Leveraging Pre-training Models for Speech Processing

Research Group @ JSALT 2022

Speaker: Fabian Ritter Gutierrez, National University of Singapore

Goal

- More efficient
- Better generalization
- Visually enhanced

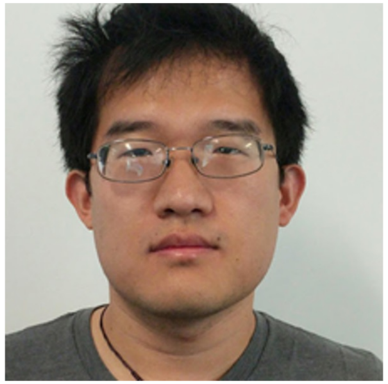
How to better use SSL models

Enhance SSL models

Push SSL models to more tasks

Toolkit

Hung-yi Lee
(NTU)

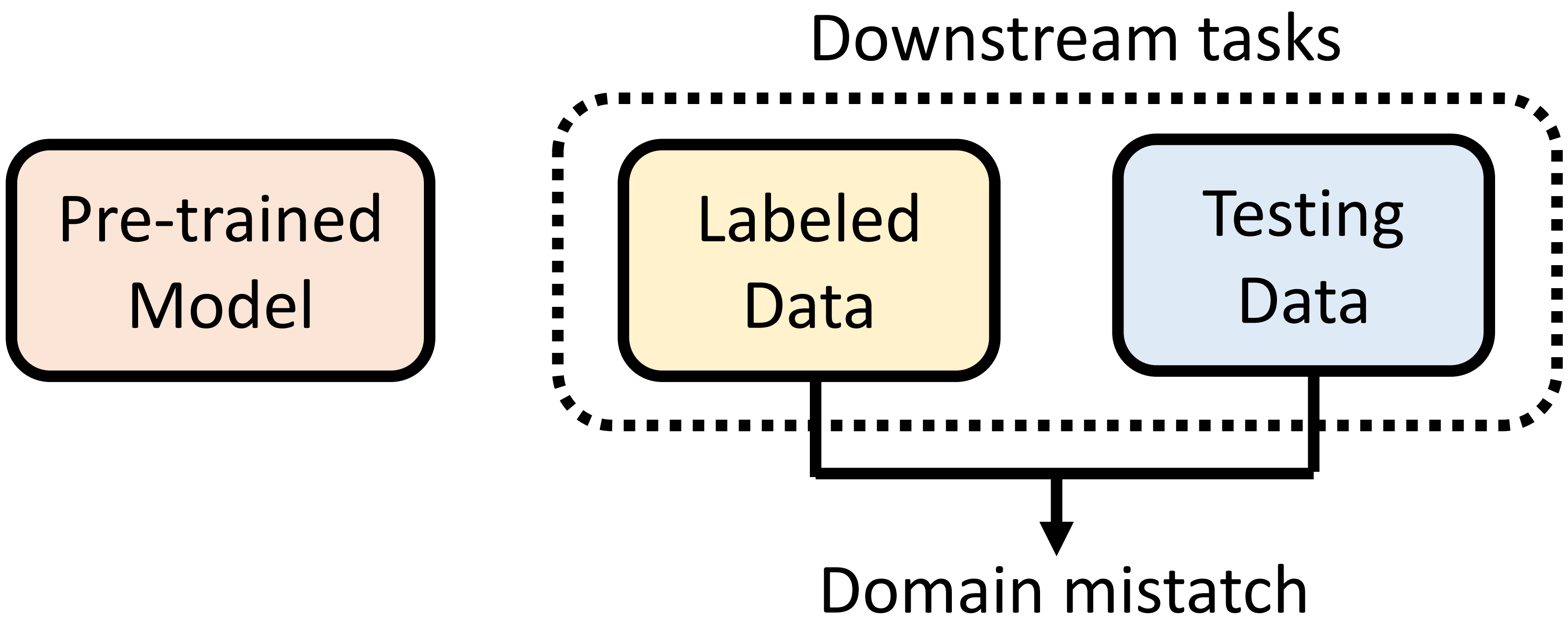


Yu Zhang
(Google)

Kuan Po Huang
(NTU)

Fabian Ritter
(NUS)

Generalization Capability of Pre-trained Model



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

Can self-supervised models maintain good performance?

2 weeks ago…

- ***We realized DistilHuBERT has poor domain generalization.***
- ***Our goal:***
  - To **reduce model size** while having **domain generalization**.

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

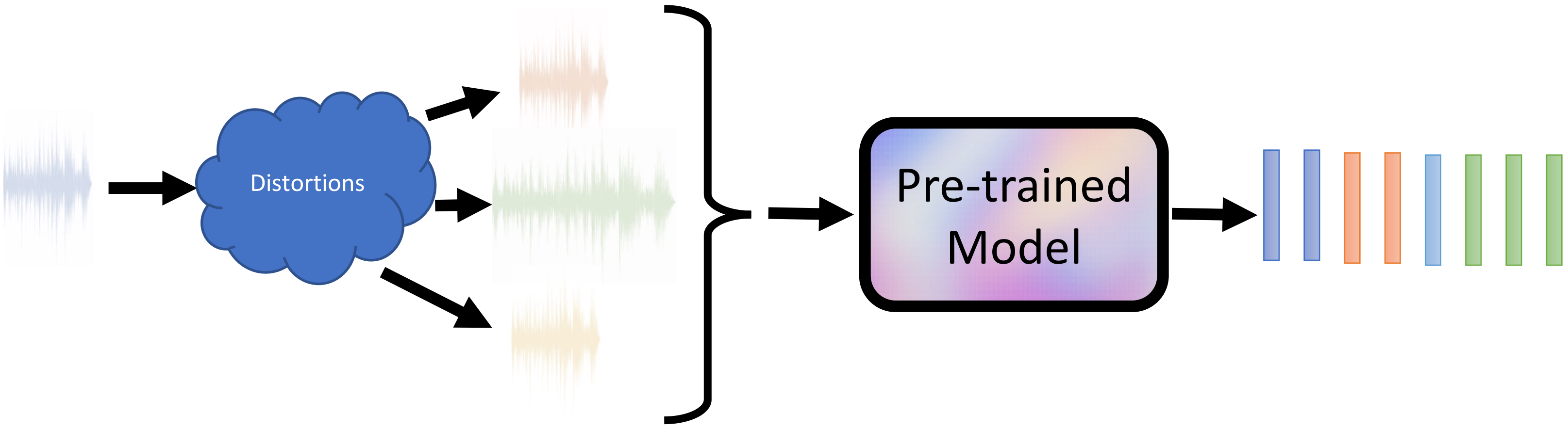
	Speaker Identification ↑		ASR (WER) ↓	
Testing Data	clean	noisy	clean	m+g+r
HuBERT	84.97	65.51	4.88	7.94
DistilHuBERT	73.02	40.42	13.77	37.59

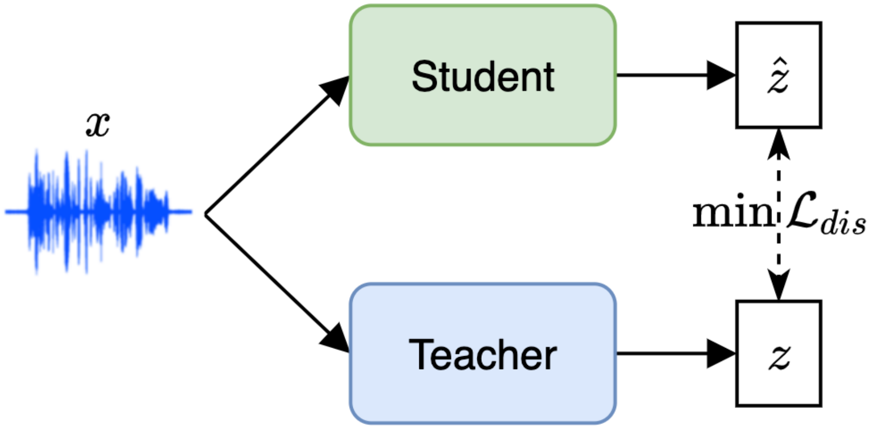
How problems are being tackled?

- Noise robustness:
  - Adding pre-defined distortions on pre-training
    - Additive noises, reverberation, time-frequency masks, speaking rate variation.

