.2
Cascaded	8.2	25.7	37.2	14.1	34.5	49.2
Parallel Large	**39.1**	**72.0**	**83.0**	**54.5**	**84.5**	**93.2**
Cascaded Large	14.7	41.2	55.1	21.8	52.0	67.7
	SpokenCOCO					
ResDAVEnet [25]	17.3	41.9	55.0	22.0	50.6	65.2
FaST-VGS$_{CO}$ [36]	31.8	62.5	75.0	42.5	73.7	84.9
FaST-VGS$_{CTF}$ [36]	**35.9**	66.3	77.9	48.8	78.2	87.0
Parallel Large	35.8	**66.5**	**78.0**	**50.6**	**80.9**	**89.1**
Cascaded Large	6.4	20.7	31.0	9.6	27.7	39.7

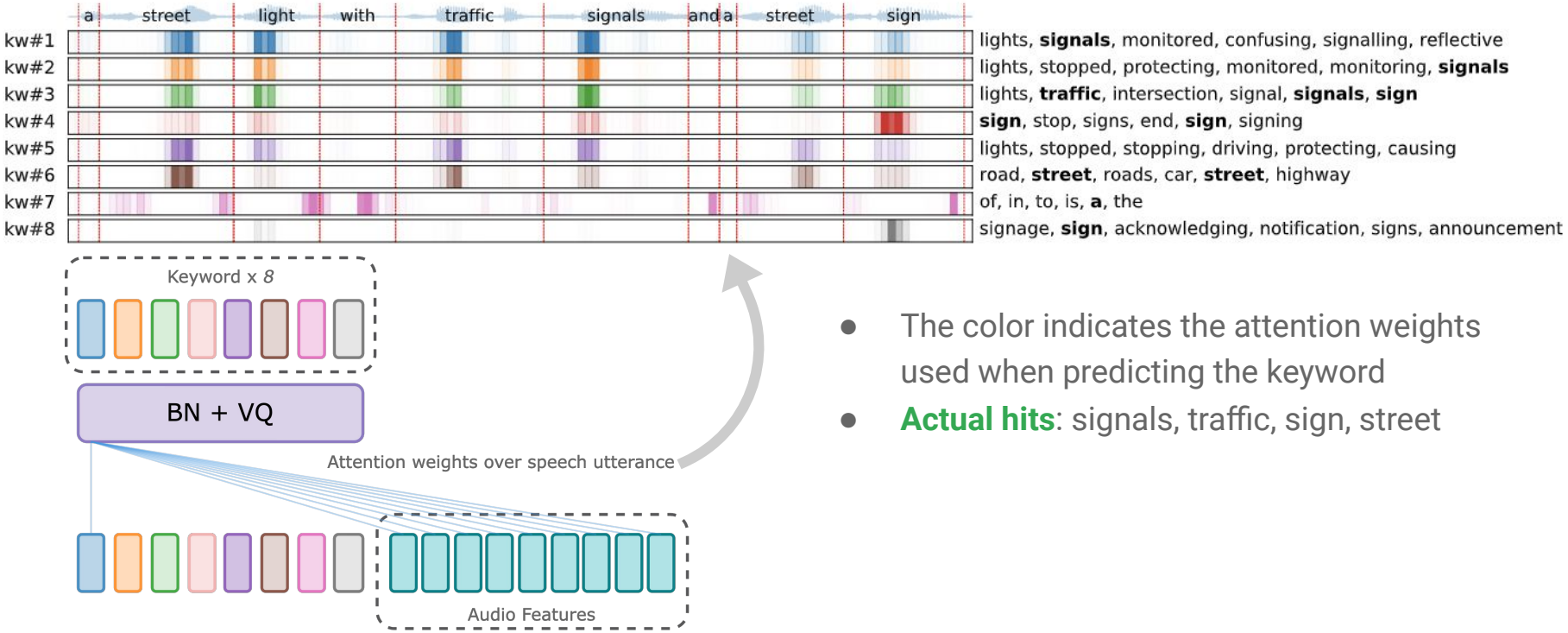
ZeroShot Speech-Text

	Speech → Text			Text → Speech		
Method	R@1	R@5	R@10	R@1	R@5	R@10
	Flickr8k					
Random	0.10	0.50	0.99	0.10	0.50	0.99
Parallel Large	19.56	44.06	58.46	22.50	44.14	54.54
Parallel Large (Sup.)	97.06	99.24	99.46	97.88	99.76	99.90
	SpokenCOCO					
Random	0.02	0.10	0.20	0.02	0.10	0.20
Parallel Large	60.32	81.81	88.18	65.45	85.82	91.27
Parallel Large (Sup.)	95.02	99.46	99.78	95.35	99.68	99.93

- **Way better** than random
- **Close to supervised method** on Large Dataset
- Could be applied to **retrieve noisy transcription of different languages**

Keyword Discovery

Attention Map for keyword retrieval on SpokenCOCO



- The color indicates the attention weights used when predicting the keyword
- **Actual hits**: signals, traffic, sign, street

Keyword Discovery (Cont'd)

Top 10 successfully retrieved subwords for each keyword on SpokenCOCO test

kw1	kw2	kw3	kw4	kw5	kw6	kw7	kw8
a	cat	bathroom	a	a	street	in	train
pizza	a	skateboard	of	tennis	bathroom	of	sign
the	room	room	in	with	kitchen	to	cake
giraffe	sheep	horse	man	eating	train	from	clock
bathroom	frisbee	elephant	woman	and	beach	for	is
skateboard	skis	motorcycle	dog	playing	bed	a	bus
living	bird	kitchen	train	the	bus	on	truck
gira	skateboard	clock	with	flying	grass	at	car
sheep	surf	tower	is	sitting	road	the	of
an	kite	bear	to	walking	room	–	signs

- **High frequent subwords**: a, of, in ...
- **Concrete objects in Images**: skateboard, street...

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

- Large model on SpokenCOCO has better hit rate

SUPERB

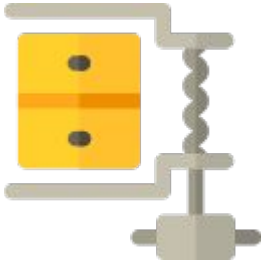
Setting	# Trm Layers	KS (↑)	PR (↓)
Parallel	1	96.46	4.95
Cascaded		96.49	4.92
Parallel	2	96.82	4.97
Cascaded		96.66	**4.90**
Parallel	4	96.66	4.94
Cascaded		**97.01**	4.92
HuBERT Base	-	96.30	4.95

- **Outperform** HuBERT Base on **KS**
- **Similar** performances on **PR**

Conclusions

- Progress during JSALT
  - Achieve SOTA on Speech-Image Retrieval
  - Achieve keywords discovery from speech utterance
  - Achieve Zeroshot Speech-Text retrieval
  - Submitted our recent results to **SLT '22**
- Takeaways
  - With the help of **large scaled pretrained Visual Language Models**, we can utilize **unparallel speech text data** with only **little amount of training required**
  - Visually Grounded method can endow SSL models with the ability to **extract speech content**
- Future work
  - Combine Parallel and Cascaded objective into a Hybrid model
  - Unsupervised ASR & Speech Translation

Outline of Part 1: Better Pre-trained Model



Compression

9:35 - 9:55

Robust

9:55 - 10:10

Visual-enhanced

10:10 - 10:25

Integration

10:25 - 10:30

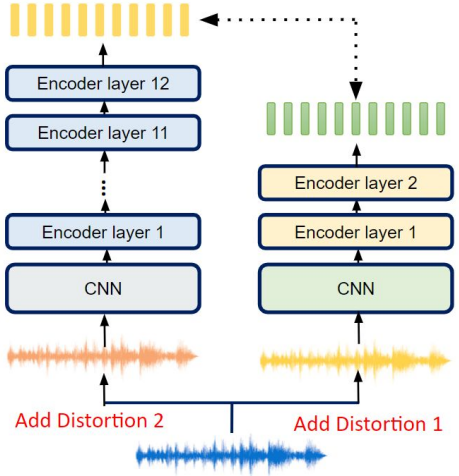
10 mins Q&A
+ 10 mins break

Chimera: Integrating all technology

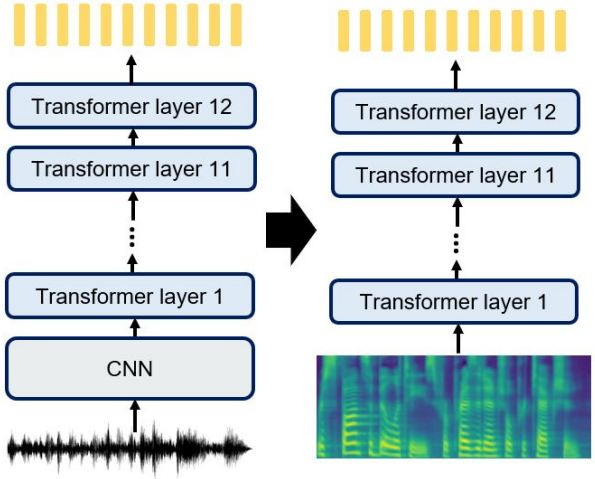
Hung-yi Lee
(NTU)

Kai-Wei Chang
(NTU)

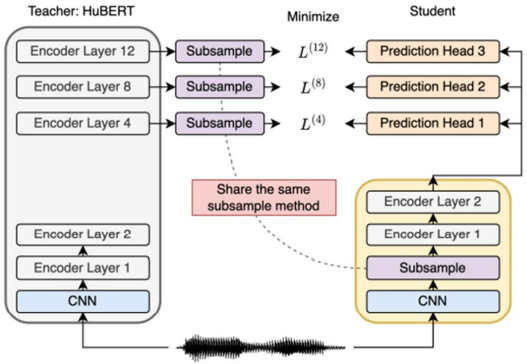
(A) DistilHuBERT
(+ Cross-Distortion Mapping)



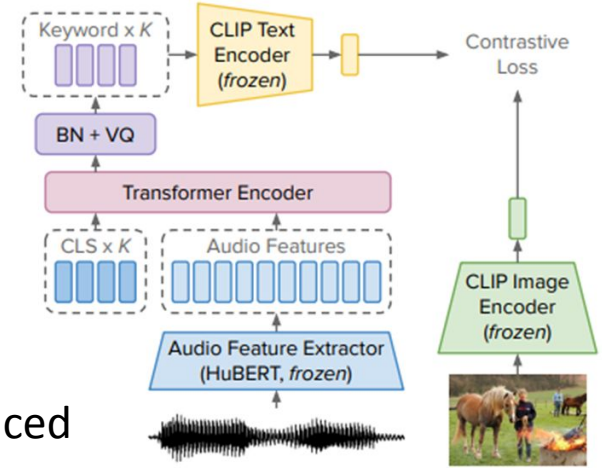
(B) MelHuBERT



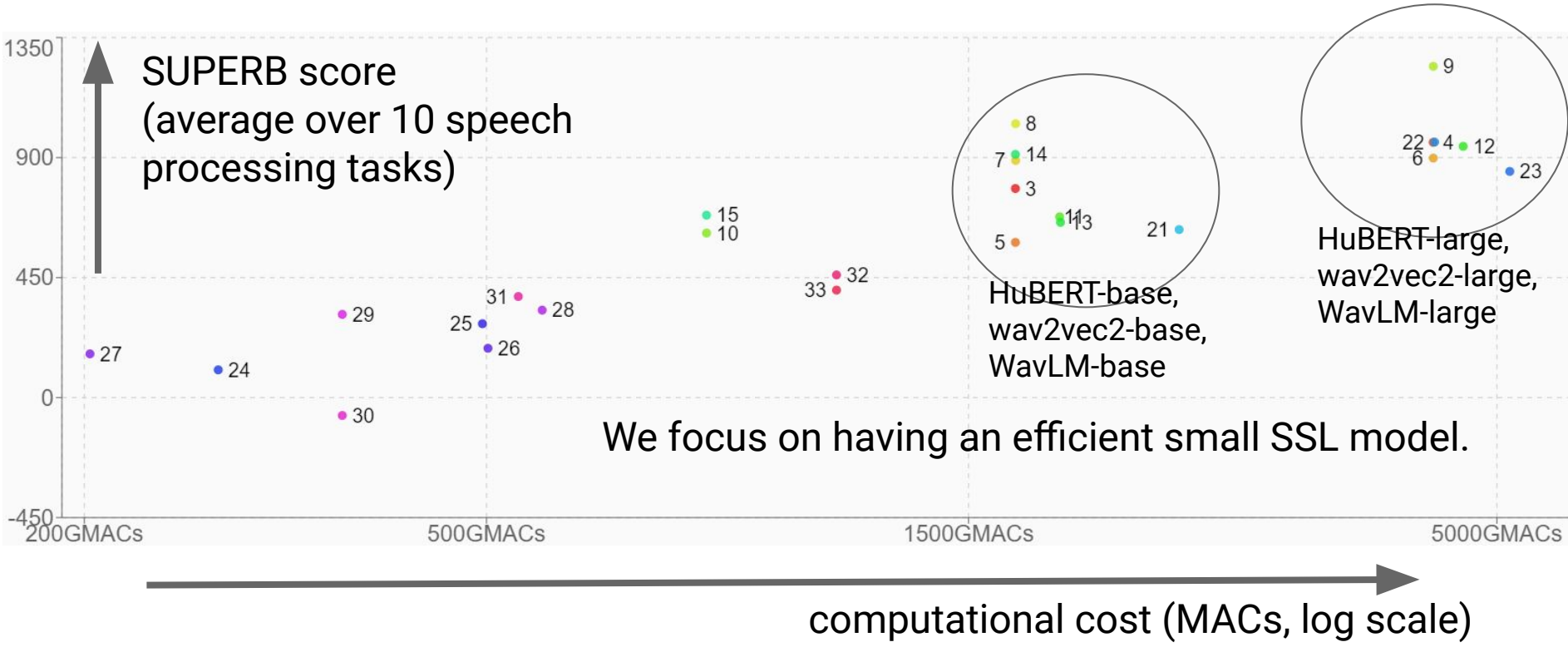
(C) Sequence Reduction



(D) Visual-enhanced

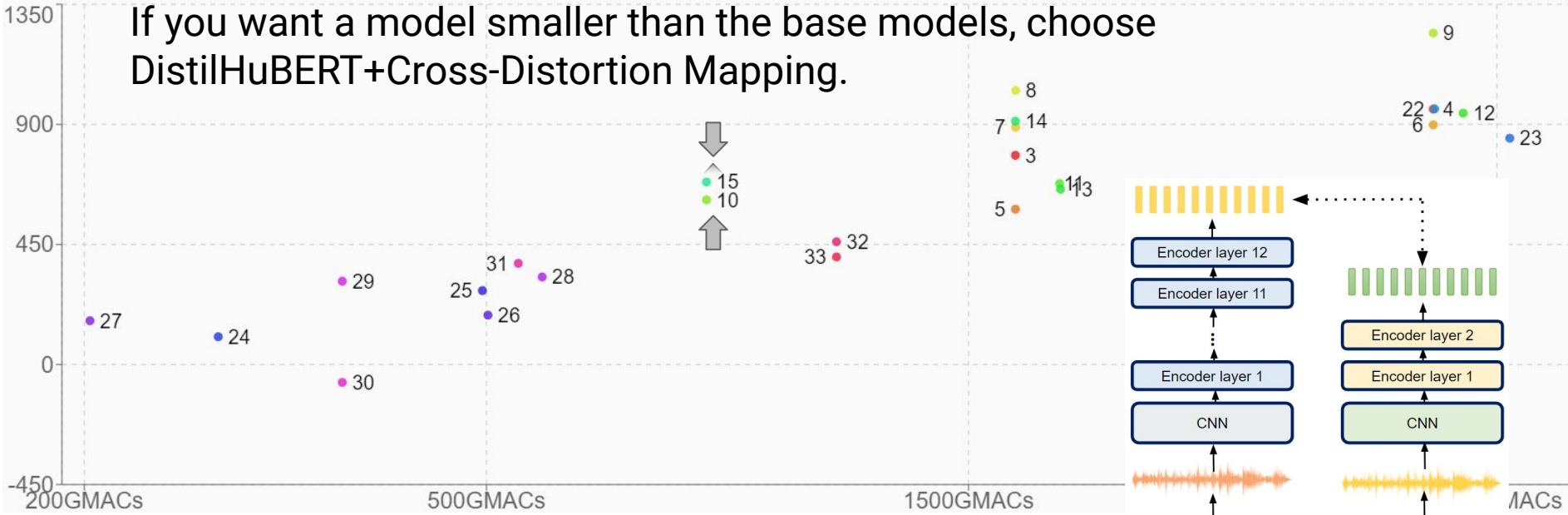


SUPERB Leaderboard - Hidden-set Track



SUPERB Leaderboard - Hidden-set Track

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



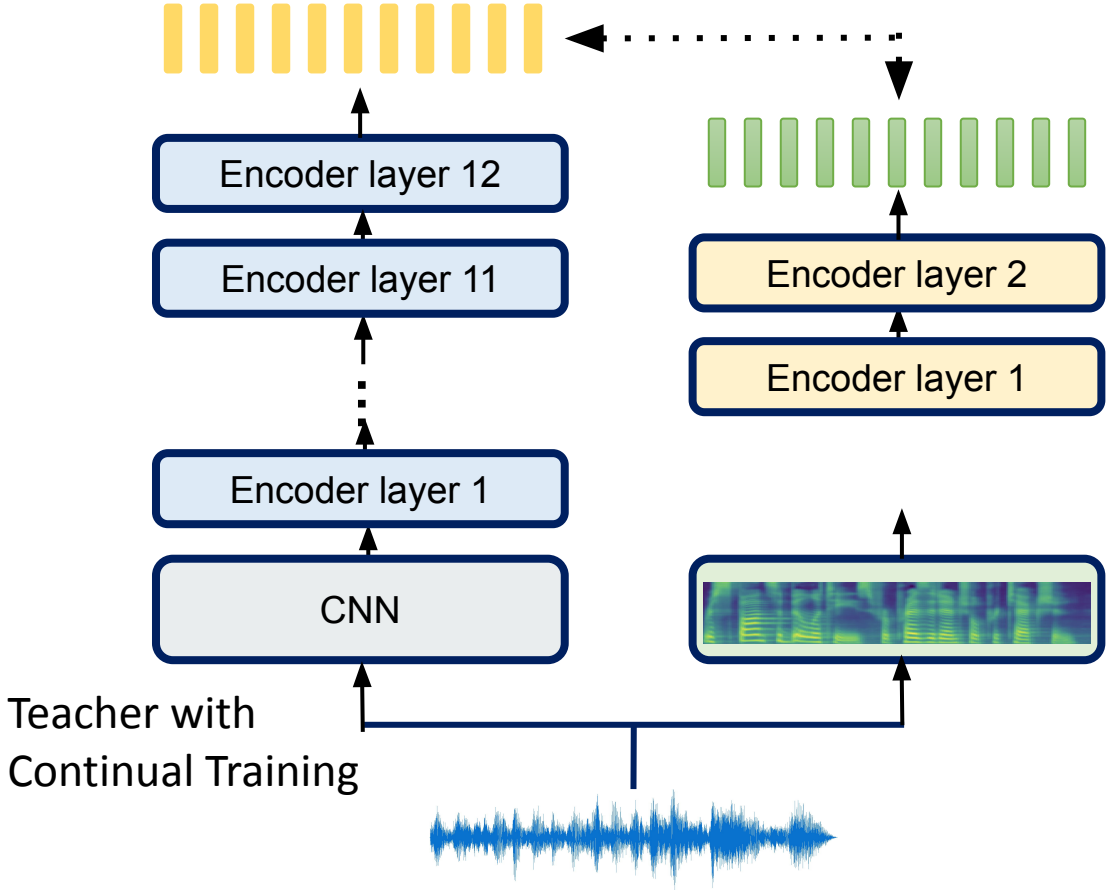
10: DistilHuBERT (CNN + 2-layer transformer)

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

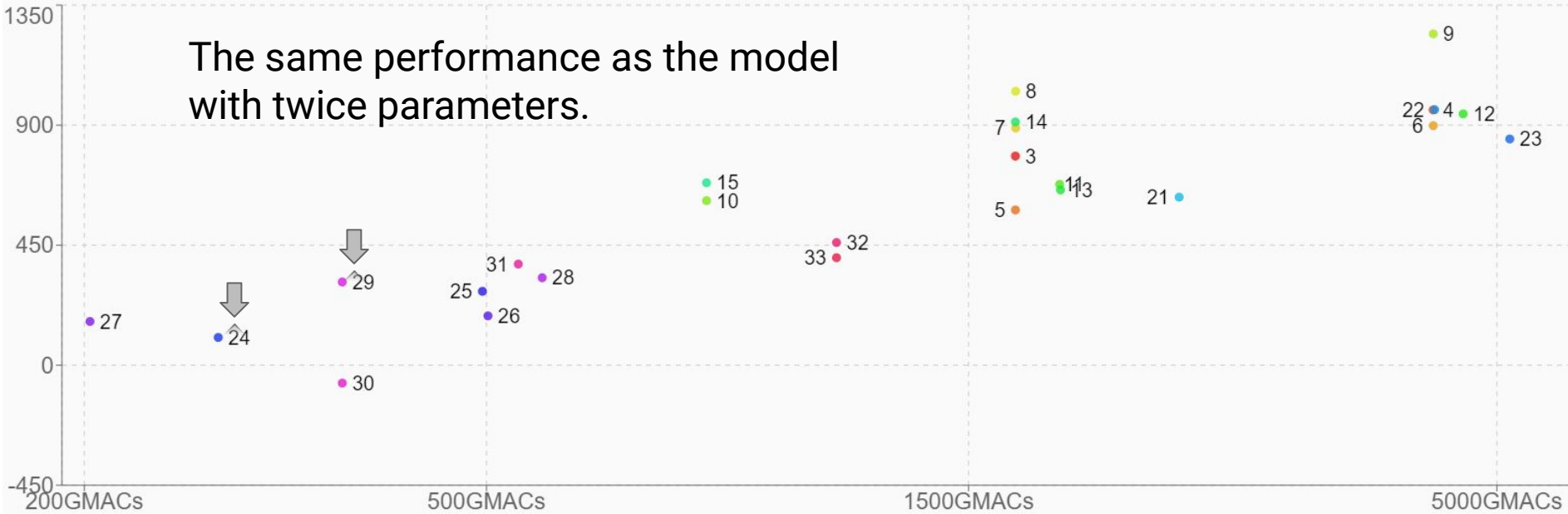
(A) DistilHuBERT

+

(B) MelHuBERT



SUPERB Leaderboard - Hidden-set Track



24: Distil + MelHuBERT (2-layer transformer)

29: Distil + MelHuBERT (3-layer transformer)

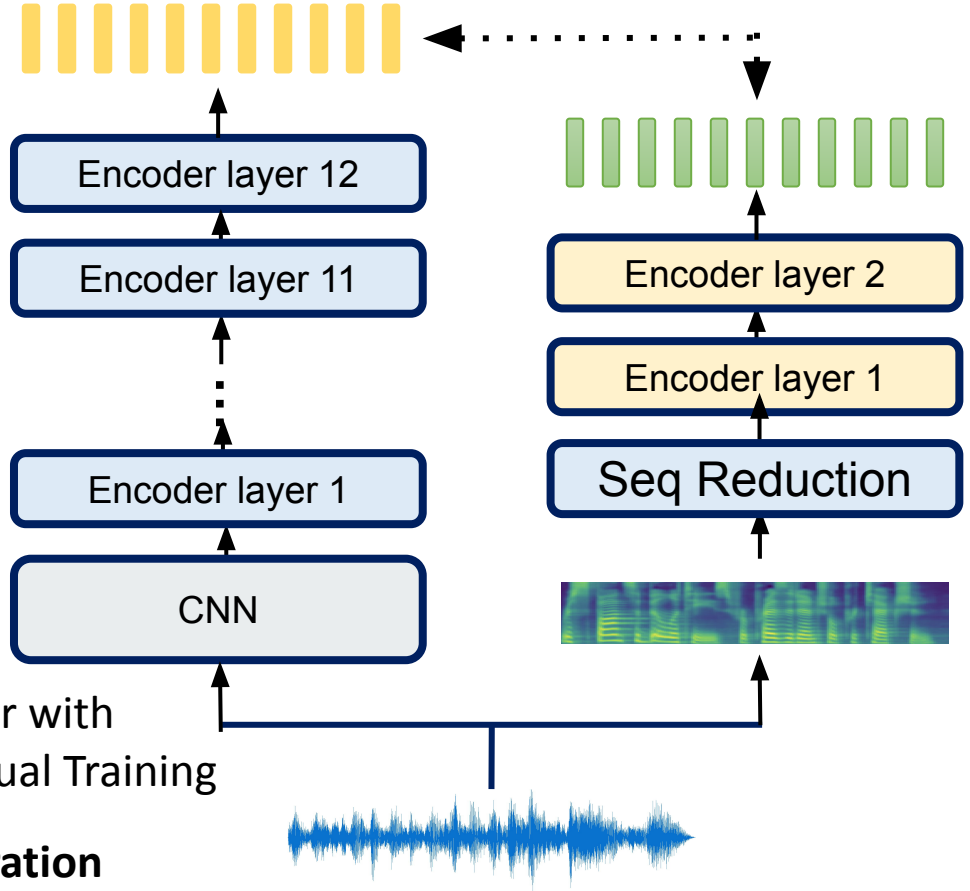
(A) DistilHuBERT

+

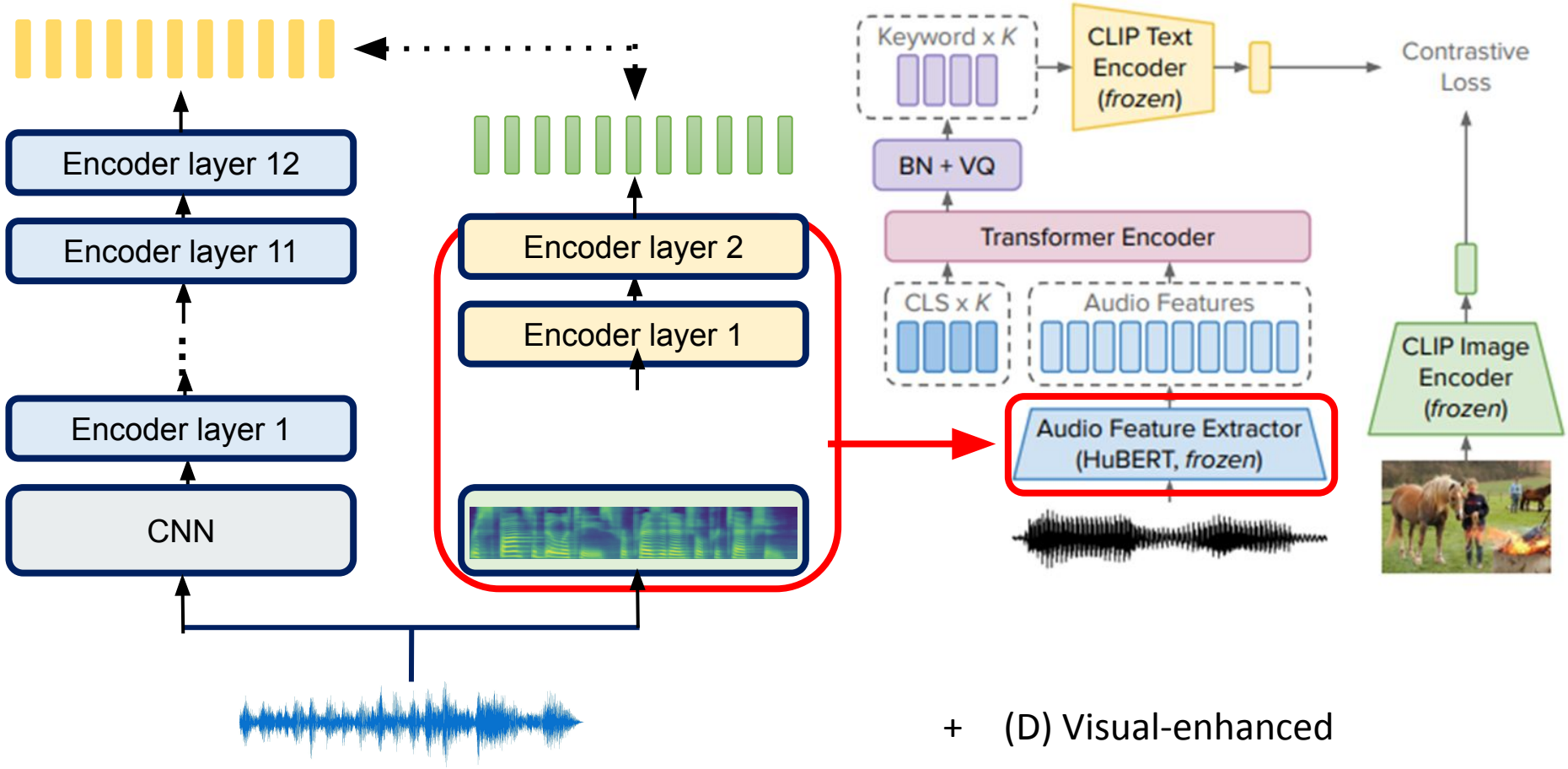
(B) MelHuBERT

+

(C) Sequence Reduction

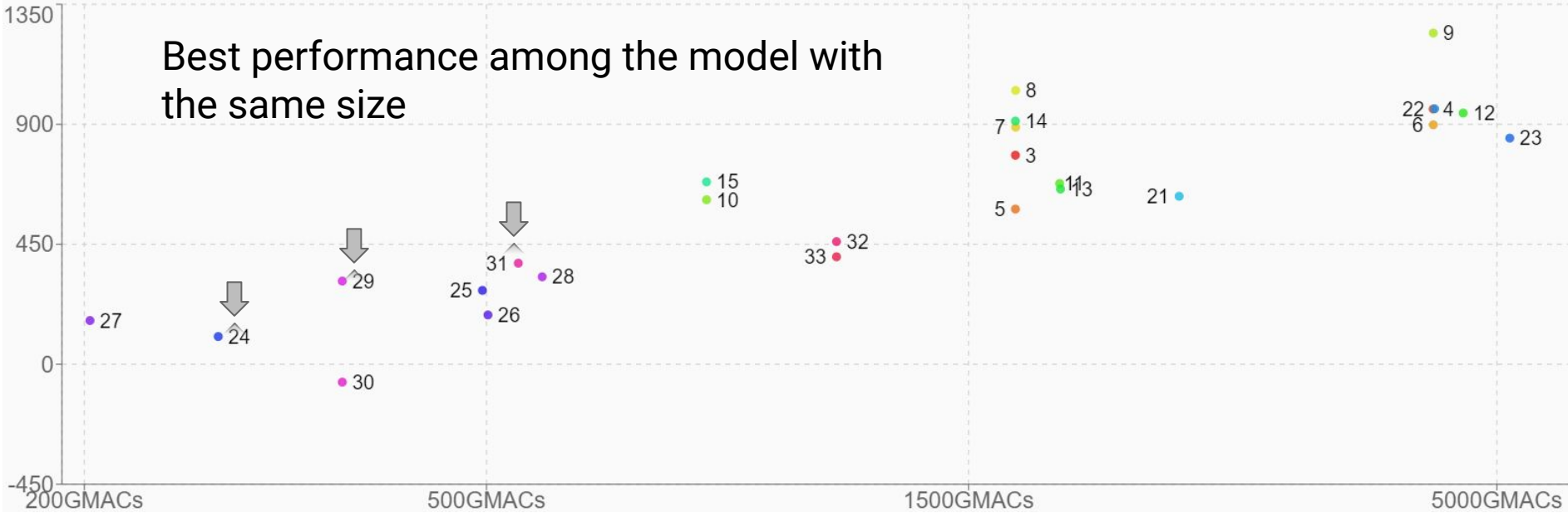


Largely degrade
speech enhancement, source separation



+ (D) Visual-enhanced

SUPERB Leaderboard - Hidden-set Track

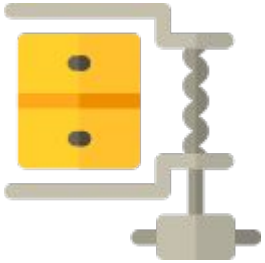


24: Distil + MelHuBERT (2-layer transformer)

29: Distil + MelHuBERT (3-layer transformer)

