Towards Universal Self-supervised Model for Speech Processing

Hung-yi Lee

https://speech.ee.ntu.edu.tw/~hylee/

National Taiwan University
國立臺灣大學

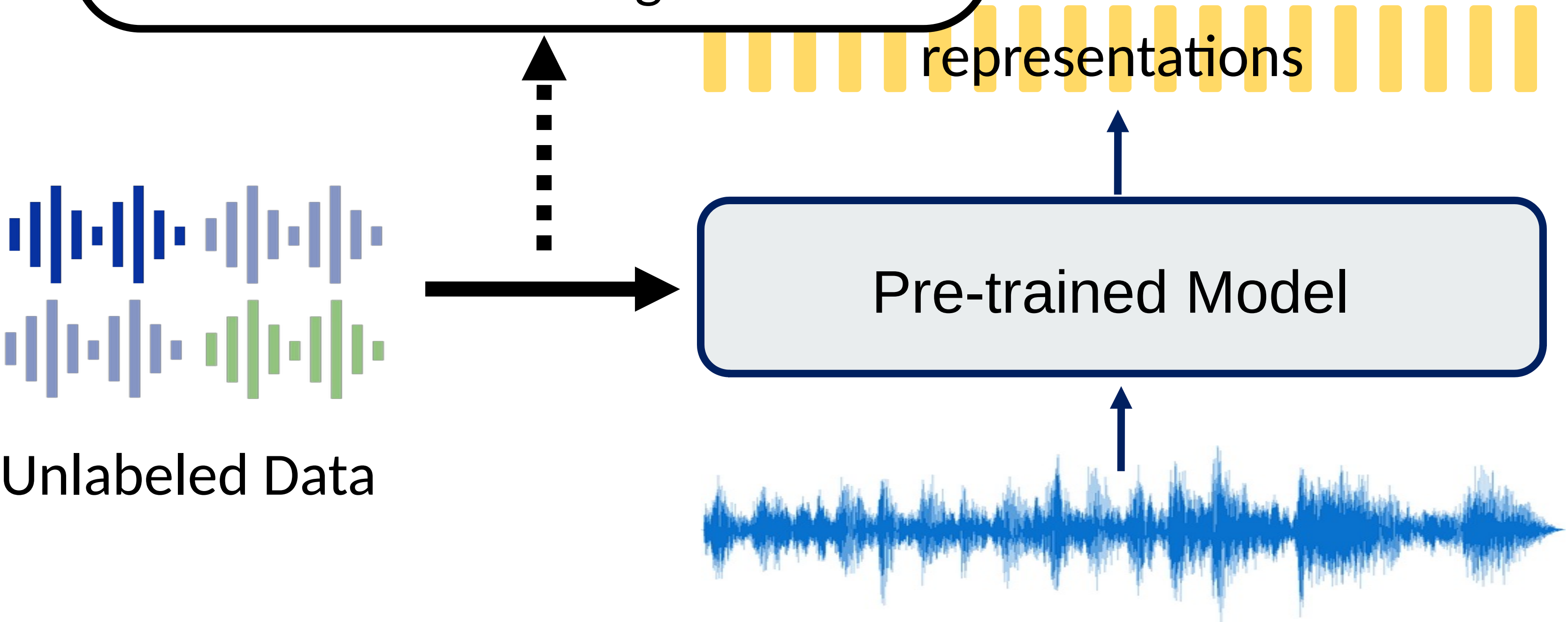
Self-supervised Learning Framework

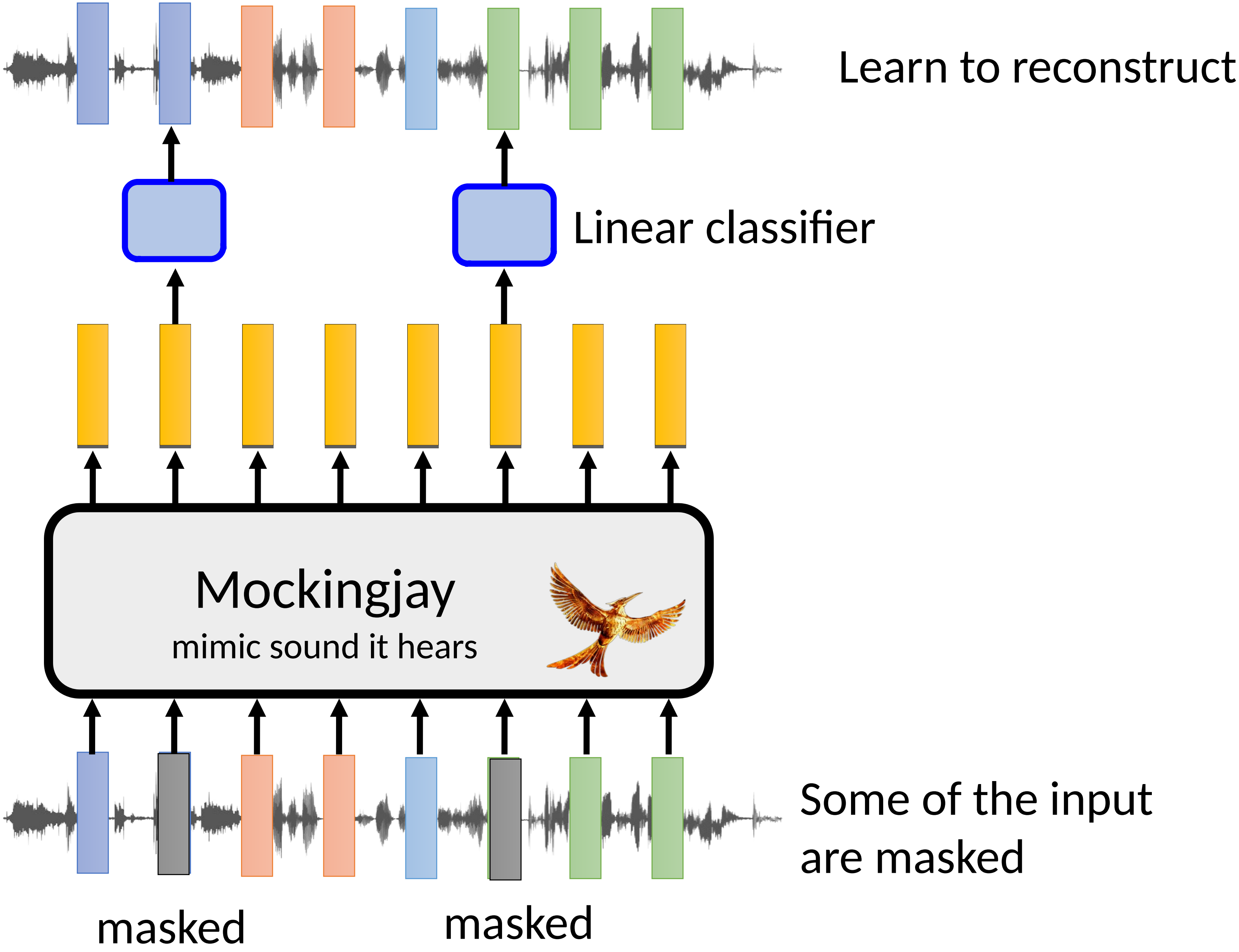
Phase 1: Pre-train

(not complete survey)

- Mask the input signals and then reconstruct them.
- Predict the targets obtained without human efforts.
- Contrastive learning