Model Size:

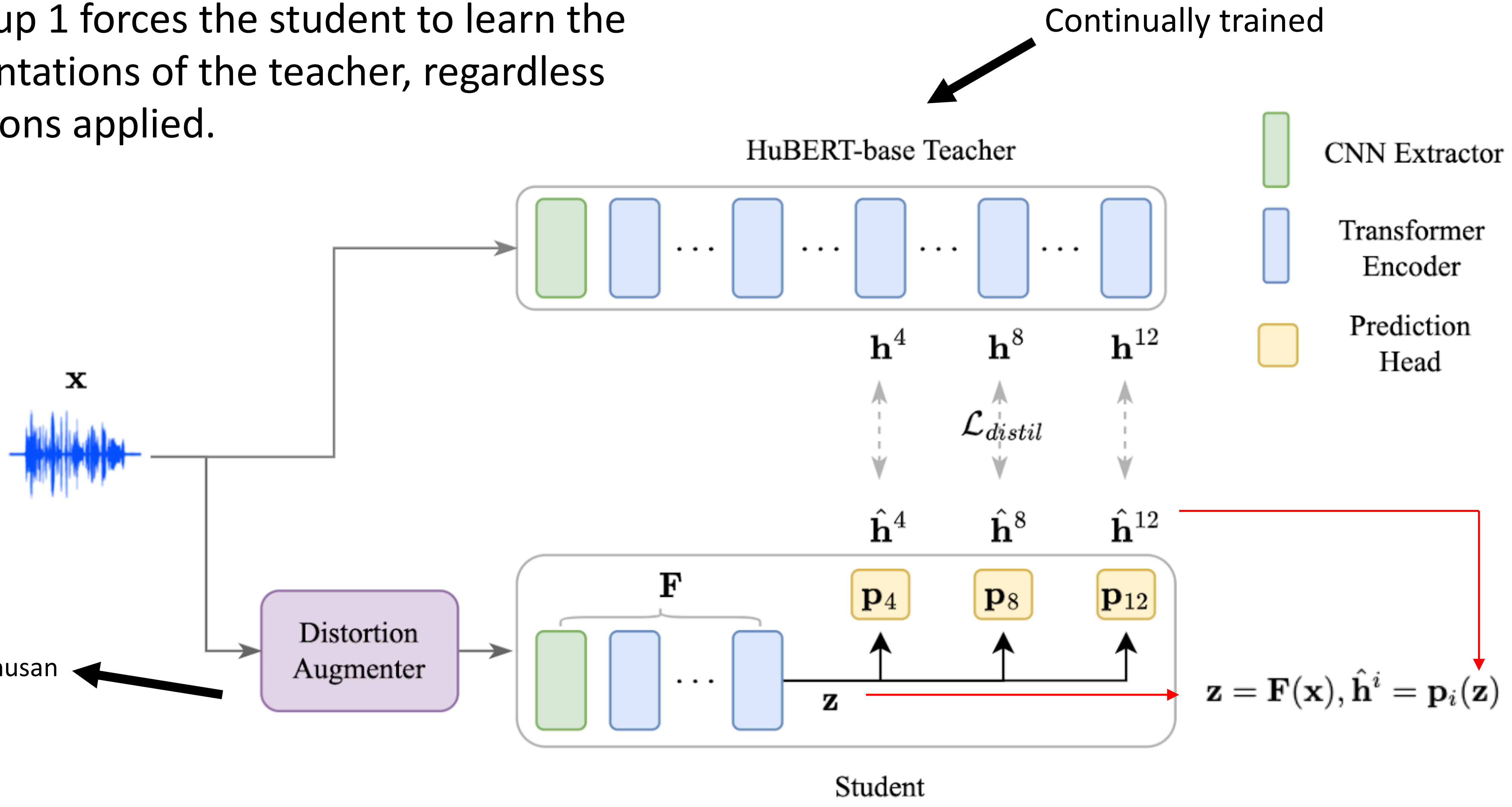
Knowledge Distillation.





Robust DistilHuBERT Setup 1

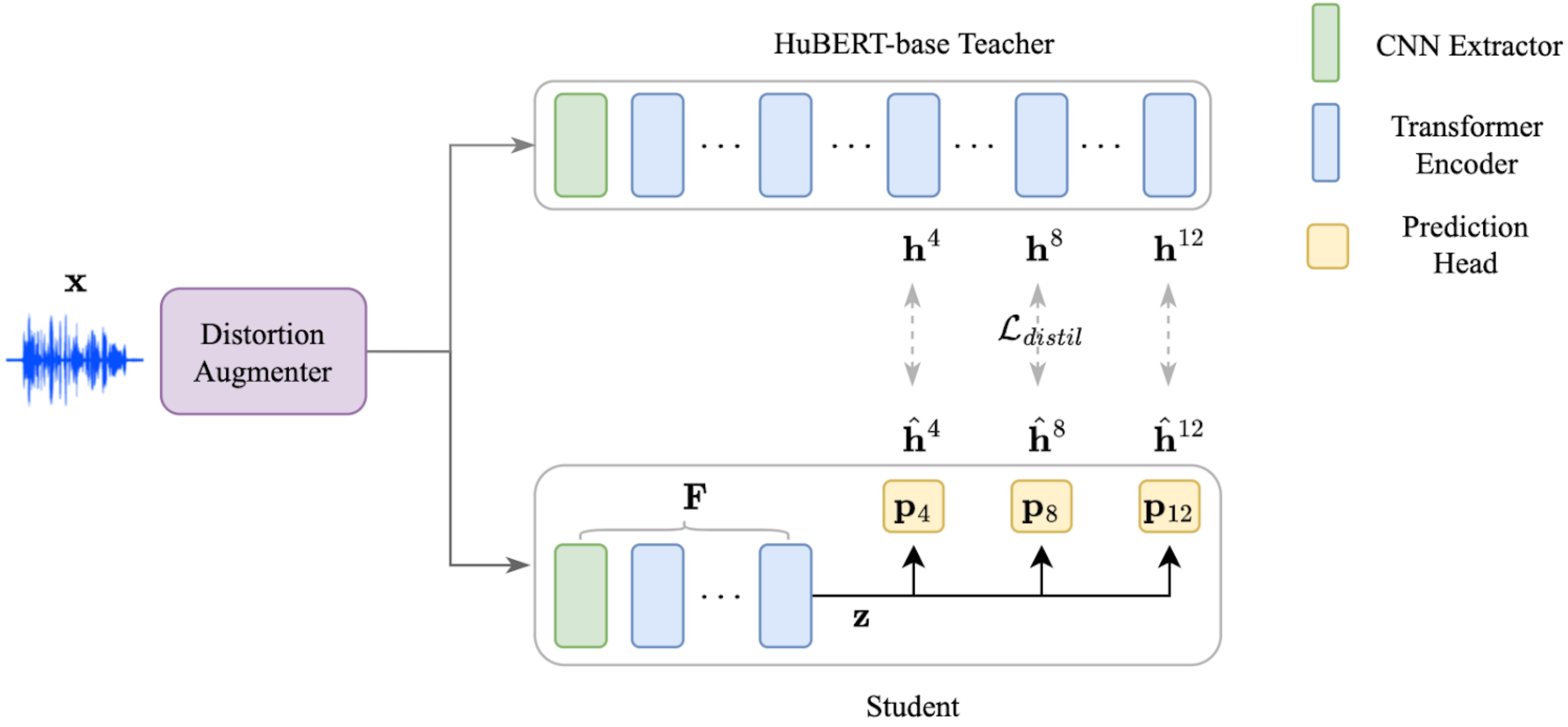
Intuition: Setup 1 forces the student to learn the clean representations of the teacher, regardless of the distortions applied.



reverberation, pitch modification, band reject filtering, gaussian noise, musan noise and wham noise.

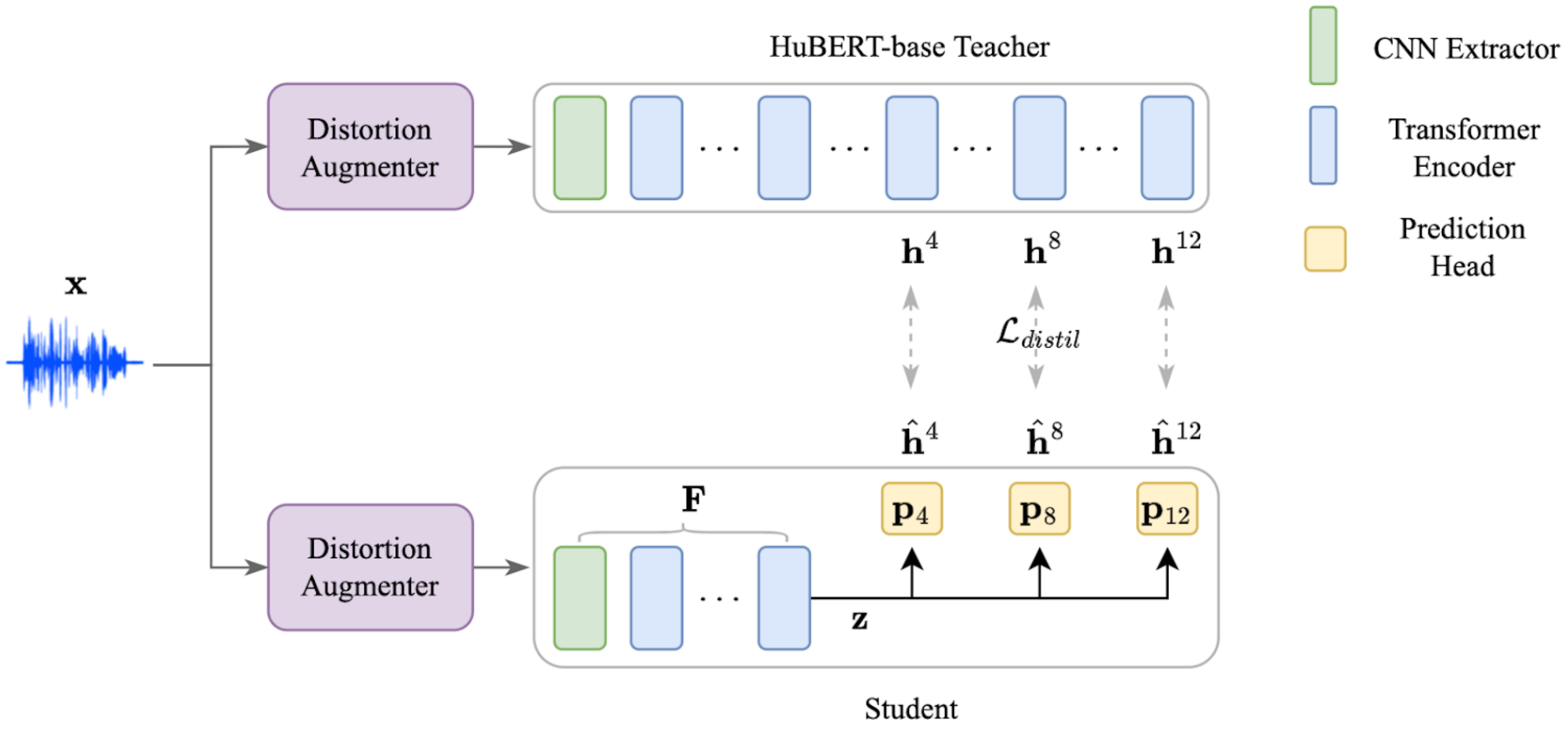
Robust DistilHuBERT Setup 2 (same)

- For setup2, we also experimented on the case where the teacher and student model take the same distorted speech as input.