31: 29 + Visual-enhanced

Outline of Part 1: Better Pre-trained Model



Compression

9:35 - 9:55

Robust

9:55 - 10:10

Visual-enhanced

10:10 - 10:25

Integration

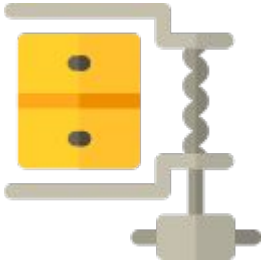
10:25 - 10:30

10 mins Q&A
+ 10 mins break

Acknowledgement

Happy Chinese Valentine's Day

Outline of Part 1: Better Pre-trained Model



Compression

9:35 - 9:55

Robust

9:55 - 10:10

Visual-enhanced

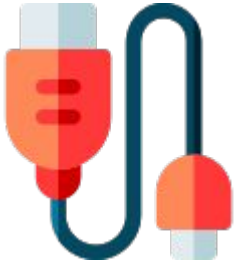
10:10 - 10:25

Integration

10:25 - 10:30

10 mins Q&A
+ 10 mins break

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

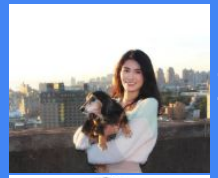
More usage

11:50 - 12:00

Toolkit for Speech Pre-training 12:00 - 12:10

Adapter & Prompt


Kai-Wei Chang (NTU)


Zih-Ching Chen (NTU)


Allen Fu (NTU)


Chih-Ying Liu (NTU)


Hua Shen (Penn State)


Fabian Ritter (NUS)

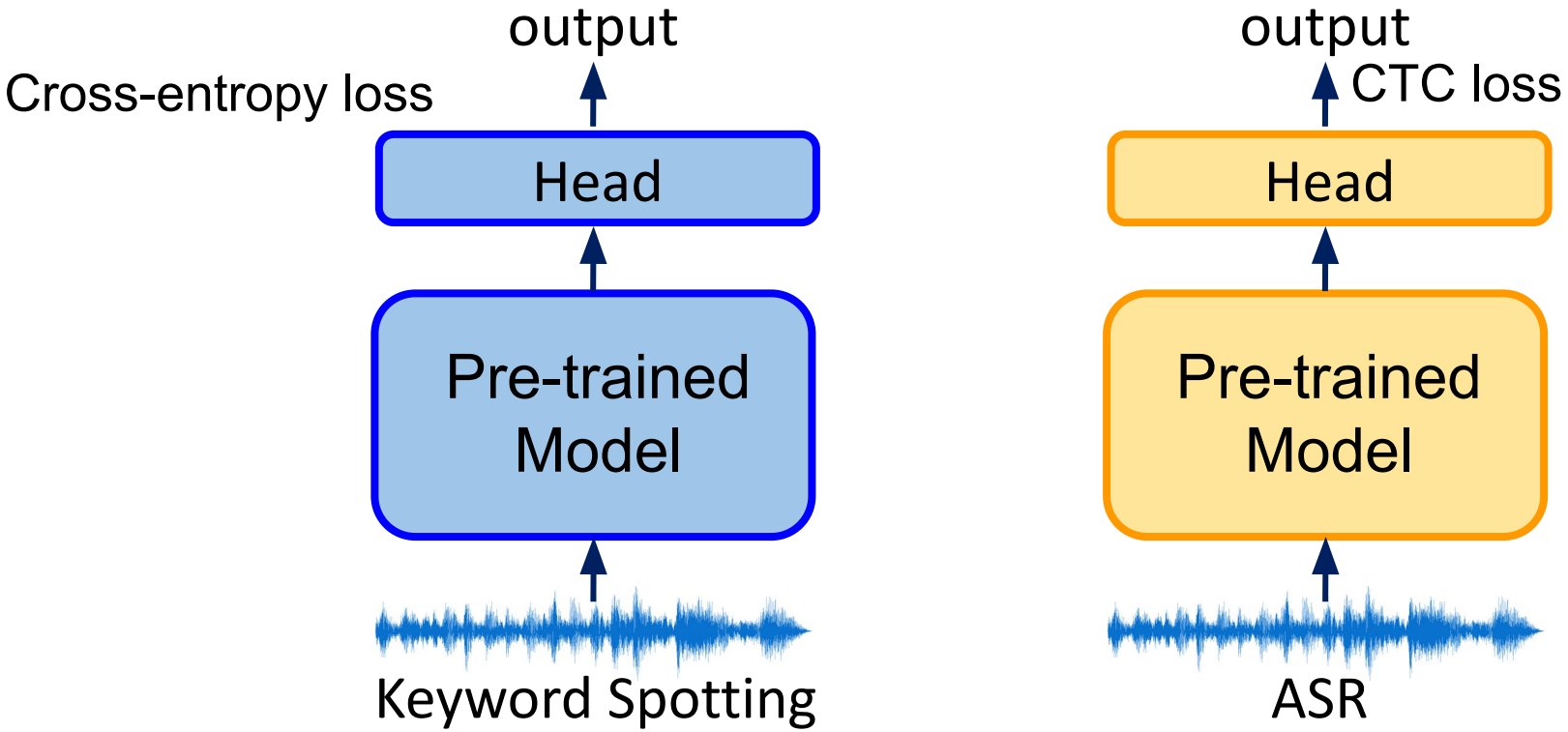

Yu-Kai Wang (NTU)


Shih-Ju Hsu (NTU)


Daniel Li (Meta)

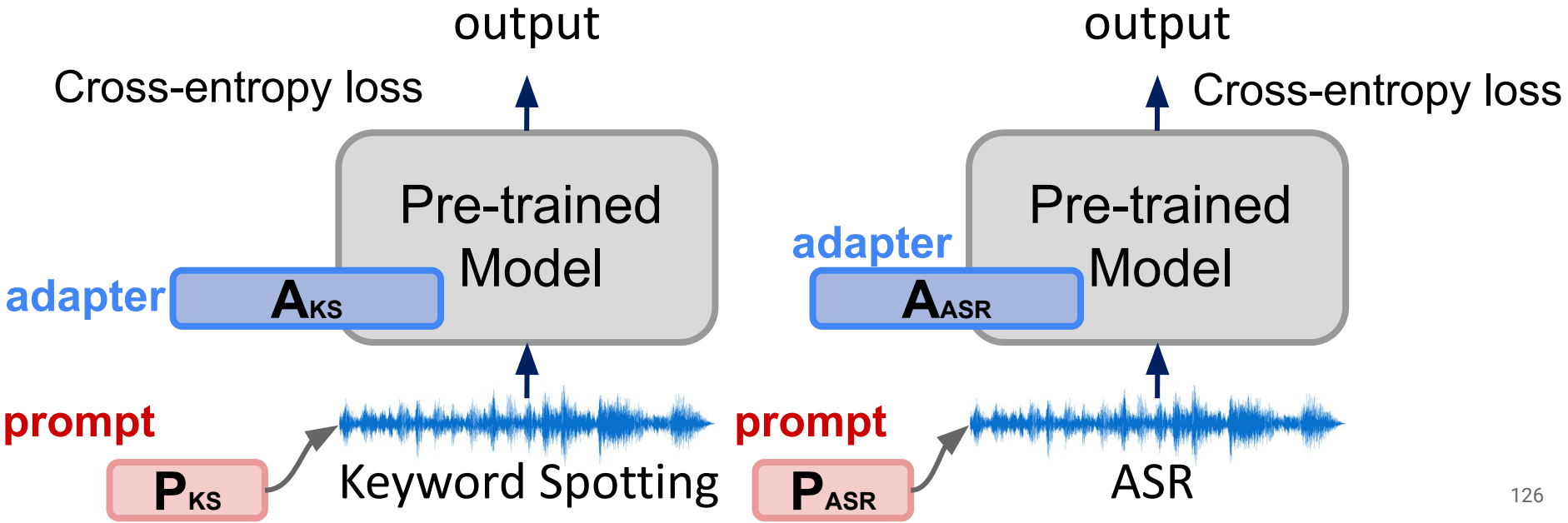

Hung-yi Lee (NTU)

Background - Pre-train, Fine-tune Paradigm

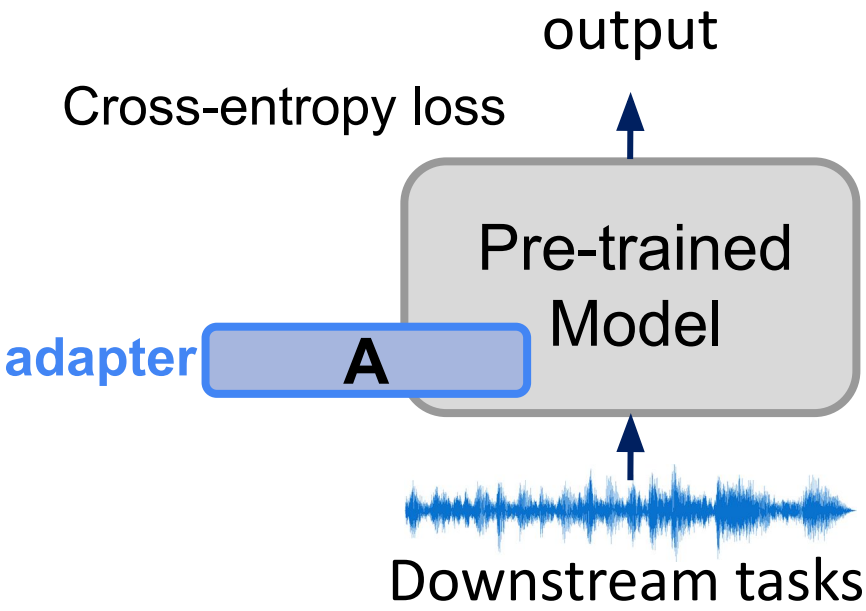


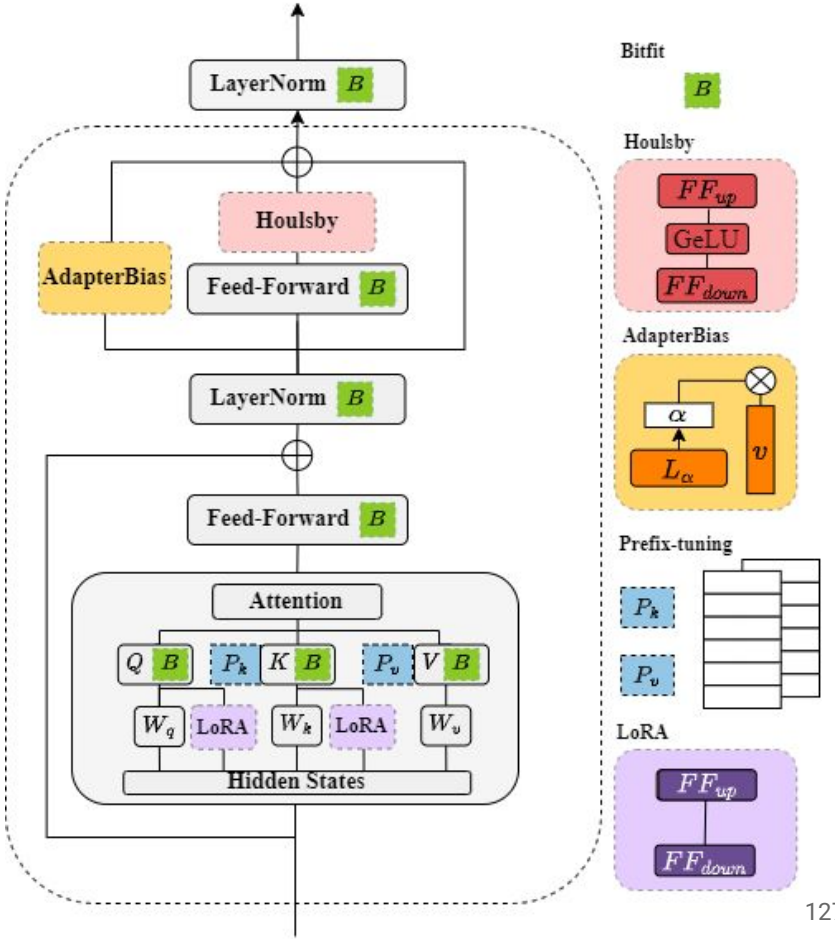
Background - Prompting / adapter Paradigm

Better performance, saving computation resources, more robust …

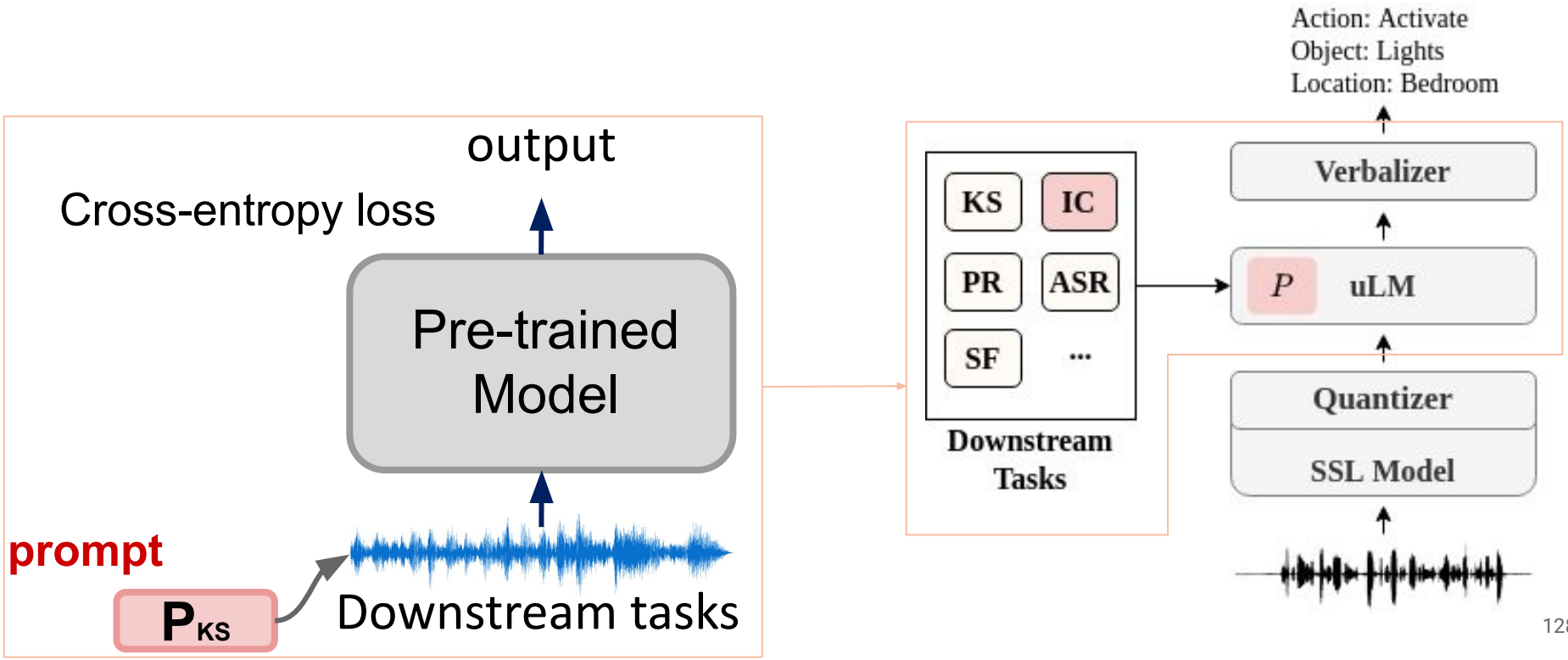


Adapter


output
Cross-entropy loss
Pre-trained Model
adapter
A
Downstream tasks


LayerNorm B
Houlsby
AdapterBias
Feed-Forward B
LayerNorm B
Feed-Forward B
Attention
Q B
P_k
K B
P_v
V B
W_q
LoRA
W_k
LoRA
W_v
Hidden States
Bitfit
B
Houlsby
FF_{up}
GeLU
FF_{down}
AdapterBias
α
L_α
v
Prefix-tuning
P_k
P_v
LoRA
FF_{up}
FF_{down}

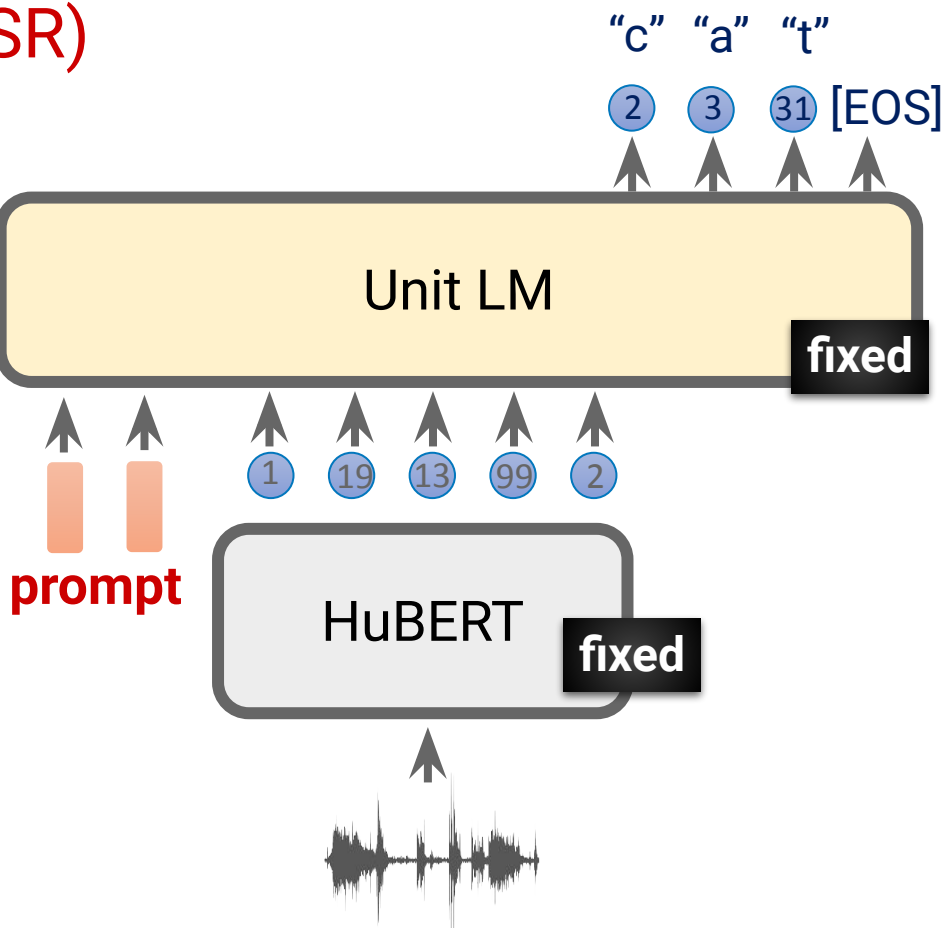
Prompting


output
Cross-entropy loss
Pre-trained Model
prompt
P_{KS}
Downstream tasks
KS
IC
PR
ASR
SF
...
Downstream Tasks
Action: Activate
Object: Lights
Location: Bedroom
Verbalizer
P
uLM
Quantizer
SSL Model

Prompting - examples (ASR)

label mapping (Verbalizer)

Unit ID	Character
1	“m”
2	“c”
3	“a”
4	“g”
…	
31	“t”

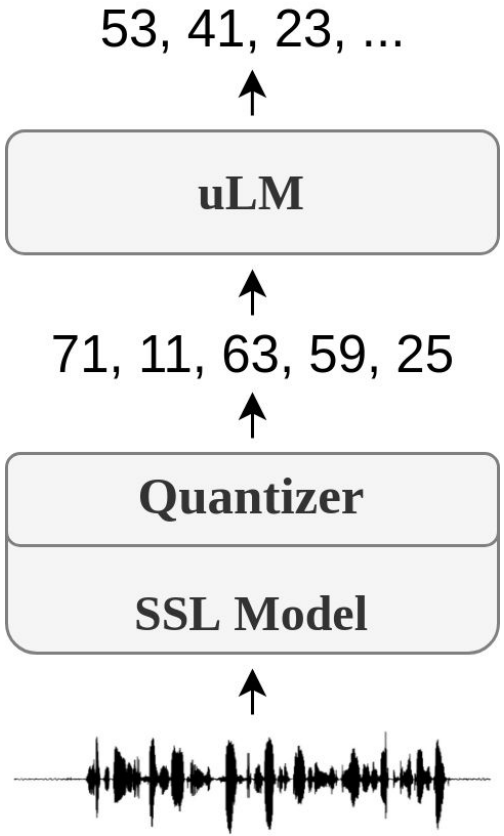


uLM / GSLM

Generative Spoken Language Model

- SSL Model: HuBERT, CPC, …
- Quantizer: K-Means
- uLM: generative unit Language Model

- First generative speech LM pre-trained on a large corpus (LibriLight- 6000hrs)



- [GSLM] Lakhotia et.al., Generative Spoken Language Modeling from Raw Audio

Adapter

- Does it work?
- What’s the benefit of using adapter?
- Any findings specific to speech or different from other areas?

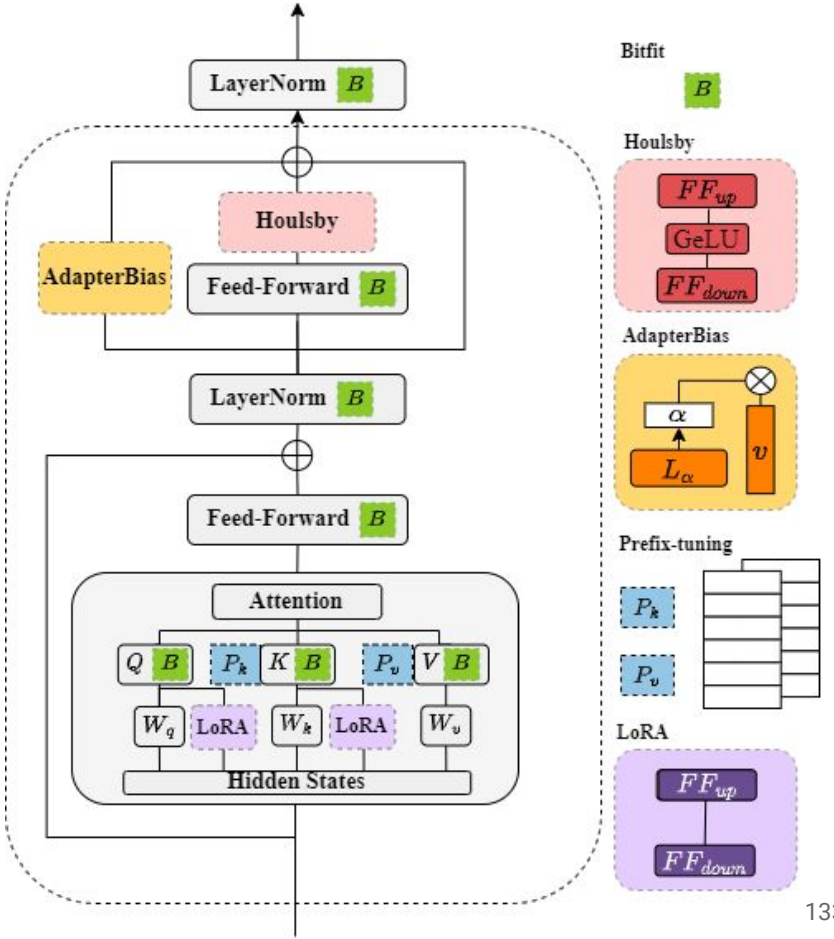
Does adapter work?

- Performance of efficient tuning methods in SUPERB benchmark

Method	Params	ASR	PR	SD	SID	SF	IC	KS
FT	94.7M	6.35	**2.45**	9.32	66.48	84.87	99.10	95.87
Baseline	0	7.09	7.74	7.05	64.78	86.25	96.39	95.32
Houlsby	0.60M	5.88	3.00	**4.00**	**87.71**	85.87	**99.60**	97.17
AdapterBias	0.02M	**5.54**	4.19	5.48	77.38	86.60	99.50	97.30
BitFit	0.10M	9.34	4.23	5.13	83.68	**87.40**	99.50	**97.33**
LoRA	0.29M	6.94	8.74	7.39	62.90	86.25	96.57	96.59
Prefix	0.10M	6.56	4.18	8.17	71.87	85.85	99.31	97.05
Weighted-sum	12	6.42	5.41	5.88	81.42	88.53	98.34	96.30

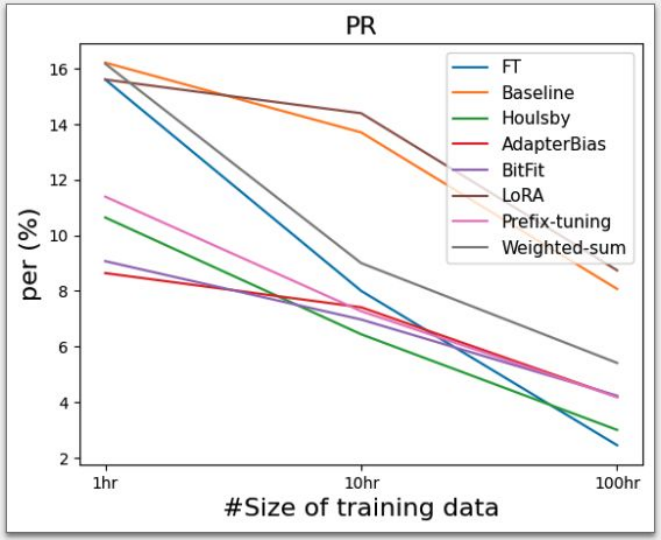
Does adapter work?

- In average, Houlsby performs best and LoRA performs worst
- Weighted-sum is powerful in Speech processing tasks



What's the benefit of using adapter?

- Adapter methods performs better especially in low-resource adaptation
- Adapter methods have more tolerance of learning rate compared to fine-tuning



Method	5×10^{-6}	5×10^{-5}	5×10^{-4}	5×10^{-3}
FT	3.03±0.1	2.81±0.4	100±0	100±0
Houlsby	6.09±0.49	3.24±0.14	2.81±0.03	3.06±0.03
AdapterBias	7.54±0.06	4.52±0.01	3.79±0.02	3.72±0.02

Speech specific findings

- Weighted-sum is surprisingly powerful in Speech tasks
  - Layer-wise: Each layer of SSL model provides different information
- Applying adapters to attention module is not effective in Speech processing task
  - LoRA does not work in SUPERB

LAYER-WISE ANALYSIS OF A SELF-SUPERVISED SPEECH REPRESENTATION MODEL

Ankita Pasad, Ju-Chieh Chou, Karen Livescu

Toyota Technological Institute at Chicago

{ankitap, jcchou, klivescu}@ttic.edu

https://arxiv.org/abs/2107.04734

Prompting

- Does it work?
- Is the gain generalizable?
- How to make prompting better?

Does prompting work?

- PT: Prompt Tuning
- FT: Fine-Tuning
- KS: Keyword Spotting - Single-label Cls
- IC: Intent Classification - Multi-label Cls

Scenarios		KS		IC	
		ACC↑	# param.	ACC↑	# param.
HuBERT	PT	95.16	0.08M	**98.40**	0.15M
	FT	**96.30**	0.2M	98.34	0.2M
CPC	PT	**93.54**	0.05M	**97.57**	0.05M
	FT	91.88	0.07M	64.09	0.07M

Is the gain generalizable?

Prompting on Different Speech Classification tasks

- LT-SCR: Lithuaniuan Speech Commands Recognition
- DM-SCR: Dysarthric Mandarin Speech Commands Recognition
- AR-SCR: Arabic Speech Commands Recognition

Scenarios	LT-SCR		DM-SCR		AR-SCR	
	ACC↑	# param.	ACC↑	# param.	ACC↑	# param.
SOTA	91.8	0.2 M	**93.5**	0.2 M	98.6	0.2 M
HuBERT - FT	70.5	151 M	12.3	151 M	88.9	151 M
HuBERT - PT	**93.2**	0.1 M	73.1	0.1 M	**99.7**	0.1 M

How to make prompting better?

Linear Verbalizer Results - Speech Classification

- Freq-Verb: Frequency-based Verbalizer
- Linear-Verb: Linear Verbalizer
- KS: Keyword Spotting - Single-label Cls
- IC: Intent Classification - Multi-label Cls

Scenarios	Verbalizer	KS		IC	
		ACC↑	# param.	ACC↑	# param.
HuBERT	Freq-Verb	94.32	0.15M	98.10	0.15M
	Linear-Verb	**94.68**	0.16M	**98.66**	0.16M
CPC	Freq-Verb	93.77	0.05M	97.89	0.05M
	Linear-Verb	**94.00**	0.06M	**97.94**	0.06M

Summary

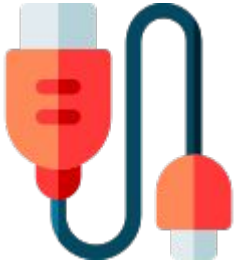
- First thorough study of adapter and prompting in speech
- Demonstrate both yield comparable or better performance with fewer parameters
- Adapter shows gain over SUPERB tasks while prompting is mainly on classification tasks
- Discover additional benefits: data efficiency (both), robust against learning rate (adapter)

  - 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", in *INTERSPEECH* 2022
  - Adapter work in submission to *SLT* 2022

Future works

- More techniques to improve performance / efficiency
  - Adapter customized to speech, where to add adapters, combined w/ weighted sum
  - Sequence compression and denoising techniques for prompting
- More domains / more challenging tasks
  - Adapter for domain adaptation or forgetting
  - Prompting for other SUPERB tasks

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

More usage

11:50 - 12:00

Toolkit for Speech Pre-training 12:00 - 12:10

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

Are pretrained models useful for prosody-intensive tasks?

Do they learn prosodic features and patterns?

Guan-Ting Lin, Chi-Luen Feng … Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li … Samuel Miller … Hung-Yi Lee, Nigel Ward

On the Utility of Self-Supervised Models for Prosody-Related Tasks*

*IEEE SLT 2022, submitted

Are pretrained models useful for prosody-intensive tasks?