Task-agnostic





Learn to reconstruct

Linear classifier

Mockingjay

mimic sound it hears

Some of the input are masked

masked

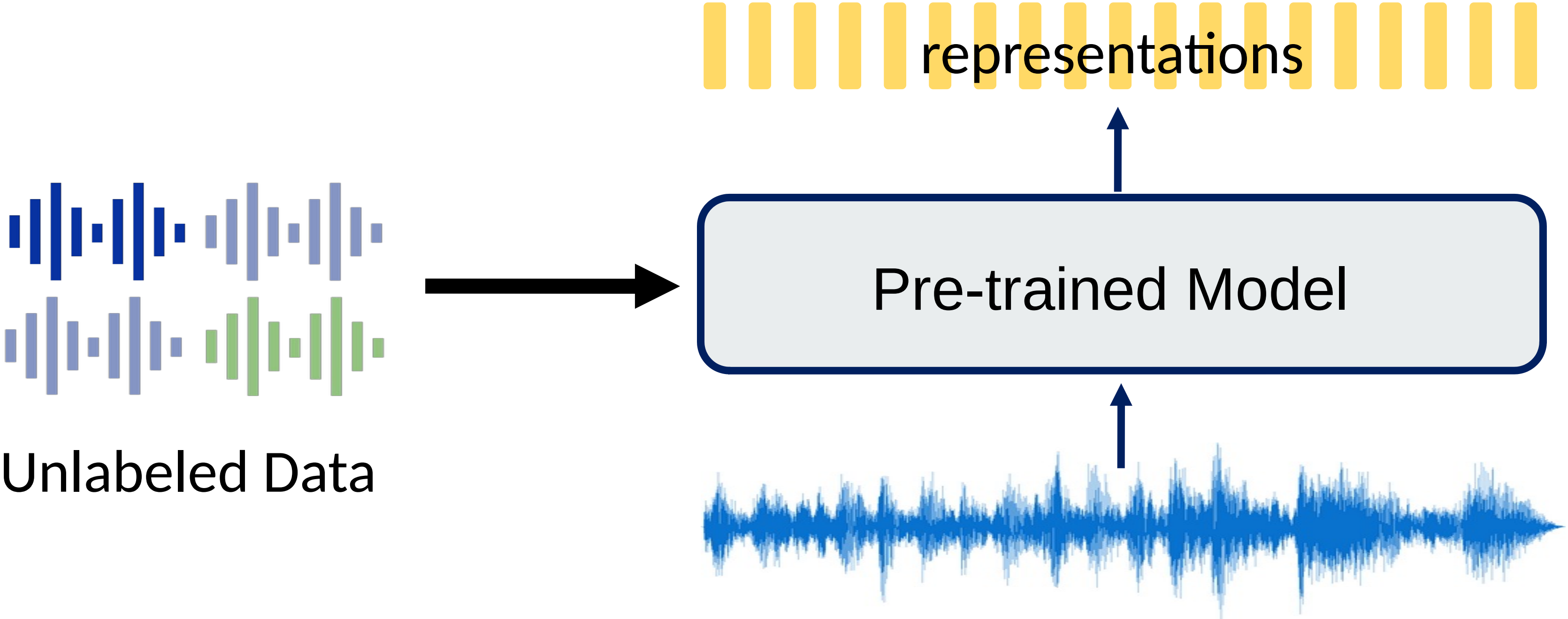
masked

Self-supervised Learning Framework

Phase 1: Pre-train

Just name a few …

PASE+ APC NPC Mockingjay DeCoAR Wav2vec HuBERT



Unlabeled Data

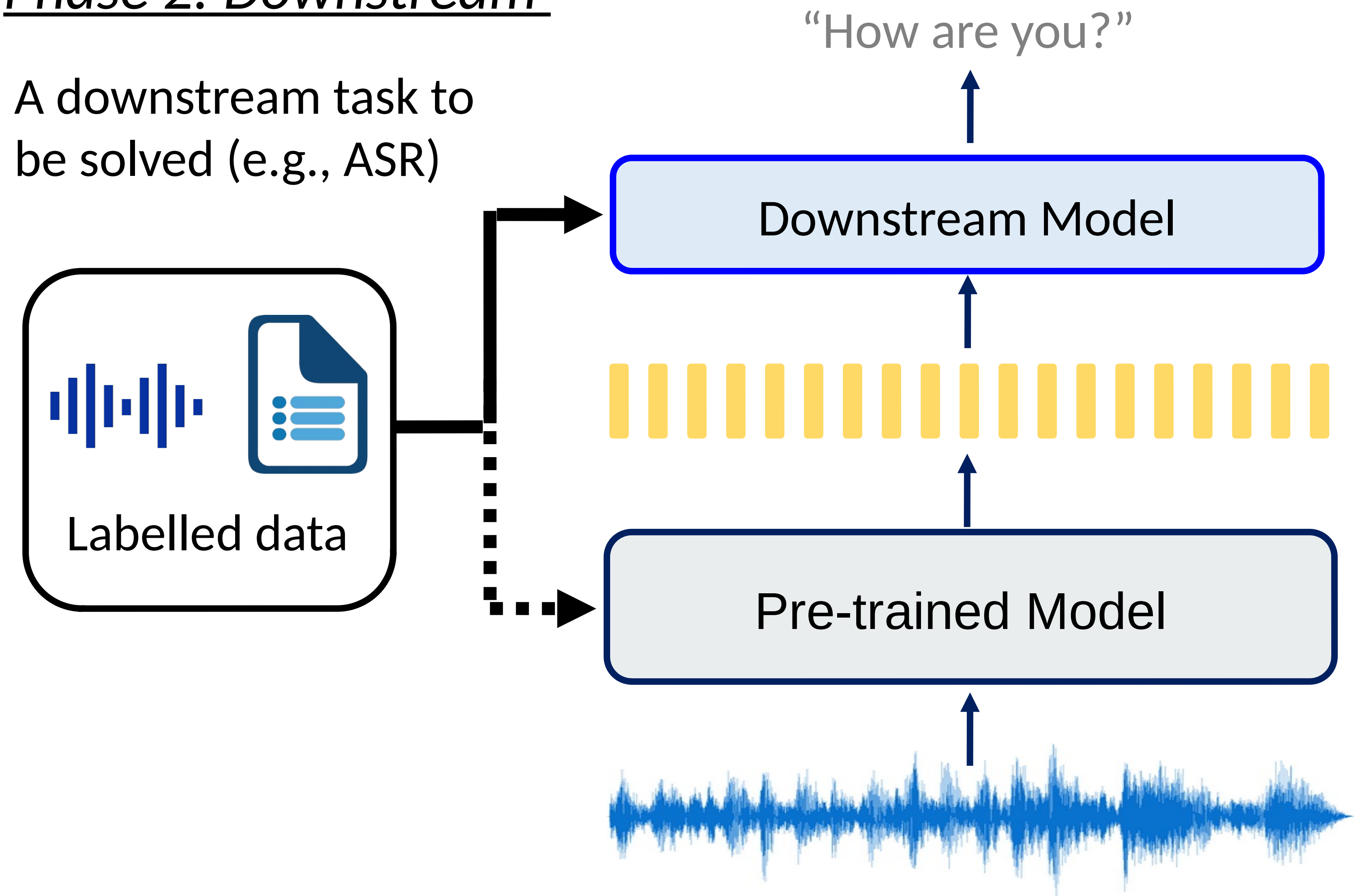
representations

Pre-trained Model

Self-supervised Learning Framework

Phase 2: Downstream

A downstream task to be solved (e.g., ASR)



Specialist? Universal?

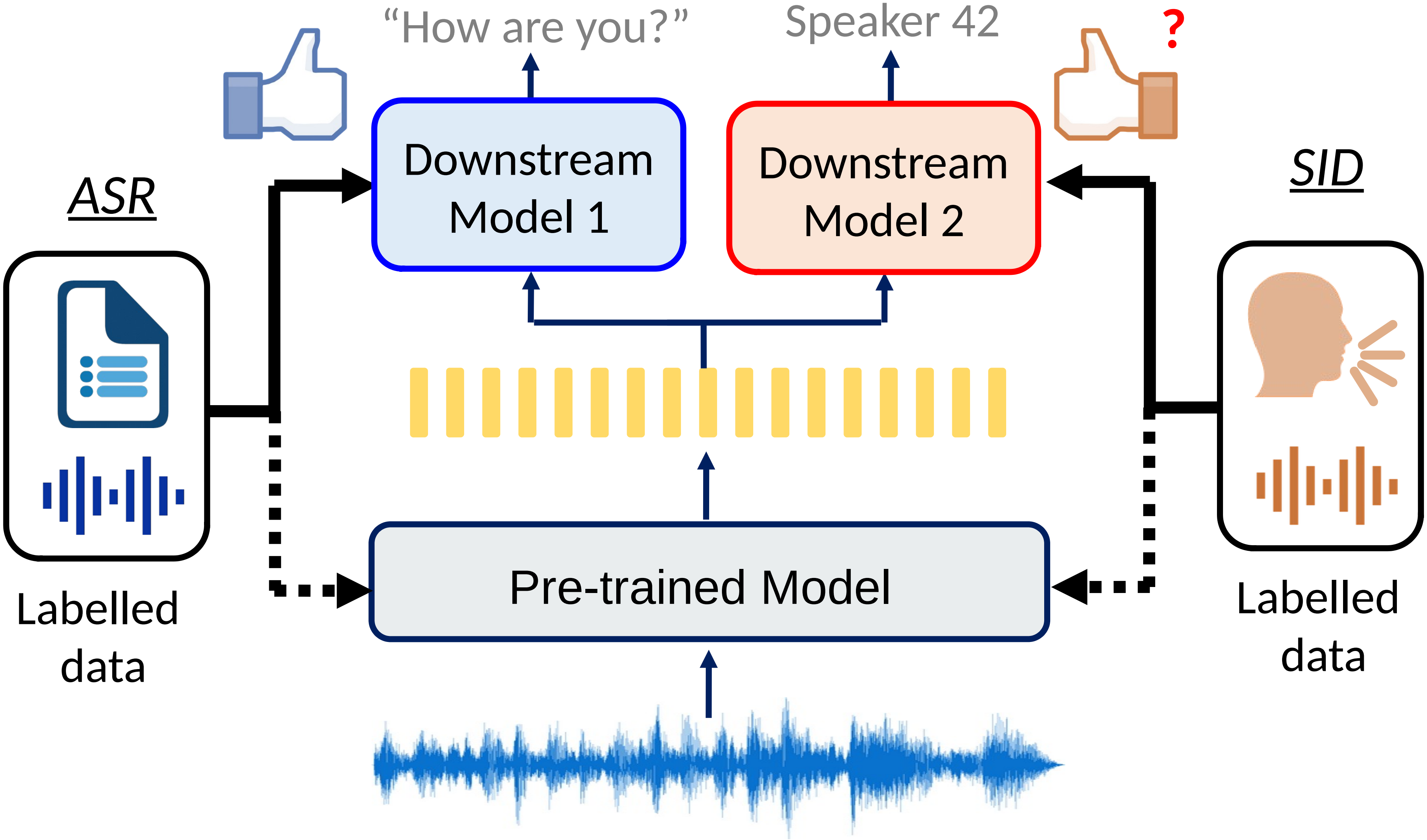
Just name a few …

PASE+ APC NPC Mockingjay DeCoAR Wav2vec HuBERT

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

Specialist? Universal?



“How are you?”

Speaker 42

?

Downstream Model 1

Downstream Model 2

ASR

SID

Pre-trained Model

Labelled data

Labelled data

Specialist? Universal?

Just name a few …

PASE+ APC NPC Mockingjay DeCoAR Wav2vec HuBERT

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

- I believe they are specialist.
- To be good at ASR, a model learns to extract content and ignore speaker.
- Hence, super good on ASR Poor performance on speaker related tasks.