Robust DistilHuBERT Setup 2 (different)

- For setup2, the student learns representations of the same speech utterance but with different distortions.



Results

		cont.	para.	KS (Acc% ↑) clean	KS 2-dist	KS fsd	KS dns	IC (Acc% ↑) clean	IC 2-dist	IC fsd	IC dns	ER (Acc% ↑) clean	ER 2-dist	ER fsd	ER dns
				Keyword Spotting				Intent Classification				Emotion Recognition			
(T1)	HuBERT [2]	X	95M	96.30	89.81	90.94	77.60	98.34	89.09	91.93	74.11	64.92	56.72	60.05	52.08
(T1')	HuBERT	V	95M	96.53	94.77	94.00	82.83	98.37	96.20	96.78	85.00	65.88	62.82	63.89	56.70
(S1)	DH [4]	X	23M	95.98	87.57	88.70	75.07	94.99	70.29	72.50	48.30	63.13	55.09	57.05	49.76
(S1')	DH	V	23M	96.14	86.86	90.56	76.47	95.65	77.99	81.73	57.50	64.01	58.89	59.06	53.14
(S2)	DH setup1	X	23M	95.52	92.92	93.44	76.66	94.17	89.53	89.61	72.11	63.51	58.11	60.17	50.66
(S2')	DH setup1	V	23M	96.17	93.61	94.09	77.44	95.57	86.11	89.03	71.26	63.72	59.62	61.42	53.69
(S3)	DH setup2 (same)	X	23M	96.11	89.84	91.69	78.42	94.62	75.40	80.33	57.92	61.87	55.72	59.41	50.27
(S3')	DH setup2 (same)	V	23M	96.33	92.57	93.48	80.04	95.68	85.16	86.84	64.46	64.25	59.62	60.93	51.78
(S4)	DH setup2	X	23M	96.27	92.99	93.96	77.47	95.91	90.72	90.77	73.87	63.77	59.89	61.62	51.25
(S4')	DH setup2	V	23M	96.53	93.61	94.38	79.10	96.57	92.25	92.67	78.41	63.08	60.38	60.89	53.38

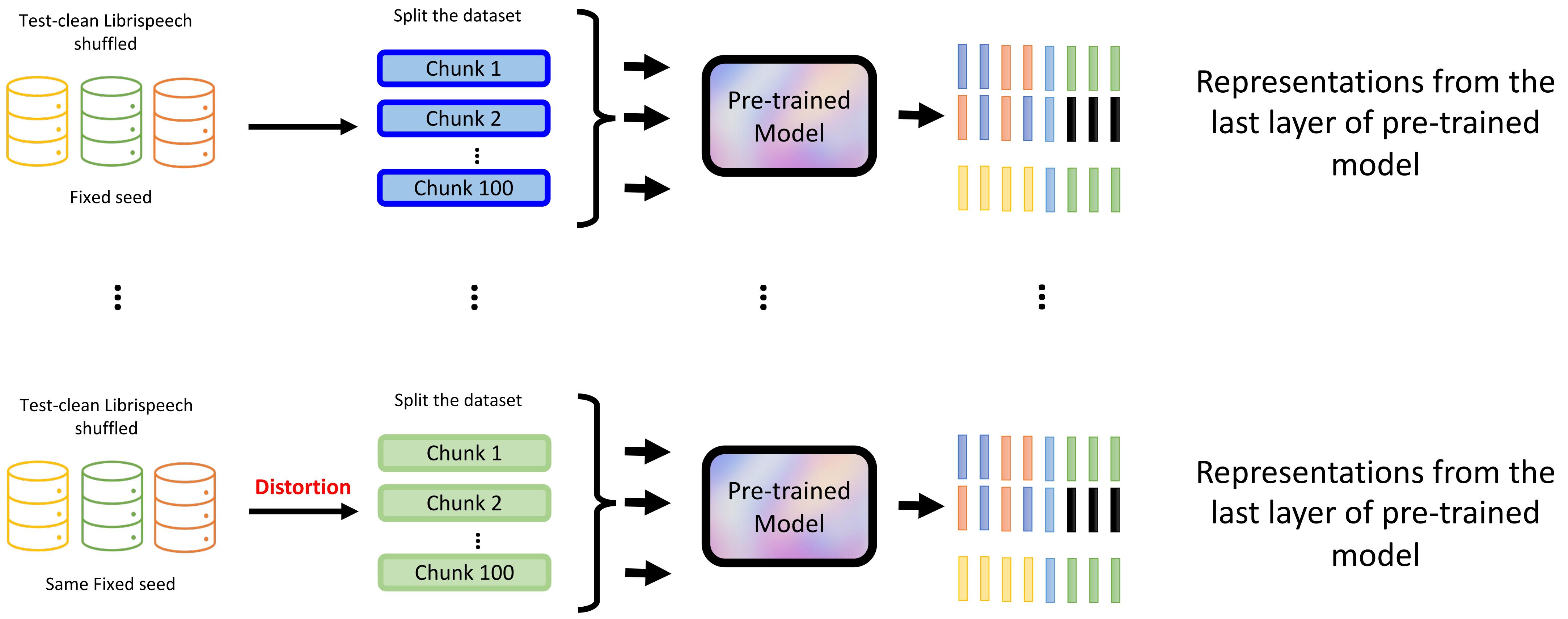
DH : DistilHuBERT , cont. stands for continually trained teacher model.

Visualization Experiments

- We want to have a sense on how the models are affected when they receive: clean speech and noisy speech.

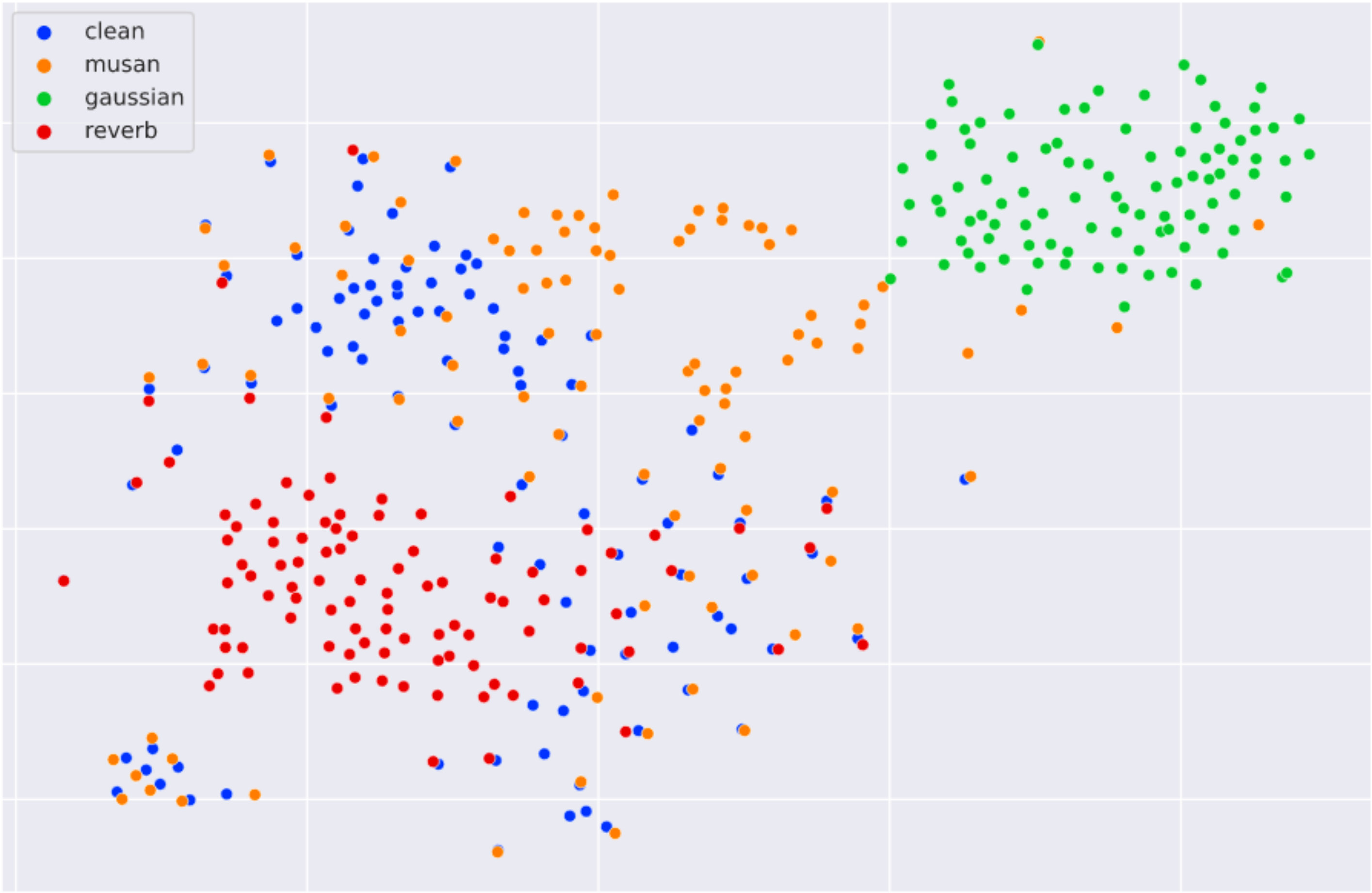
- Hence we plot the last layer of the SSL representations into a low dimensional space.