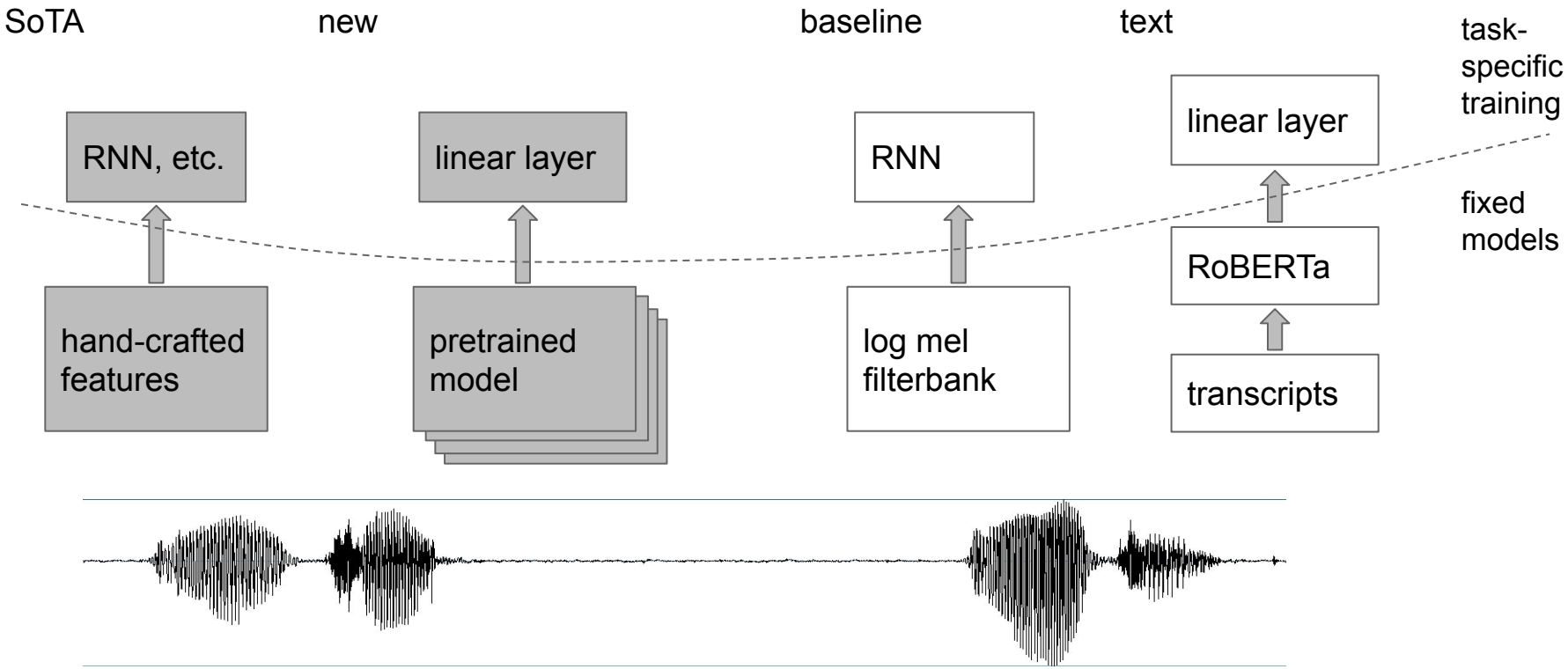
Do they learn prosodic features and patterns? **YES, YES**

Guan-Ting Lin, Chi-Luen Feng … Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li … Samuel Miller … Hung-Yi Lee, Nigel Ward

On the Utility of Self-Supervised Models for Prosody-Related Tasks*

*IEEE SLT 2022, submitted

Overview


SoTA
new
baseline
text
task-specific training
fixed models
RNN, etc.
linear layer
RNN
linear layer
RoBERTa
transcripts
hand-crafted features
pretrained model
log mel filterbank

On the Utility of Self-Supervised Models for Prosody-Related Tasks

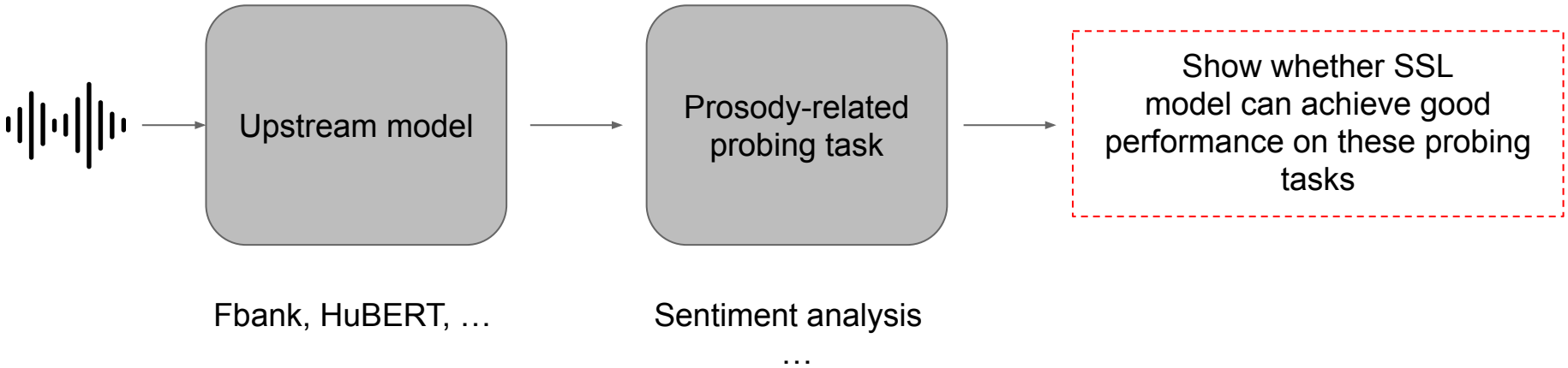
Outline

- Five Prosody-Related Tasks
- Experiment Setup
- Performance Results

- Layerwise Analysis
- Cross-language Generality

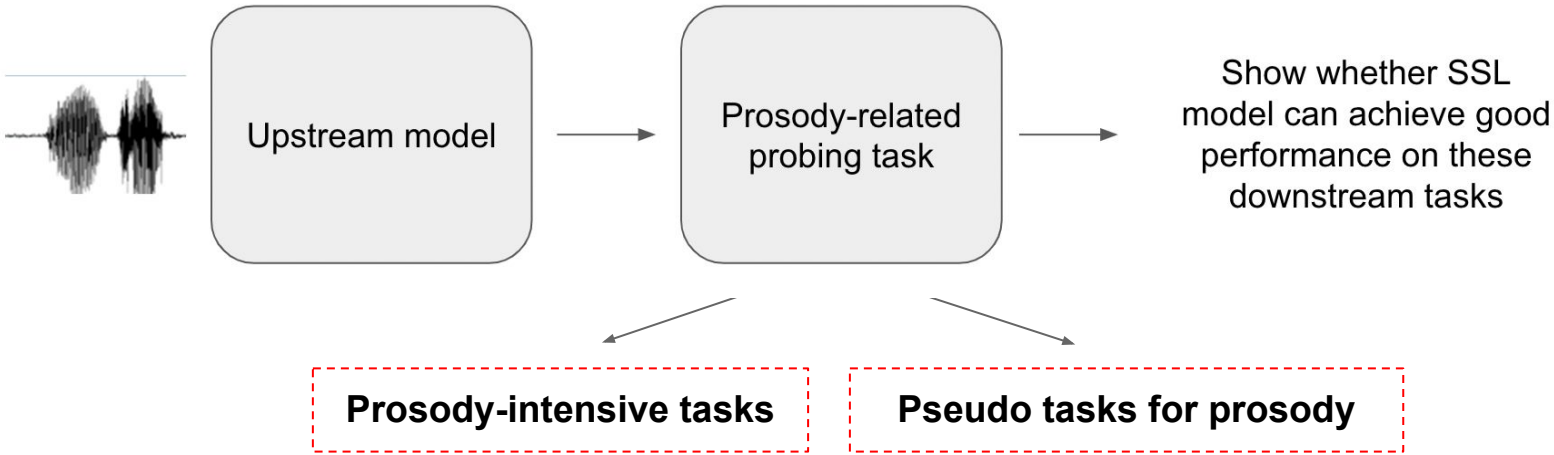
Introduction - The overall framework

- Use probing task to measure whether the output of SSL model have prosody information
- Use **16** upstream models with **5** probing tasks for probing prosody information



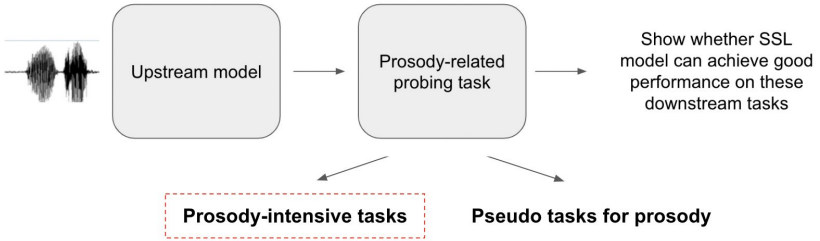
Probing task introduction

- Two categories of probing task:
  Prosody-intensive tasks and **Pseudo tasks for prosody**



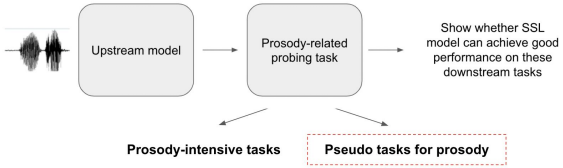
Probing task introduction - Prosody intensive

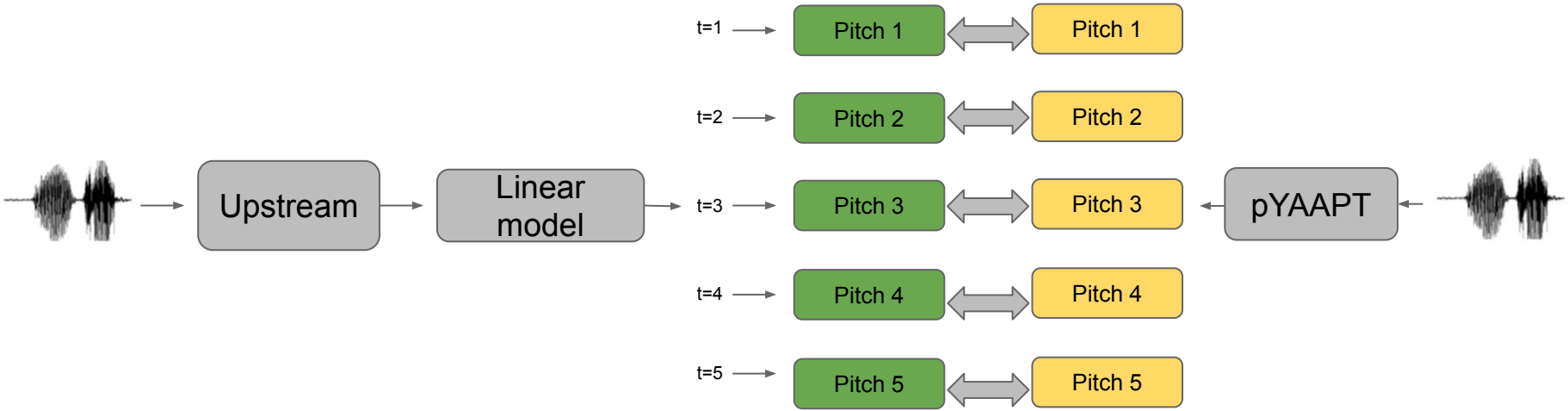
- **Sarcasm Detection**
  - Given an utterance, predict whether it's sarcastic
  - Dataset: MUStARD
  - Binary classification task
- **Sentiment Analysis**
  - Given an utterance, predict the sentiment of it
  - Has multiple sentiment. Ex: happy, angry, …
  - Dataset: CMU-MOSEI
  - Multiclass classification task
- **Persuasiveness Prediction**
  - Given an utterance, predict whether it's persuasive
  - Dataset: POM
  - Binary classification task



Probing task introduction - Pseudo tasks for prosody

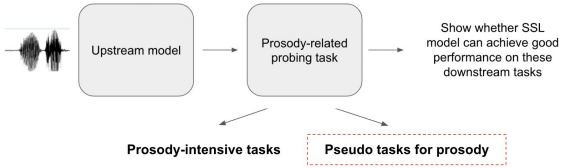
- **Prosody Reconstruction**
  - Given an utterance, predict pitch/energy for each frame
  - Pitch/energy label: generate by pYAAPT
  - Dataset: LibriTTS
  - Regression task

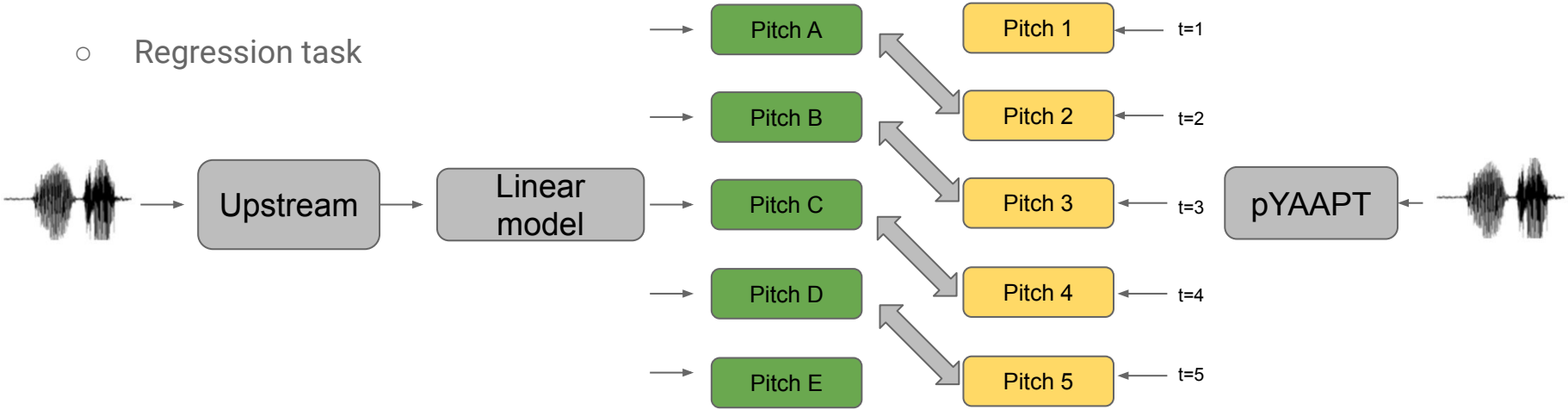




Probing task introduction - Pseudo tasks for prosody

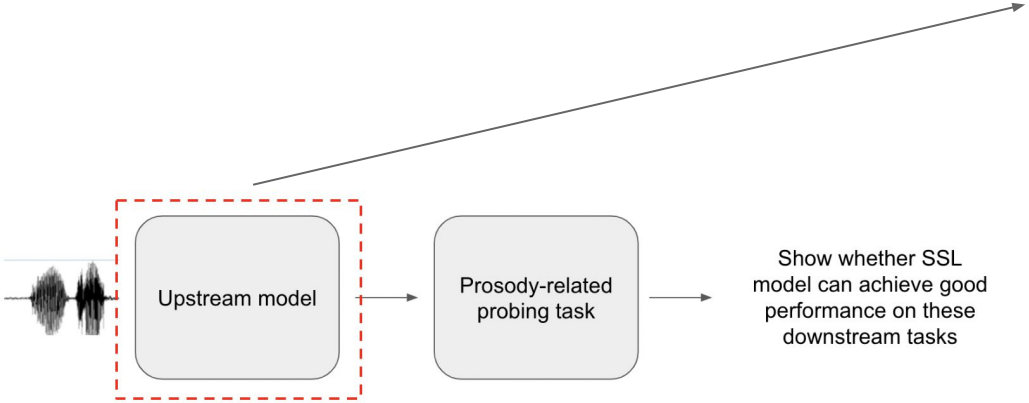
- **Future Value Prediction**
  - Given an utterance, predict future pitch/energy for each frame
  - Dataset: LibriTTS
  - Regression task





Experiment setup

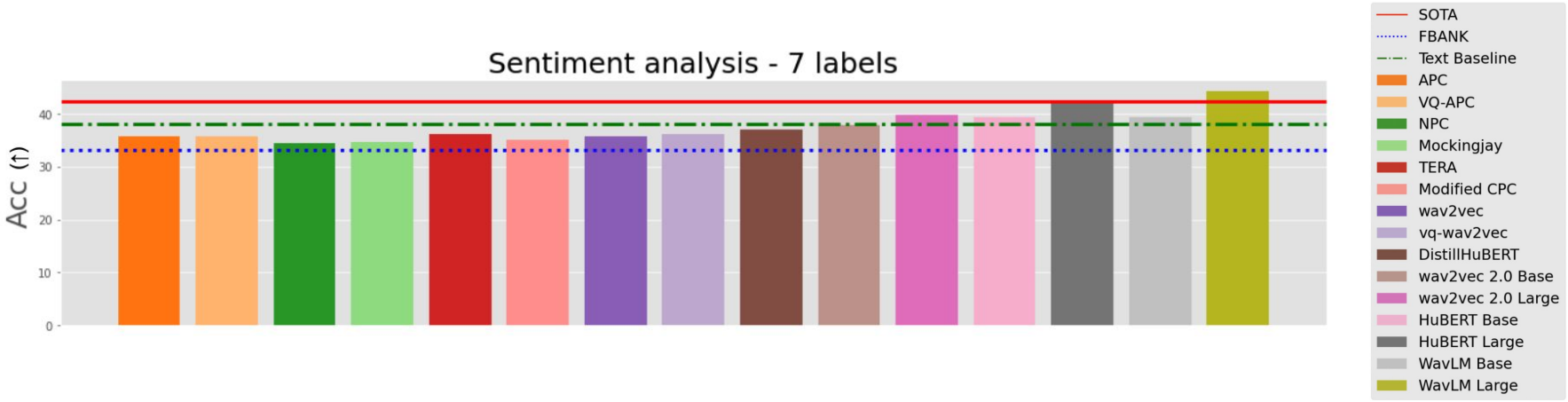
- Upstream models selection
- 16 models, including traditional method and SSL models



FBANK	vq-wav2vec
APC	DistilHuBERT
VQ-APC	wav2vec 2 Base
NPC	wav2vec 2 Large
Mockingjay	HuBERT Base
TERA	HuBERT Large
modified CPC	WavLM Base
wav2vec	WavLM Large

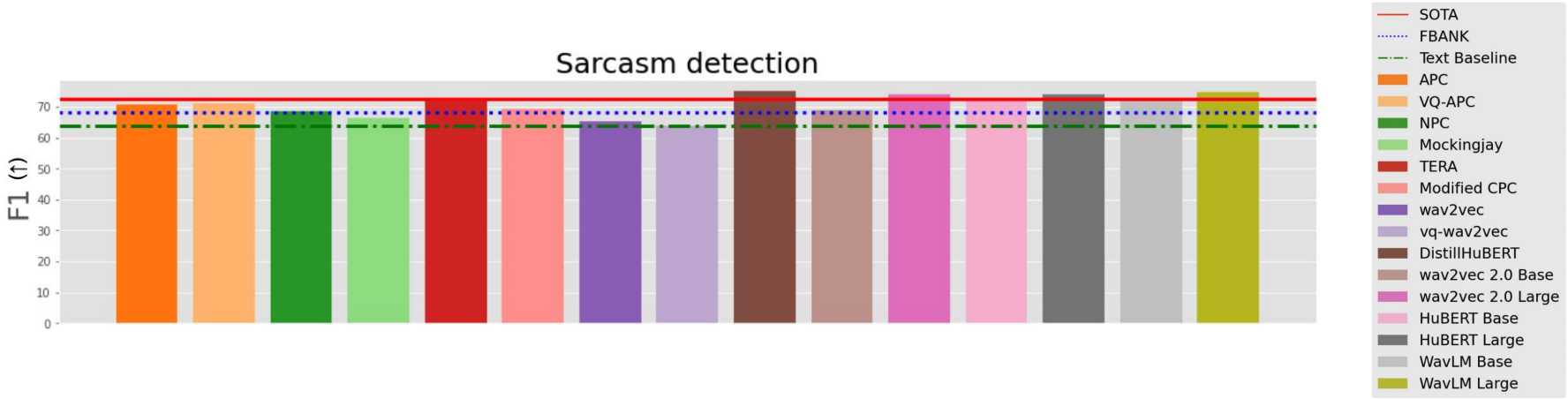
Main results

SSL models perform well on prosody-intensive tasks



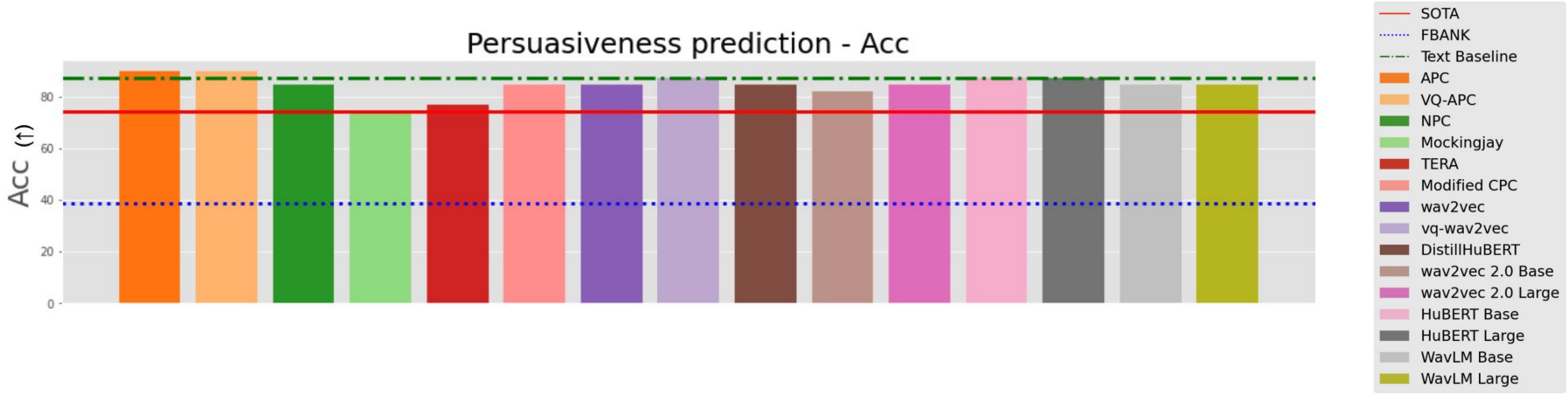
- All SSL models outperform the baseline FBANK feature
- WavLM and HuBERT Large even yield performance better than SOTA and text-only baseline

SSL models perform well on prosody-intensive tasks (Cont'd)



- Text-only baseline has the worst performance
- Although some SSL models show inferior performance to the baseline FBANK, other SSL models (DistillHuBERT, HuBERT, and WavLM) achieve better F1 score than the previous SOTA

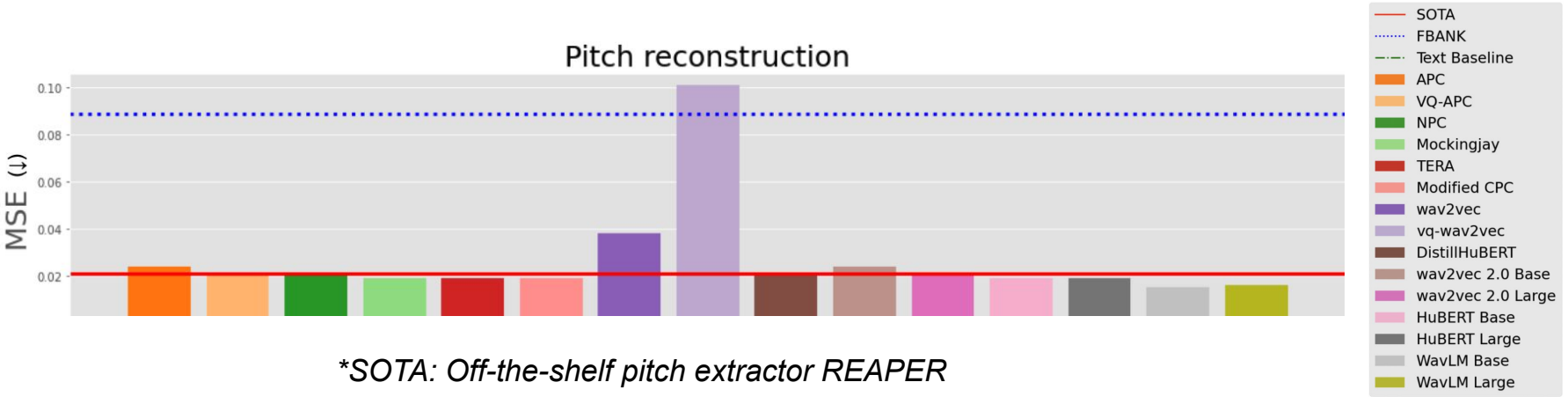
SSL models perform well on prosody-intensive tasks (Cont'd)



- SSL models significantly yield better performance than the FBANK and previous SOTA
- APC, VQ-APC, vq-wav2vec, and HuBERT obtain superior accuracy compared to text-only baseline

SSL models truely encode prosodic information

Result of Prosody Reconstruction (PR) on pitch



**SOTA: Off-the-shelf pitch extractor REAPER*

- Several SSL models surpass REAPER and FBANK
- The best-performed model is WavLM

SSL models can predict future prosody

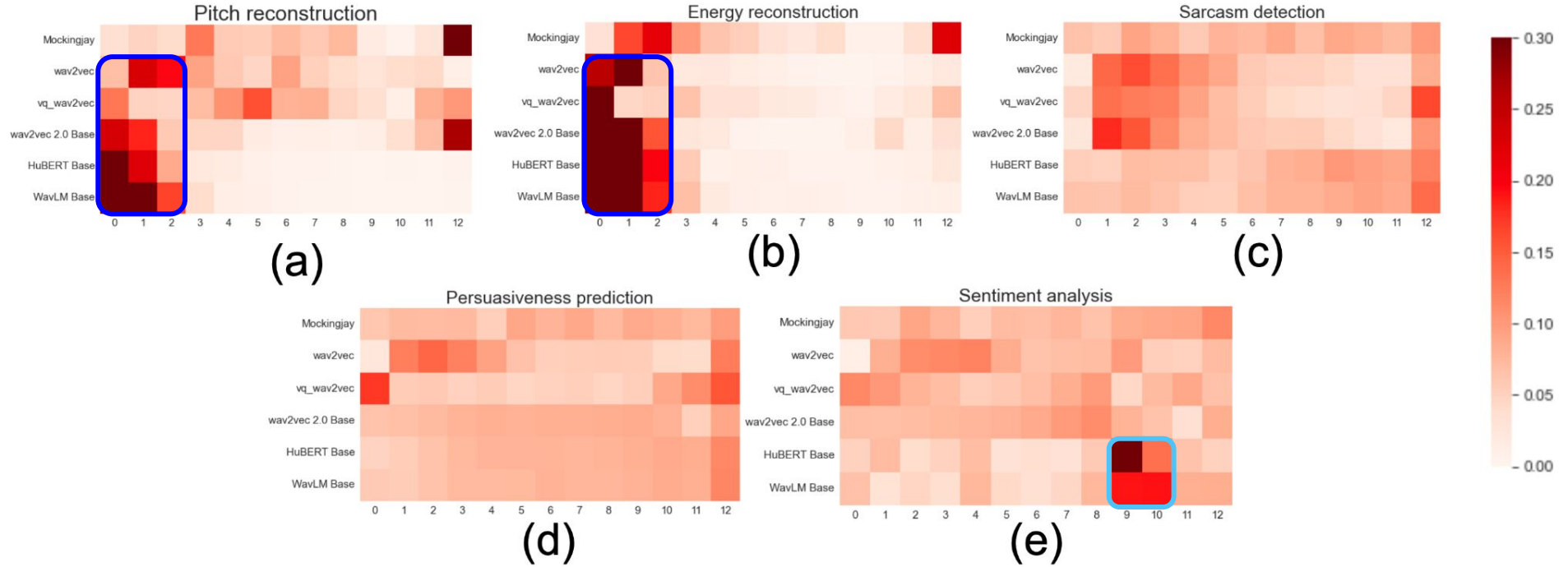
MSE of Future Value Prediction (FVP) on pitch

Method	Pitch w/ Prediction Horizon (s)			
	0.12	0.24	0.50	1.00
FBANK + RNN	0.049	0.104	0.142	0.157
APC	0.033	0.043	0.052	0.053
modified CPC	0.038	0.051	0.062	0.065
wav2vec	0.053	0.064	0.075	0.075
Mockingjay	0.069	0.077	0.081	0.099
wav2vec 2.0 Base	0.038	0.047	0.047	0.049
wav2vec 2.0 Large	0.035	0.039	0.045	0.046
HuBERT Base	0.029	0.036	0.041	0.042
HuBERT Large	**0.025**	**0.027**	**0.028**	**0.037**

- HuBERT Large significantly outperforming other SSL models and baseline FBANK+RNN in all four horizons
- Although some SSL models' pre-training objectives is future generation/discrimination, they are not good as HuBERT

Further Analysis

Layerwise Contribution Analysis



Feature Integration

- Setup 1 (**Low-level enhanced**) : the first two layers and the best layer we discovered
- Setup 2 (**Neighbor layers**): the best layer with its two neighbor layers

Low-level prosodic information improves the modeling of SD and PP

	SD		PP		SA	
Layer selection	(0,1,12)	(10,11,12)	(0,1,12)	(10,11,12)	(0,1,8)	(7,8,9)
wav2vec 2.0 Base	**70.6**	66.3	**81.3**	81.3	73.6	**74.0**
HuBERT Base	**72.0**	70.0	**85.8**	83.8	75.2	**76.2**

The best layer is determined by the contribution analysis, which obtains the largest contribution, marked with underline.

SA highly rely on content instead of low-level prosodic information

Case Study: Cross-Lingual Transfer of SSL Models

Do Models Pre-Trained on English Data Learn Generalizable Methods to Extract Pitch?

Sam Miller

FLEURS Dataset

2009

n-way parallel sentences

Total of 350GB / 1,400 hours

384,060 audio files (avg of 2.3 recordings / sentence)

Read speech from Wikipedia

Different speakers for train, dev, and test sets

102

languages from across Eurasia and Africa

No Indigenous representation from the Americas or Australia

25 West Europe, 16 East Europe, 12 MENA, 20 Sub-Saharan Africa, 14 South Asia, 11 Southeast Asia, 4 East Asia

Labels

transcription, language id, and speaker gender

Also id, path, raw transcription, language group id, audio array

https://huggingface.co/datasets/google/fleurs

Data Example

the winter can be deceptively chilly temperatures rarely go below freezing but the wind and humidity combine to make it feel colder than what the thermometer says

English

冬 天 可 能 会 比 看 上 去
显 得 更 冷 温 度 很 少 低
于 零 度 但 当 风 和 湿 度
结 合 在 一 起 会 让 人 感
觉 比 温 度 计 显 示 的 温
度 要 冷

Mandarin

زستان دەکرێت زۆر سارد بێت
پلەی گەرمی زۆر بە دەگمەن بۆ
بەستان دادەبەزێت بەڵام با و
شێداری یەک دەگرن بۆ ئەوەی
هەست بکەیت ساردترە لەوەی
گەرمیێو نیشانی دەدات

Sorani Kurdish

ubusika bungaba nzima kakhulu amazing okushisa angase adlulele ngale kweqhwa kodwa umoya kanye nomswakama sekuhlangene kwenza kubande ngaphezu kwalokho isikali sokushisa esikushoyo

Zulu

Data Example

the winter can be deceptively chilly temperatures rarely go below freezing but the wind and humidity combine to make it feel colder than what the thermometer says

English

冬 天 可 能 会 比 看 上 去
显 得 更 冷 温 度 很 少 低
于 零 度 但 当 风 和 湿 度
结 合 在 一 起 会 让 人 感
觉 比 温 度 计 显 示 的 温
度 要 冷

Mandarin

زستان دەکرێت زۆر سارد بێت
پلەی گەرمی زۆر بە دەگمەن بۆ
بەستان دادەبەزێت بەڵام با و
شێداری یەک دەگرن بۆ ئەوەی
هەست بکەیت ساردترە لەوەی
گەرمیپێو نیشانی دەدات

Sorani Kurdish

ubusika bungaba nzima kakhulu
amazing okushisa angase
adlulele ngale kweqhwa kodwa
umoya kanye nomswakama
sekuhlangene kwenza kubande
ngaphezu kwalokho isikali
sokushisa esikushoyo

Zulu

Data Example

the winter can be deceptively chilly temperatures rarely go below freezing but the wind and humidity combine to make it feel colder than what the thermometer says

冬 天 可 能 会 比 看 上 去
显 得 更 冷 温 度 很 少 低
于 零 度 但 当 风 和 湿 度
结 合 在 一 起 会 让 人 感
觉 比 温 度 计 显 示 的 温
度 要 冷

زستان دەکرێت زۆر سارد بێت
پلەی گەرمی زۆر بە دەگمەن بۆ
بەستان دادەبەزێت بەڵام با و
شێداری یەک دەگرن بۆ ئەوەی
هەست بکەیت ساردترە لەوەی
گەرمیێو نیشانی دەدات

ubusika bungaba nzima kakhulu
amazing okushisa angase
adlulele ngale kweqhwa kodwa
umoya kanye nomswakama
sekuhlangene kwenza kubande
ngaphezu kwalokho isikali
sokushisa esikushoyo

English

Mandarin

Sorani Kurdish

Zulu

Data Example

the winter can be deceptively chilly temperatures rarely go below freezing but the wind and humidity combine to make it feel colder than what the thermometer says

English

冬 天 可 能 会 比 看 上 去
显 得 更 冷 温 度 很 少 低
于 零 度 但 当 风 和 湿 度
结 合 在 一 起 会 让 人 感
觉 比 温 度 计 显 示 的 温
度 要 冷

Mandarin

زستان دەکرێت زۆر سارد بێت
پلەی گەرمی زۆر بە دەگمەن بۆ
بەستان دادەبەزێت بەڵام با و
شێداری یەک دەگرن بۆ ئەوەی
هەست بکەیت ساردترە لەوەی
گەرمیێو نیشانی دەدات

Sorani Kurdish

ubusika bungaba nzima kakhulu
amazing okushisa angase
adlulele ngale kweqhwa kodwa
umoya kanye nomswakama
sekuhlangene kwenza kubande
ngaphezu kwalokho isikali
sokushisa esikushoyo

Zulu

Data Example

the winter can be deceptively chilly temperatures rarely go below freezing but the wind and humidity combine to make it feel colder than what the thermometer says

English

冬 天 可 能 会 比 看 上 去
显 得 更 冷 温 度 很 少 低
于 零 度 但 当 风 和 湿 度
结 合 在 一 起 会 让 人 感
觉 比 温 度 计 显 示 的 温
度 要 冷

Mandarin

زستان دەکرێت زۆر سارد بێت
پلەی گەرمی زۆر بە دەگمەن بۆ
بەستان دادەبەزێت بەڵام با و
شێداری یەک دەگرن بۆ ئەوەی
هەست بکەیت ساردترە لەوەی
گەرمیپێو نیشانی دەدات

Sorani Kurdish

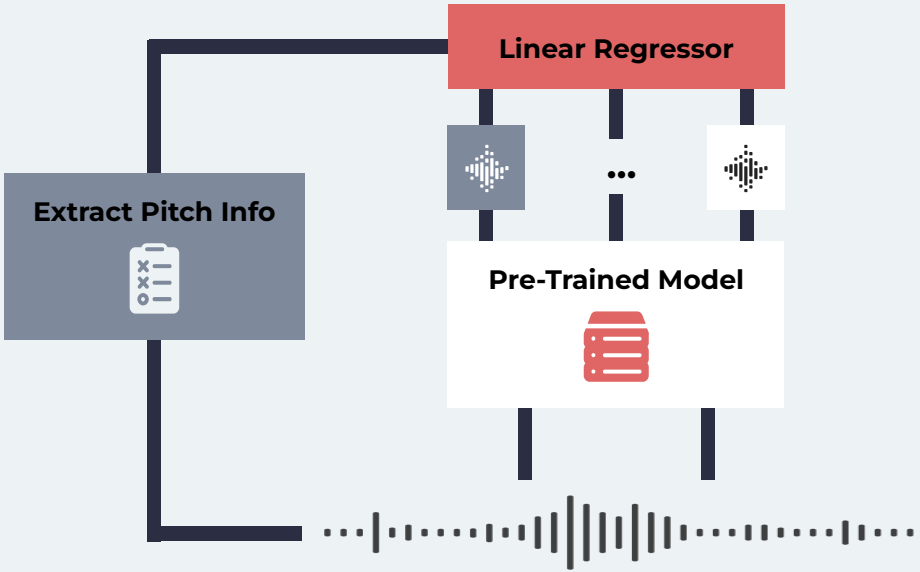
ubusika bungaba nzima kakhulu
amazing okushisa angase
adlulele ngale kweqhwa kodwa
umoya kanye nomswakama
sekuhlangene kwenza kubande
ngaphezu kwalokho isikali
sokushisa esikushoyo

Zulu

Main Objective

Cross-Lingual Transfer of Prosodic Information

Pitch Reconstruction



Label Gen. – pYAAPT

Pitch Tracking

Pitch is calculated based on pitch information extracted frame by frame

Language Indep.