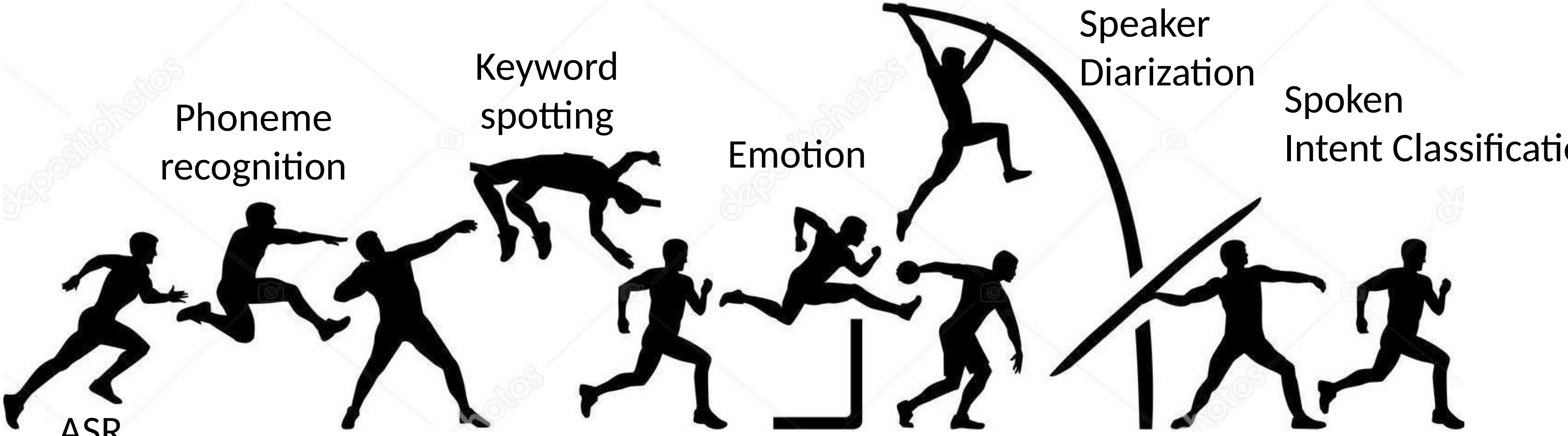
My two cents

(one year ago)