- Visualizations were done using test-clean partition of Librispeech.

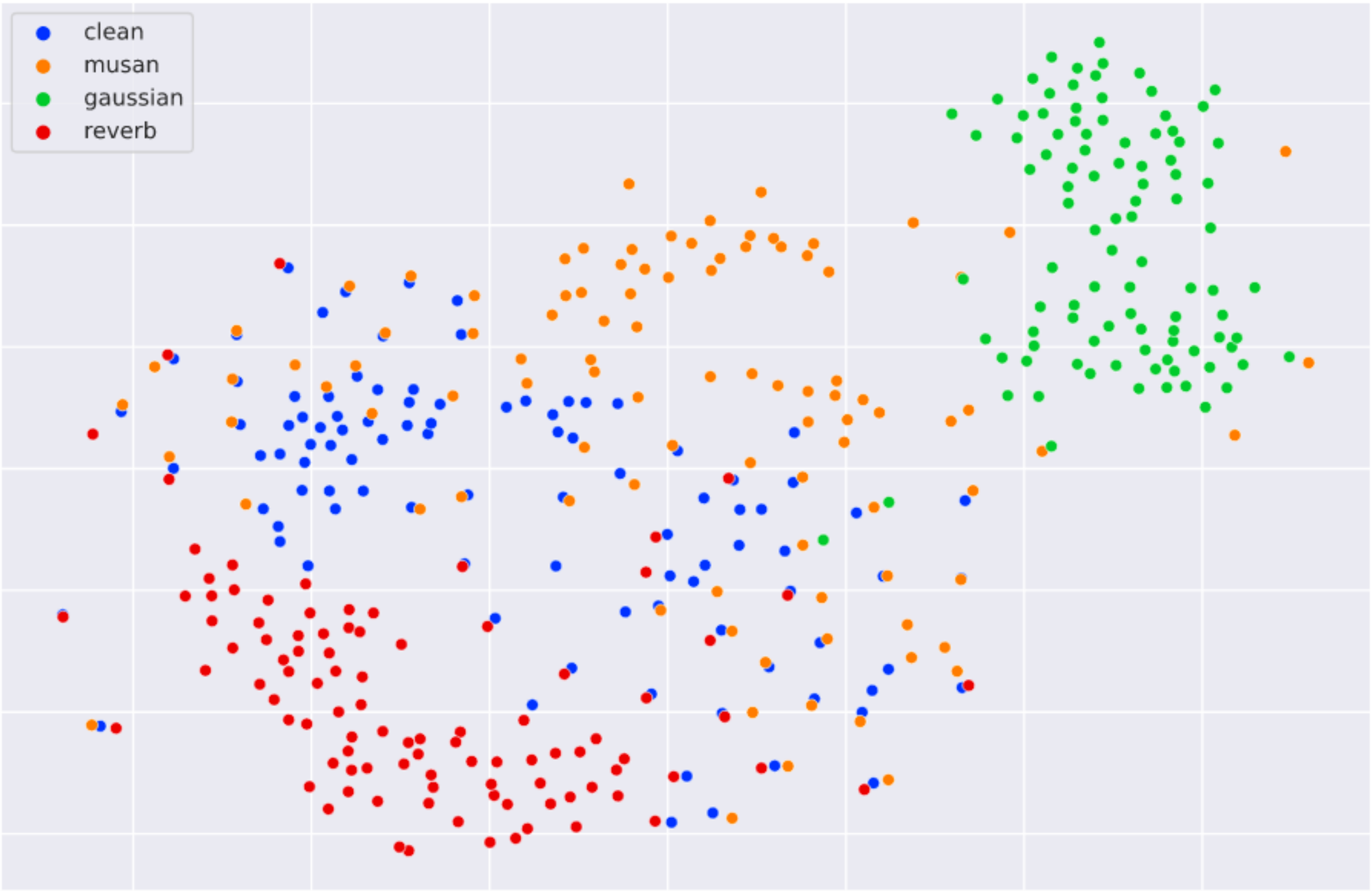


The final 100 points portray the last layer representation averages for test-clean, and perturbed versions of musan noise, Gaussian noise and reverberation.