Initially evaluated on English and Japanese

Algorithm Brief

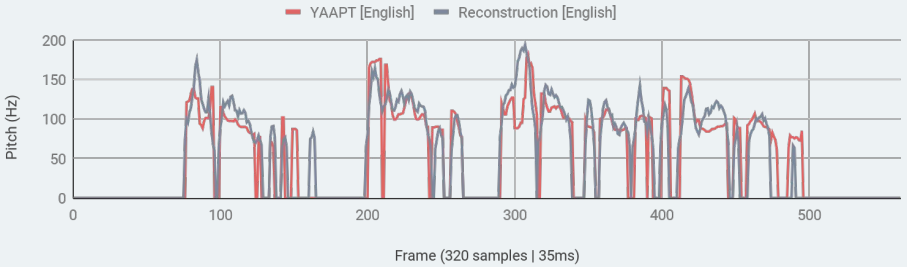
Calculate F0 from spectrogram of audio generated with signal processing. Approx. F0 using a spectral harmonics correlation

Refinement

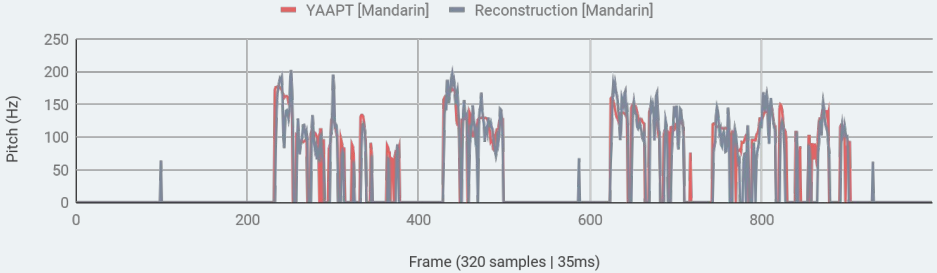
Many options are generated, to then be refined, selecting one

Sample Result

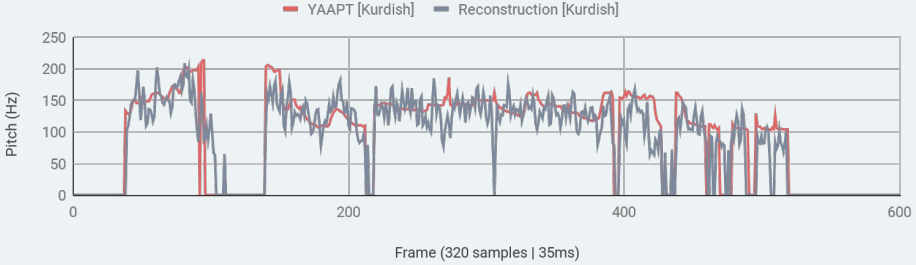
English



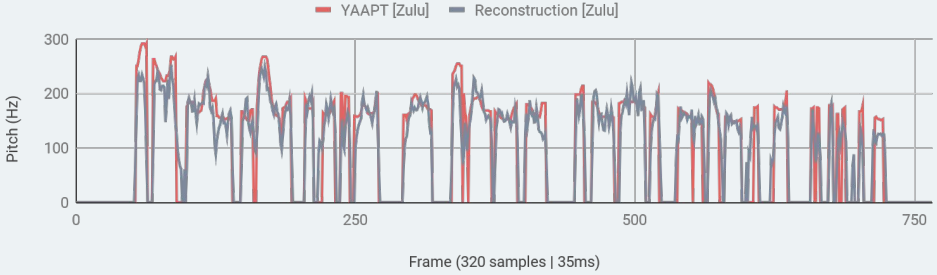
Mandarin



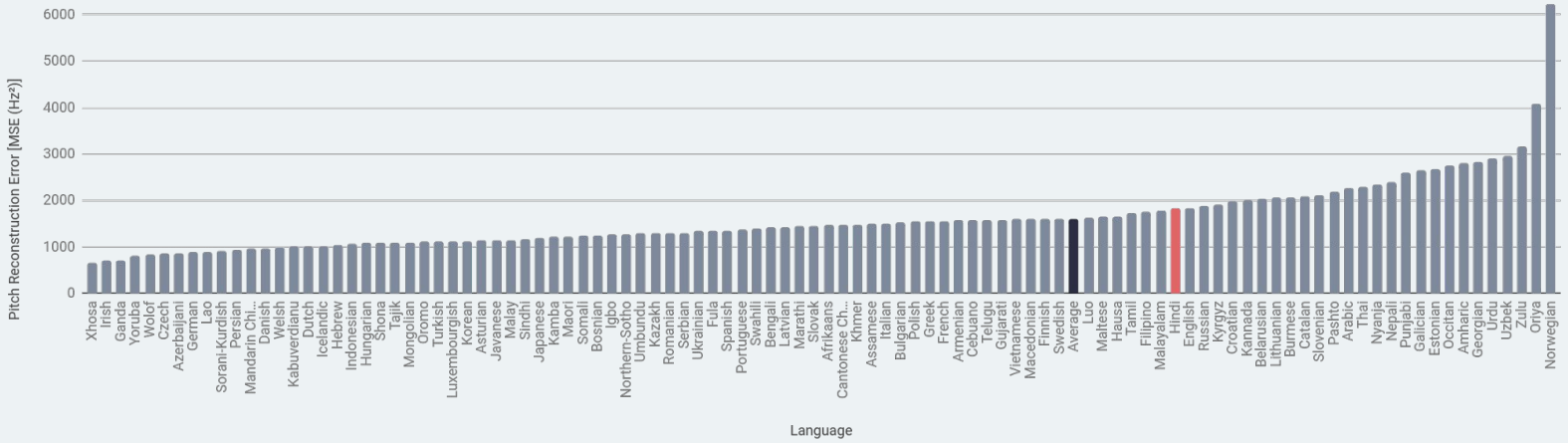
Sorani Kurdish



Zulu



Results by Language



Pitch Reconstruction Error [MSE (Hz^2)] for HuBERT, average of scores for all layers, by language

English

xx.xx

Average across layers for English

All

xx.xx

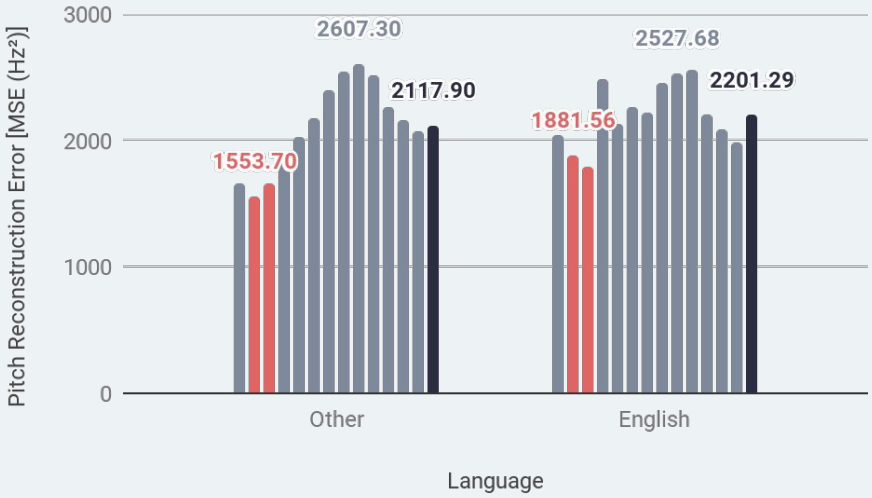
Average across layers for each language

Avg

xx.xx

Average across layers and languages

Results by Layer



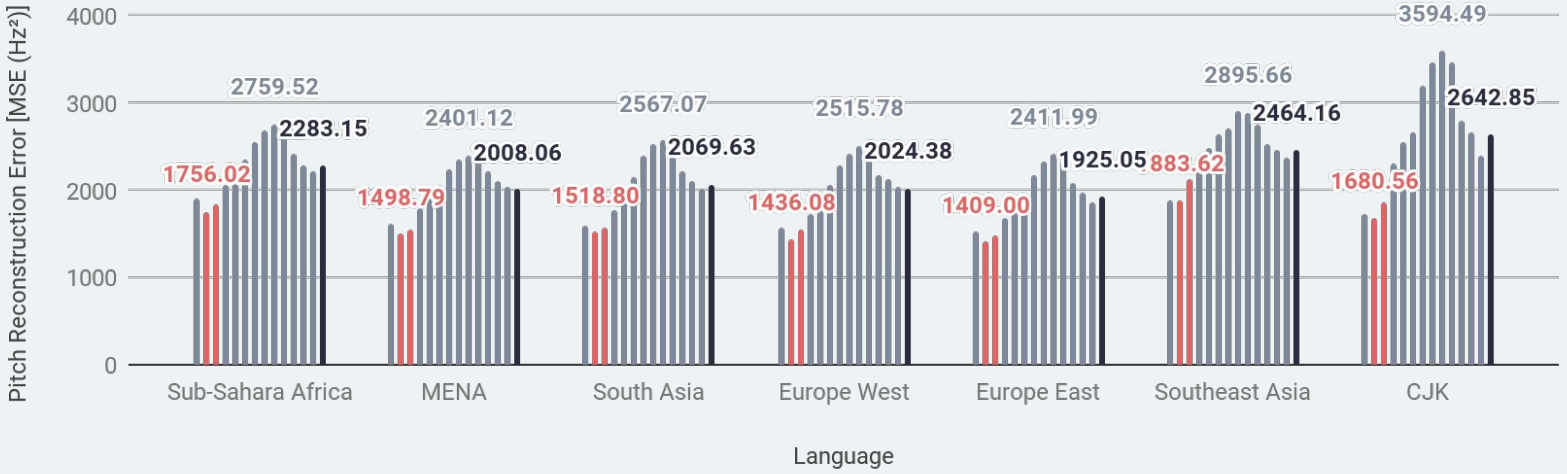
Pitch Reconstruction Error [MSE (Hz^2)] for HuBERT, average of scores for all languages, by layer + avg

Worst
9th Layer
Average across languages for each layer

Best
2nd & 3rd
For Other: 2
For English: 3

Avg
All Layers
Average across languages for each layer

Trends: Lang. Group



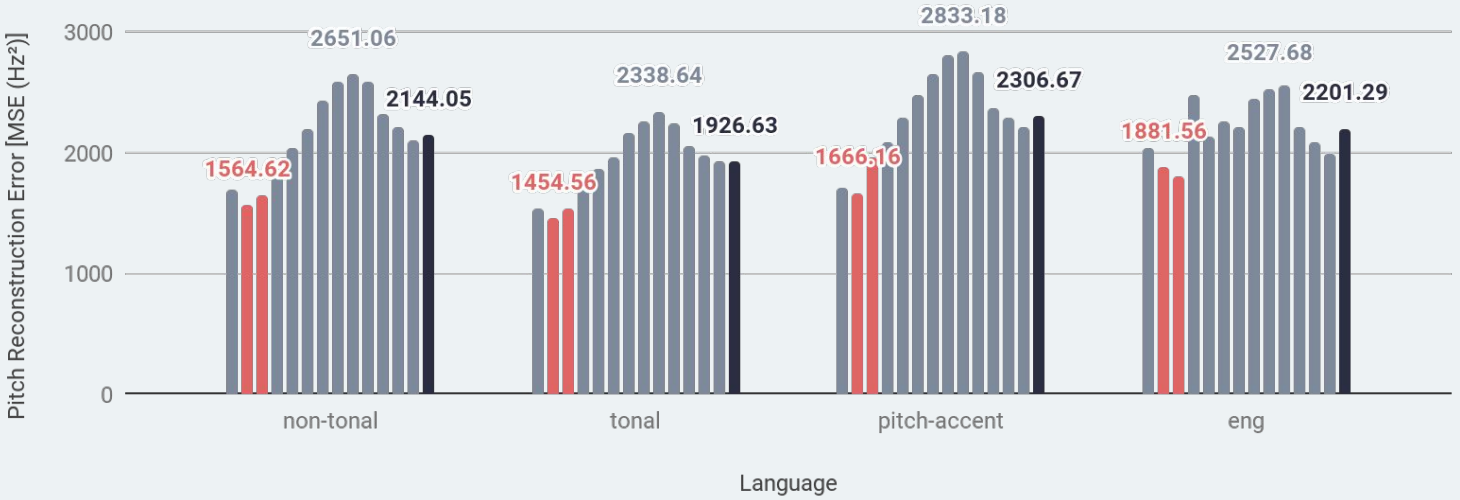
Pitch Reconstruction Error [MSE (Hz^2)] for HuBERT, for all layers + average, by language group

Layer 2
Best
For all

Worst
Layer 9
Average within groups for each layer

Avg
All Layers
Average within groups for each layer

Trends: Lang. Type



Pitch Reconstruction Error [MSE (Hz^2)] for HuBERT, for all layers + average, by language type

Best
Tonal

Best Language Type

All

All Layers

Average across languages for each layer

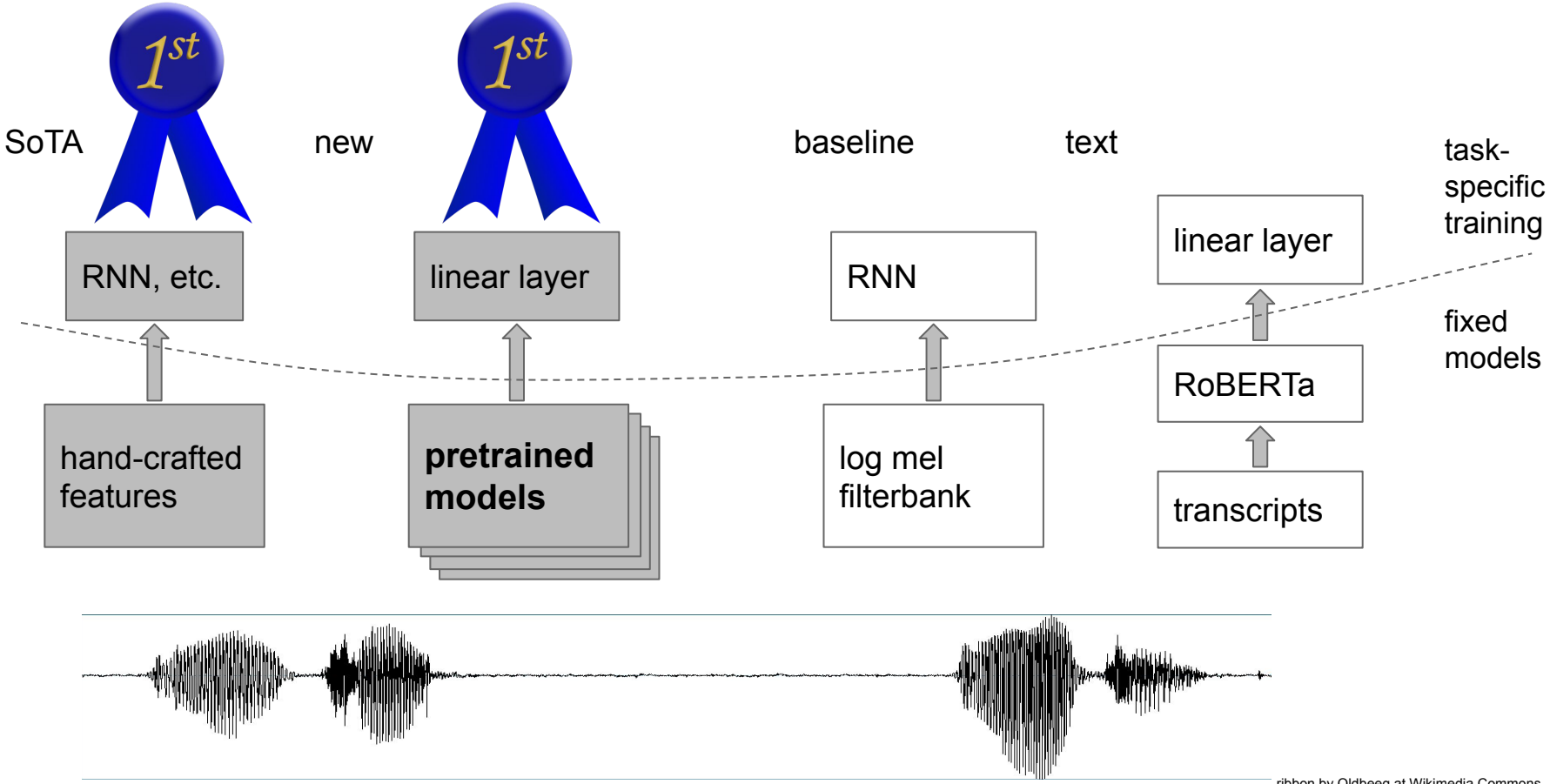
Worst
Pitch-Acc.

For some reason these are the worst

Conclusions

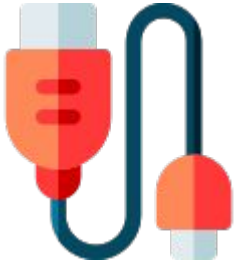
- Shows that training on one language can be enough to teach a model transferrable methods to extract pitch information
- Indicates that pre-trained models can be highly effective for tasks involving pitch information in a multilingual setting
- Specifically, indicates they can be very useful for performing such tasks on low resource languages

Overview



ribbon by Oldbeeg at Wikimedia Commons

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

More usage

11:50 - 12:00

Toolkit for Speech Pre-training 12:00 - 12:10

SSL for code-switching ASR

Léa-Marie LAM-YEE-MUI (LISN & Vocapia Research), Ondřej KLEJCH (CSTR), Lucas ONDEL (LISN), Hao TANG (CSTR), Hung-yi LEE (NTU), Ahmed ALI (QCRI)

On the difficulties of doing ASR for code-switching speech

- Code-switching (CS) occurs in everyday life for many of us.
- It can take many forms and can also be infrequent in speech and texts.
- These characteristics make it hard to collect some accurate CS data.

Can we apply typical ASR approaches for low-resource languages on CS data? Would state-of-the-art SSL models be useful for CS?

Code-switching data: soap operas

Code-switched South African languages [1]

- sesotho-english (3h)
- setswana-english (3h)
- xhosa-english (3h)
- zulu-english (5.5h)

Labeled audio: 15h of soap operas

Unlabeled audio: ~200 hours, all languages mixed, also soap operas

Few monolingual texts

[1] Barnard, Etienne, et al. "The NCHLT speech corpus of the South African languages." Workshop Spoken Language Technologies for Under-resourced Languages (SLTU), 2014.

Examples

Examples

- sesotho-english JA JA WELL I MEAN HO HONA HO TLA MO LERATONG
- tetswana-english DILO TSE NKA GO BOLELLANG TSONE KA FAMILY ELE
- xhosa-english NDAMXELELA NAY'USIBUSISO KODWA KE UDINEO ZANGA AFUN'UKU PRESS THE CHARGES SO
- zulu-english ODWA NEMVUNULO NAYO NJE IVEZA YONKE INTO OBALA

Using pretrained models for features extraction

Baseline: MFCC with HMM/GMM-TDNN

SSL models:

- XLSR-53: multilingual training data
- HuBERT: english training data

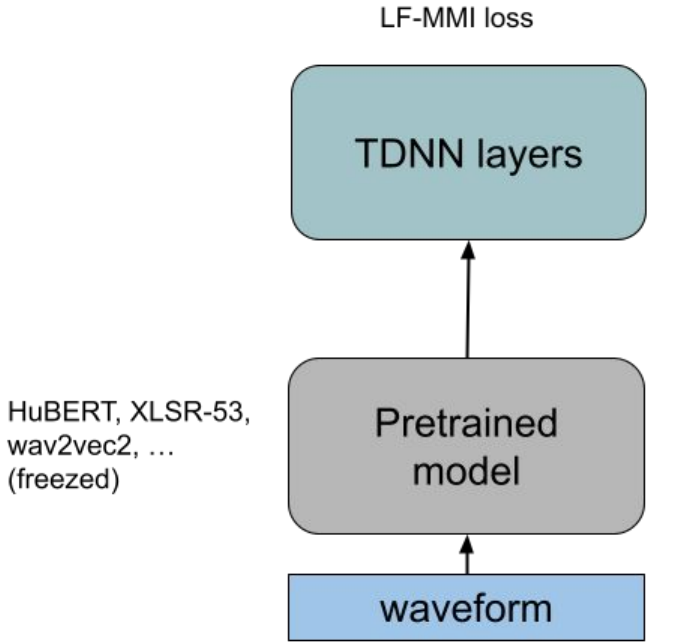
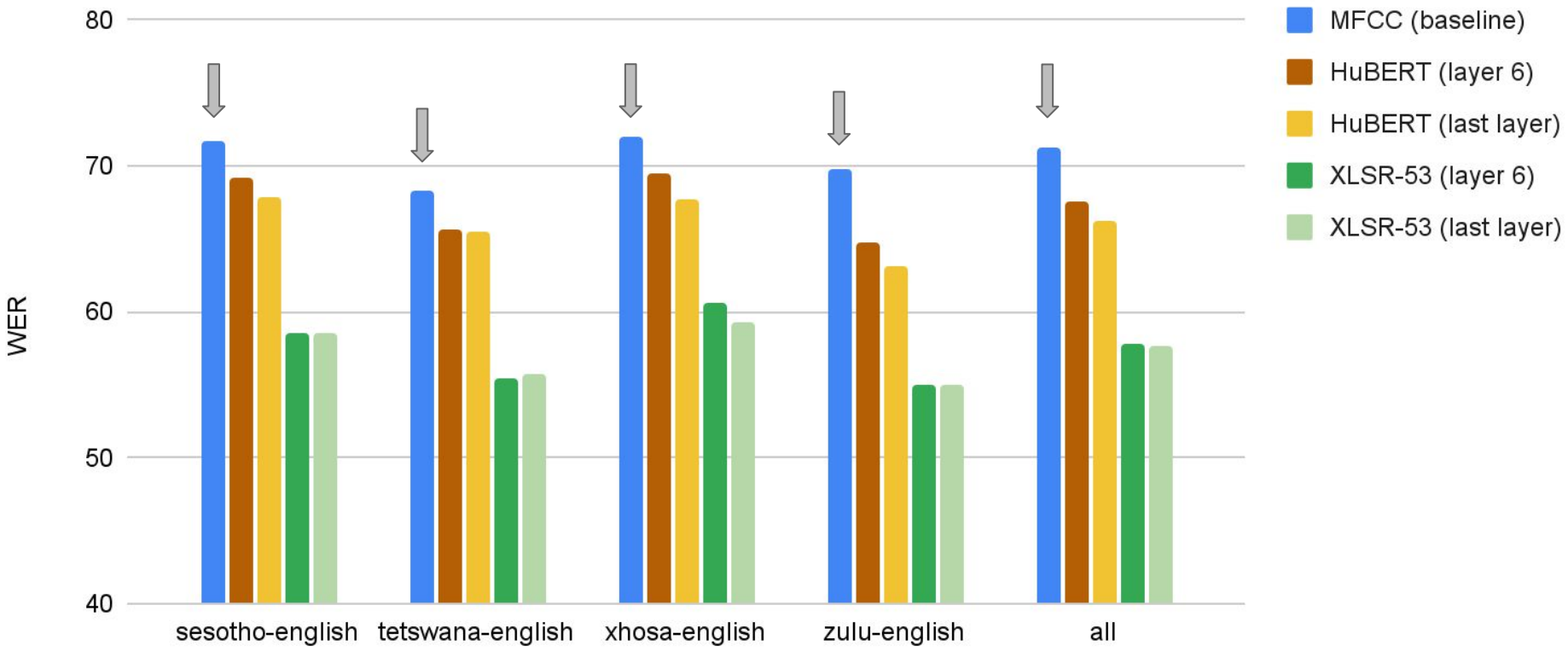
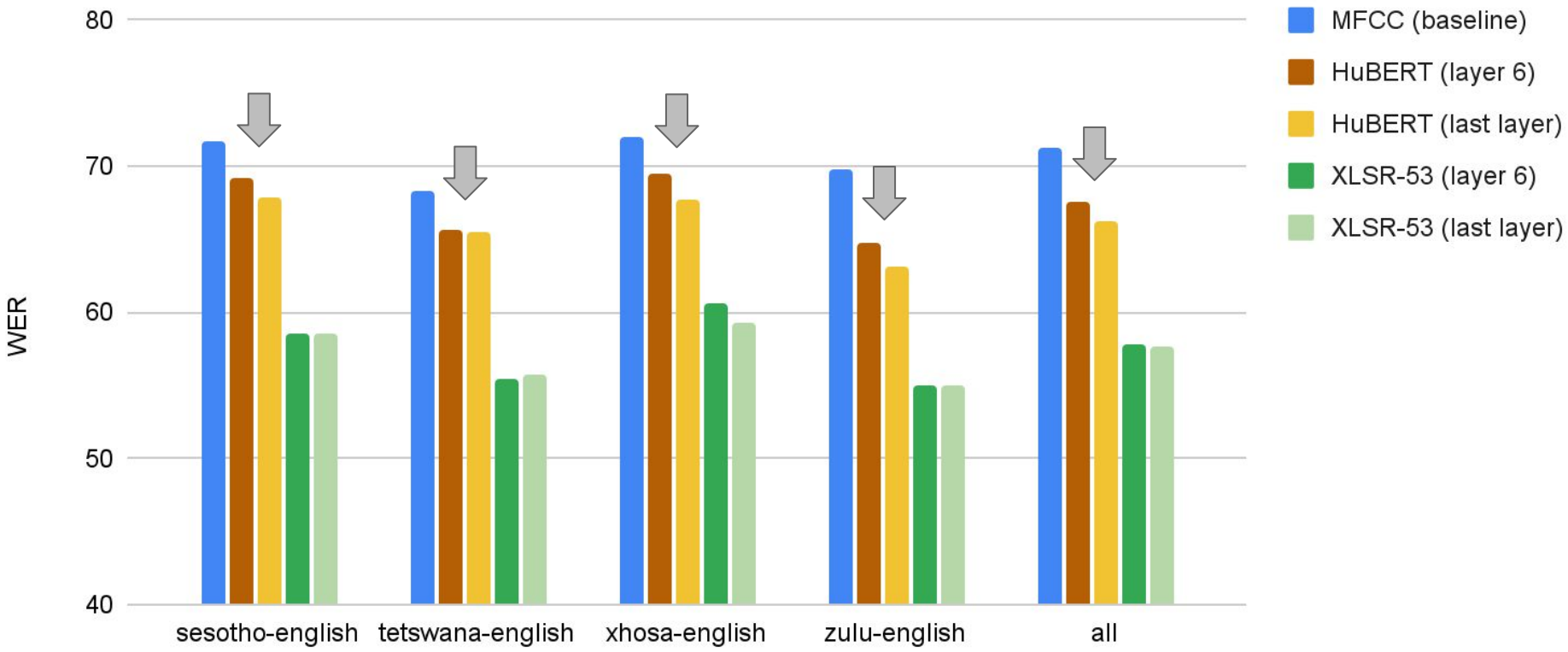


Fig: Proposed neural acoustic model

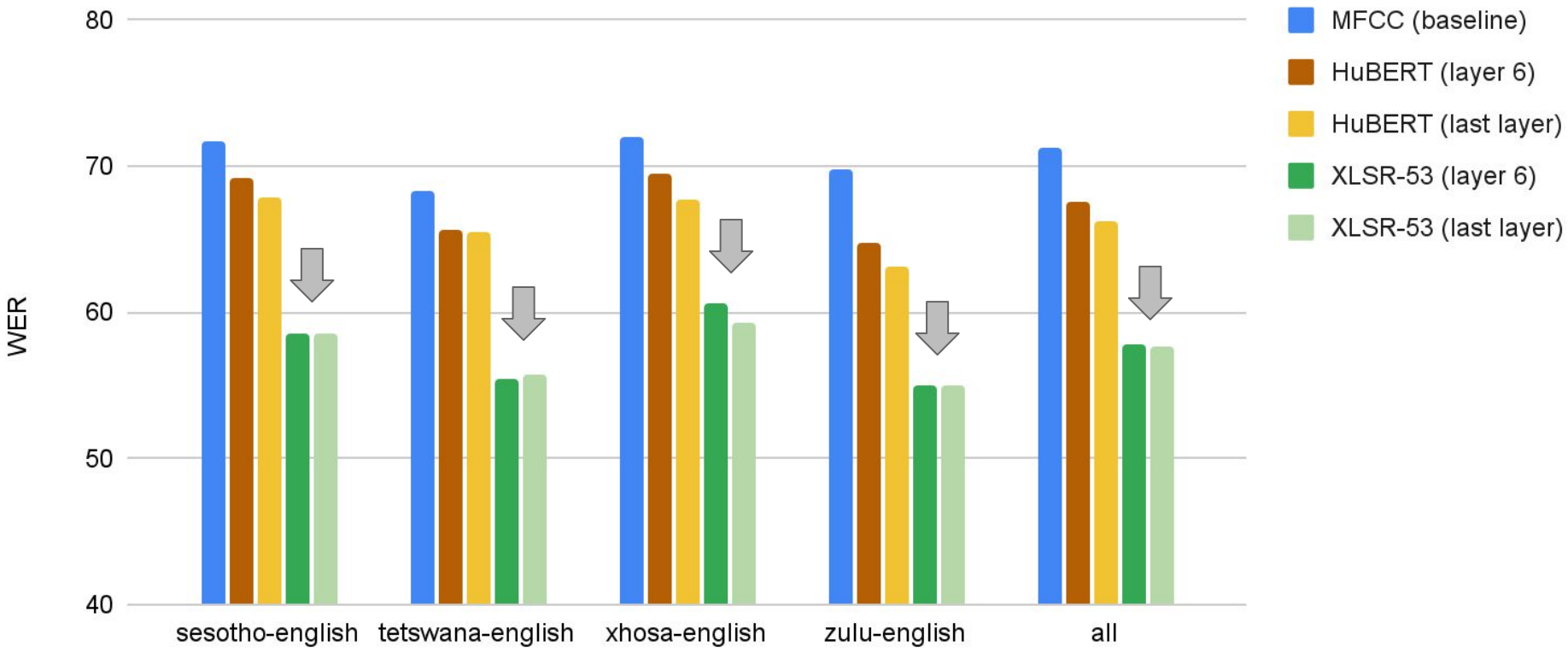
Comparison of features from pretrained models