SUPERB

Speech processing Universal PERformance Benchmark

PASE+ APC NPC Mockingjay DeCoAR Wav2vec HuBERT



Keyword spotting

Phoneme recognition

Emotion

Speaker Diarization

Spoken Intent Classificati

ASR

Query-by-example

Speaker Identification

Speaker verification

Spoken Slot Filling

https://arxiv.org/abs/2105.01051

SUPERB

Speech processing Universal PERformance

National Taiwan University 國立臺灣大學

Carnegie Mellon University

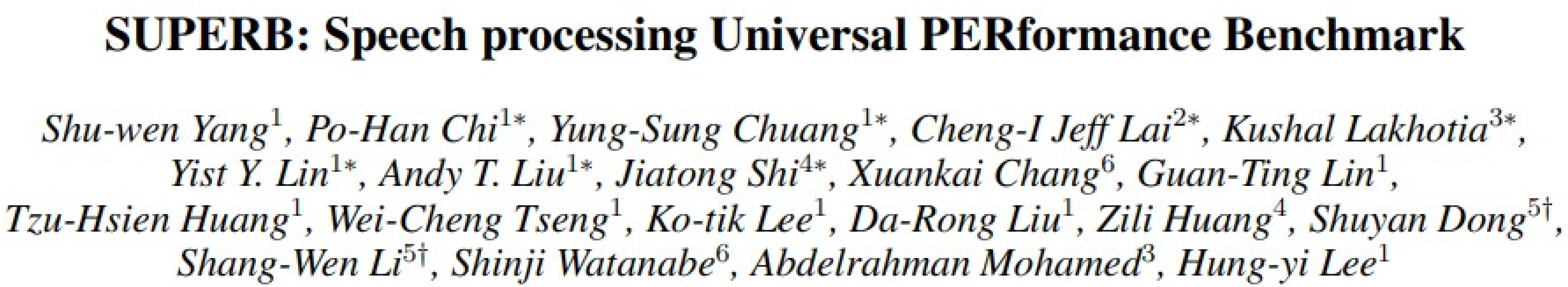
Massachusetts Institute of Technology

JOHNS HOPKINS UNIVERSITY

FACEBOOK AI

LXT

Hugging Face

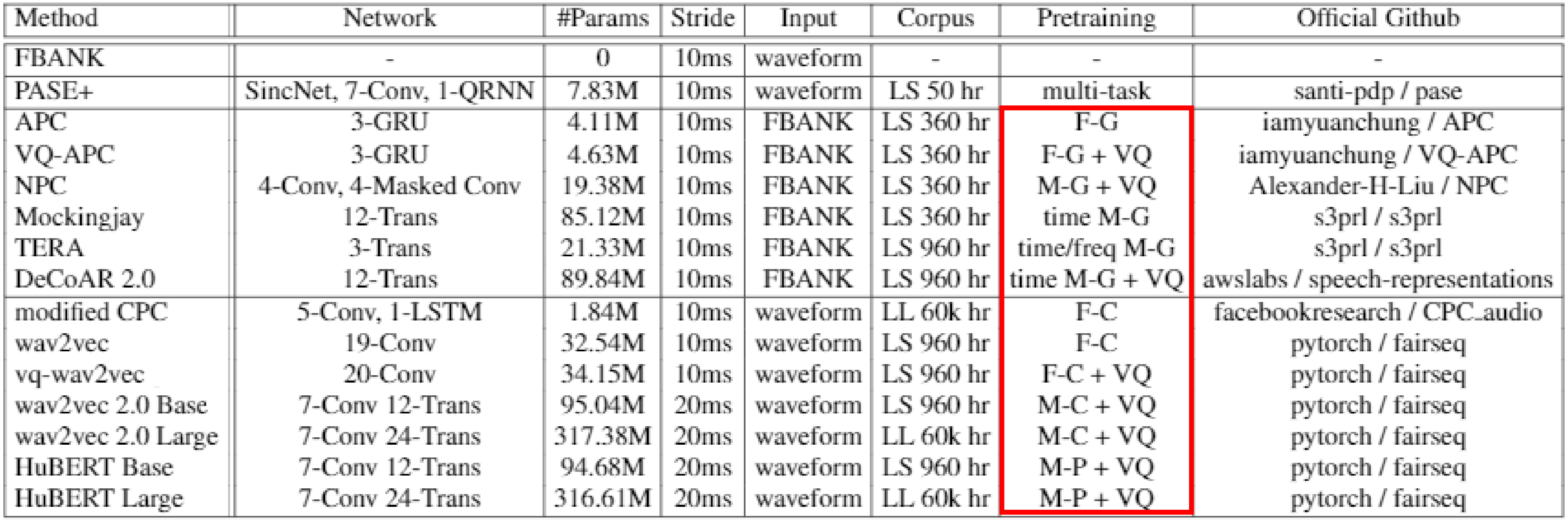


SUPERB: Speech processing Universal PERformance Benchmark

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

Presented at INTERSPEECH 2021

Introduction of Contestants

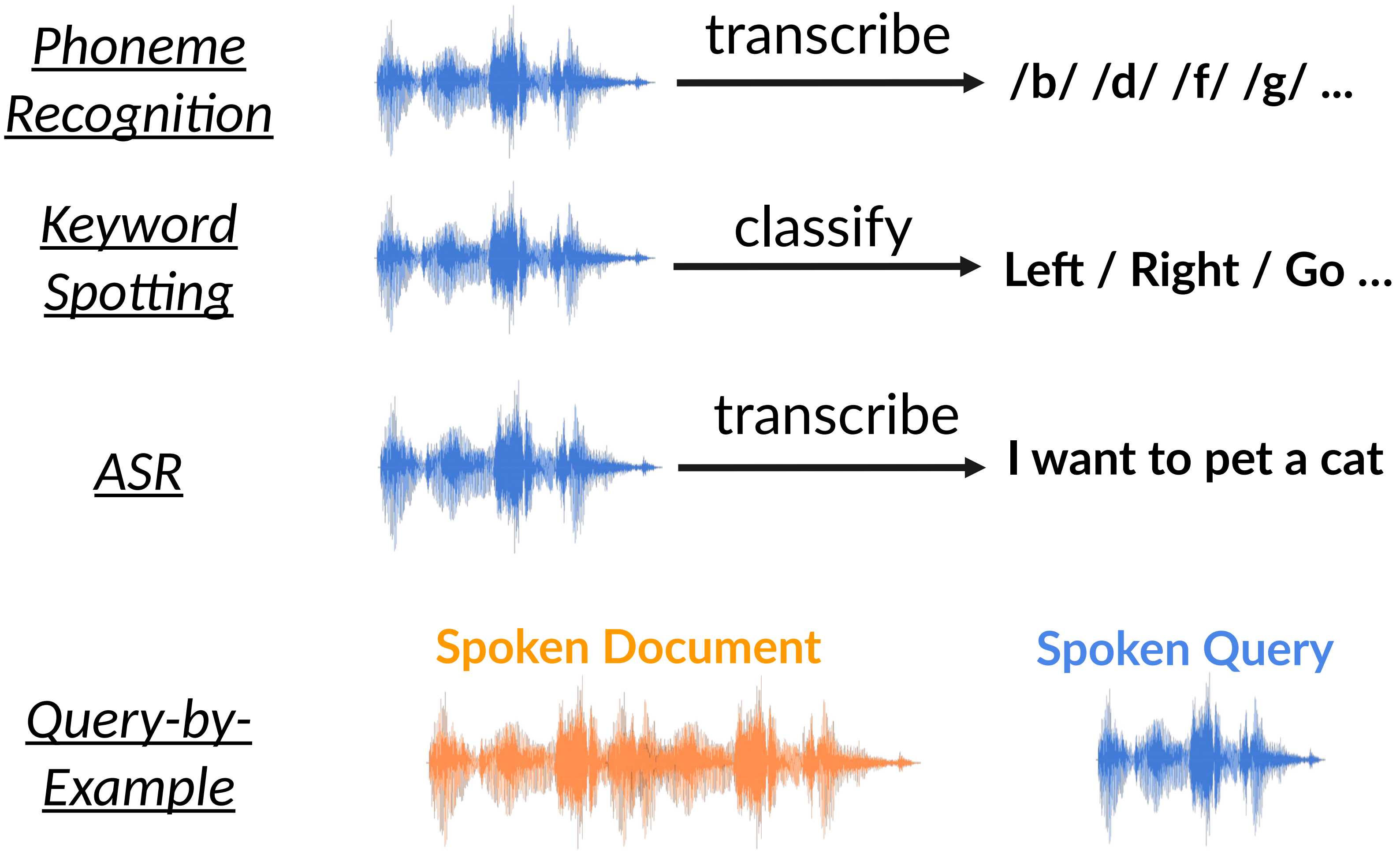


Method	Network	#Params	Stride	Input	Corpus	Pretraining	Official Github
FBANK	-	0	10ms	waveform	-	-	-
PASE+	SincNet, 7-Conv, 1-QRNN	7.83M	10ms	waveform	LS 50 hr	multi-task	santi-pdp / pase
APC	3-GRU	4.11M	10ms	FBANK	LS 360 hr	F-G	iamyuanchung / APC
VQ-APC	3-GRU	4.63M	10ms	FBANK	LS 360 hr	F-G + VQ	iamyuanchung / VQ-APC
NPC	4-Conv, 4-Masked Conv	19.38M	10ms	FBANK	LS 360 hr	M-G + VQ	Alexander-H-Liu / NPC
Mockingjay	12-Trans	85.12M	10ms	FBANK	LS 360 hr	time M-G	s3prl / s3prl
TERA	3-Trans	21.33M	10ms	FBANK	LS 960 hr	time/freq M-G	s3prl / s3prl
DeCoAR 2.0	12-Trans	89.84M	10ms	FBANK	LS 960 hr	time M-G + VQ	awslabs / speech-representations
modified CPC	5-Conv, 1-LSTM	1.84M	10ms	waveform	LL 60k hr	F-C	facebookresearch / CPC_audio
wav2vec	19-Conv	32.54M	10ms	waveform	LS 960 hr	F-C	pytorch / fairseq
vq-wav2vec	20-Conv	34.15M	10ms	waveform	LS 960 hr	F-C + VQ	pytorch / fairseq
wav2vec 2.0 Base	7-Conv 12-Trans	95.04M	20ms	waveform	LS 960 hr	M-C + VQ	pytorch / fairseq
wav2vec 2.0 Large	7-Conv 24-Trans	317.38M	20ms	waveform	LL 60k hr	M-C + VQ	pytorch / fairseq
HuBERT Base	7-Conv 12-Trans	94.68M	20ms	waveform	LS 960 hr	M-P + VQ	pytorch / fairseq
HuBERT Large	7-Conv 24-Trans	316.61M	20ms	waveform	LL 60k hr	M-P + VQ	pytorch / fairseq

- G: reconstructing the input
- P: token prediction
- C: contrastive learning

- VQ: quantization
- F: predicting future information
- M: input masking

Tasks – Content



Phoneme Recognition

transcribe → **/b/ /d/ /f/ /g/ ...**

Keyword Spotting

classify → **Left / Right / Go ...**

ASR

transcribe → **I want to pet a cat**

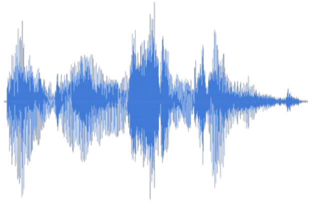
Query-by-Example

Spoken Document

Spoken Query

Tasks – Speaker

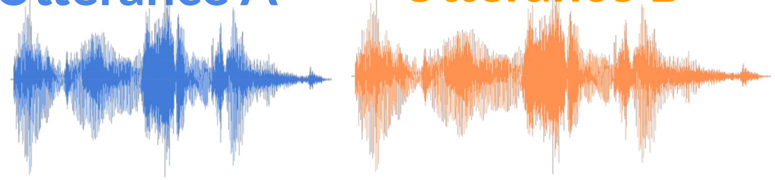
Speaker Identification



classify → **Speaker ID**

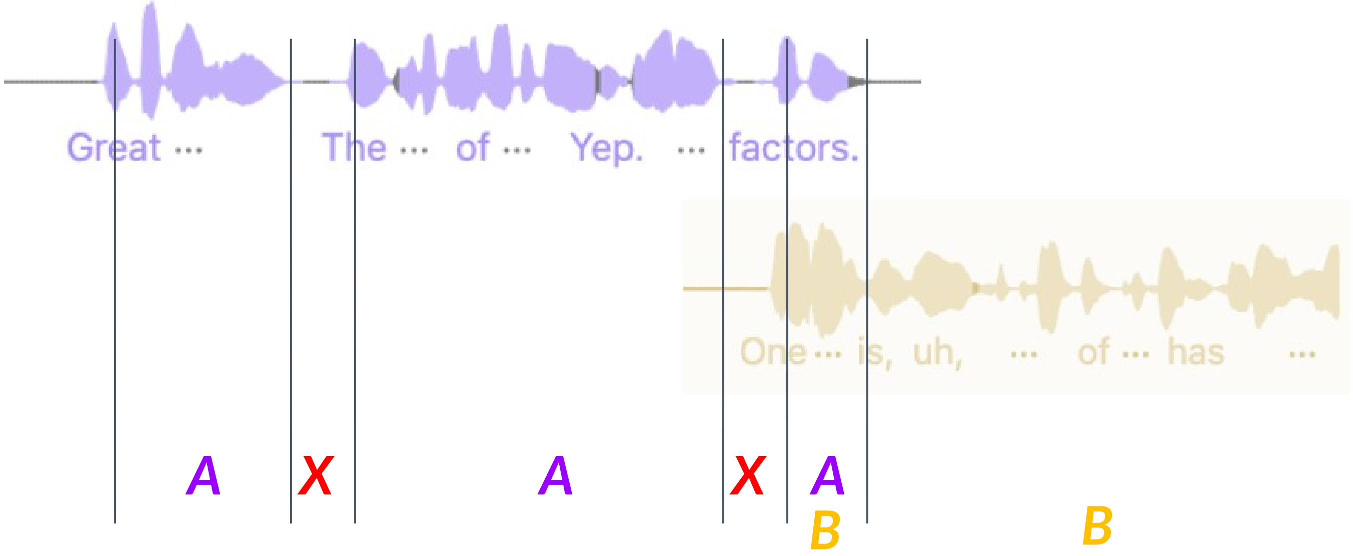
Speaker Verification

<span style="color:#4a7fd4">Utterance A</span> Utterance B



A & B
***same speaker*?**
Yes / No

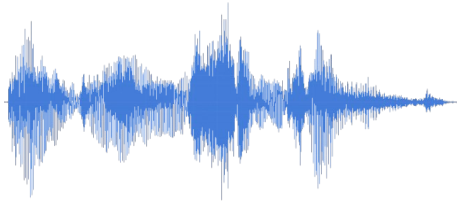
Speaker Diarization



Great ···
The ··· of ··· Yep. ··· factors.
One ··· is, uh, ··· of ··· has ···
A
X
A
X
A
B
B