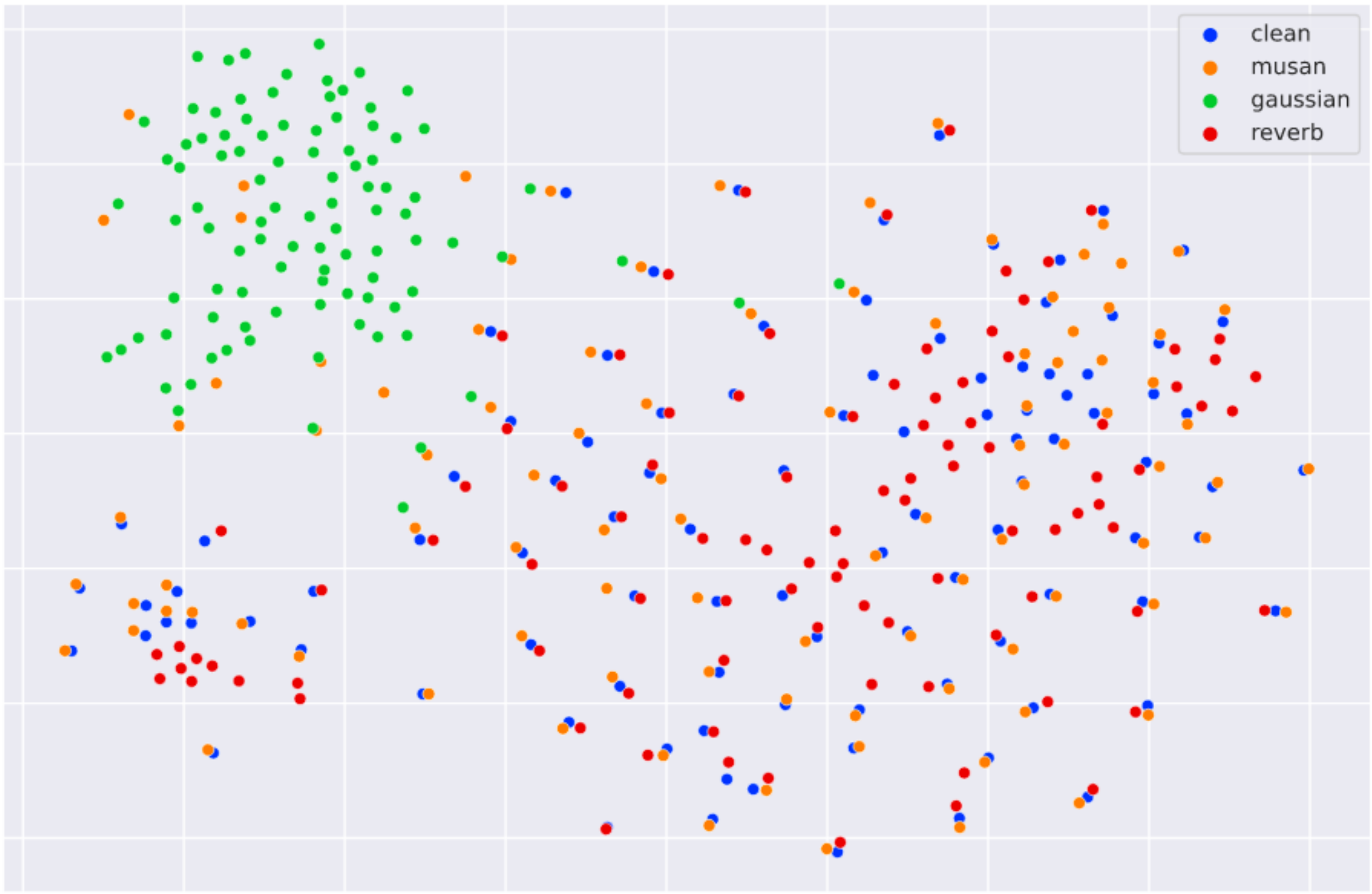
HuBERT



DistilHuBERT

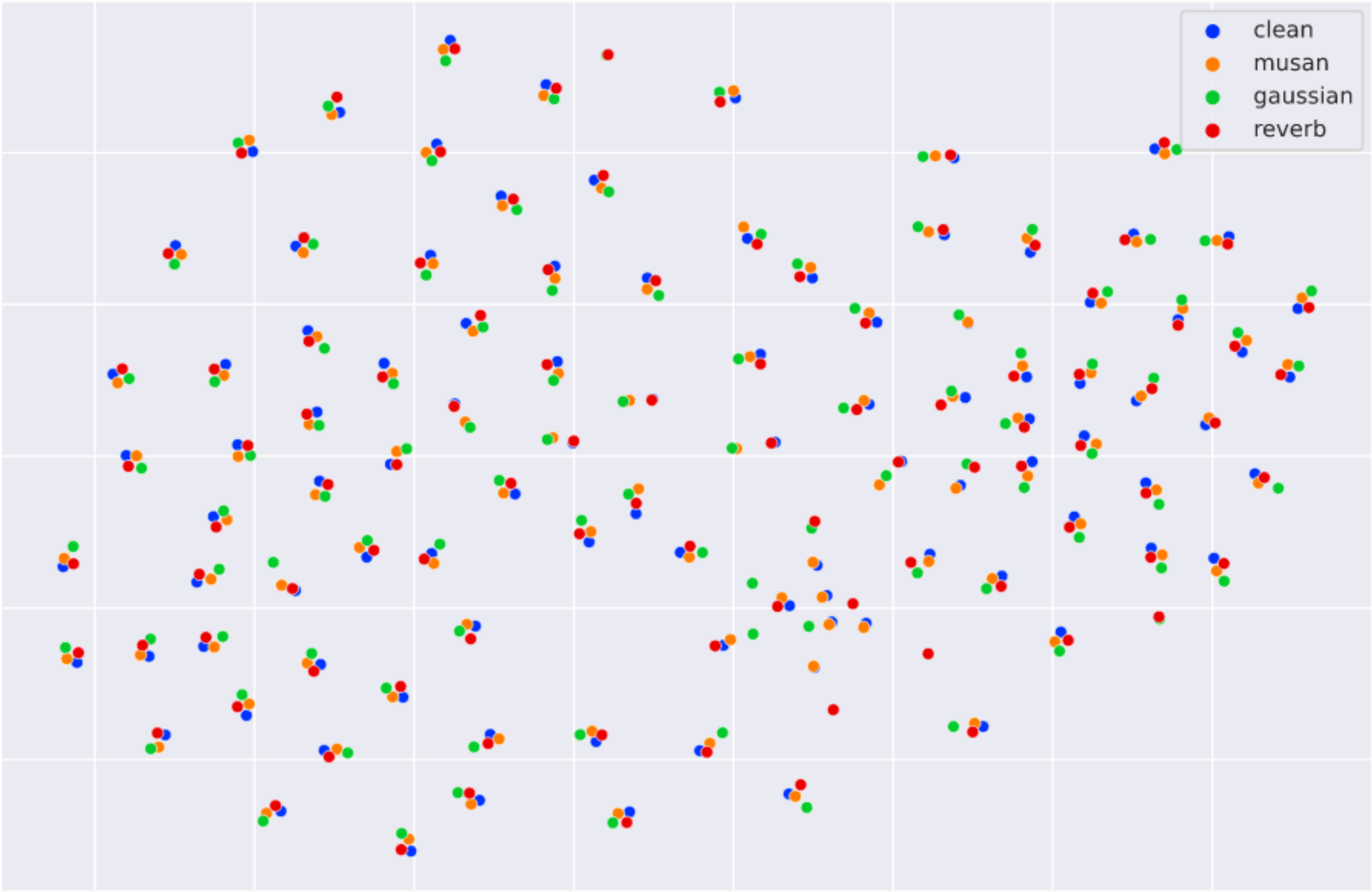


WavLM Base+



Hubert_base_robust_mgr

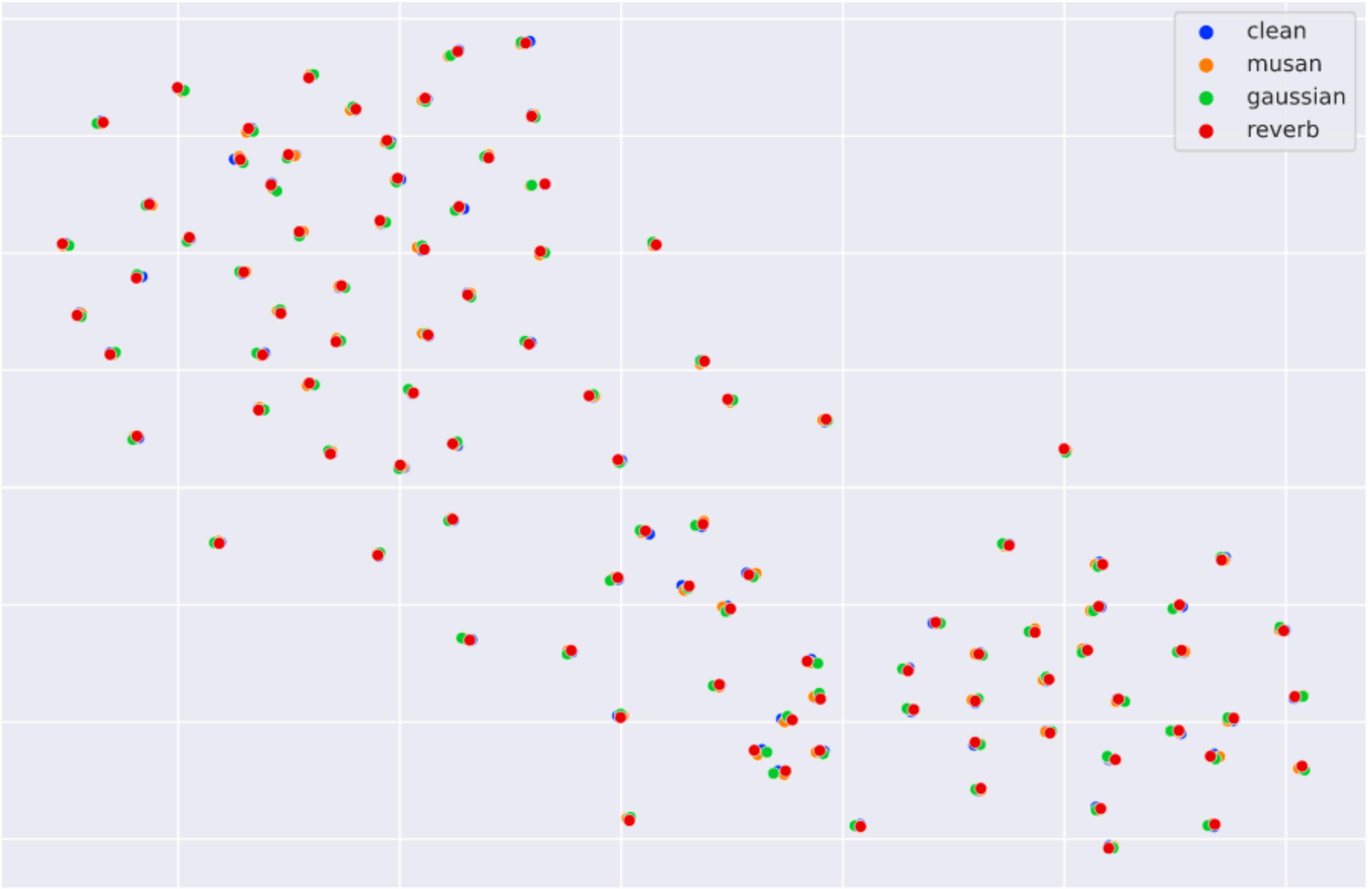
Model corresponds to
A continually trained
HuBERT on musan noise,
gaussian noise and
reverberation.



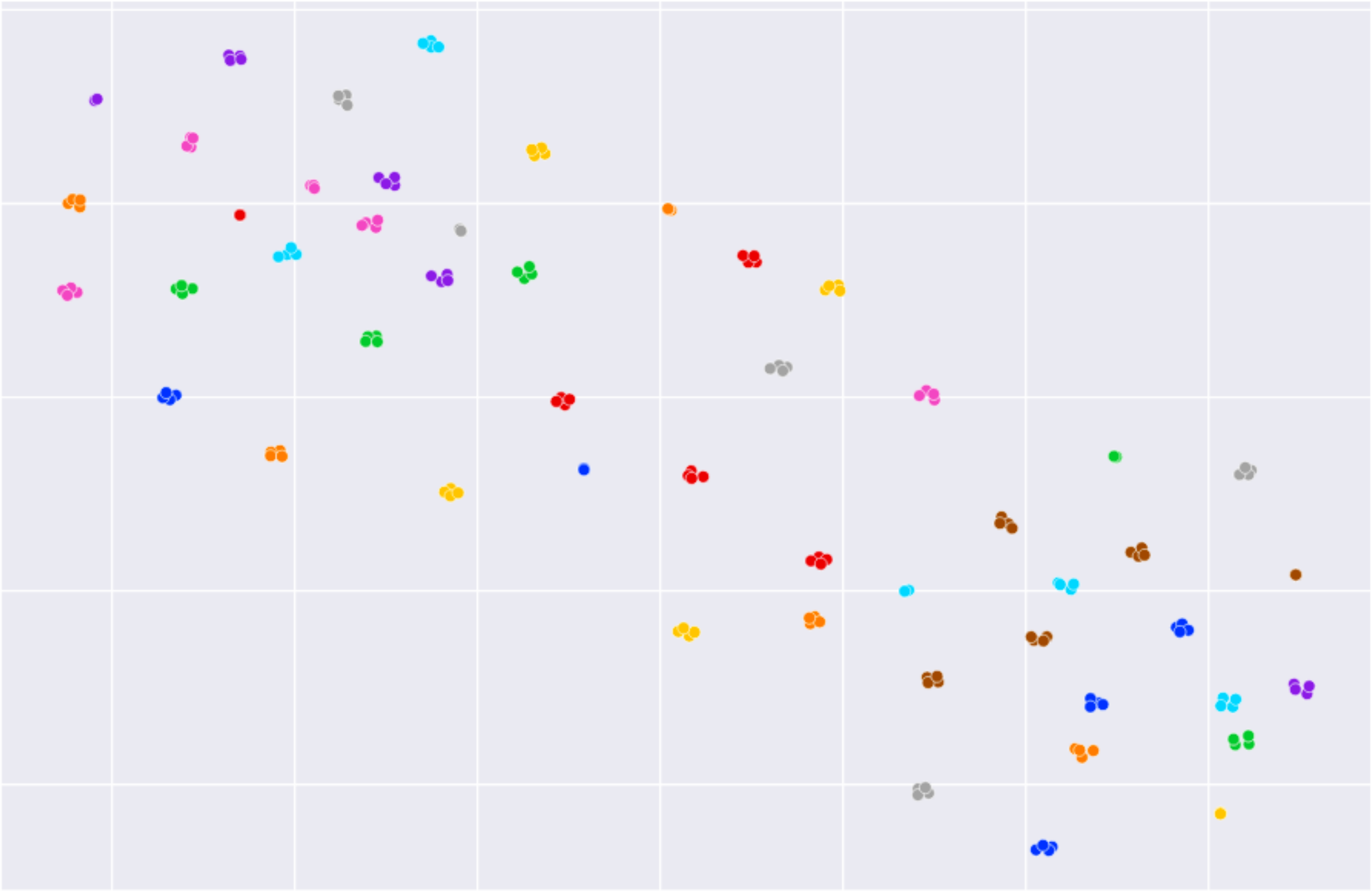
Preliminary results have been accepted by INTERSPEECH 2022.