Comparison of features from pretrained models



Comparison of features from pretrained models



Cross-lingual learning with pretrained models

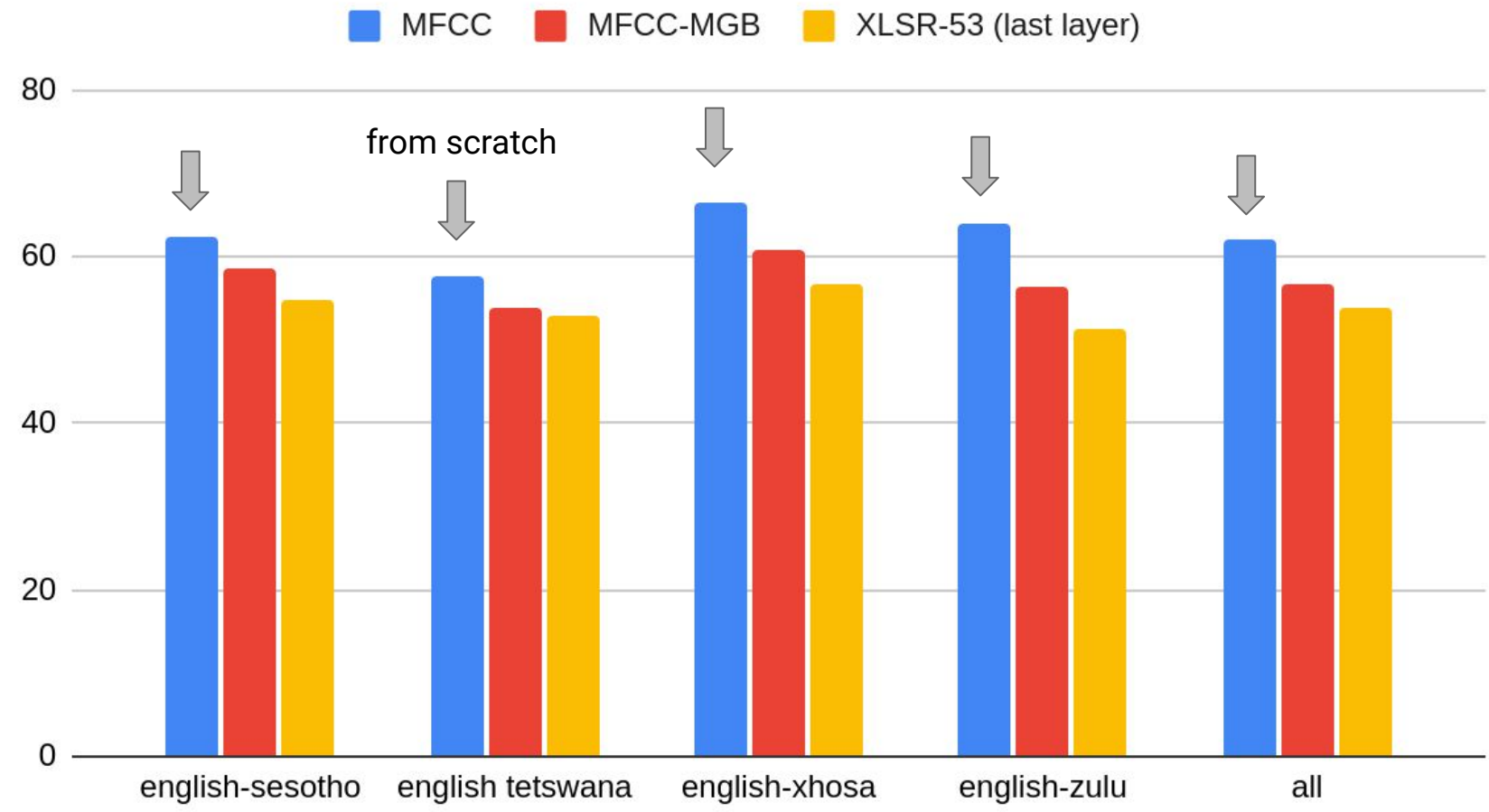
MFCC-MGB model:

- MFCC features
- CNN-TDNN trained on MGB (british english BBC) labeled data
- continue supervised training with labeled soap operas data

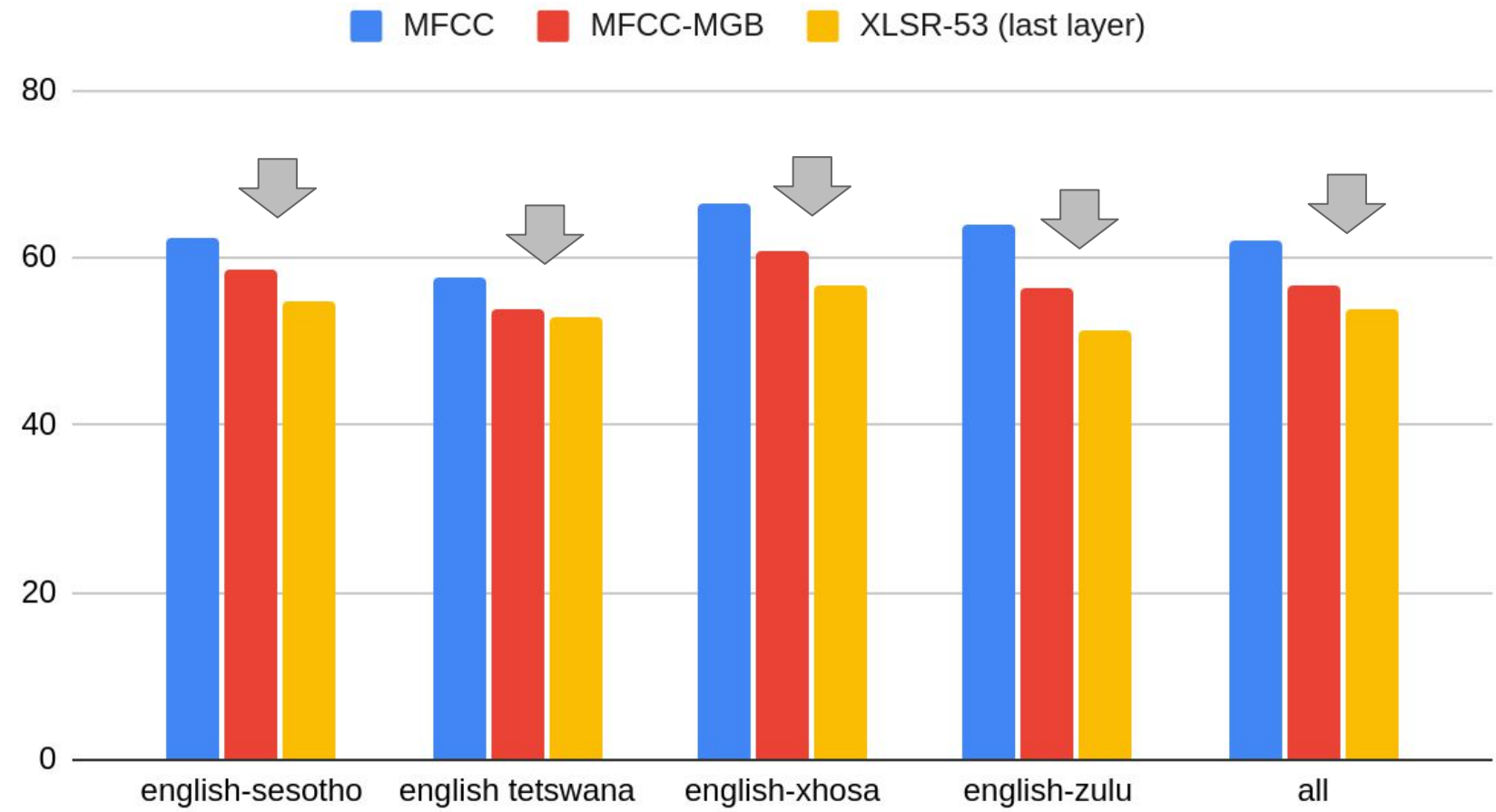
XLSR-53 model:

- waveform
- unsupervised learning

Unsupervised and supervised pretraining



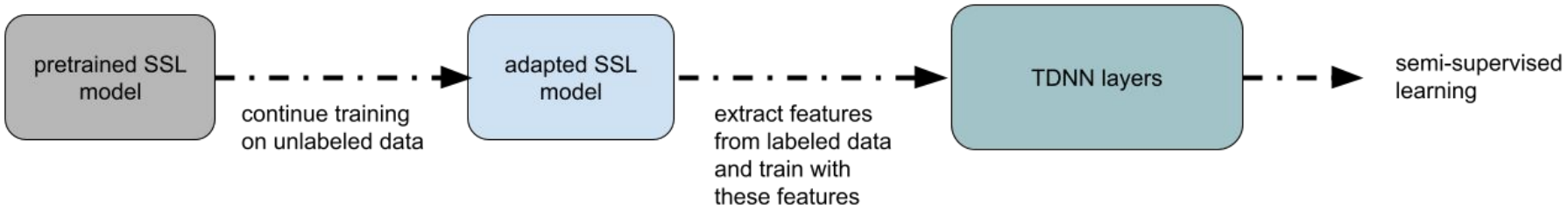
Unsupervised and supervised pretraining



Future work

- features extracted from SSL models are better than MFCC
- using XLSR-53 features is slightly better than cross-lingual transfer with MFCC features for low-resource languages

Can we have better results by using the unlabeled data?



SSL with Unsupervised ASR

Ann Lee (Meta)

Paola Garcia (JHU)

David Harwath (UT)

Shinji Watanabe (CMU)

Hung-yi Lee (NTU)

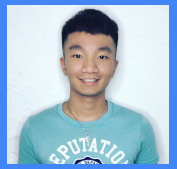
Dongji Gao (JHU)

Virginia Layne Berry (UT)

Jiatong Shi (CMU)

Yen Meng (NTU)

Hsuan-Jui Chen (NTU)



Andy. Liu (NTU)

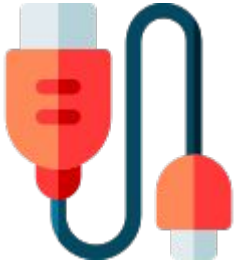
SSL with Unsupervised ASR (Overview)

- Unsupervised ASR (Dongji Gao)
  - From supervised learning to unsupervised learning
  - ESPnet - UASR

- Unsupervised ASR and visual grounding (Layne Berry)
  - Motivation
  - Aligned SpeechCLIP
  - Obstacles in unsupervised ASR + visual grounding

- Usage extension of unsupervised ASR (Jiatong Shi)
  - As a self-supervised model
  - As a segmenter
  - As a connector

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

More usage

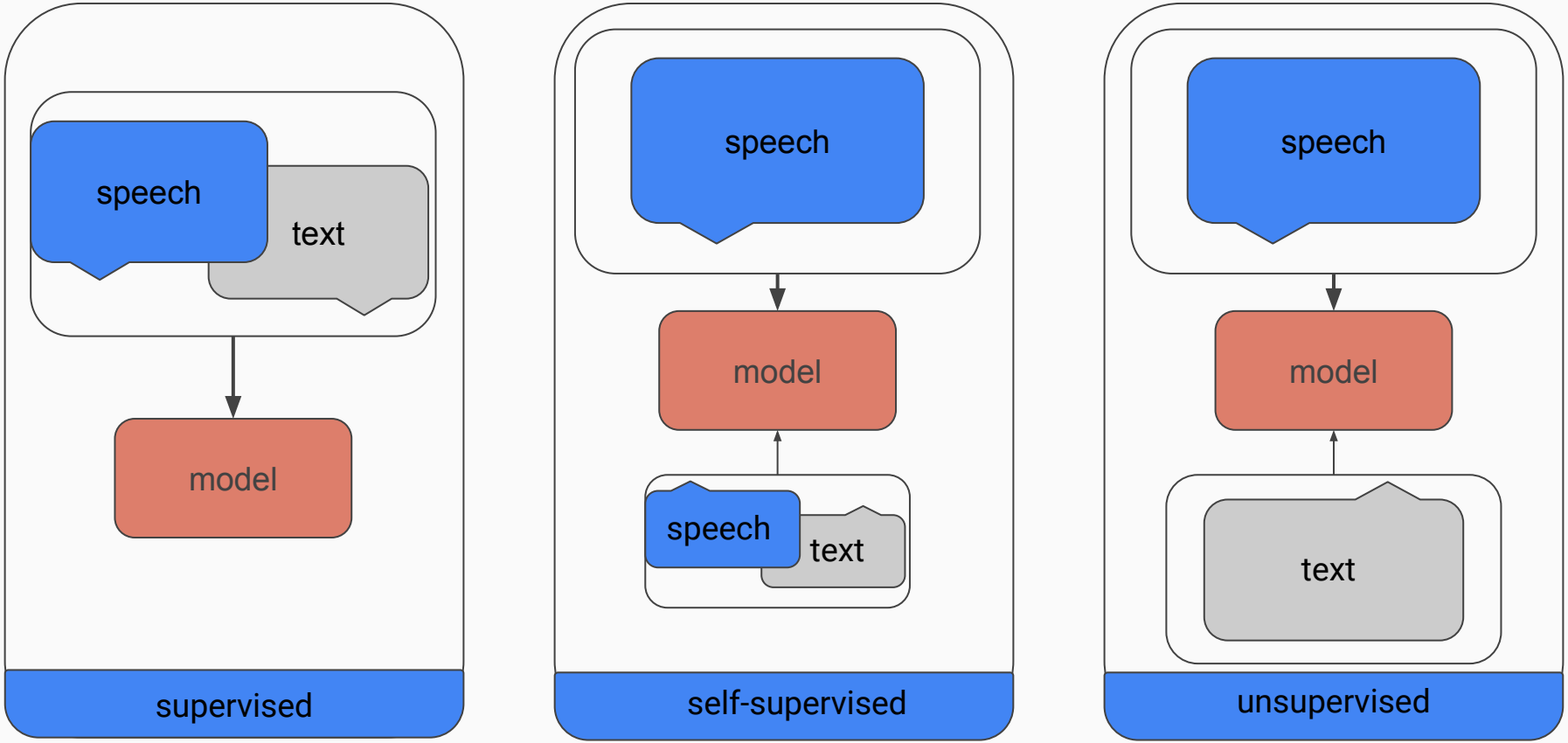
11:50 - 12:00

Toolkit for Speech Pre-training 12:00 - 12:10

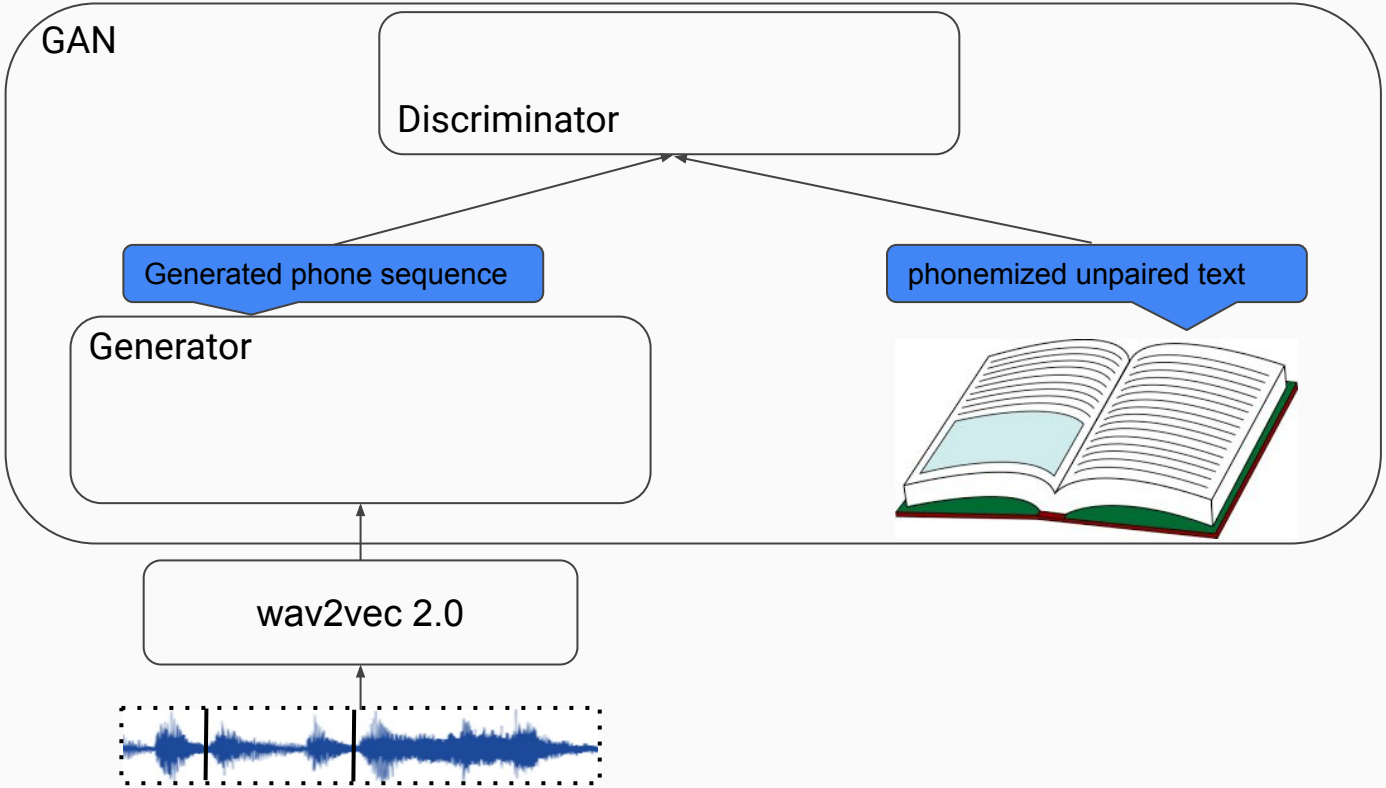
Unsupervised Automatic Speech Recognition

Dongji Gao

Supervised -> self-supervised -> unsupervised

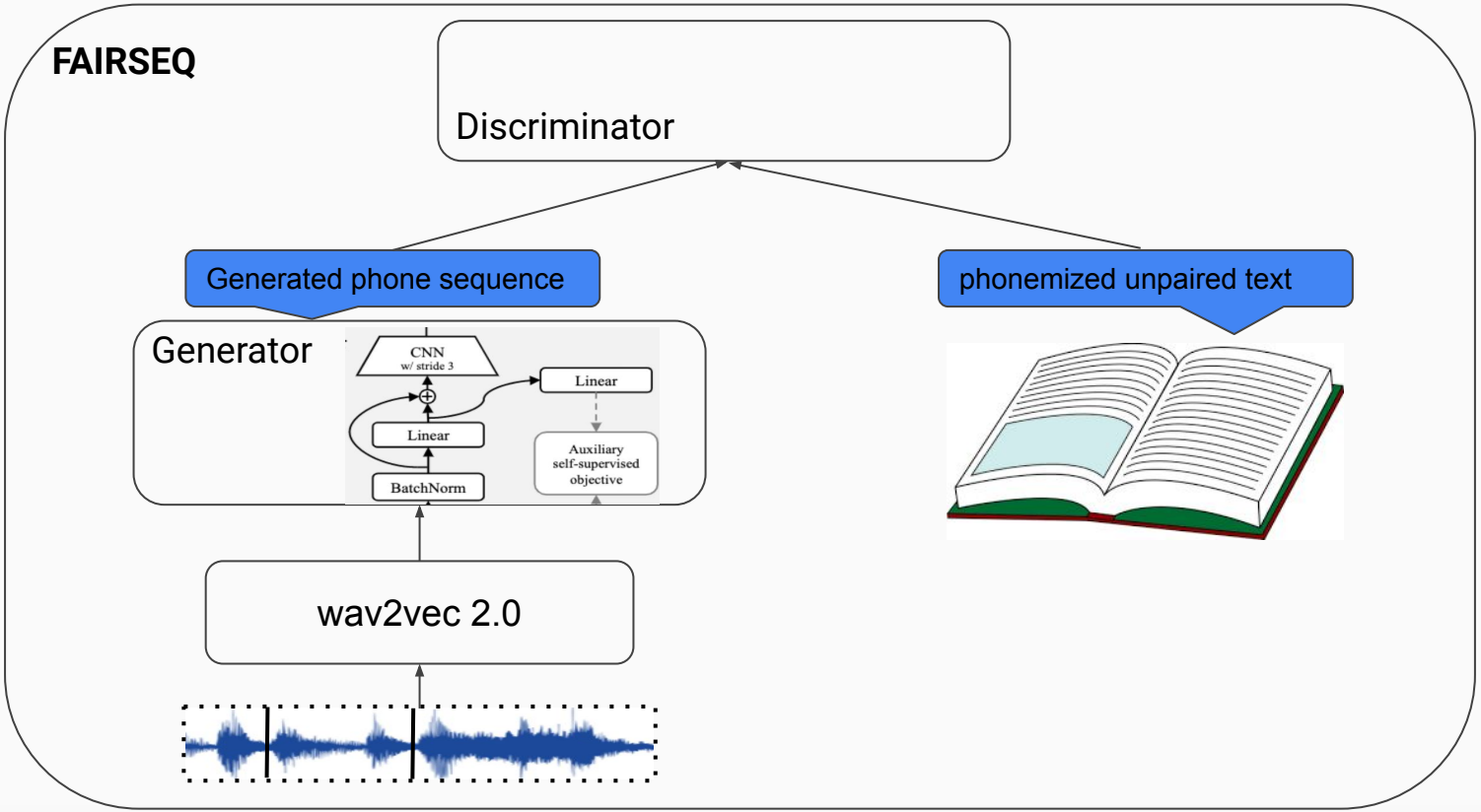


wav2vec–u2



Liu, Alexander H., Wei-Ning Hsu, Michael Auli, and Alexei Baevski. "Towards End-to-end Unsupervised Speech Recognition." (2022).

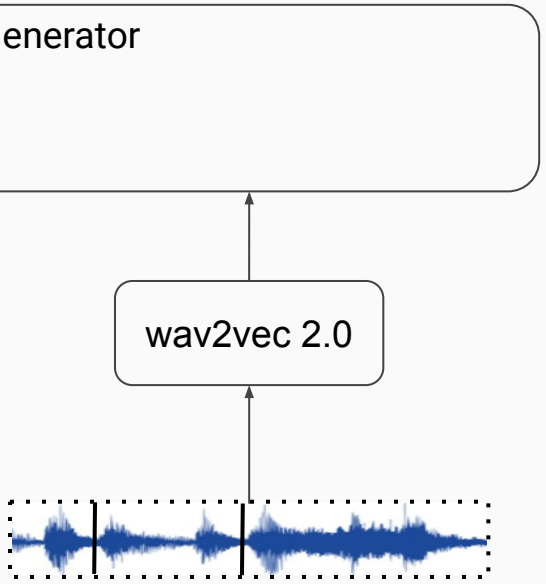
wav2vec–u2



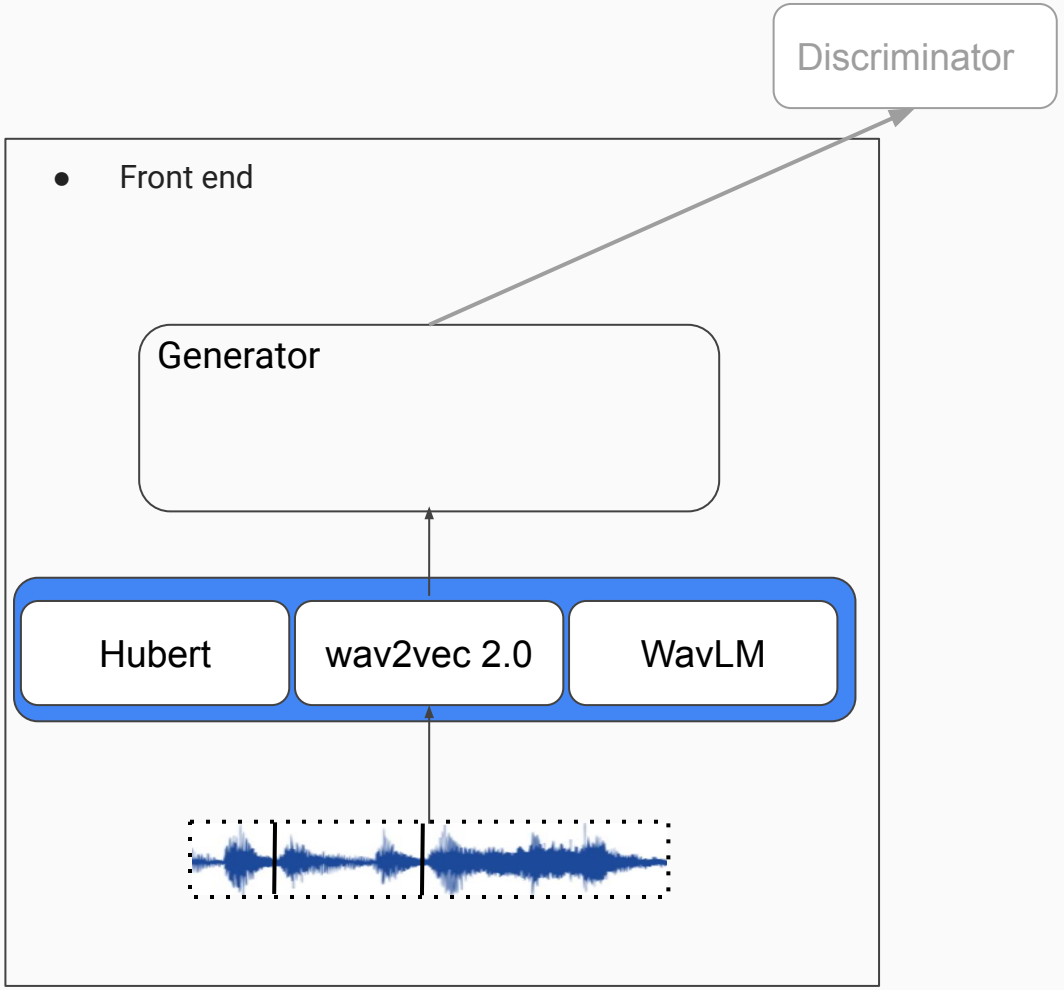
Liu, Alexander H., Wei-Ning Hsu, Michael Auli, and Alexei Baevski. "Towards End-to-end Unsupervised Speech Recognition." (2022).

ESPnet

- Front end

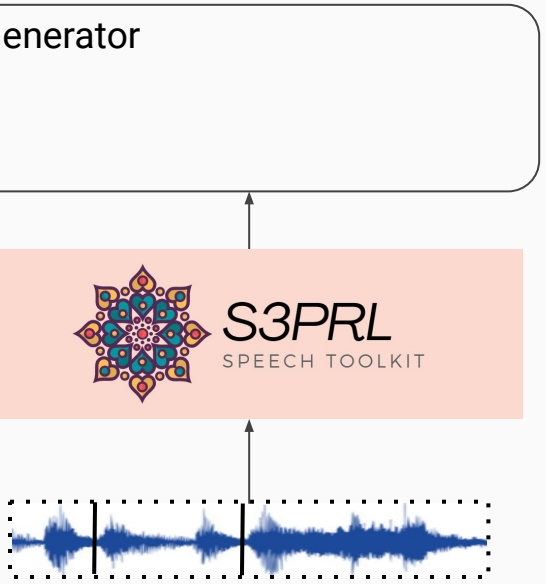


ESPnet


Discriminator
Front end
Generator
Hubert
wav2vec 2.0
WavLM

ESPnet

- Front end



ESPnet

- Front end

- Faster data preprocessing
  - Parallel
    - VAD
    - Remove silence
    - MFCC clustering
  - On-the-fly feature extraction
    - Trainable weighted sum of features from different layer

ESPnet

- Front end

- Faster data preprocessing
  - Parallel
    - VAD
    - Remove silence
    - MFCC clustering
  - On-the-fly feature extraction
    - Trainable weighted sum of features from different layer
- Training

ESPnet

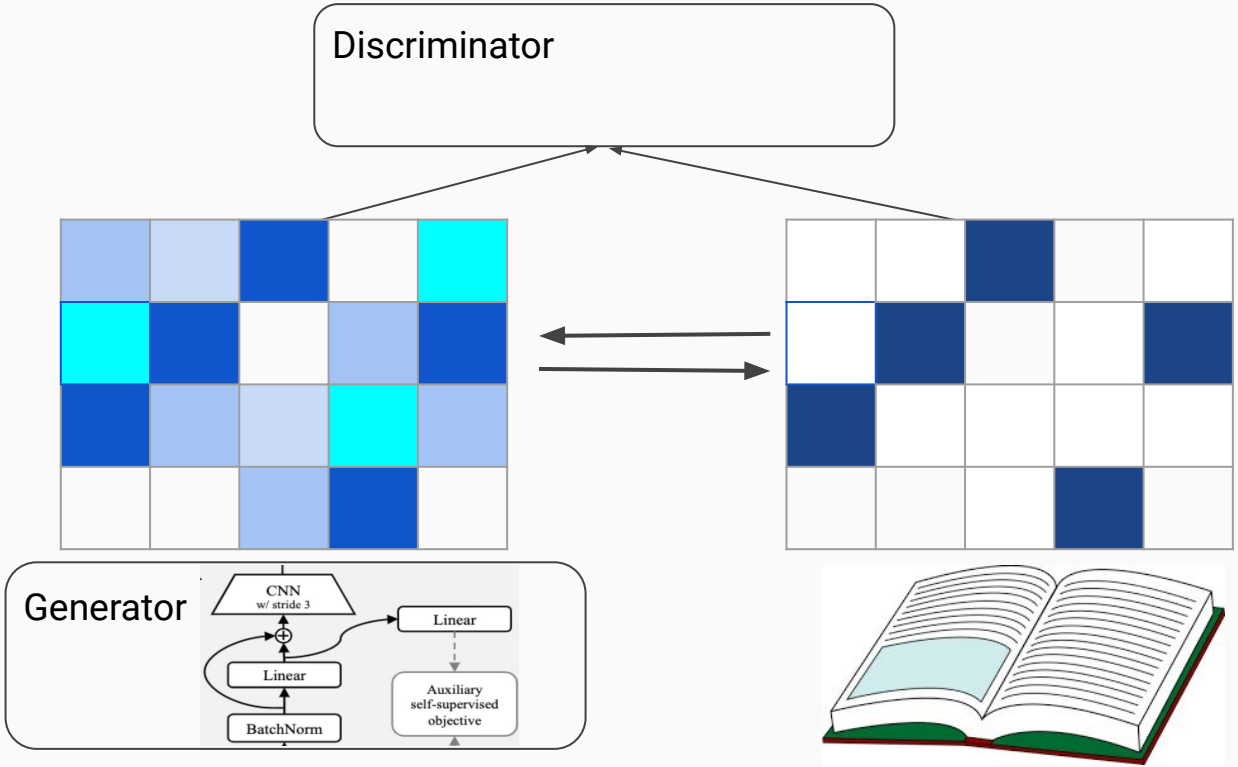
- Front end

- Faster data preprocessing
  - Parallel
    - VAD
    - Remove silence
    - MFCC clustering
  - On-the-fly feature extraction
    - Trainable weighted sum of features from different layer
- Training
- More on Training
  - Reproducibility
  - Efficiency
  - Performance

Finished

Ongoing

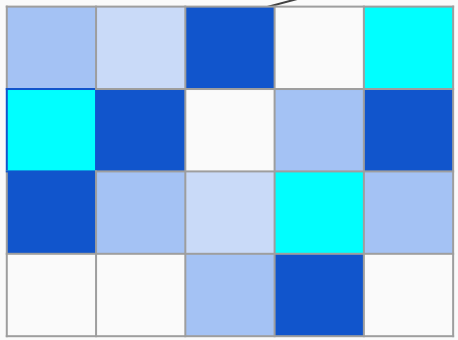
Text: “C B A D B”, phoneme set {A, B, C, D}