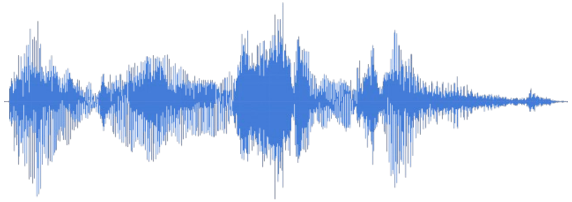
Tasks – Semantic

Intent Classification



classify → **Intent classes**

Slot Filling



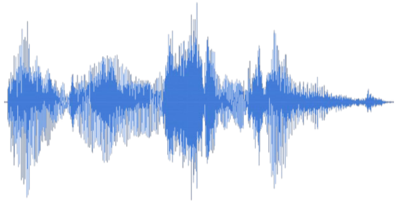
I fly from **Taipei** to **New York**

extract →

Slot type	***Slot value***
from_location	Taipei
to_location	New York

Tasks – Emotion

Emotion Recognition



classify → **Happy/Angry/Sad …**

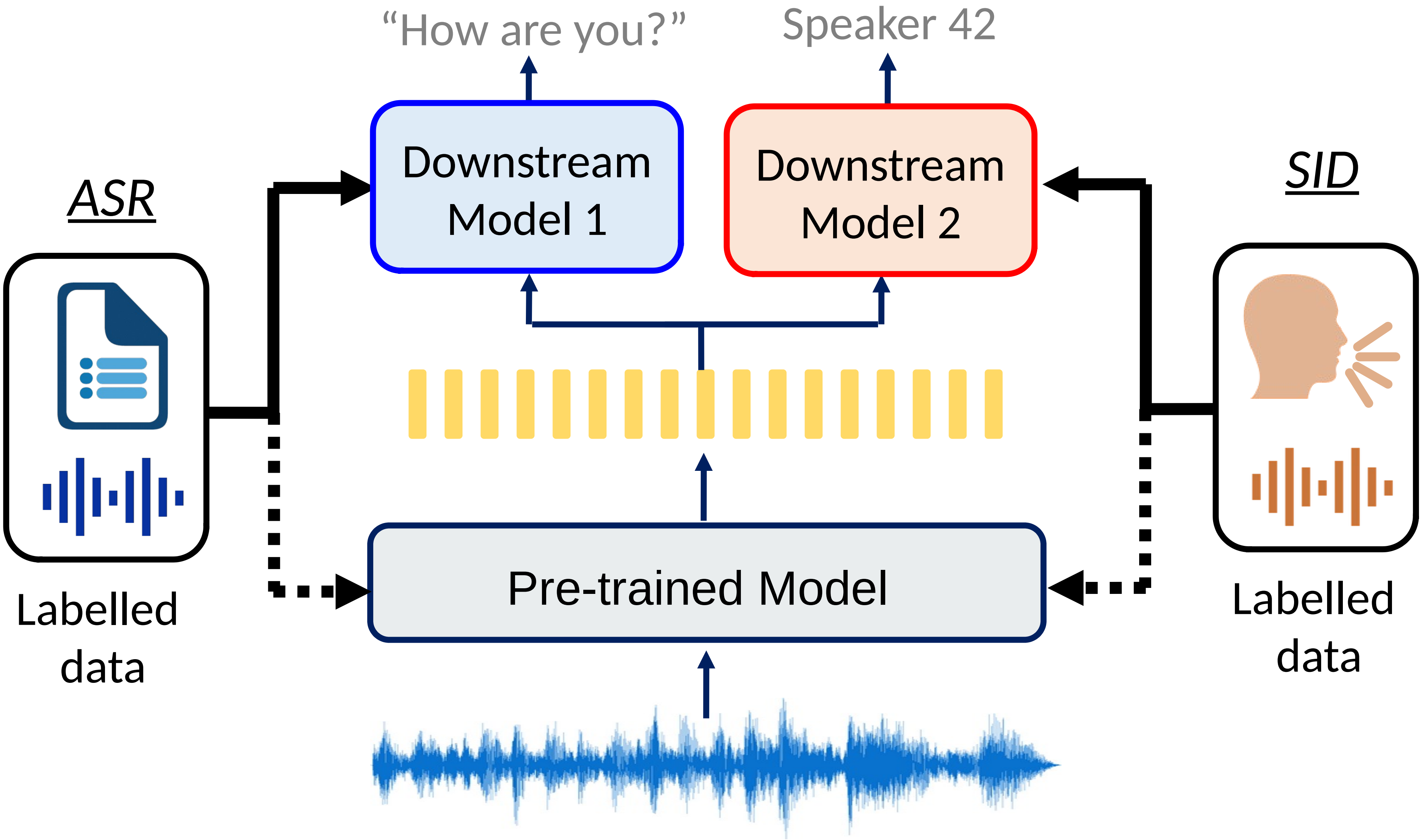
Please refer to the paper for more details.

Game Start!

Round 1

Rules in Round 1

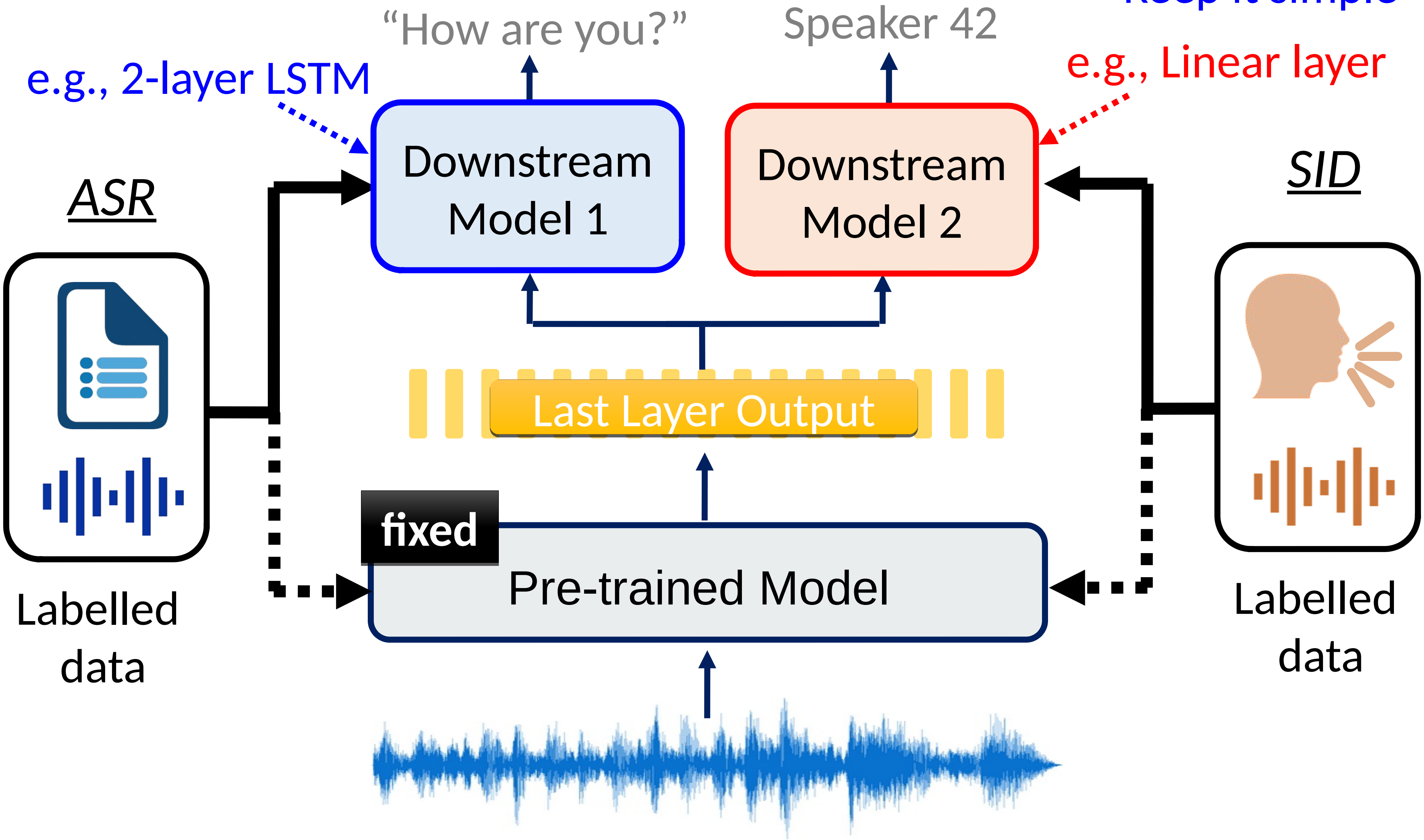
I only put two out of ten downstream models for simplicity.



Rules in Round 1

The network architecture of a downstream model is predefined.

Keep it simple



e.g., 2-layer LSTM

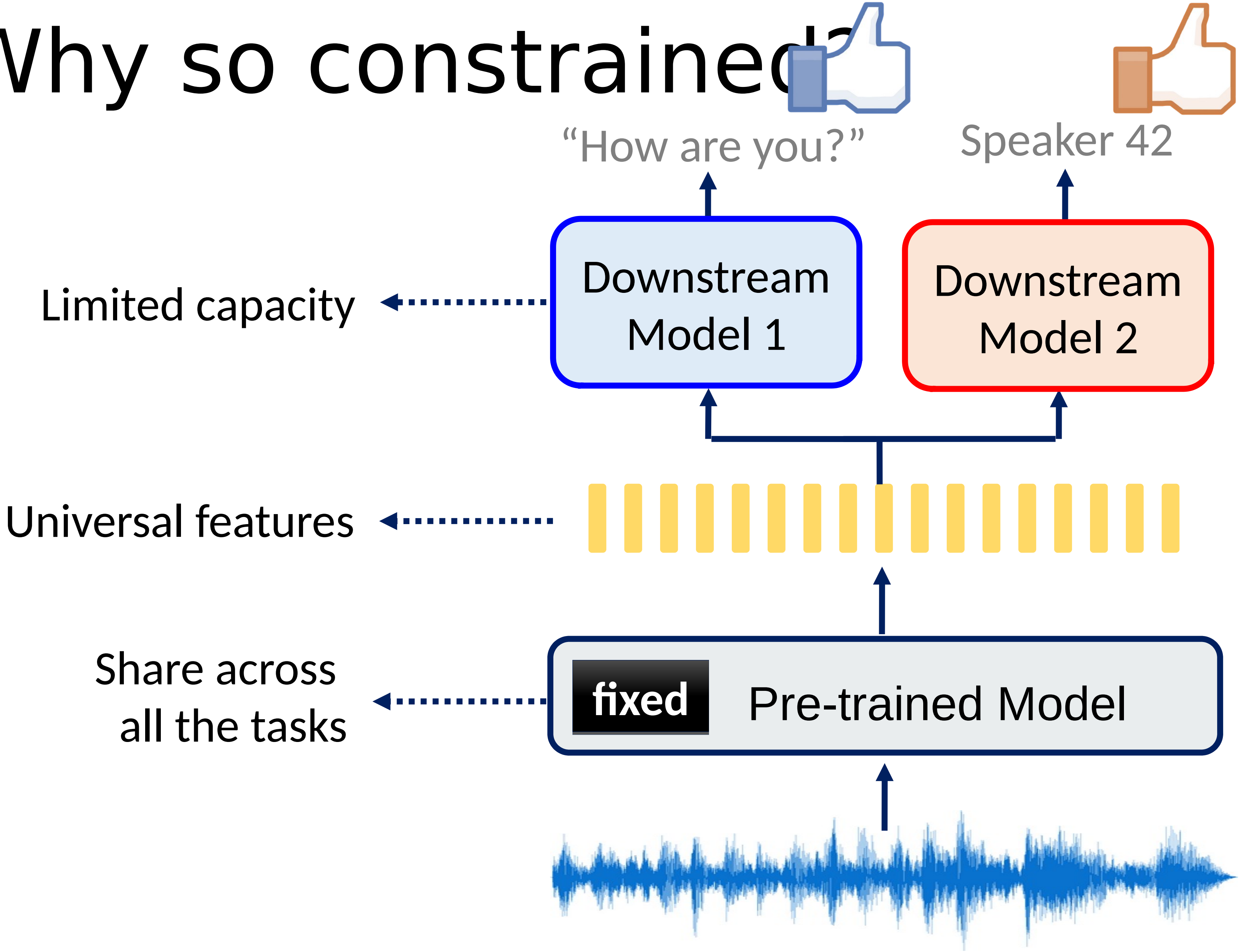
e.g., Linear layer

Rules in Round 1 – Downstream

- Phoneme Recognition: linear layer
- Keyword Spotting: linear layer
- Speech Recognition: 2-layer LSTM
- Query-by-example: none
- Speaker Identification: linear layer
- Speaker Verification: the same as x-vector
- Speaker Diarization: 1-layer LSTM
- Intent Classification: linear layer
- Slot Filling: 2-layer LSTM

Keep it simple

Why so constrained?