https://arxiv.org/abs/2203.16104

Robust DistilHubert setup1 (S2)

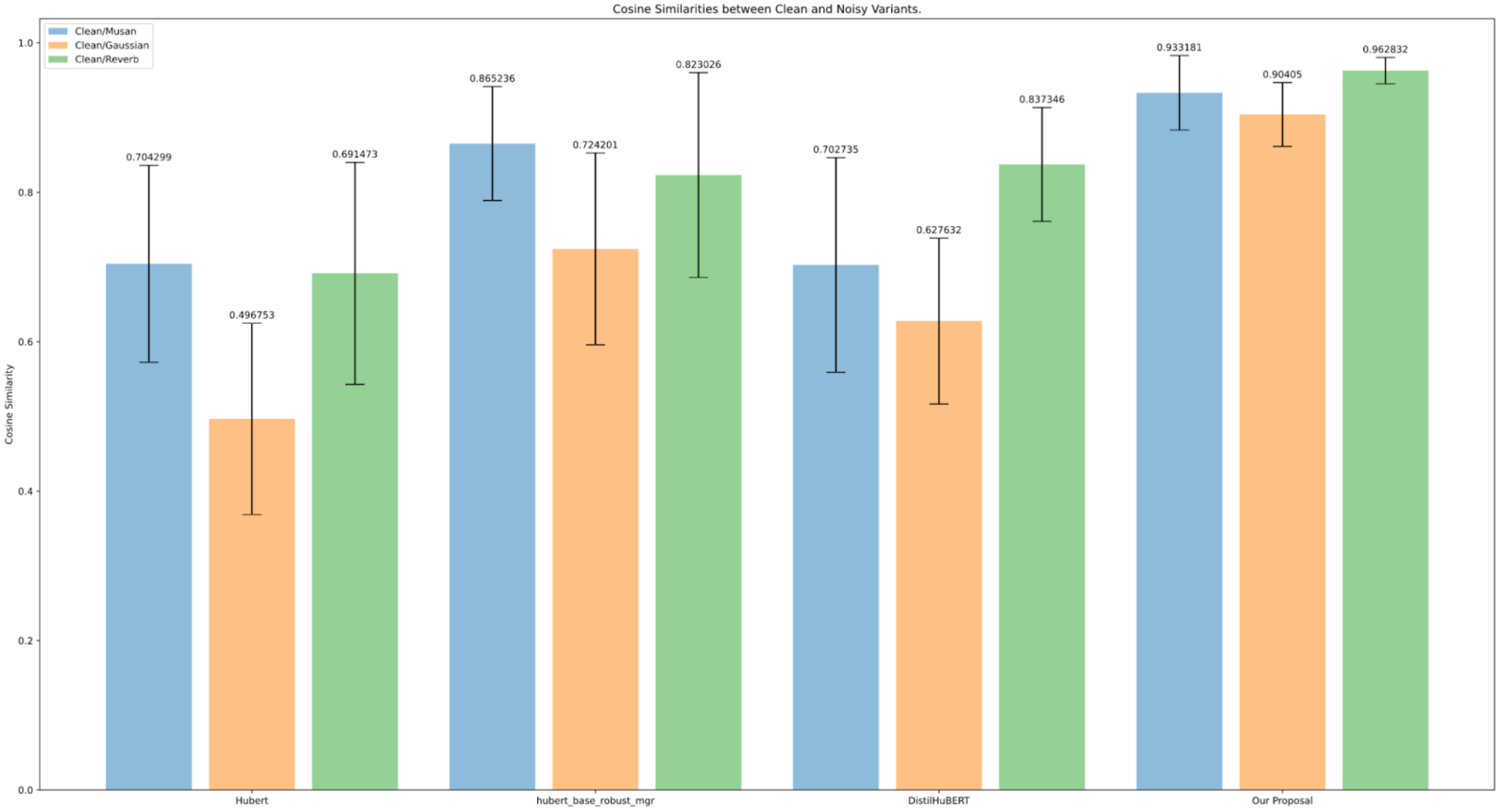


This plot correspond to the last model, but by grouping the labels by embedding ID. This plot shows that same speech IDs but with distortions lies in the same space. Thus, features are noise invariant!



Cosine Similarities between clean and noisy embeddings

- Aiming at having a concrete objective number to assess how invariant the feature representations are, here we show the cosine similarities between the clean embeddings and its noisy variants for each of the previous models. Bigger similarity means more robustness.



Cosine Similarities between Clean and Noisy Variants.
Clean/Musan
Clean/Gaussian
Clean/Reverb
Cosine Similarity
1.0
0.8
0.6
0.4
0.2
0.0
0.704299
0.496753
0.691473
0.865236
0.724201
0.823026
0.702735
0.627632
0.837346
0.933181
0.90405
0.962832
Hubert
hubert_base_robust_mgr
DistilHuBERT
Our Proposal

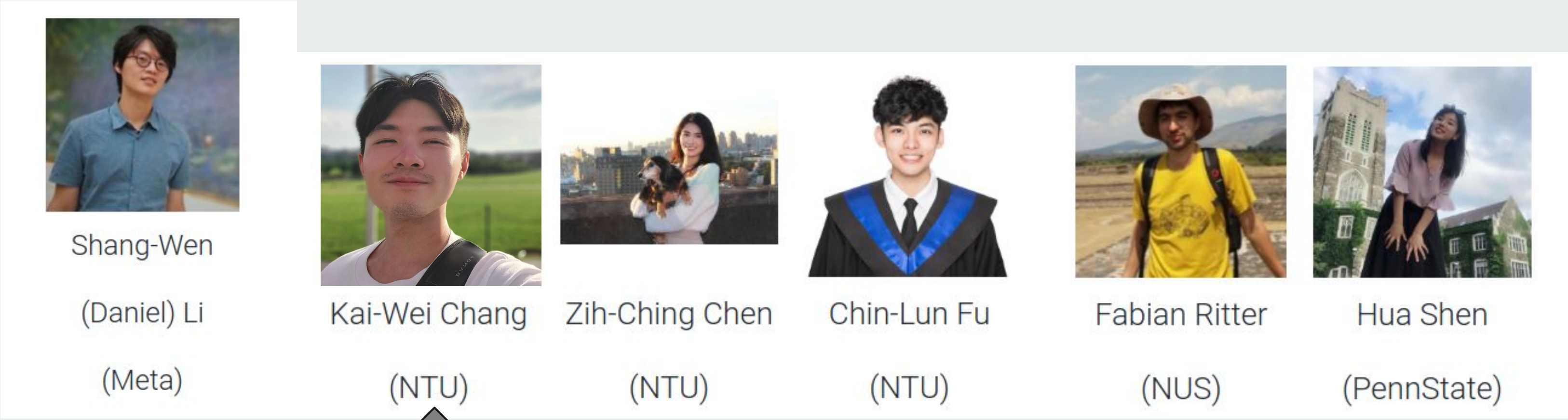
Conclusion

- Our proposed method proves to be robust while keeping a low number of parameters.
- Visualization corroborates noise invariant characteristics of our model.