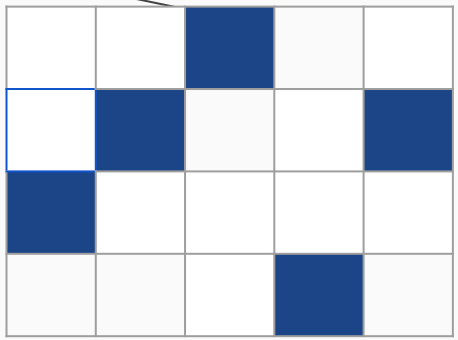


Text: “C B A D B”, phoneme set {A, B, C, D}

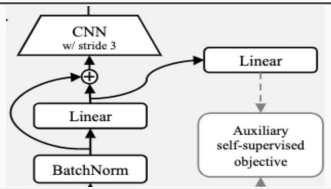
Guidance

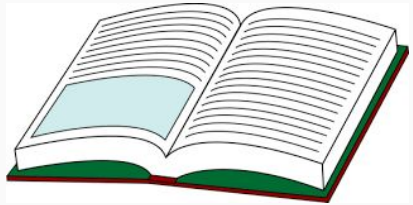
Discriminator





Generator



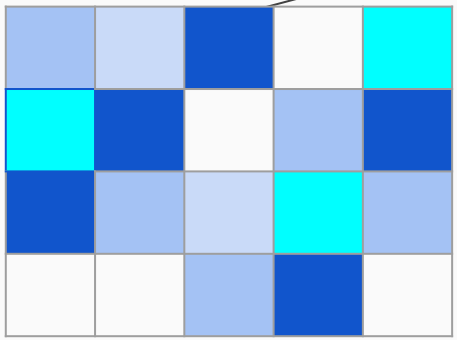


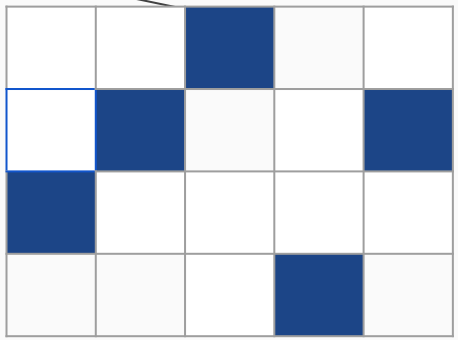
Text: “C B A D B”, phoneme set {A, B, C, D}

Guidance

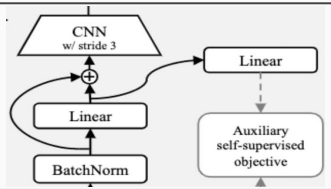
Language model

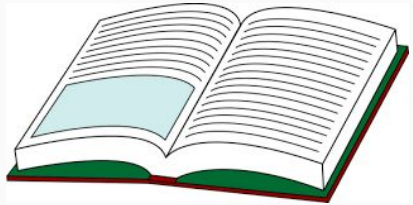
Discriminator





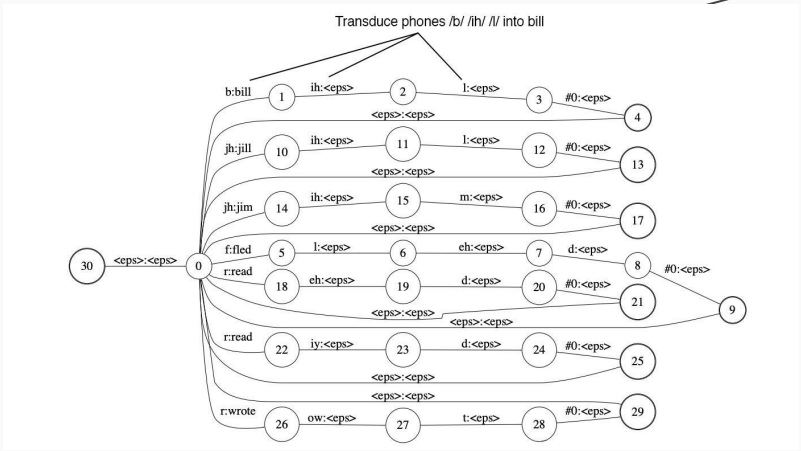
Generator

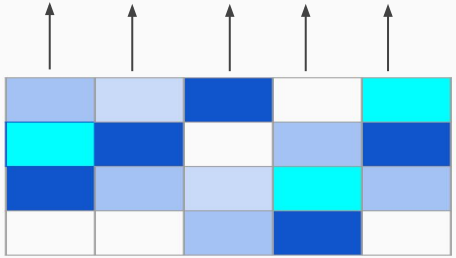


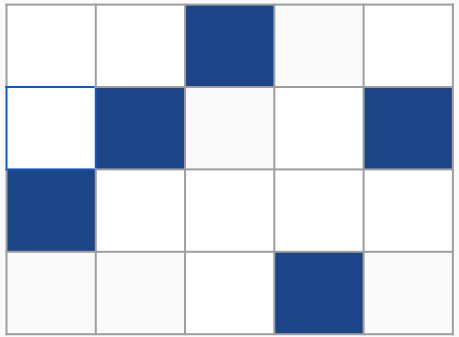


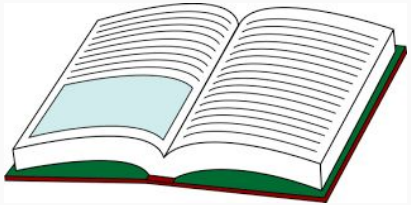
Text: “C B A D B”, phoneme set {A, B, C, D}

Discriminator









ESPnet

ESPnet

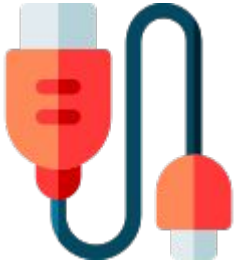
- Front end

S3PRL SPEECH TOOLKIT

- Faster data preprocessing
  - Parallel
    - VAD
    - Remove silence
    - MFCC clustering
  - On-the-fly feature extraction
- Training

k2

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

More usage

11:50 - 12:00

Toolkit for Speech Pre-training 12:00 - 12:10

Incorporating Visual Grounding into Unsupervised ASR

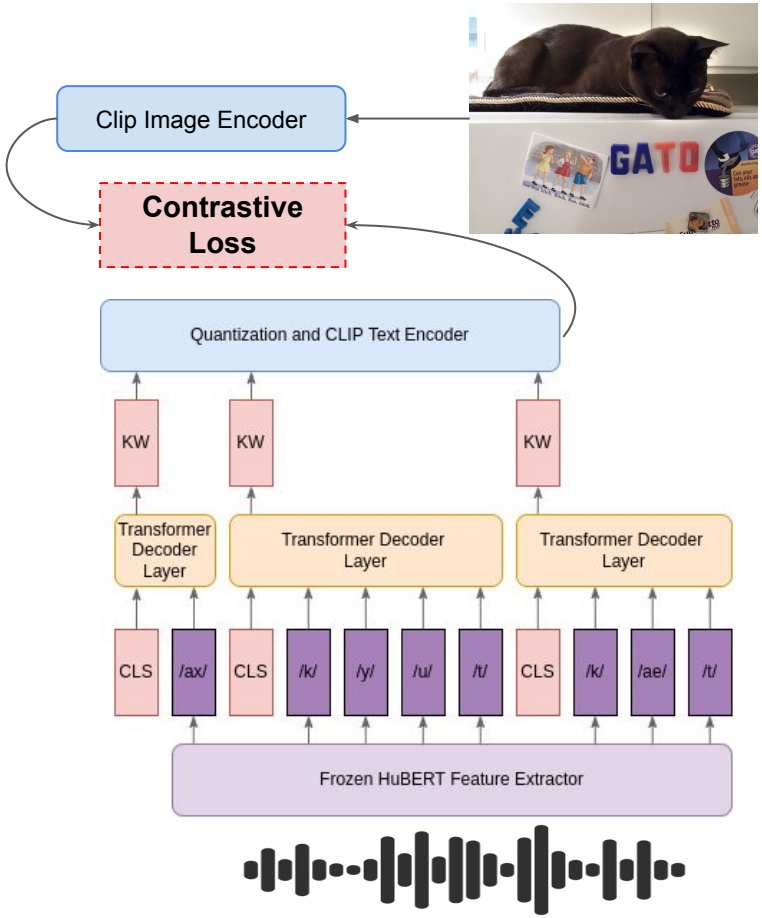
Layne Berry, University of Texas at Austin

Why Would Visual Grounding Benefit Unsupervised ASR?

- Bypass Phonemization
  - Pronunciation Lexicons are difficult and expensive to produce
  - Visual grounding increases word-level information encoded by SSL models (Peng and Harwath, 2022)
- Improve Stability
  - wav2vec-U 2.0 convergence depends on random seed, multiplying training cost
  - Contrastive loss is more stable than GAN loss
- Increase Robustness
  - wav2vec-U 2.0 non-parallel text and speech come from the same dataset
  - Semantics are less impacted by domain shift than form

Adapting Cascaded SpeechCLIP for Unsupervised ASR

- Aligned SpeechCLIP
  - Predict one keyword per segment
- Requires input segmentation
  - Ground-truth or VG-HuBERT
- Learned segment labeler predicts keywords
  - Perfect segment labeling would perform ASR with no speech-text pairs



Can Aligned SpeechCLIP Recognize Words?

- Generated texts are not well-formed:

Correct: a man feeding a giraffe food with his mouth

Predicted: shirt hand giving have animals pizza taking giving have

- Predicted keywords are semantically related to ground truth words:

$$\text{Consistency Score}(w) = \max_{t \in \text{CLIP tokens}} \frac{\text{count}(G(w) = t))}{\text{count}(w)}$$

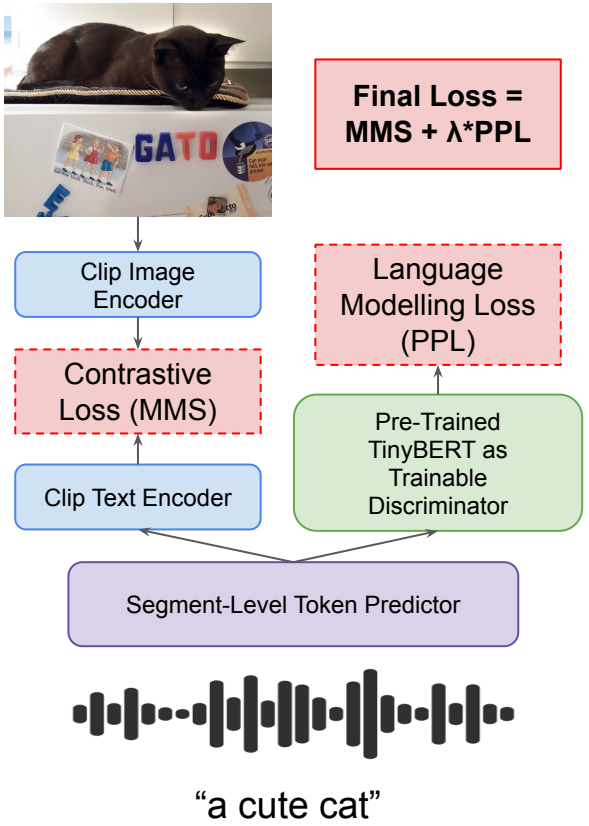
- Average Consistency Score: **61.46%**
- **31.15%** of words get a score > 75%

Pattern Type	Examples
Perfect Match	bathroom→bathroom (93.75%) **kitchen→kitchen (100%)** skateboard→skateboard (95.38%) vegetables→vegetables (82.76%) truck→truck (97.62%)
Semantically Related	elephants→cattle (89.29%) dark→seen (100%) **parked→stopped (92.68%)** rock→forest (100%)
Bucketing	street, train, traffic, intersection → cars meat, food, plate, sandwich → food soccer, tennis, court → frisbee woman, men, women, girls, guy → players **skis, snow, skier, skiing, ski → skiing**
Default Token 'into'	a, on, **the**, has, white, in, his, **and**, at, with, is, their, its, up, **to**, from, that, **green**, next, front, various, **one**, some, …. → into (total: 66!)
Unexpected	benches→appears (76.92%) **snowy→following (85.71%)** glass→together (78.57%) player, baseball, bat, batter → toothbrush

Regularizing Aligned SpeechCLIP Output

- Add TinyBERT model with LM head perplexity loss term
- Problem: pretrained LMs are trickable
- Solution: adversarial training
  - SpeechCLIP or wav2vec-U 2.0 as the generator

Candidate Caption	TinyBERT Perplexity	BERT-Base Perplexity
a man holding a cake that says happy birthday	13.9381	16.4122
group students enjoying among organized garden outside kitchen together together tracks	10.0006	11.4861

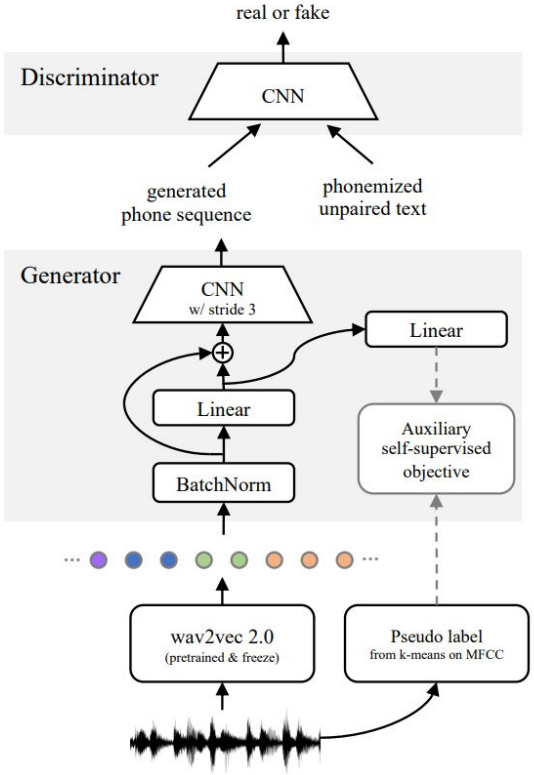


Adding CLIP Loss to wav2vec-U 2.0

Step 1: Replace LibriSpeech audio with SpokenCOCO audio
Step 2: Replace phonemization with tokenization
Step 3: Increase kernel size and stride
- Stride: 3 → 9
- Size: 9 → 27

Step 4: SpeechCLIP Quantization Strategy
- Predict token embedding
- Batch normalization
- Cosine similarity with codebook
- Softmax over similarity scores

Step 5: Add contrastive loss term to generator training steps



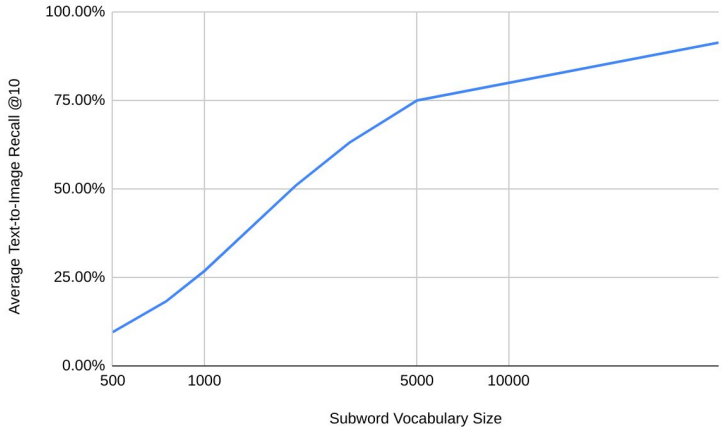
wav2vec-U 2.0 is Brittle to Domain Shift

- Replacing LibriSpeech audio with SpokenCOCO leads to failure to converge
  - 12 models trained for one week, not improving over time
  - PER of two best checkpoints: 100.215% and 79.959%
- LibriSpeech and SpokenCOCO domain mismatch
  - Both datasets consist of read English speech
  - LibriSpeech comes from LibriVox audiobooks
  - SpokenCOCO comes from MSCOCO image captions
- Utterance-Initial “the” (DH AH) vs “a” (AH)
  - SpokenCOCO: “A…” is 16.3x more common than “The…”
  - LibriSpeech: “The…” is 5.6x more common than “A…”
  - wav2vec-U 2.0 hallucinates a “DH” before any utterance-initial “AH”

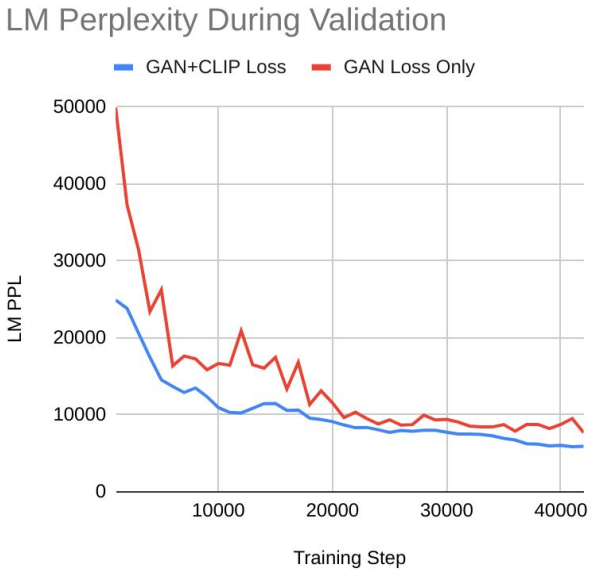
CLIP is Brittle to Subword Vocab Size

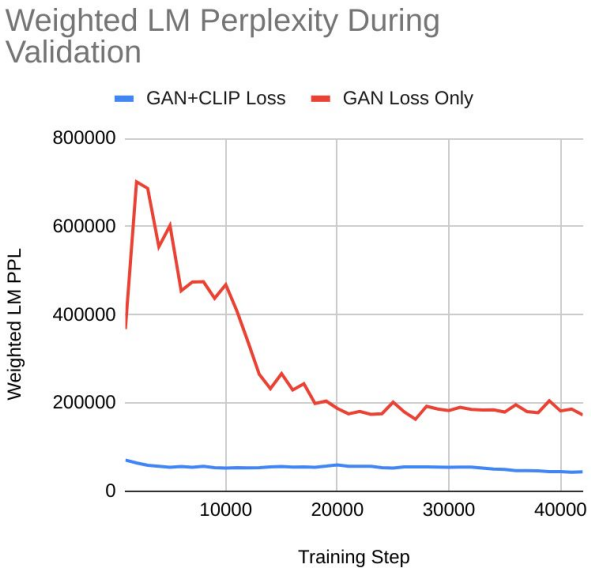
- CLIP Vocab Size: 49,408 tokens with long tail
- Reducing vocab size hurts retrieval performance

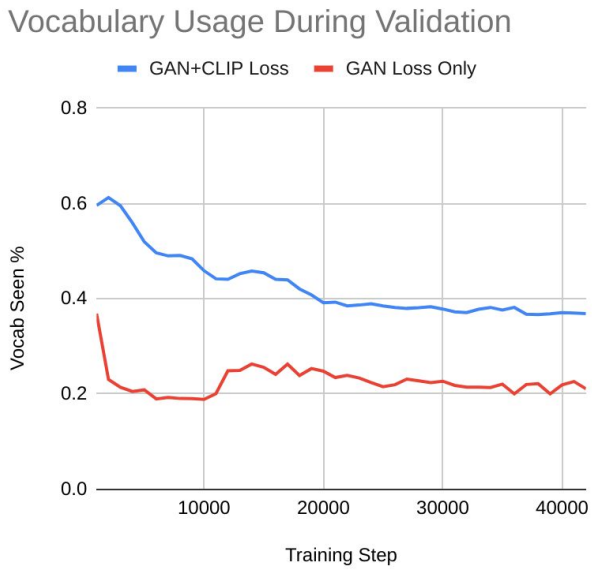
Vocab Size	# Tokens in Tail (< 10)	# Tokens in Tail (<100)	CLIP Retrieval R@10 (Averaged over 5 batches of 1000 candidates)
500	1	8	9.54%
750	4	31	18.26%
1000	13	85	26.8%
2000	108	461	51.0%
3000	250	957	63.1%
4000	486	1583	69.78%
5000	796	2272	75.02%
49k	23925	38078	91.36%



CLIP Loss Stabilizes wav2vec-U 2.0 Training







Ongoing Work and Next Steps

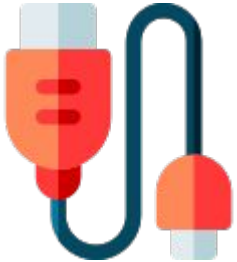
Aligned SpeechCLIP with Language Model:

- Dropout segments with the strongest signal
- Pseudo-label reconstruction auxiliary task
- Gradient penalty to stabilize GAN training

wav2vec-U 2.0 with CLIP Loss:

- Use reserved SpokenCOCO captions as non-parallel text
- Fine-tune CLIP on a smaller vocabulary (49k → 5k or 2k)
- Add CLIP loss to phoneme-level wav2vec-U 2.0 with differentiable mapping from phonemes to tokens

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

More usage

11:50 - 12:00

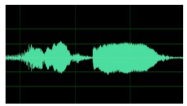
Toolkit for Speech Pre-training 12:00 - 12:10

Usage Extension of Unsupervised ASR

Jiatong Shi

After Unsupervised ASR?

Of course, use for ASR!



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

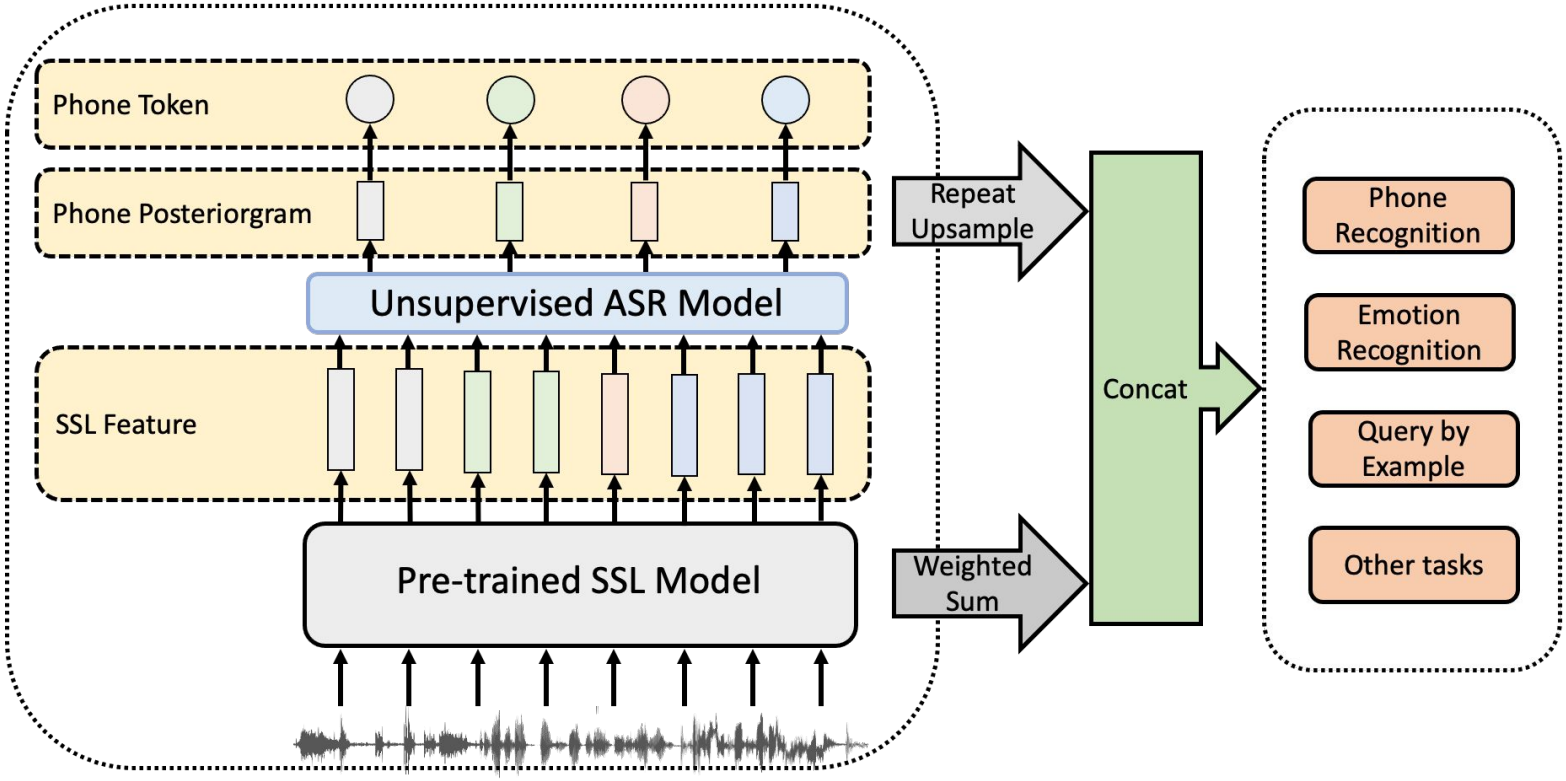
I'm not you

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

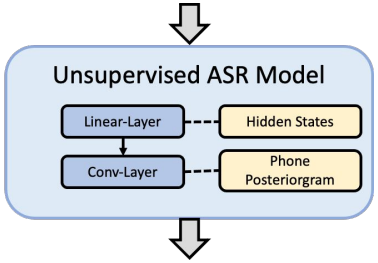
Next?

- Use as **a self-supervised model**
  - No supervised data needed
- Use as **a segmenter**
  - Unsupervised phone segmentation
- Use as **a connector**
  - Connecting Speech SSL with Text SSL

Unsupervised ASR as an SSL Model



Unsupervised ASR as an SSL Model (SUPERB Public Leaderboard)



Upstream model		Param (M)	PR (↓)	PR-10h (↓)	ASR (↓)
Wav2vec2 (Large)		317.39	5.51	7.09	3.79
UASR	Hidden states	320.18	4.57	7.50	**3.76**
	Phone posteriorgram (PPG)	320.18	**4.53**	**6.26**	3.83

Hubert (Large)	316.61	3.53	5.15	3.56

- **Better** performances in **PR**
- **Similar** performances in **ASR**
- Still **cannot fill the gap** between **Hubert**

- Phone Recognition (PR) - SUPERB public set (Librispeech-100)
- Phoneme Recognition (PR-10h) - Librilight 10h split
- Automatic speech recognition (ASR) - SUPERB public set (Librispeech-100)

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

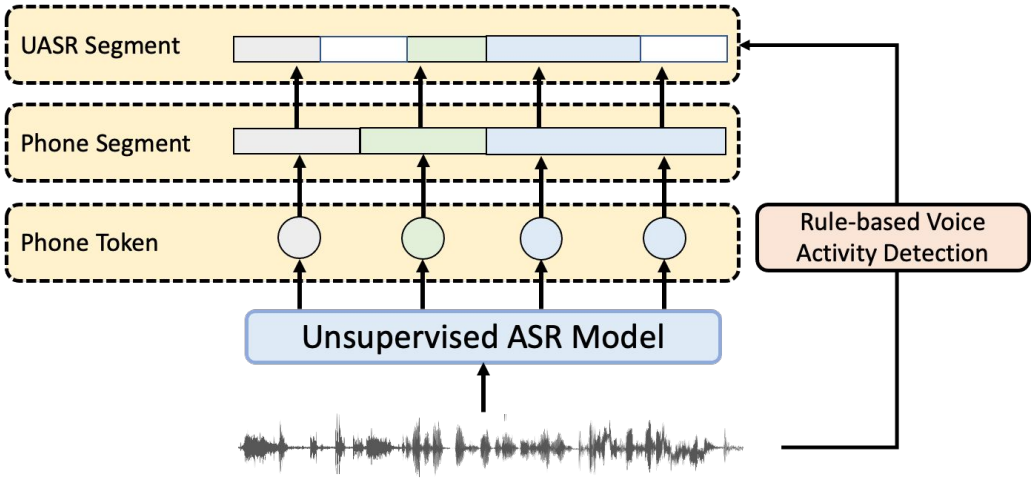
Models	Phone Recognition (↓)	Speech Recognition (↓)	Emotion Recognition (↑)	Query by Example (↓)	SUPERB Score (↑)
Wav2vec2	22.55	23.58	60.99	22.48	902
Hubert	18.22	**22.03**	64.84	33.05	959
UASR (PPG)	**17.22**	23.75	**65.11**	**21.99**	**962**

- **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)

Unsupervised ASR as a Segmenter

- **Step1**: Combine frames with same predicted phone tokens
- **Step2**: Apply rule-based voice activity detection to refine the segmentation*



*: The voice activity detection is applied in preprocessing of unsupervised ASR training.

Unsupervised ASR as a Segmenter (Cont'd)

Apply segmentation from unsupervised ASR as a guidance for sequence-reduction training

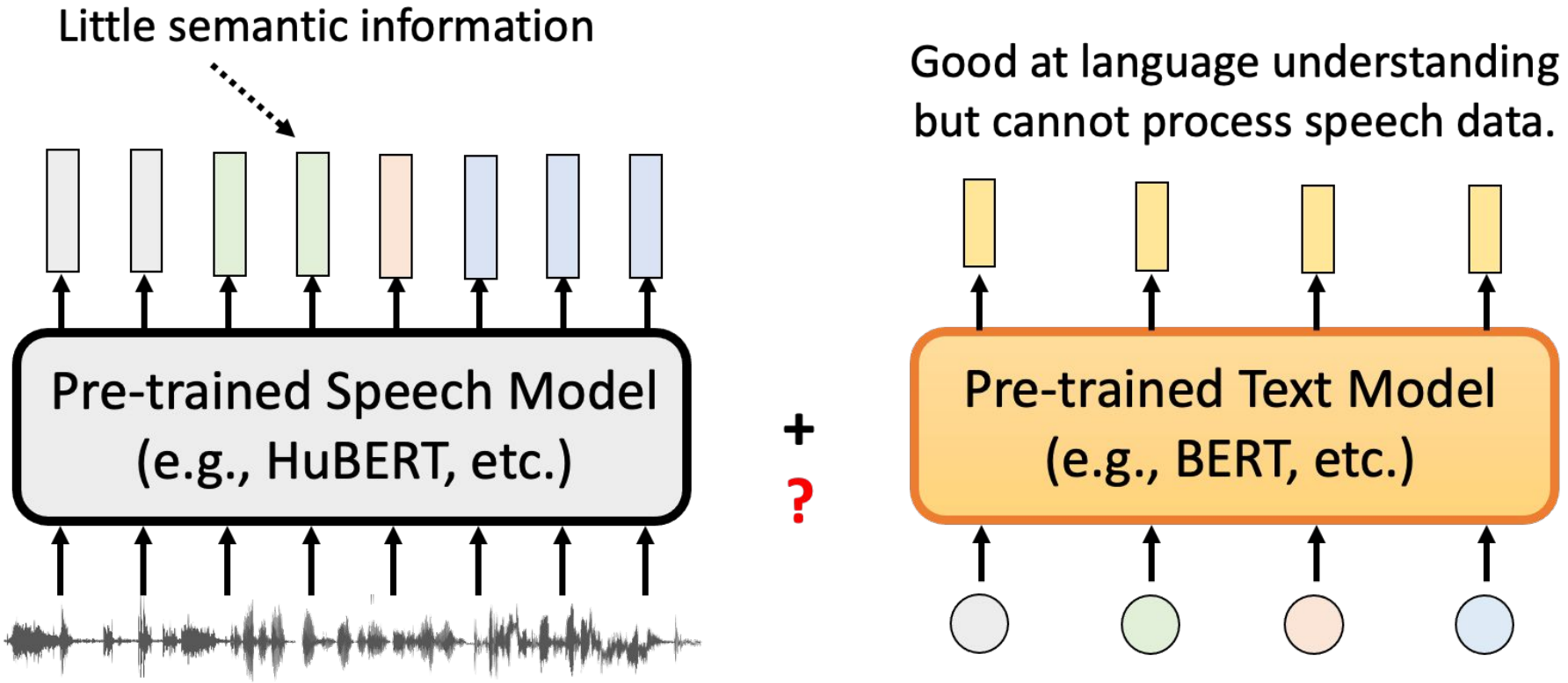
Sequence Reduction Guidance	Avg. Frame Length	Phone Recognition (↓)	ASR (↑)
Hubert token*	90ms	31.73	16.66
UASR boundary	**90ms**	**23.37**	**15.71**

Force-aligned Boundary	100ms	12.33	11.58

- **UASR boundary** is **better than Hubert-token-based boundary** in the same frame rate (90ms)
- Still have **a large gap** with **supervised** segmentation

*: SSL cluster-based guidance is based on "Herman Kamper and Benjamin van Niekerk, '*Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks*,' in Interspeech, 2021. " Note that we use lambda as 35 to get similar frame-rate with unsupervised ASR.