“How are you?”

Speaker 42

Limited capacity

Downstream Model 1

Downstream Model 2

Universal features

Share across all the tasks

fixed

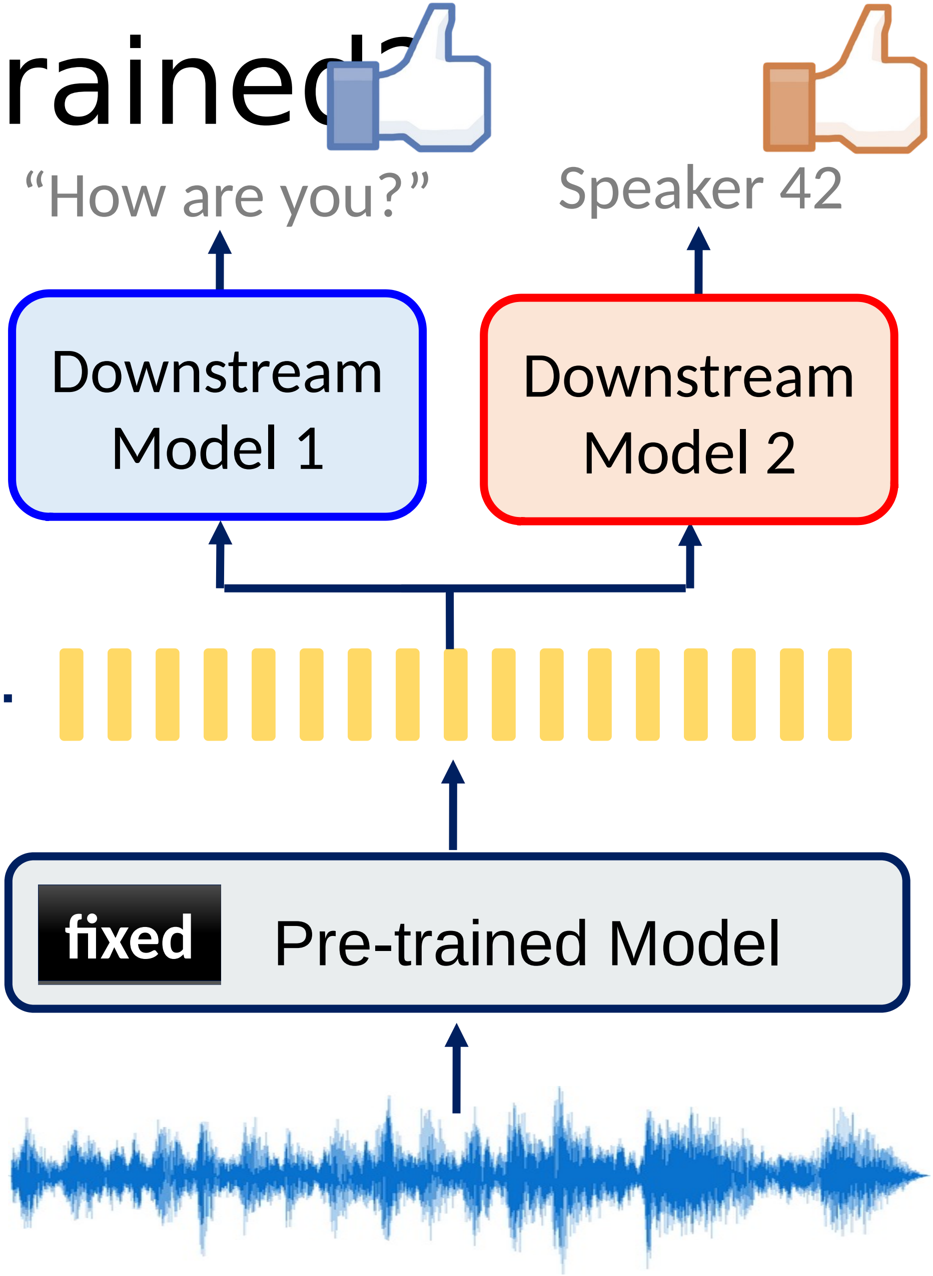
Pre-trained Model

Why so constrained?

Easy to build new applications!

Universal features

This sounds too good to be true ……



Results of Round 1

	Content				Speaker			Semantic		Emotion
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.88	82.37	16.61	7.00E-04	35.84	10.91	8.52	30.29	60.41	57.64
APC	41.85	91.04	15.09	0.0268	59.79	8.81	10.72	74.64	71.26	58.84
VQ-APC	42.86	90.52	15.37	0.0205	49.57	9.29	10.49	70.52	69.62	58.31
NPC	52.67	88.54	14.69	0.022	50.77	10.28	9.59	64.04	67.43	59.55
Mockingjay	80.01	82.67	15.94	3.10E-10	34.5	23.22	11.24	28.87	60.83	45.72
TERA	47.53	88.09	12.44	8.70E-05	58.67	16.49	9.54	48.8	63.28	54.76
modified CPC	41.66	92.02	13.57	0.0061	42.29	9.67	11.00	65.01	74.18	59.28
wav2vec	32.39	94.09	11.3	0.0307	44.88	9.83	10.79	78.91	77.52	58.17
vq-wav2vec	53.49	92.28	12.69	0.0302	39.04	9.50	9.93	59.4	70.57	55.89
wav2vec 2.0 base	28.37	92.31	6.32	8.80E-04	45.62	9.69	7.48	58.34	79.94	56.93
HuBERT base	6.85	95.98	4.93	0.0759	64.84	7.22	6.76	95.94	86.24	62.94

Pre-trained Models

Results of Round 1

	Content				Speaker			Semantic		Emotion
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.88	82.37	■	■	35.84	■	8.52	30.29	■	57.64
APC	41.85	91.04	15.09	0.0268	59.79	8.81	■	74.64	71.26	58.84
VQ-APC	42.86	90.52	■	0.0205	49.57	9.29	■	70.52	■	58.31
NPC	52.67	88.54	14.69	0.022	50.77	■	9.59	64.04	■	59.55
Mockingjay	80.01	82.67	■	■	34.5	■	■	28.87	■	45.72
TERA	47.53	88.09	12.44	■	58.67	■	9.54	48.8	■	54.76
modified CPC	41.66	92.02	13.57	0.0061	42.29	■	■	65.01	74.18	59.28
wav2vec	32.39	94.09	11.3	0.0307	44.88	■	■	78.91	77.52	58.17
vq-wav2vec	53.49	92.28	12.69	0.0302	39.04	9.50	9.93	59.4	70.57	55.89
wav2vec 2.0 base	28.37	92.31	6.32	■	45.62	■	7.48	58.34	79.94	56.93
HuBERT base	6.85	95.98	4.93	0.0759	64.84	7.22	6.76	95.94	86.24	62.94

worse than fbank

- Pre-trained models outperform fbank across many tasks.
- But they are not good at automatic speaker verification (ASV)?

Results of Round 1

	Content				Speaker			Semantic		Emotion
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.88	82.37			35.84		8.52	30.29		57.64
APC	41.85	91.04	15.09	0.0268	59.79	8.81		74.64	71.26	58.84
VQ-APC	42.86	90.52		0.0205	49.57	9.29		70.52		58.31
NPC	52.67	88.54	14.69	0.022	50.77		9.59	64.04		59.55
Mockingjay	80.01	82.67			34.5			28.87		45.72
TERA	47.53	88.09	12.44		58.67		9.54	48.8		54.76
modified CPC	41.66	92.02	13.57	0.0061	42.29			65.01	74.18	59.28
wav2vec	32.39	94.09	11.3	0.0307	44.88			78.91	77.52	58.17
vq-wav2vec	53.49	92.28	12.69	0.0302	39.04	9.50	9.93	59.4	70.57	55.89
wav2vec 2.0 base	28.37	92.31	6.32		45.62		7.48	58.34	79.94	56.93
HuBERT base	6.85	95.98	4.93	0.0759	64.84	7.22	6.76	95.94	86.24	62.94

- We do not show the results of wav2vec 2.0 **large** and HuBERT **large** here because they do not perform well in round 1.
- In round 1, we have not released the power of pre-trained models.