JSALT Progress Report 20 July 2022

Leveraging Pre-training Models for Speech Processing

Speech Prompt

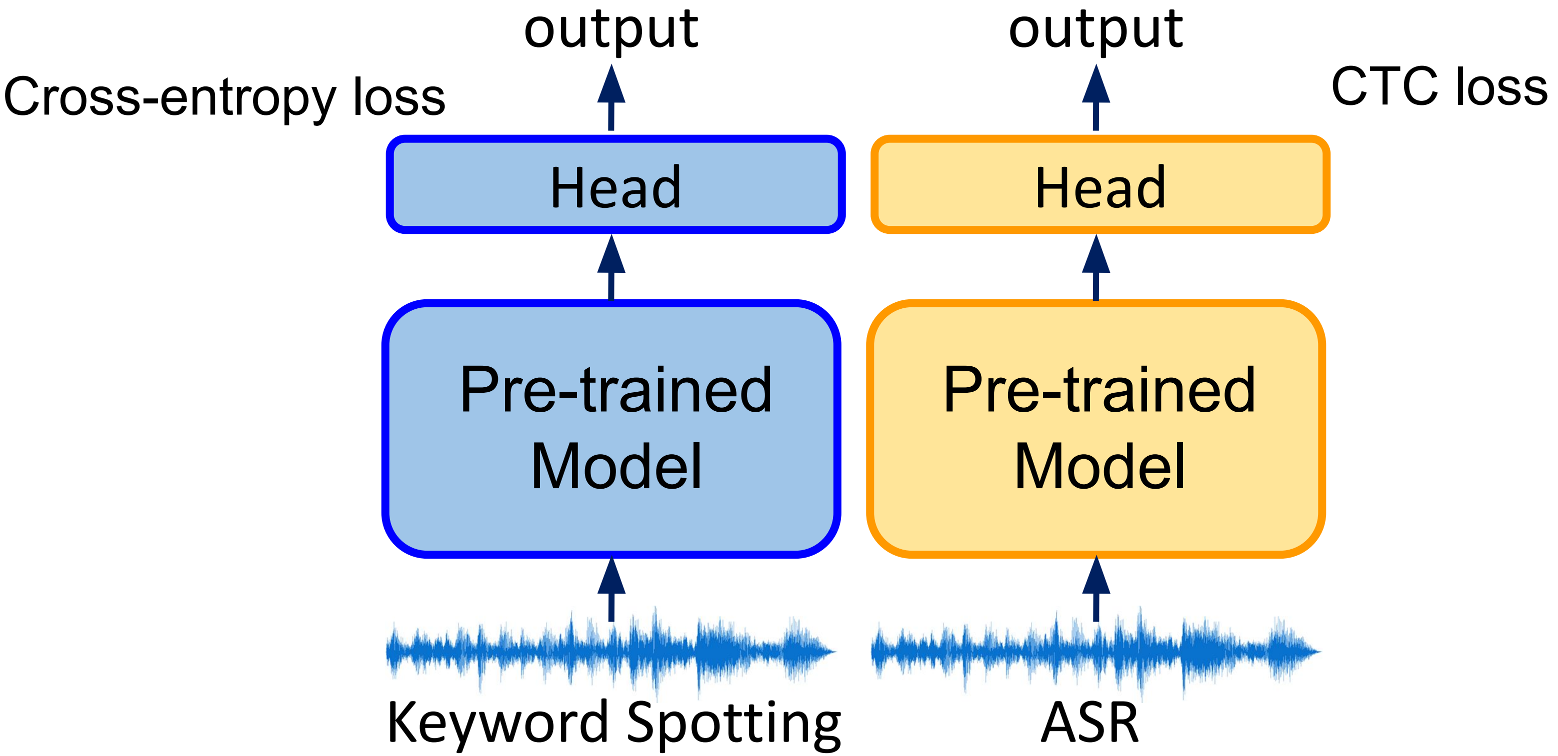


Presenter: Kai-Wei Chang, National Taiwan University

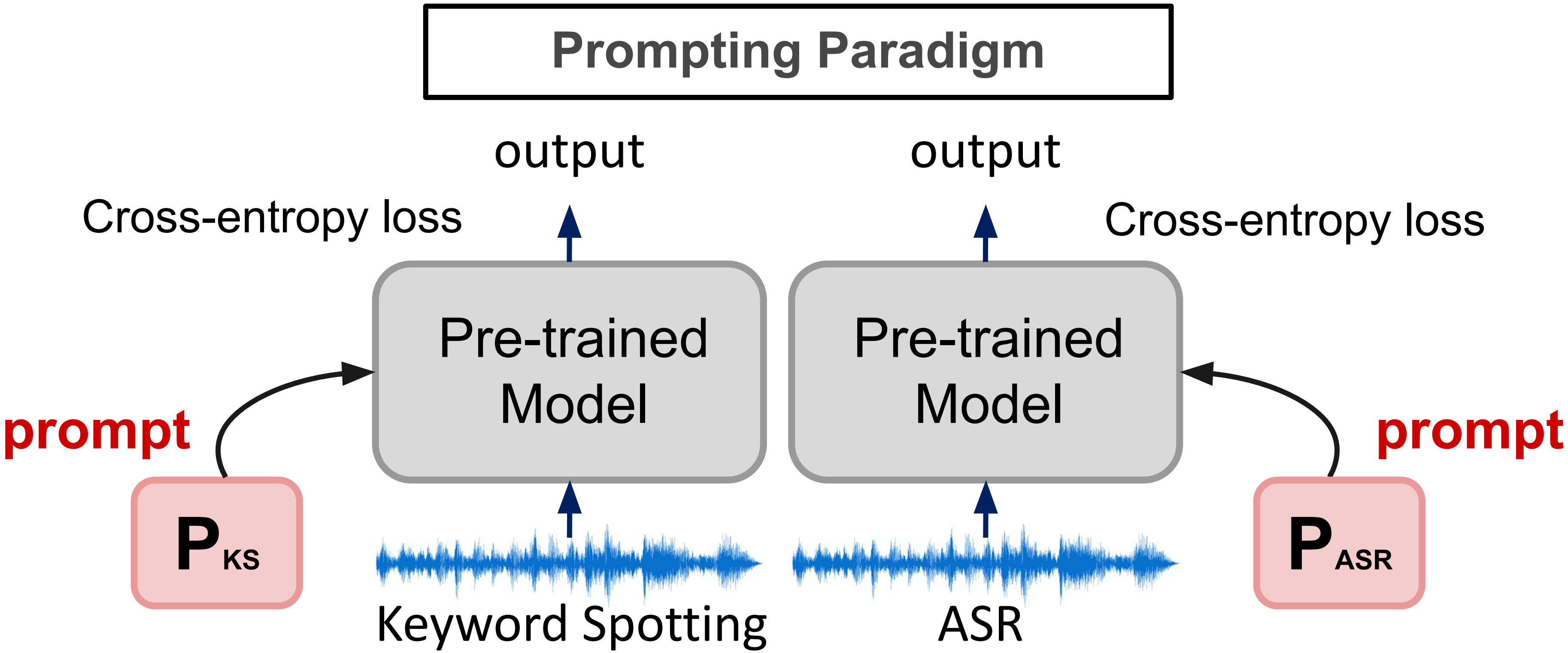
Motivation

Motivation

Pre-train, Fine-tune Paradigm



Motivation



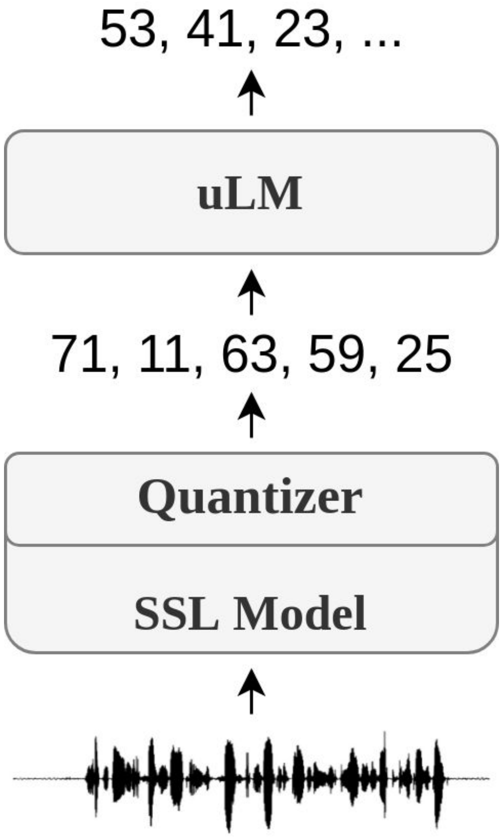
Prompting Paradigm
output
output
Cross-entropy loss
Cross-entropy loss
Pre-trained Model
Pre-trained Model
prompt
prompt
P KS
P ASR
Keyword Spotting
ASR

Method

Background - GSLM

Generative Spoken Language Model

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



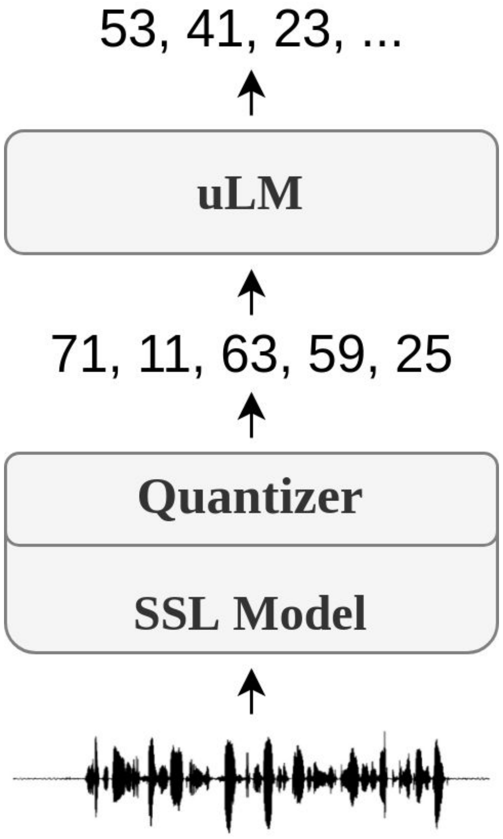
- [GSLM] Lakhotia et.al., Generative Spoken Language Modeling from Raw Audio