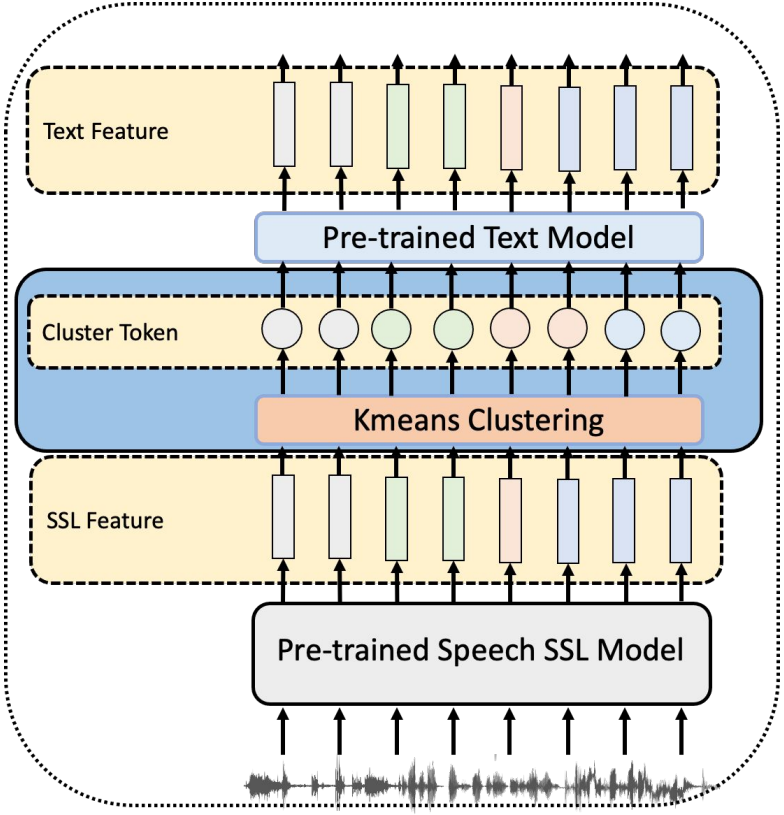
Unsupervised ASR as a Connector



Unsupervised ASR as a Connector (Cont'd)

Existing method* to connect speech-SSL and text-SSL

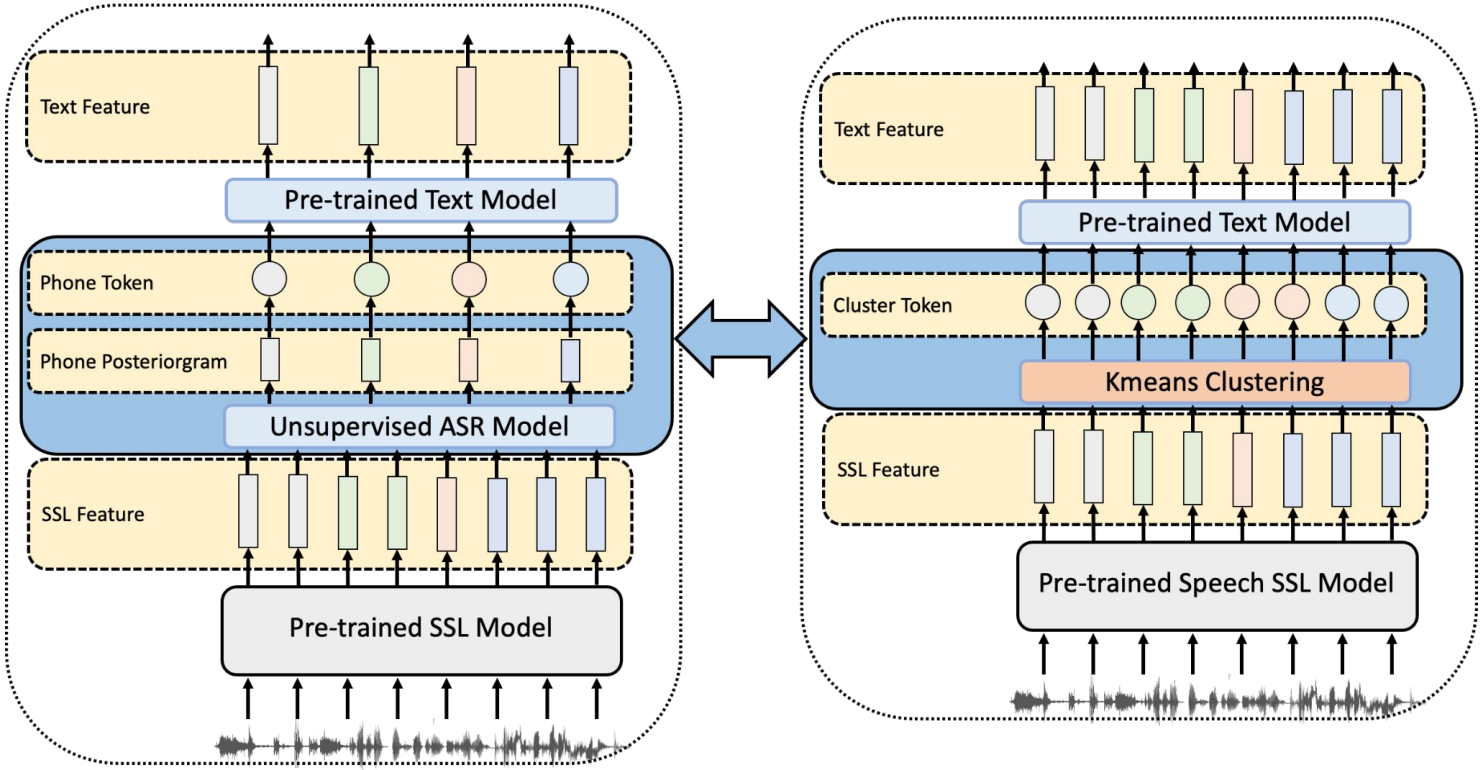
- Method: Use speech-SSL feature clusters
- Domain is still mismatched
  - Acoustic v.s. Semantic



*: Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu-wen Yang, Hsuan-Jui Chen, Shuyan Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, Lin-shan Lee. "DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering" in Interspeech 2022

Unsupervised ASR as a Connector (Cont'd)

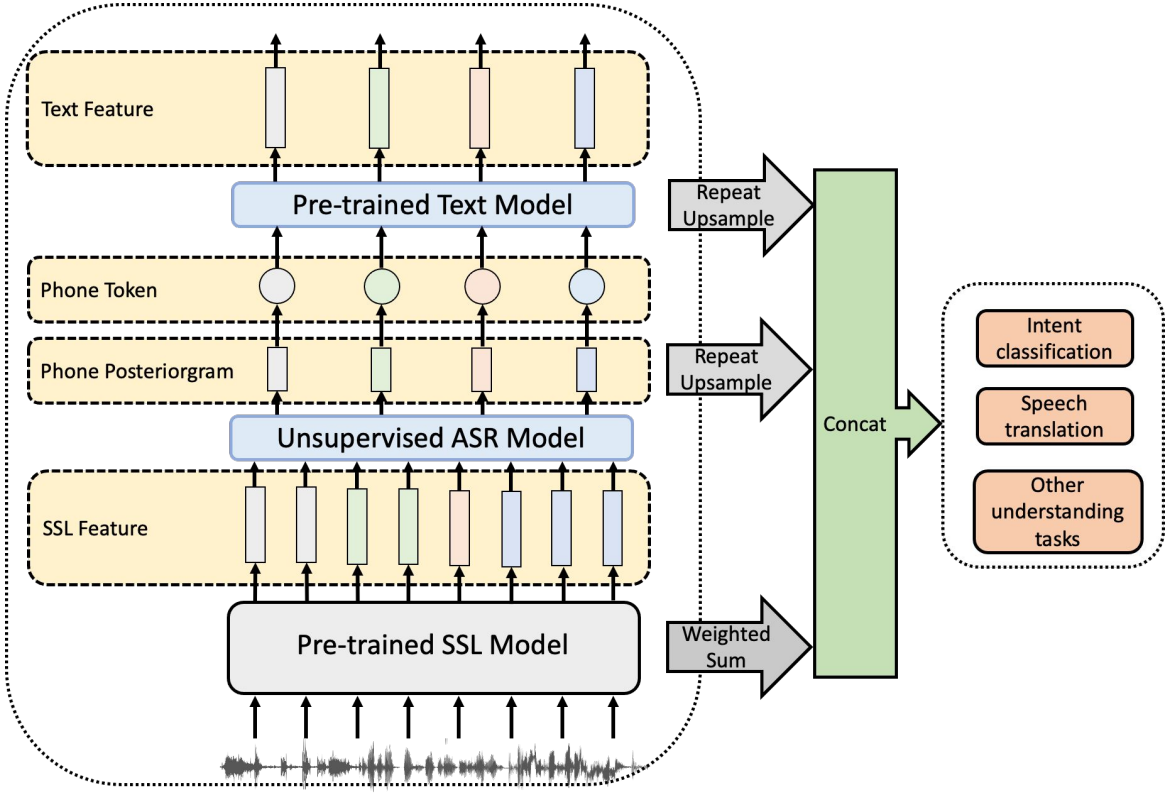
- Close the Domain Mismatch



Unsupervised ASR as a Connector (Cont'd)

Mainly focus on understanding tasks

(e.g., intent classification, speech translation, etc.)



Experimental Settings for UASR as a Connector

- Speech SSL model: wav2vec2-large (LibriLight 60k pre-trained)
- Connector:
  - **KM**: Kmeans clusters (Pre-trained Kmeans at https://dl.fbaipublicfiles.com/textless_nlp/gslm/w2v2/km50/km.bin)
  - **UASR**: UASR token ID
- Pre-trained text model:
  - **PT5** - Phoneme ByteT5, used as **default** setting (https://huggingface.co/voidful/phoneme_byt5_v2)*
- Training Options
  - **Fixed** feature extractor — **Intent Classification (IC): FSC** (fluent speech commands), 19h, SUPERB public benchmark
  - **Fine-tuning** feature extractor — **Intent Classification (IC)**: **SLURP**, 58h, ESPnet recipe

*: PT5 has a mismatched phone set, so we random mapping our UASR token ID to it while training

Unsupervised ASR as a Connector (Connector Options)

Tasks	Fixed - FSC (↑)	Fine-tuning - SLURP (↑)
Baseline (wav2vec2)	94.38	82.82
KM	93.69	85.31
UASR	**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

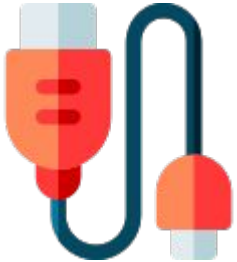
Take-home messages

- **UASR as a SSL model** can get better performances in phoneme recognition as well as some improvements on other tasks

- **UASR as a segmenter** has better quality than Hubert cluster-based segmentation

- **UASR as a connector** can connect speech SSL model and text pre-trained model. Better results could be achieved either with or without fine-tuning.

Outline of Part 2: Use of Pre-trained Model



Efficient way to use pre-trained model

10:50 - 11:05

Prosody-related Tasks

11:05 - 11:20

Code-switching ASR

11:20 - 11:30

Unsupervised ASR 11:30 - 11:40

Visual-enhanced

11:40 - 11:50

More usage

11:50 - 12:00

Toolkit for Speech Pre-training 12:00 - 12:10

Toolkit

S3PRL and its recent upgrade in JSALT

 Leo Yang (NTU)

 Andy T. Liu (NTU)

 Harry Chang (MIT)

 Haibin Wu (NTU)

 Cheng Liang (NTU)

s3prl

s3prl

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

youtu.be/PkMFnS6cjAc

1.4k stars 315 forks

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

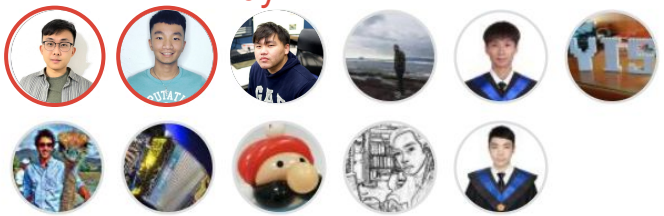
Used by 14

Contributors 38

Creators

Leo

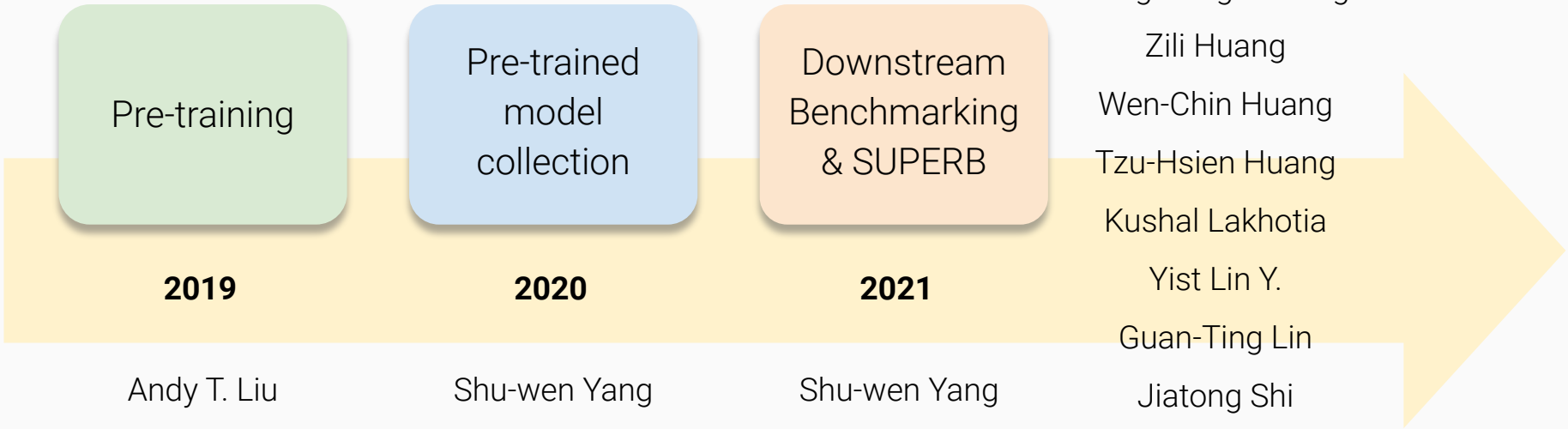
Andy



+ 27 contributors

Prof. Hung-yi Lee, Advisor & Sponsor

Major functionality



Andy T. Liu
Shu-wen Yang
Po-Han Chi

Shu-wen Yang
Andy T. Liu

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

Pre-training

Pre-training

Mockingjay

TERA

NPC

DistilHubert

AudioAlbert

APC

VQ-APC

Pre-trained model collection

Pre-trained model collection

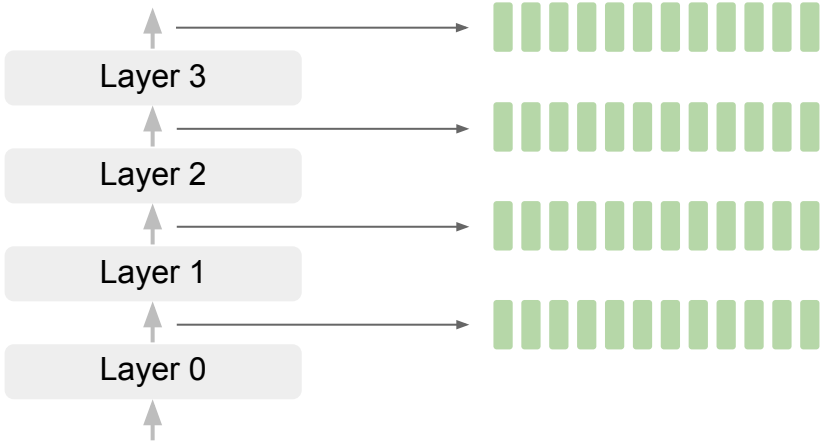
Isn't wav2vec 2.0 / HuBERT / WavLM / data2vec always the best?

No! Different task, different story, like VC

Generative	Mockingjay	TERA	AudioAlbert	APC	
	VQ-APC	NPC	DeCoAR	DeCoAR 2.0	
Contrastive	Modified CPC	wav2vec	vq-wav2vec		
	discreteBERT	wav2vec 2.0			
Predictive	HuBERT	Unispeech-SAT	WavLM	data2vec	
Multi-task	PASE+				
Distillation	DistilHuBERT	LightHuBERT	FitHuBERT		

Pre-trained model collection

- Unify retrieval & I/O interface for all the SSL models
- Extract all the hidden states for all models




```
model = s3prl.hub.wav2vec2().cuda()

wav1 = torchaudio.load("your audio path").view(-1).cuda()
wav2 = ...

all_representations = model([wav1, wav2])
# padding and masking is done automatically in the correct way
```

Downstream benchmarking & SUPERB

Downstream Benchmarking & SUPERB

Tasks in Mockingjay, TERA

- phoneme frame classification
- MOSEI sentiment
- speaker frame/utterance classification

SUPERB

- phoneme recognition
- speaker ID
- emotion recognition
- intent classification
- keyword spotting
- ASR
- ASV
- query by example
- slot filling
- speaker diarization

SUPERB-SG

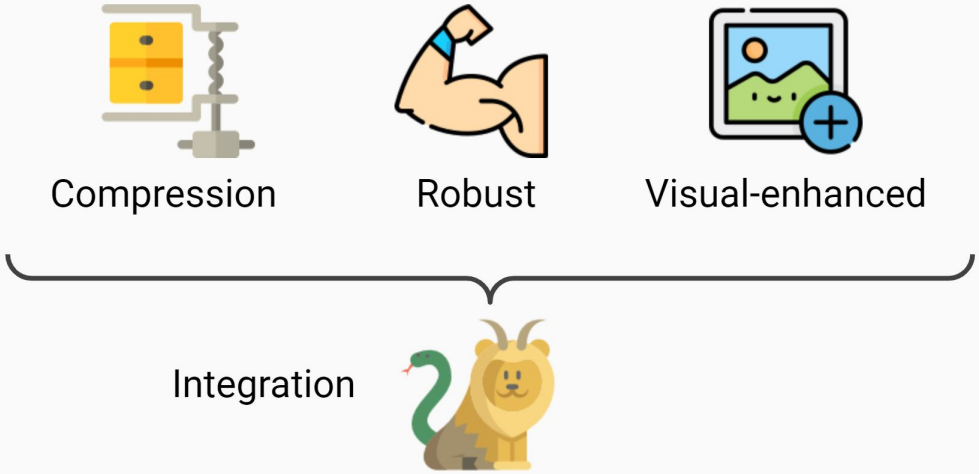
- speech translation
- speech enhancement
- source separation
- out-of-domain ASR
- voice conversion

others

- atis
- audio snips

Feedback ⟷ Improvement during JSALT workshop

- It was intensively used in the JSALT pre-training team for evaluating new techniques
  - Bug report
  - Feature request



Feedback ⟷ Improvement during JSALT workshop

- Feature requests
  - How to change the corpus for XXX task?
  - How to change the probing model for the XXX task?
  - The steps to benchmark a new SSL model is too complicated
  - Connection to the HuggingFace models
  - How to benchmark with just a subset of the corpus?
  - The latest SSL models?
- Bug report
  - Unstable loss curve…
  - Fairseq installation issues…

Major Updates in JSALT

- **More** & **stabler** upstream
- **Modularized** & **Customizable**
- **Sound** SSL & Tasks

More upstream

In SUPERB

Mockingjay | TERA | HuBERT | APC | VQ-APC | NPC | DeCoAR

DeCoAR 2.0 | Modified CPC | wav2vec | vq-wav2vec | wav2vec 2.0 | PASE+

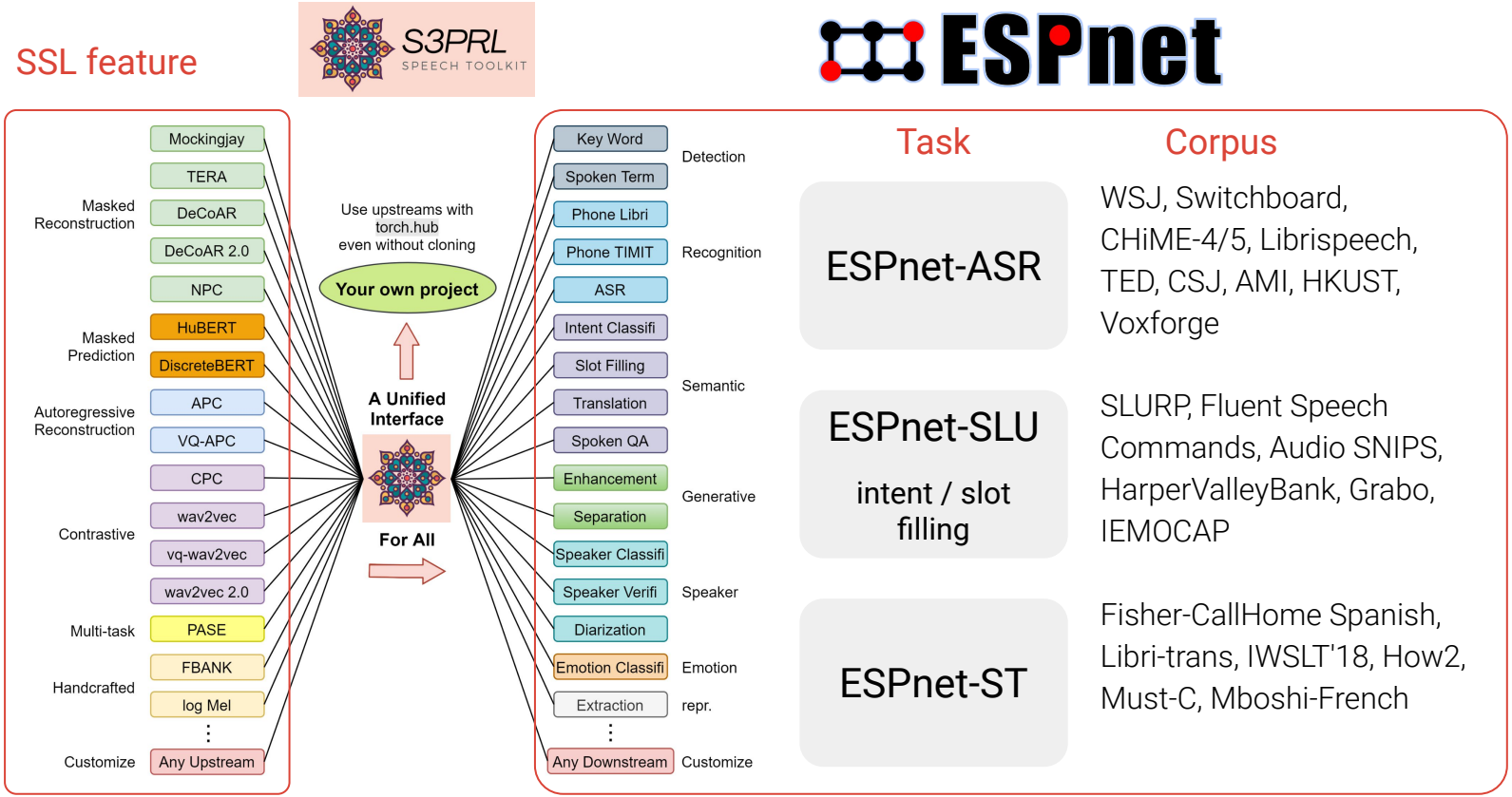
New models

discreteBERT | HuBERT-MGR | LightHuBERT | FitHuBERT

BYOL-S | XLS-R | data2vec | W2v-conv | XLSR-53

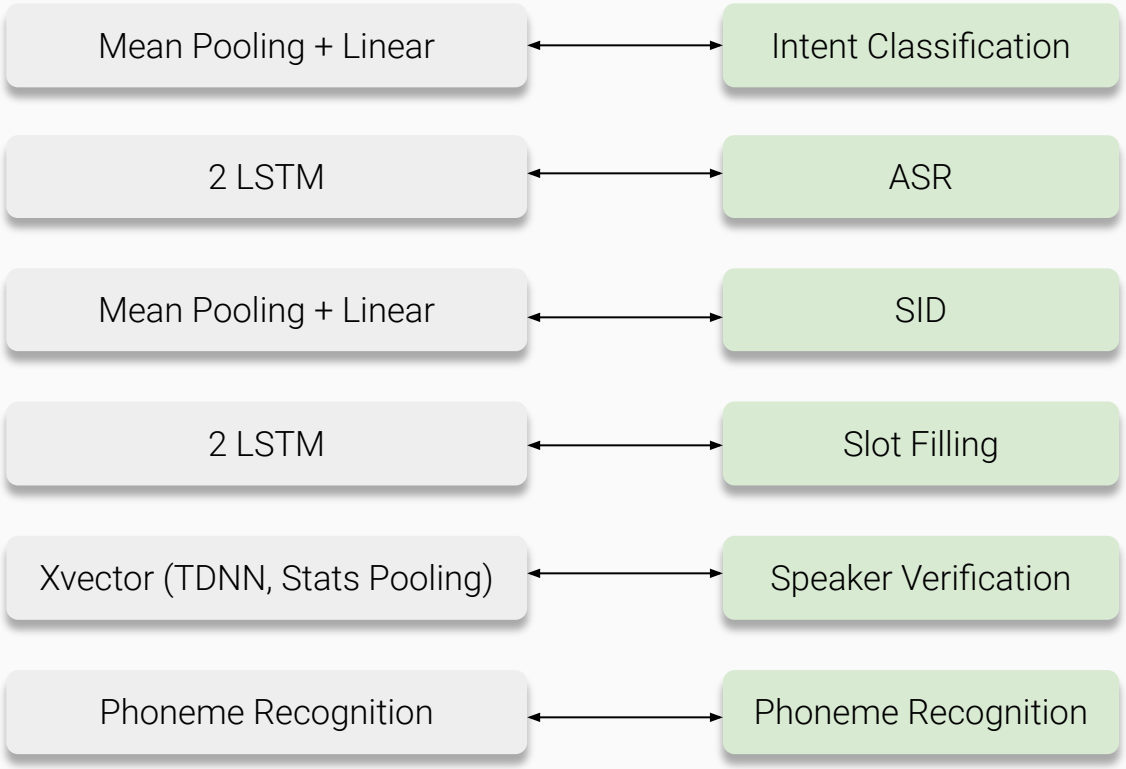
Stabler upstream

- **Remove fairseq dependencies**
- **Numerical tests for all upstreams' forward & backward**