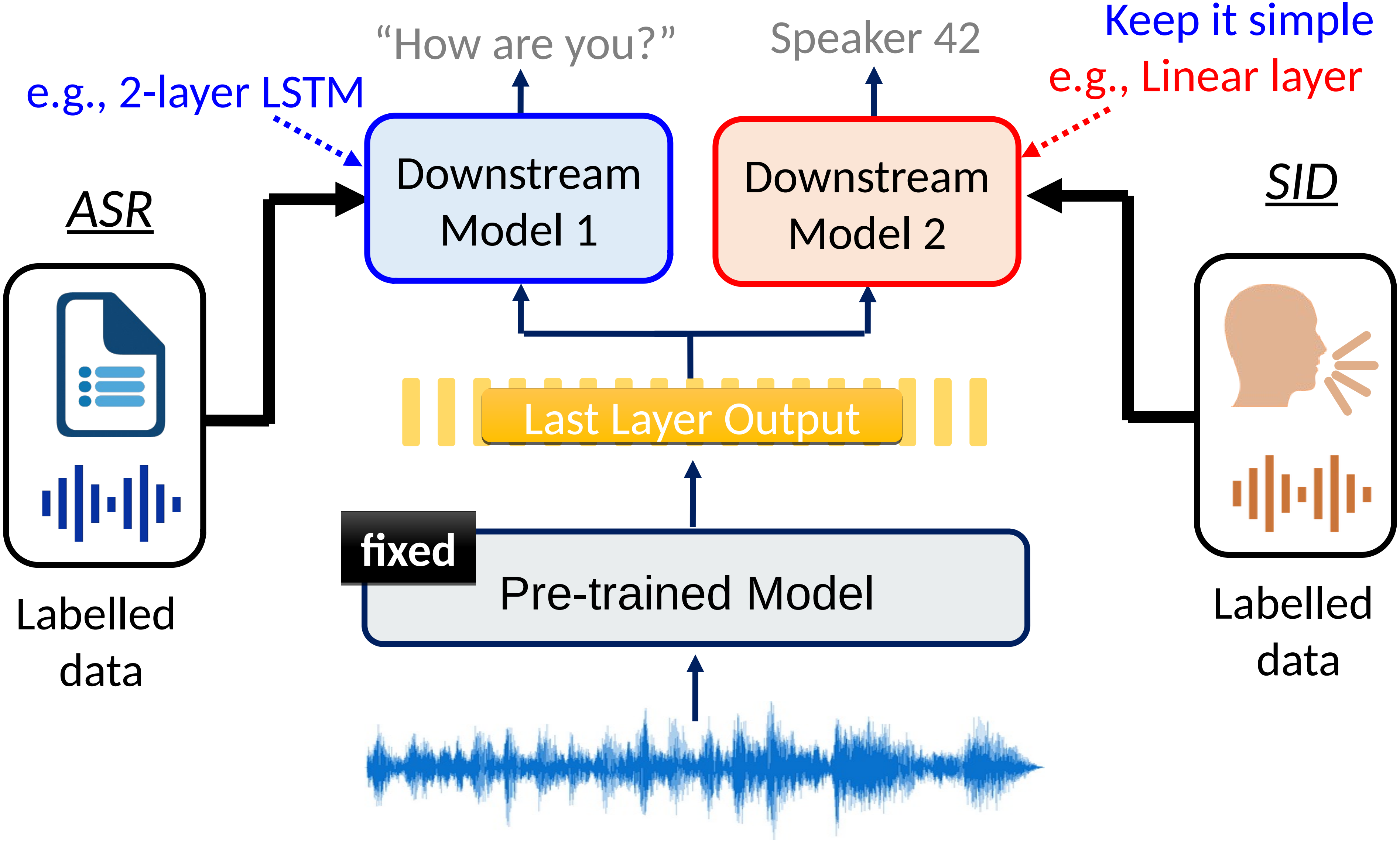
Game Start!

Round 2

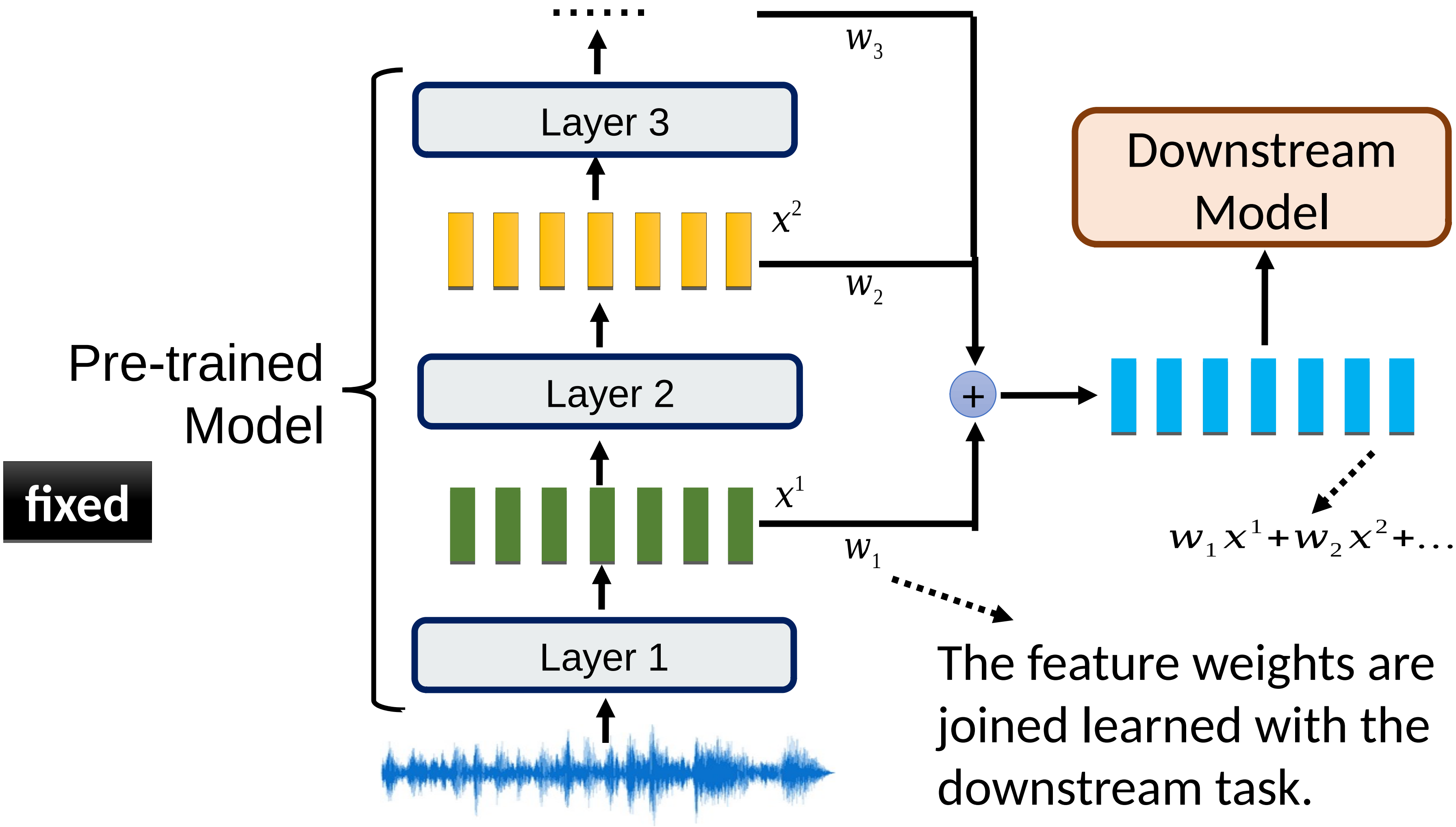
Rules in Round 2

All the upstream models use the same downstream models.

Keep it simple



Rules in Round 2



The feature weights are joined learned with the downstream task.

Results of Round 2

	Content				Speaker			Semantic		Emotion
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.87	82.54	16.62	0.0072	37.99	11.61	8.68	29.82	62.14	57.86
APC	41.98	91.01	14.74	0.0310	60.42	8.56	10.53	74.69	70.46	59.33
VQ-APC	41.08	91.11	15.21	0.0251	60.15	8.72	10.45	74.48	68.53	59.66
NPC	43.81	88.96	13.91	0.0246	55.92	9.40	9.34	69.44	72.79	59.08
Mockingjay	70.19	83.67	15.48	6.60E-04	32.29	11.66	10.54	34.33	61.59	50.28
TERA	49.17	89.48	12.16	0.0013	57.57	15.89	9.96	58.42	67.50	56.27
DeCoAR 2.0	14.93	94.48	9.07	0.0406	74.42	7.16	6.59	90.80	83.28	62.47
modified CPC	42.54	91.88	13.53	0.0326	39.63	12.86	10.38	64.09	71.19	60.96
wav2vec	31.58	95.59	11.00	0.0485	56.56	7.99	9.90	84.92	76.37	59.79
vq-wav2vec	33.48	93.38	12.80	0.0410	38.80	10.38	9.93	85.68	77.68	58.24
wav2vec 2.0 base	5.74	96.23	4.79	0.0233	75.18	6.02	6.08	92.35	88.30	63.43
wav2vec 2.0 large	4.75	96.66	3.10	0.0489	86.14	5.65	5.62	95.28	87.11	65.64
HuBERT base	5.41	96.30	4.79	0.0736	81.42	5.11	5.88	98.34	88.53	64.92
HuBERT large	3.53	95.29	2.94	0.0353	90.33	5.98	5.75	98.76	89.81	67.62

Results of Round 2

	Content				Speaker			Semantic		Emotion
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.87	82.54		0.0072	37.99		8.68	29.82		57.86
APC	41.98	91.01	14.74	0.0310	60.42	8.56		74.69	70.46	59.33
VQ-APC	41.08	91.11		0.0251	60.15	8.72		74.48		59.66
NPC	43.81	88.96	13.91	0.0246	55.92	9.40	9.34	69.44	72.79	59.08
Mockingjay	70.19	83.67			32.29			34.33		50.28
TERA	49.17	89.48	12.16		57.57		9.96	58.42		56.27
DeCoAR 2.0	14.93	94.48	9.07	0.0406	74.42	7.16	6.59	90.80	83.28	62.47
modified CPC	42.54	91.88	13.53	0.0326	39.63			64.09	71.19	60.96
wav2vec	31.58	95.59	11.00	0.0485	56.56	7.99	9.90	84.92	76.37	59.79
vq-wav2vec	33.48	93.38	12.80	0.0410	38.80		9.93	85.68	77.68	58.24
wav2vec 2.0 base	5.74	96.23	4.79	0.0233	75.18	6.02	6.08	92.35	88.30	63.43
wav2vec 2.0 large	4.75	96.66	3.10	0.0489	86.14	5.65	5.62	95.28	87.11	65.64
HuBERT base	5.41	96.30	4.79	0.0736	81.42	5.11	5.88	98.34	88.53	64.92
HuBERT large	3.53	95.29	2.94	0.0353	90.33	5.98	5.75	98.76	89.81	67.62