Background - GSLM

Generative Spoken Language Model

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

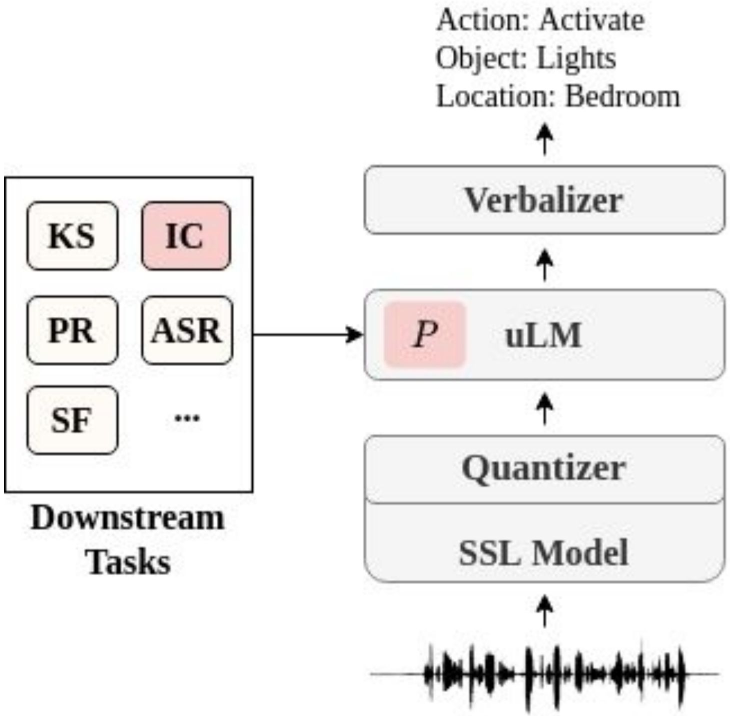
- Autoregressive LM
(flexible output length)



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

Prompting for GSLM: Framework

- Prompt: a small set of trainable parameters for each task
- uLM: generate units conditioned on the prompts
- Verbalizer: map the units back to task labels.

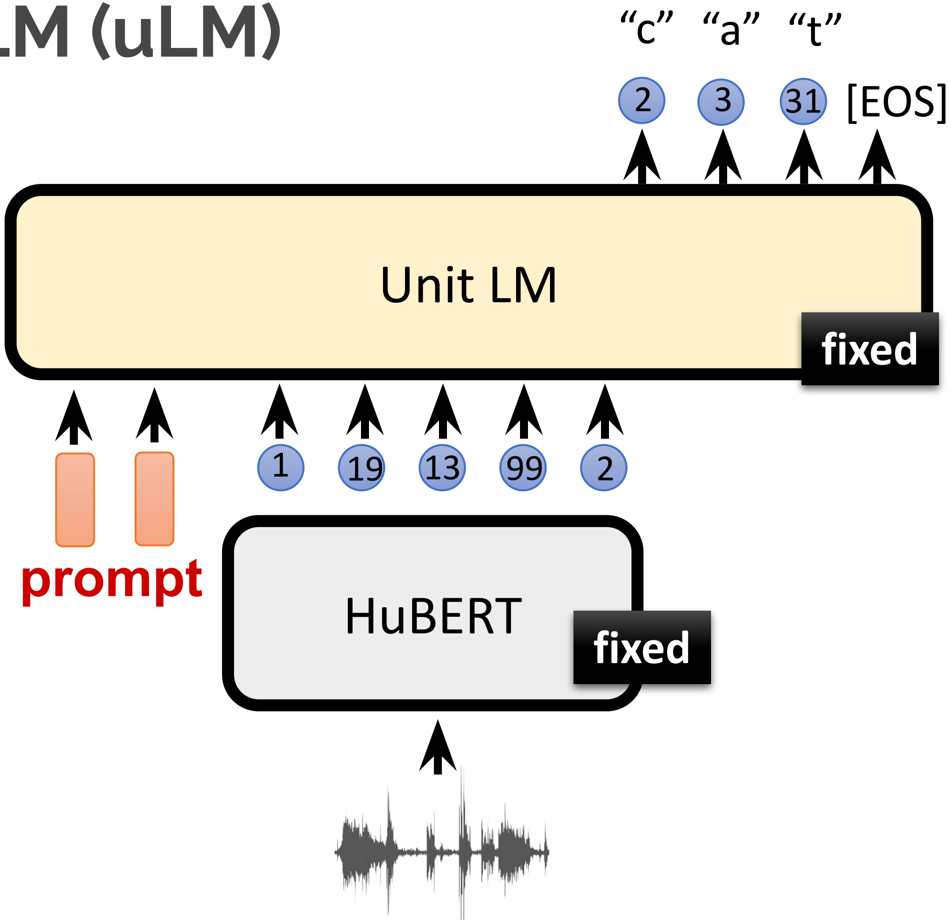


Prompting on Unit LM (uLM)

Speech Recognition (ASR)

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

label mapping (Verbalizer)



Experiments

Experiment Results - Speech Classification

- 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
HuBERT-FT	**96.30**	0.2M	98.34	0.2M

Experiment Results - Speech Classification

- 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.
CPC-PT	**93.54**	0.05M	**97.57**	0.05M
CPC-FT	91.88	0.07M	64.09	0.07M

Experiment Results - Sequence Generation

- PT: Prompt Tuning
- FT: Fine-Tuning

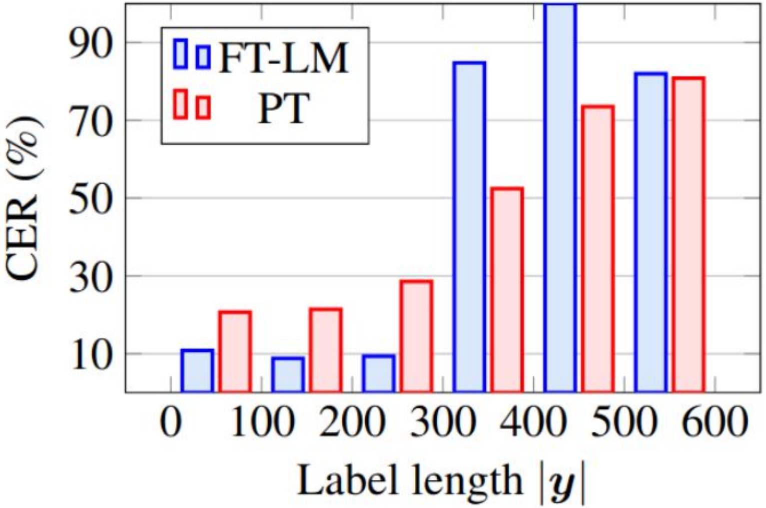
- ASR: Automatic Speech Recognition
- SF: Slot Filling

Scenarios	ASR		SF	
	WER↓	# param.	F1↑	# param.
HuBERT-PT	34.17	4.5M	66.90	4.5M
HuBERT-FT	**6.42**	43M	**88.53**	43M

Analysis - The Curse of Long Sequences

Task	Type	Avg. label length
KS	CLS	1
IC	CLS	3
ASR	SG	173
SF	SG	54

ASR



The performance suffers from long sequences severely!

Future Works

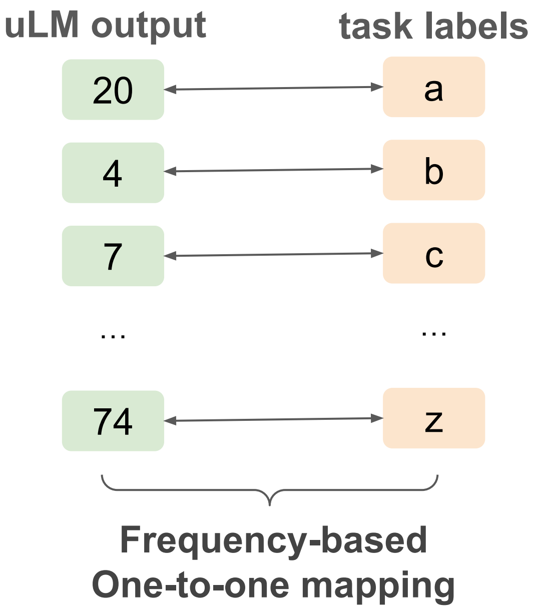
Future works

For sequence generation task, the performance suffers from “long sequences”

- Sequence compression technique

For now, we’re using a heuristic verbalizer

- Learnable verbalizer (a small neural network)



Conclusion

Conclusion

1. The first exploration of prompting for speech processing tasks
2. It shows high efficiency on speech classification tasks
3. We’re trying different methods to improve the performance on sequence generation tasks

SpeechPrompt: An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks

Kai-Wei Chang[1], *Wei-Cheng Tseng*[1], *Shang-Wen Li*[2], *Hung-yi Lee*[1]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]Amazon AI, USA

kaiwei.chang.tw@gmail.com, r09942094@ntu.edu.tw, swdanielli@gmail.com, hungyilee@ntu.edu.tw

Accepted at Interspeech 2022

https://arxiv.org/abs/2203.16773

Q&A