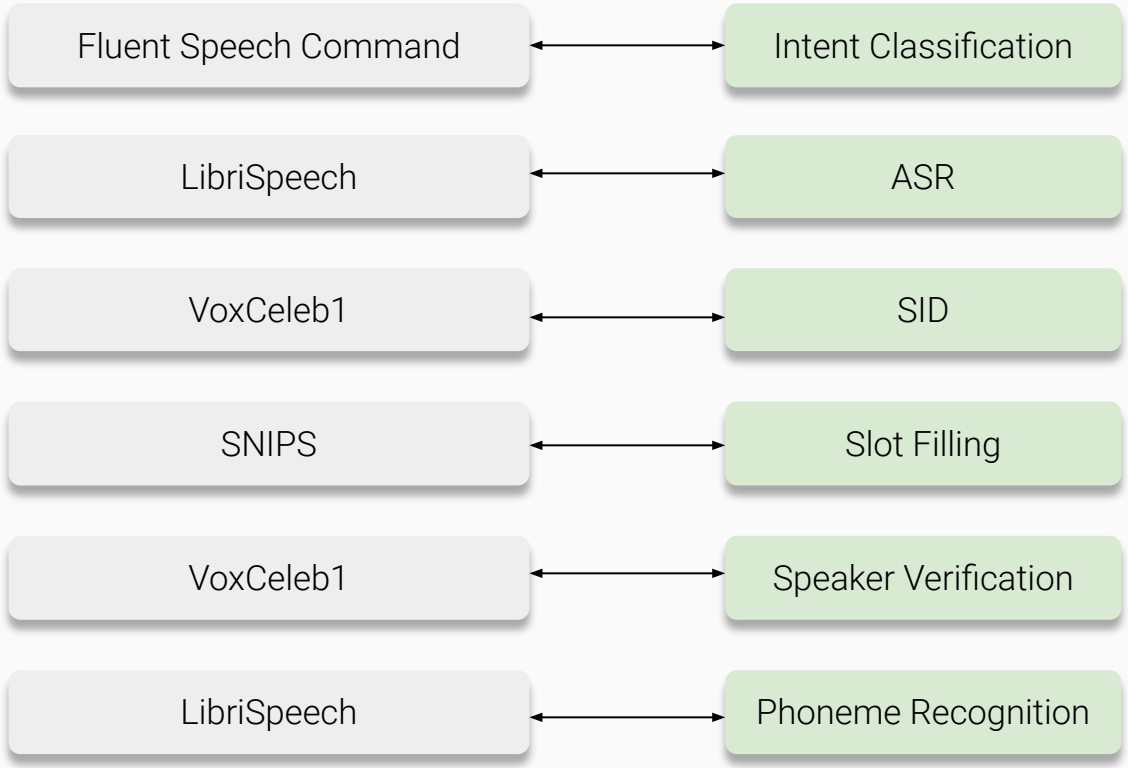
Task was tied to a specific model

A specific model

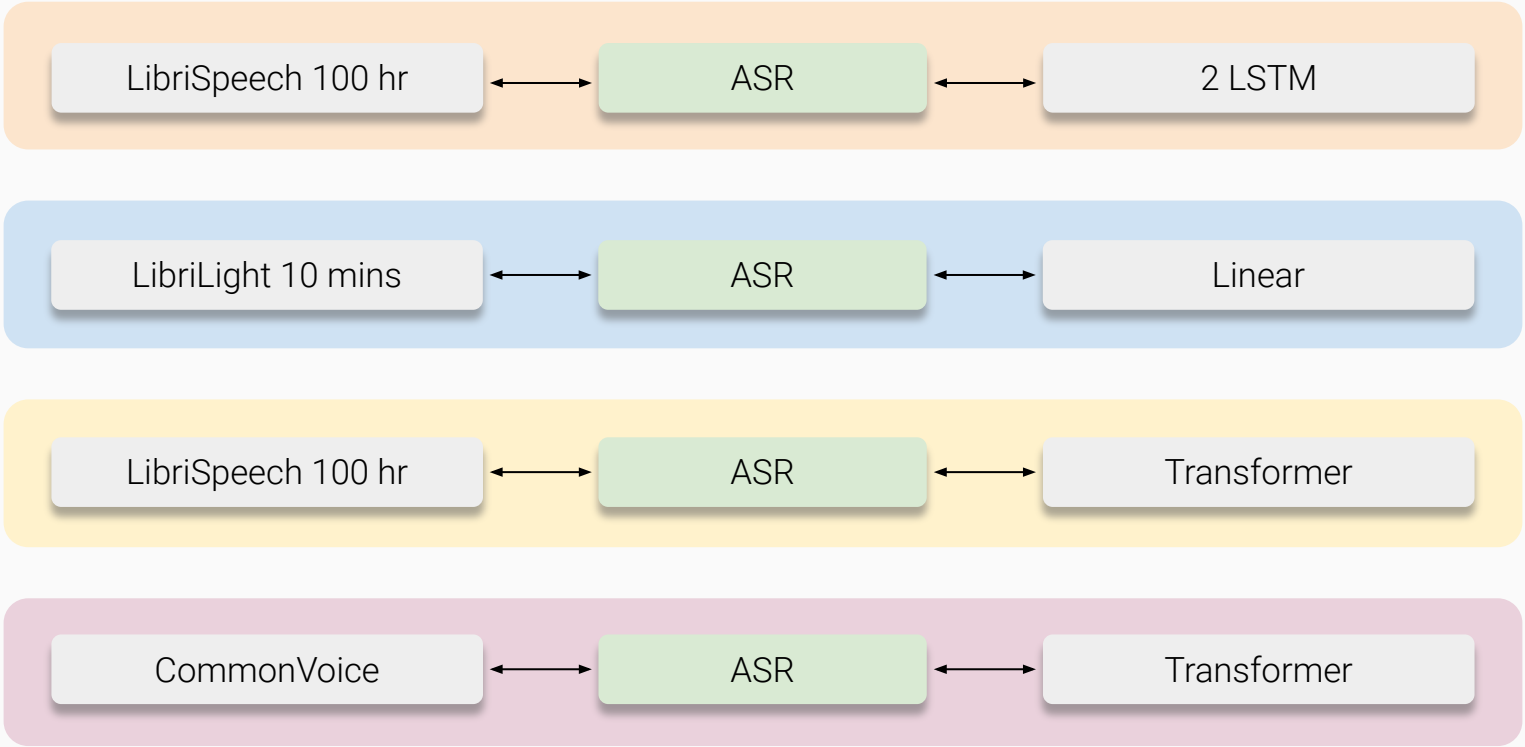


Task was tied to a specific corpus

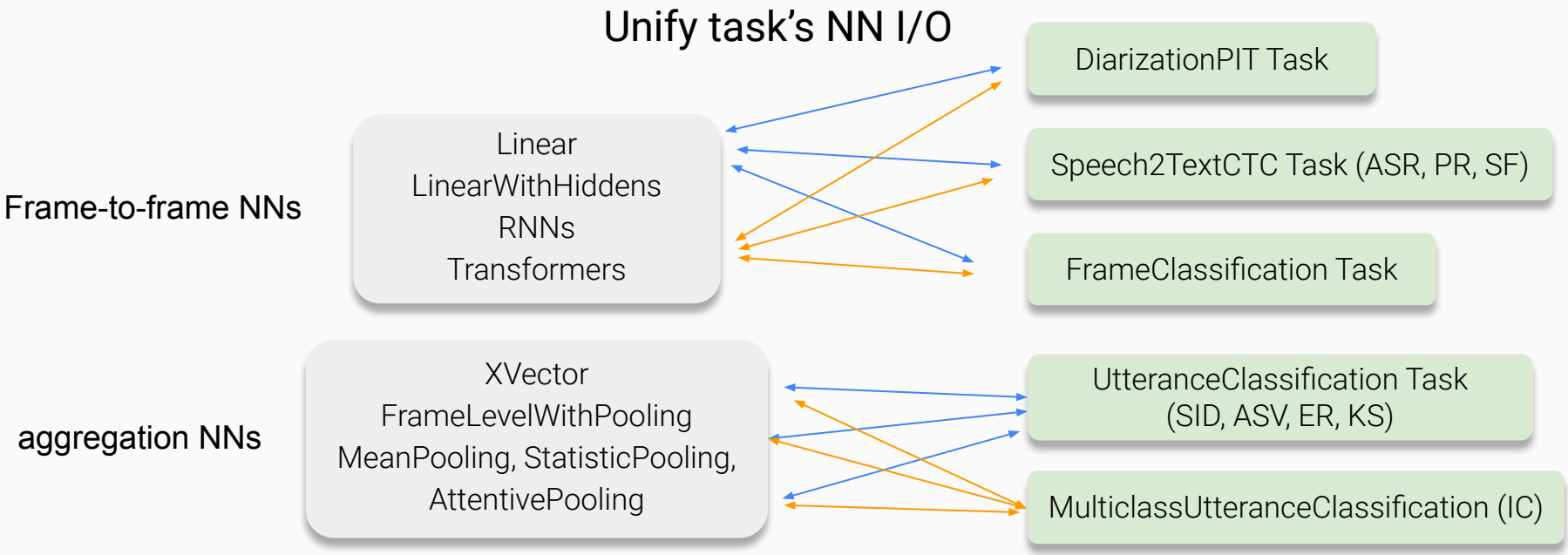
A specific corpus



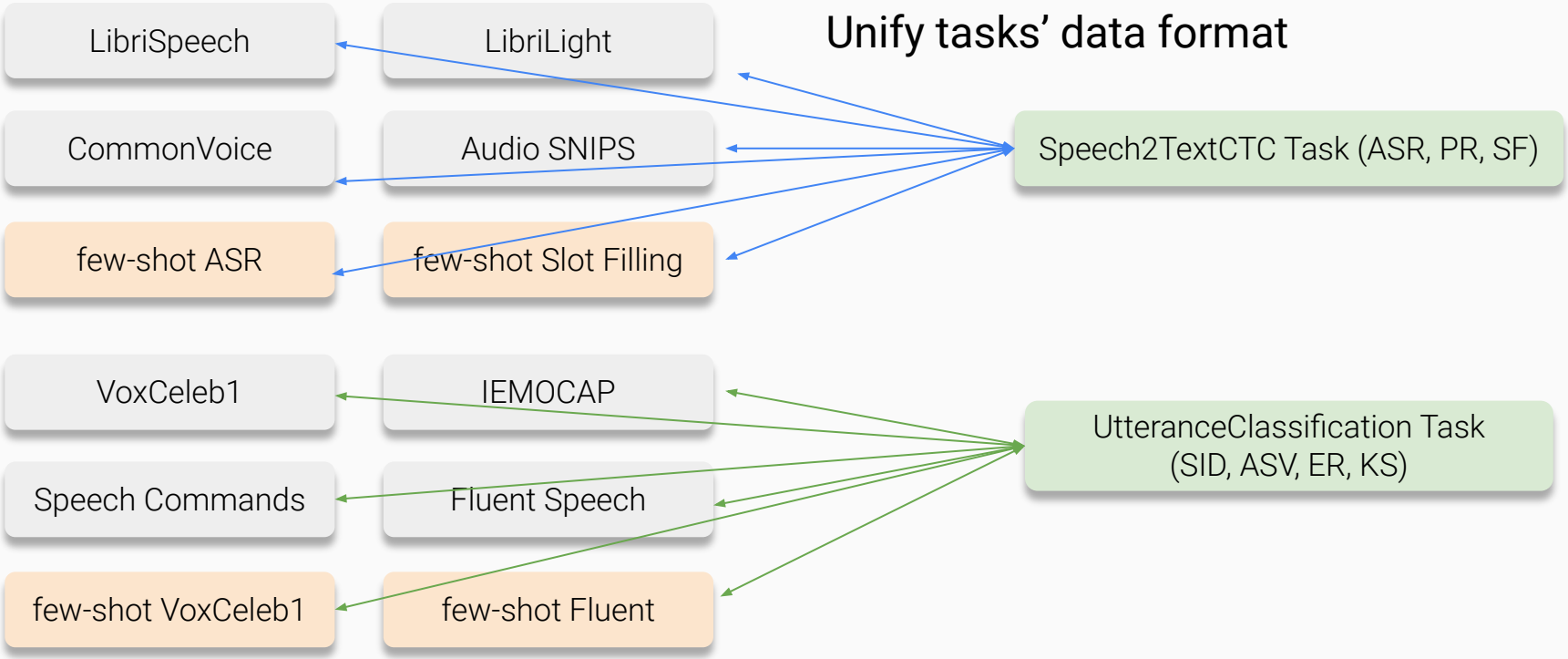
Every slightly change requires library code change



Disentangle model from the task



Disentangle corpus from the task

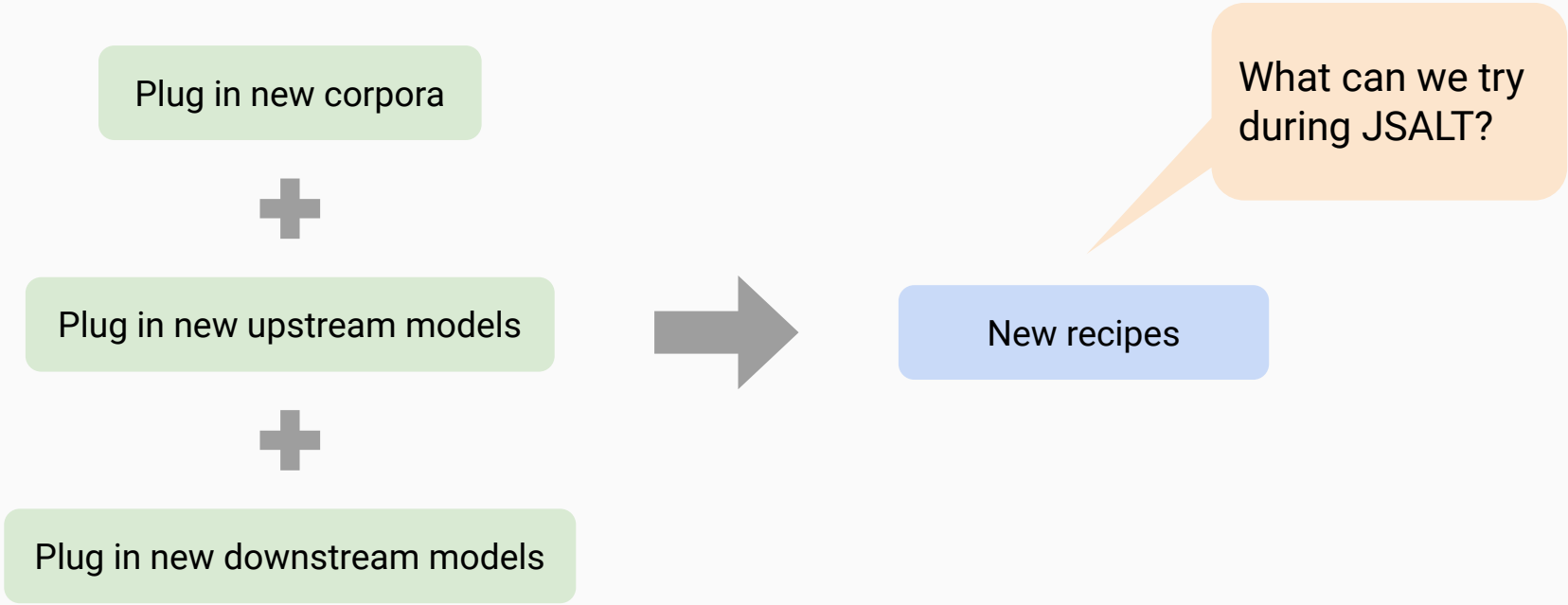


New codebase, yet comparable results

Hubert results

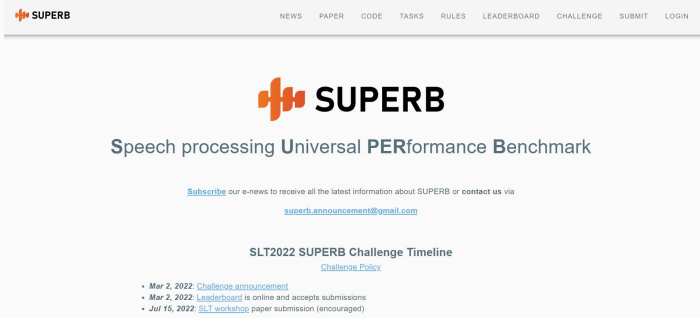
Task	PR	IC	SID	KS	ER	ASR	QBE	SD	SF	SV
Metric	PER	ACC	ACC	ACC	ACC	WER	MTWV	DER	CER	EER
Old	**5.41**	**98.34**	**81.19**	**96.3**	**64.92**	**6.11**	**7.37**	**5.88**	**25.2**	**5.11**
New	**5.483**	**98.207**	**80.69**	**96.62**	**64.76**	**6.14**	**7.37**	**5.8**	**24.22**	**5.15**
Relative	**-1%**	**-0.1%**	**-0.6%**	**0.3%**	**-0.2%**	**-0.4%**	**0%**	**1.3%**	**3%**	**-0.7%**

With the new codebase, we can easily



What we have

Speech pre-training & Benchmark



Pretrained models

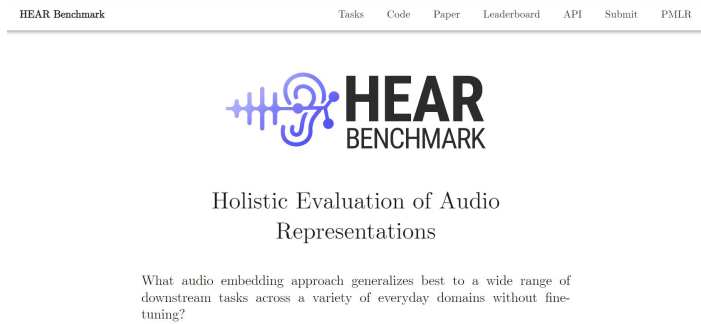
- Mockingjay
- TERA
- HuBERT
- DeCoAR 2.0
- Modified CPC
- APC
- wav2vec
- vq-wav2vec
- VQ-APC
- wav2vec 2.0
- NPC
- DeCoAR
- PASE+

Tasks (SUPERB, 10 + 5 tasks)

- Keyword
- ASR
- ASV
- Speaker ID
- Query-by-example
- Emotion
- Diarization
- Intent
- Slot filling
- Speech Translation
- Voice Conversion
- Source Separation

What about other sounds ?

Sound pre-training & Benchmark

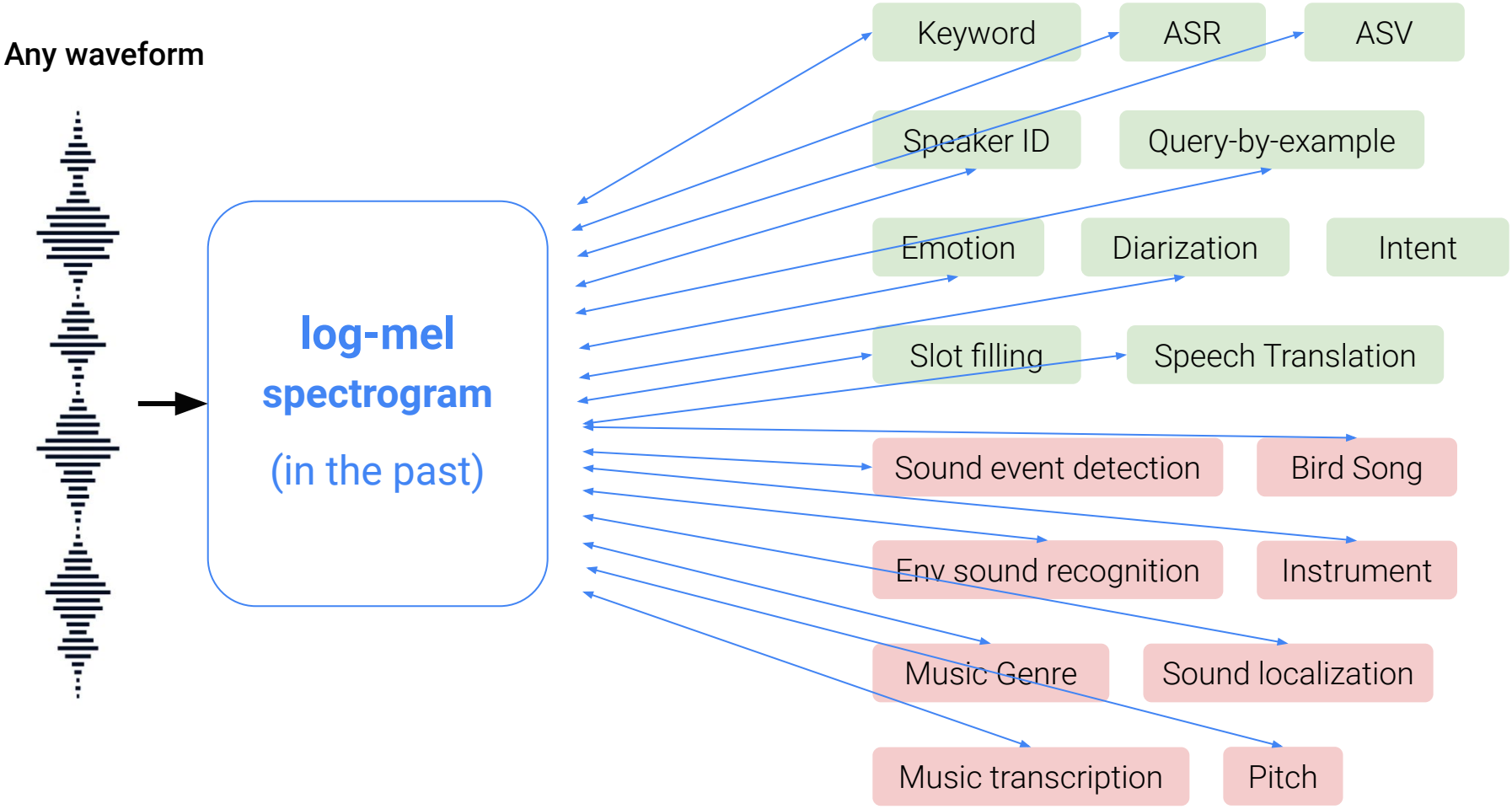


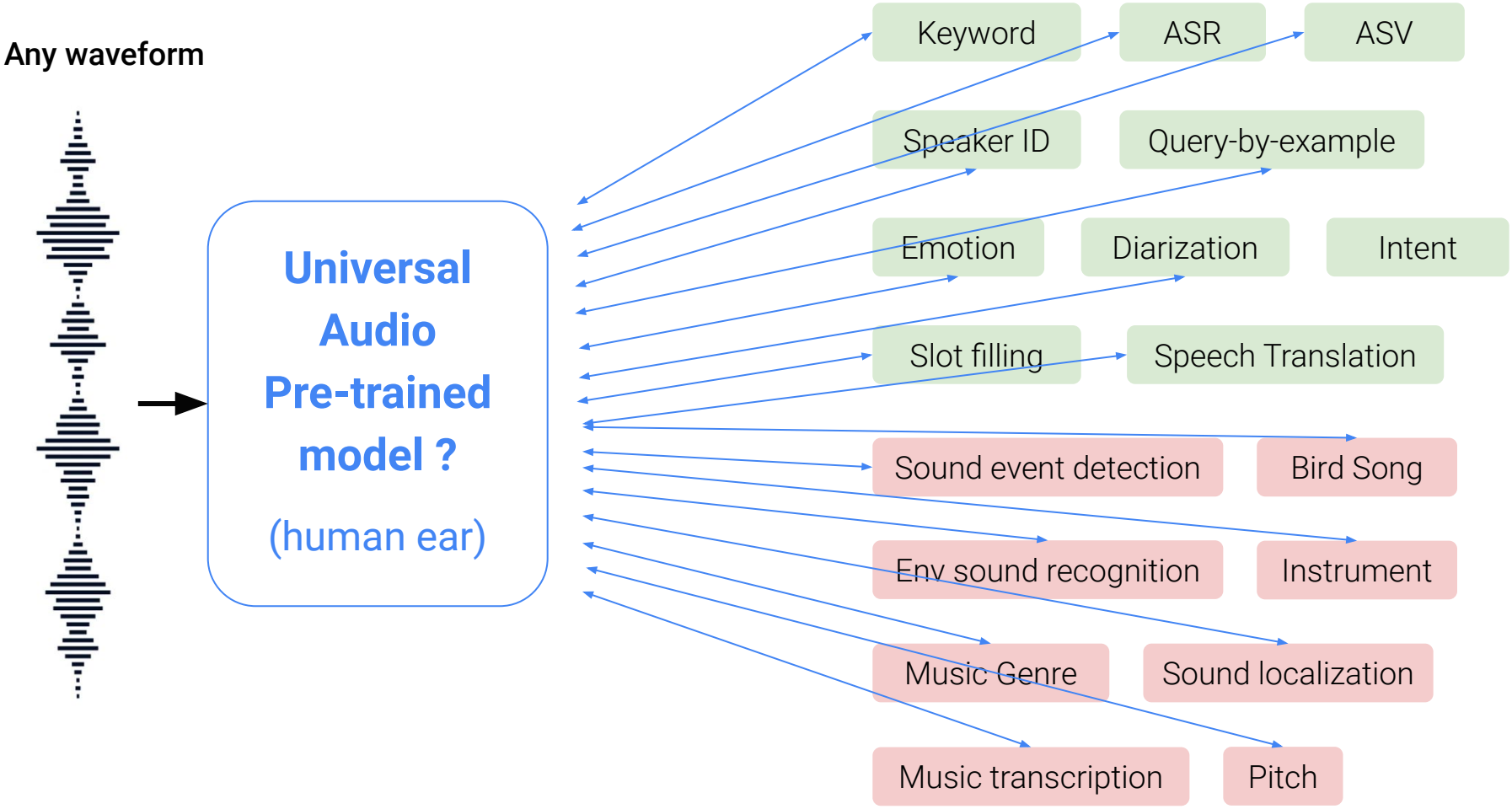
Pretrained models

- PANN
- VGGish
- AST
- SSAST
- MAE-AST
- COLA
- Byol-A
- Byol-S
- PaSST
- AudioMAE (by Meta, to be released)

Tasks (HEAR, 19 tasks)

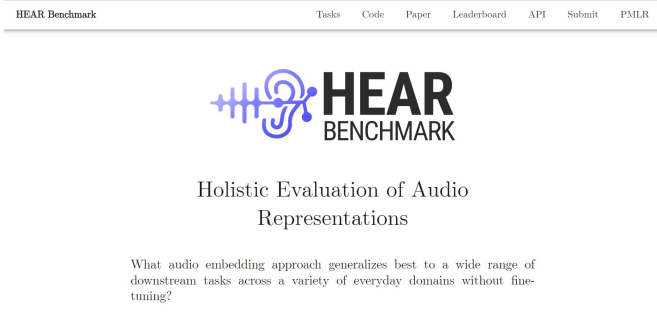
- Sound event detection
- Bird Song
- Env sound recognition
- Instrument
- Music Genre
- Sound localization
- Music transcription
- Pitch
- Some overlapped speech tasks: KS, ER


Any waveform
log-mel
spectrogram
(in the past)
Keyword
ASR
ASV
Speaker ID
Query-by-example
Emotion
Diarization
Intent
Slot filling
Speech Translation
Sound event detection
Bird Song
Env sound recognition
Instrument
Music Genre
Sound localization
Music transcription
Pitch


Any waveform
Universal Audio Pre-trained model ?
(human ear)
Keyword
ASR
ASV
Speaker ID
Query-by-example
Emotion
Diarization
Intent
Slot filling
Speech Translation
Sound event detection
Bird Song
Env sound recognition
Instrument
Music Genre
Sound localization
Music transcription
Pitch

Toward universal audio pretraining

- Need a handy evaluation codebase for all kinds of audio
- We already have lots of evaluation tools for **speech** in S3PRL
  - baseline models
  - downstream tasks



Toward universal audio pretraining

- **19** HEAR tasks are supported by the existing tasks

Simple classification

Keyword

Speaker ID

Change corpus
Change downstream

Pitch

Speech Commands (5hr)

K-fold cross-validation classification

Emotion

CREMA-D

Music Genre

ESR (ESC-50)

FSD-50K

Gunshot Triangulation

Frame-level classification (segmentation)

Diarization

Sound event detection

Music transcription

Pretrained models

Mockingjay | TERA | HuBERT

DeCoAR 2.0 | Modified CPC | APC

wav2vec | vq-wav2vec | VQ-APC

wav2vec 2.0 | NPC | DeCoAR

PANN | VGGish | AST

SSAST | MAE-AST | COLA

Byol-A | Byol-S | PaSST

SUPERB Tasks

Keyword | ASR | ASV

Speaker ID | Query-by-example

Emotion | Diarization | Intent

Slot filling | Speech Translation

Sound event detection | Bird Song

Env sound recognition | Instrument

Music Genre | Sound localization

Music transcription

HEAR BENCHMARK

Correctness & More possibilities

- On-the-fly feature extraction
- Slower, yet support more options:
  - weighted-sum
  - finetune upstream
  - Adaptor…

Hear Benchmark	Task	Frame level start/end segmentation		muticlass	multilabel
	Corpus	Maestro	Dcase 2016	ESC-50	FSD50k
Model	Recipe	Music Transcription	Sound Event Detection	ESR	ESR
	Metric Codebase	F1	F1	ACC	MAP
wav2vec2 large vox (last layer)	Official	3.29	66.3	56.1	34.17
wav2vec2 large vox (last layer)	S3PRL	3.292	65.8	55.7	34.23
wav2vec2 large vox **(weighted-sum)**	S3PRL	**27.92**	**90.204**	**66.85**	**41.08**

Results - Speech v.s. Sound

			Speech		**Sound**	
		Benchmark	SUPERB		HEAR	
	Model	Task	KS	SID	ESC-50	FSD50k
		Metric / Pre-training data	ACC	ACC	ACC	MAP
Speech	Wav2vec2 large vox	Voxpopuli 100K hr	**97.47**	**80.68**	66.85	41.08
Sound	Byol Audio	AudioSet 5K hr	93.1	57.6	**83.2**	**44.8**

Results - Speech + Sound

			Speech		Sound			
		Benchmark	SUPERB		HEAR			
		Task	KS	SID	Music Transcription	Sound Event Detection	ESC-50	FSD
			ACC	ACC	F1	F1	ACC	MAP
baseline	Log-mel		8.63	8.5e-4	47.48	52.544	26.3	13.29
SOTA	Leaderboard SOTA		**97.86**	**95.49**	46.91	92.54	**94.75**	**64.09**
Speech	wav2vec2 large vox	Voxpopuli 100K hr	97.47	80.68	27.92	90.204	66.85	41.08
Sound	Byol Audio	AudioSet 5K hr	93.1	57.6	-	-	83.2	44.8
Speech + Sound	**MAE-AST**	**LibriSpeech 1K hr AudioSet 5K hr**	97.3	63.3	**49.898**	**93.073**	88.9	45.66

Take-away

1. **A new modularized & customizable codebase**
   - Quickly adapt all the HEAR Benchmark tasks during JSALT

✅	DCASE 2016 Task 2	T	L	TVT	120.0	72	Onset FMS
✅	NSynth Pitch 5hr	S	C	TVT	4.0	5000	Pitch Acc.
✅	NSynth Pitch 50hr	S	C	TVT	4.0	49060	Pitch Acc.
✅	Speech Commands 5hr	S	C	TVT	1.0	22890	Accuracy
✅	Speech Commands Full	S	C	TVT	1.0	100503	Accuracy
✅	Beehive States	S	C	TVT	600.0	576	AUCROC
✅	Beijing Opera Percussion	S	C	5-fold	4.77	236	Accuracy
✅	CREMA-D	S	C	5-fold	5.0	7438	Accuracy
✅	ESC-50	S	C	5-fold	5.0	2000	Accuracy
✅	FSD50K	S	L	TVT	0.3-30.0	51185	mAP
✅	Gunshot Triangulation	S	C	7-fold	1.5	88	Accuracy
✅	GTZAN Genre	S	C	10-fold	30.0	1000	Accuracy
✅	GTZAN Music Speech	S	C	10-fold	30.0	128	Accuracy
✅	LibriCount	S	C	5-fold	5.0	5720	Accuracy
✅	MAESTRO 5hr	T	L	5-fold	120.0	185	Onset FMS
✅	Mridangam Stroke	S	C	5-fold	0.81	6977	Accuracy
✅	Mridangam Tonic	S	C	5-fold	0.81	6977	Accuracy
✅	Vocal Imitations	S	C	3-fold	11.26	5601	mAP

Take-away

2. More & stabler upstream

- Scheduled to integrate with ESPNet after JSALT

data2vec

WavLM

Hubert MGR

XLS-R

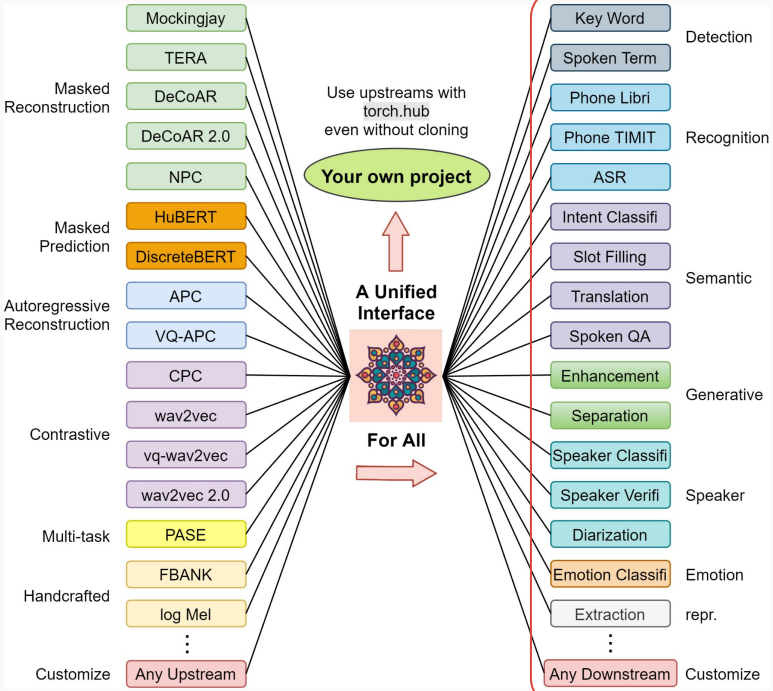
Byol-A

AST

SSAST

MAE-AST

PANN

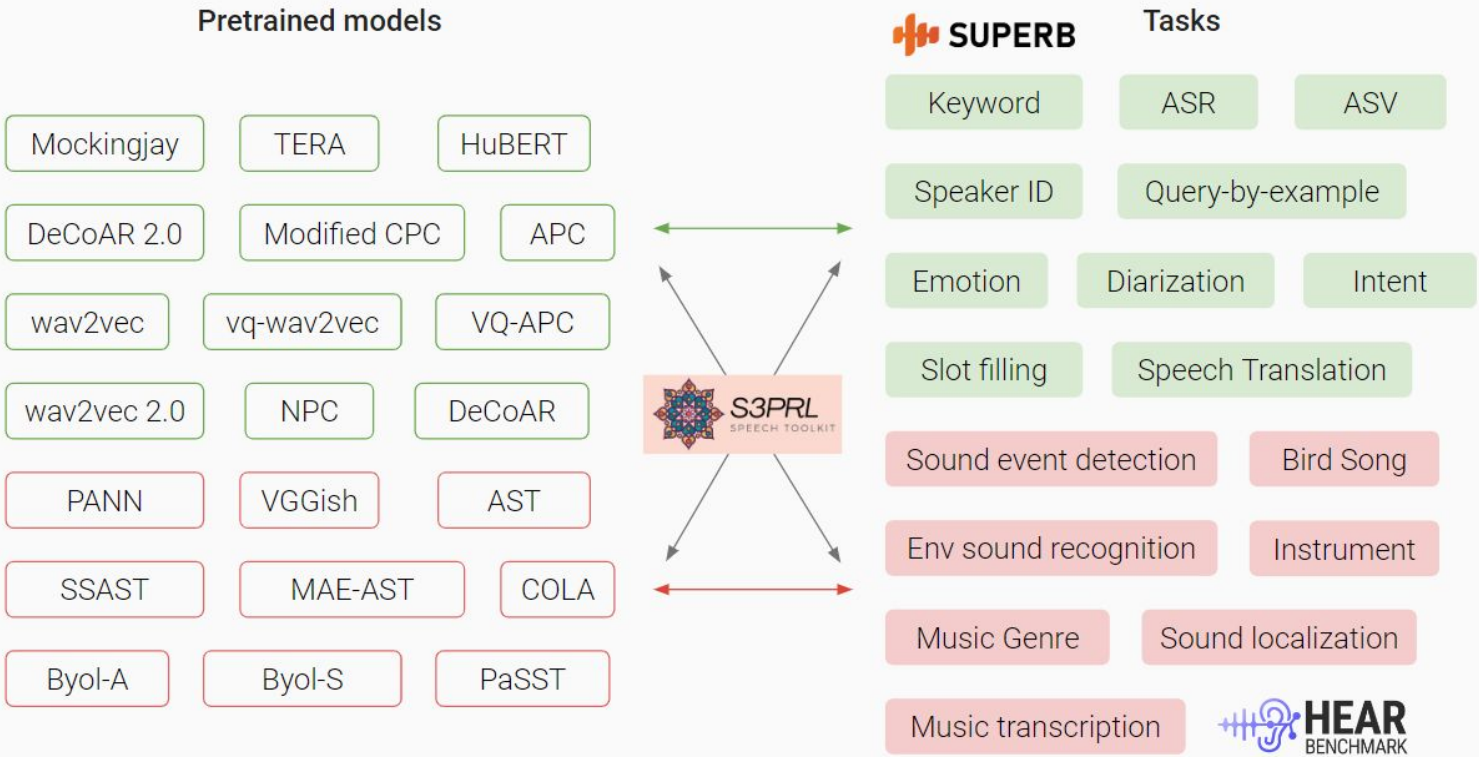


Task	Corpus
ESPnet-ASR	WSJ, Switchboard, CHiME-4/5, Librispeech, TED, CSJ, AMI, HKUST, Voxforge
ESPnet-SLU intent / slot filling	SLURP, Fluent Speech Commands, Audio SNIPS, HarperValleyBank, Grabo, IEMOCAP
ESPnet-ST	Fisher-CallHome Spanish, Libri-trans, IWSLT'18, How2, Must-C, Mboshi-French

Take-away

3. Toward unifying SSL evaluation on all waveforms

- Build the evaluation codebase for Speech SSL + Sound SSL



Concluding Remarks

Hung-yi Lee
(NTU)

Concluding Remarks

- Network compression for Speech SSL:
  - Replace CNN with Melspectrogram
  - Sequence Reduction
- Generalization: The knowledge distilled speech SSL models are less robust, but we can fix the issue.
- Speech CLIP:
  - The visual-language model can improve speech SSL models.
  - Has the potential to enhance unsupervised ASR.

Concluding Remarks

- Adapter and prompt yield comparable or better performance with fewer parameters.
- SSL speech models are good at extracting prosody information.
- SSL models even outperform supervised pre-training on code-switching ASR.
- Unsupervised ASR has the potential to be a connecter of speech SSL models and text SSL models.
- Refactor toolkit for speech pre-training - S3PRL