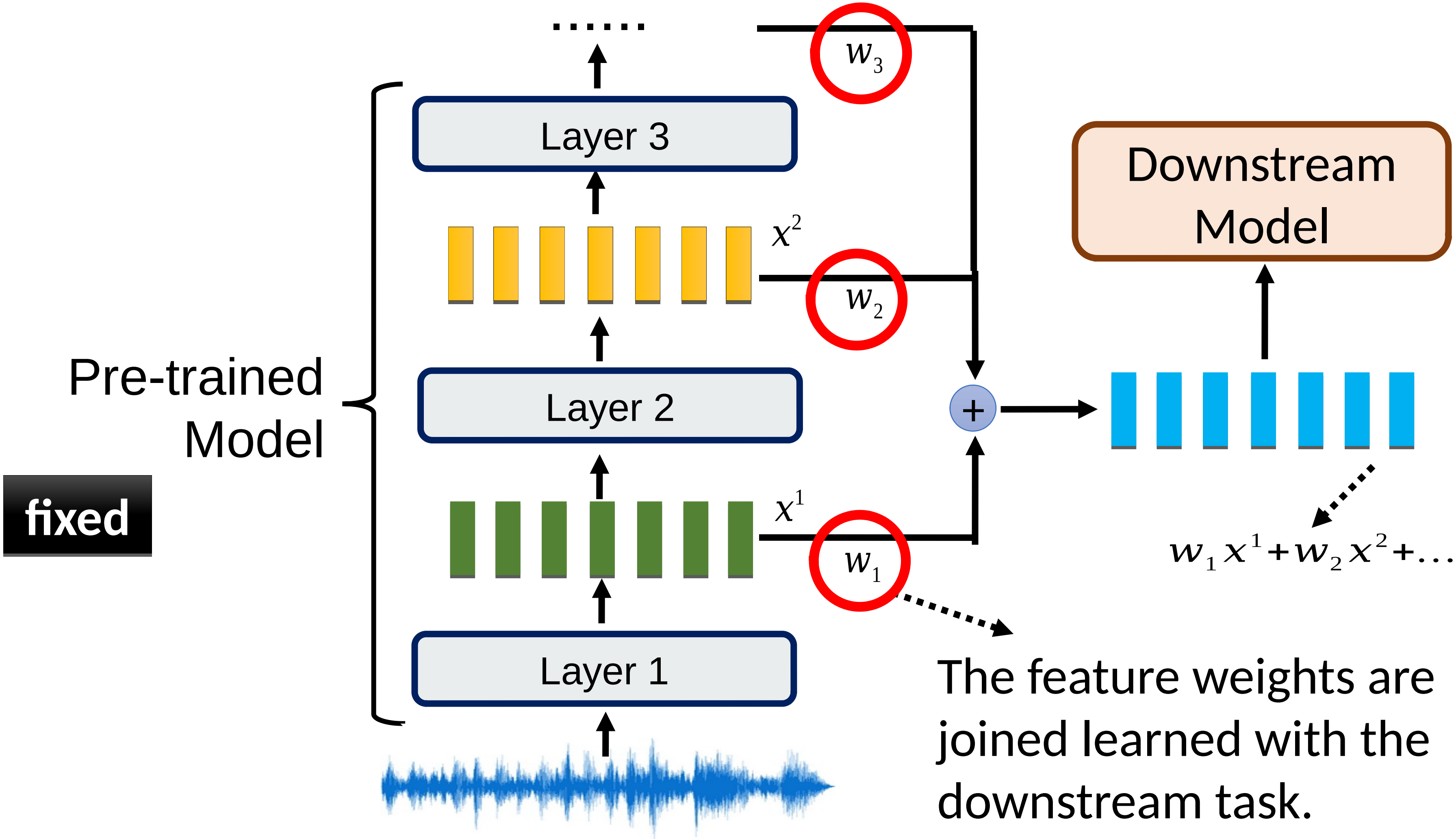
- Pre-trained learning outperforms fbank in most cases.

Results of Round 2

	Content				Speaker			Semantic		Emotion
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.87	82.54		0.0072	37.99		8.68	29.82		57.86
APC	41.98	91.01	14.74	0.0310	60.42	8.56		74.69	70.46	59.33
VQ-APC	41.08	91.11		0.0251	60.15	8.72		74.48		59.66
NPC	43.81	88.96	13.91	0.0246	55.92	9.40	9.34	69.44	72.79	59.08
Mockingjay	70.19	83.67			32.29			34.33		50.28
TERA	49.17	89.48	12.16		57.57		9.96	58.42		56.27
DeCoAR 2.0	14.93	94.48	9.07	0.0406	74.42	7.16	6.59	90.80	83.28	62.47
modified CPC	42.54	91.88	13.53	0.0326	39.63			64.09	71.19	60.96
wav2vec	31.58	95.59	11.00	0.0485	56.56	7.99	9.90	84.92	76.37	59.79
vq-wav2vec	33.48	93.38	12.80	0.0410	38.80		9.93	85.68	77.68	58.24
wav2vec 2.0 base	5.74	96.23	4.79	0.0233	75.18	6.02	6.08	92.35	88.30	63.43
wav2vec 2.0 large	4.75	96.66	3.10	0.0489	86.14	5.65	5.62	95.28	87.11	65.64
HuBERT base	5.41	96.30	4.79	0.0736	81.42	5.11	5.88	98.34	88.53	64.92
HuBERT large	3.53	95.29	2.94	0.0353	90.33	5.98	5.75	98.76	89.81	67.62

- Several pre-trained models are all-around.

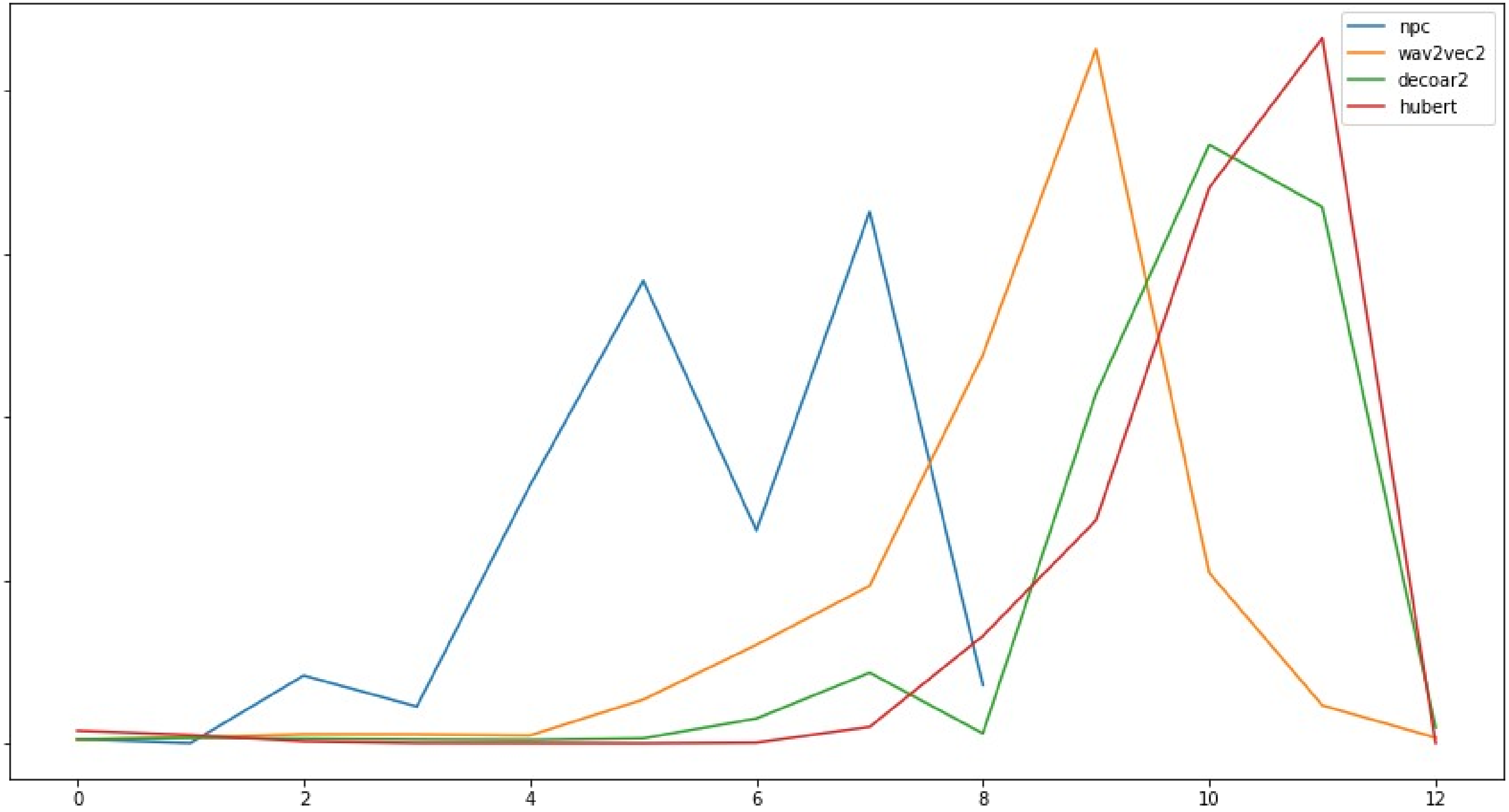
Analysis of the Weights



The feature weights are joined learned with the downstream task.

Layer Weights – Phoneme Recognition

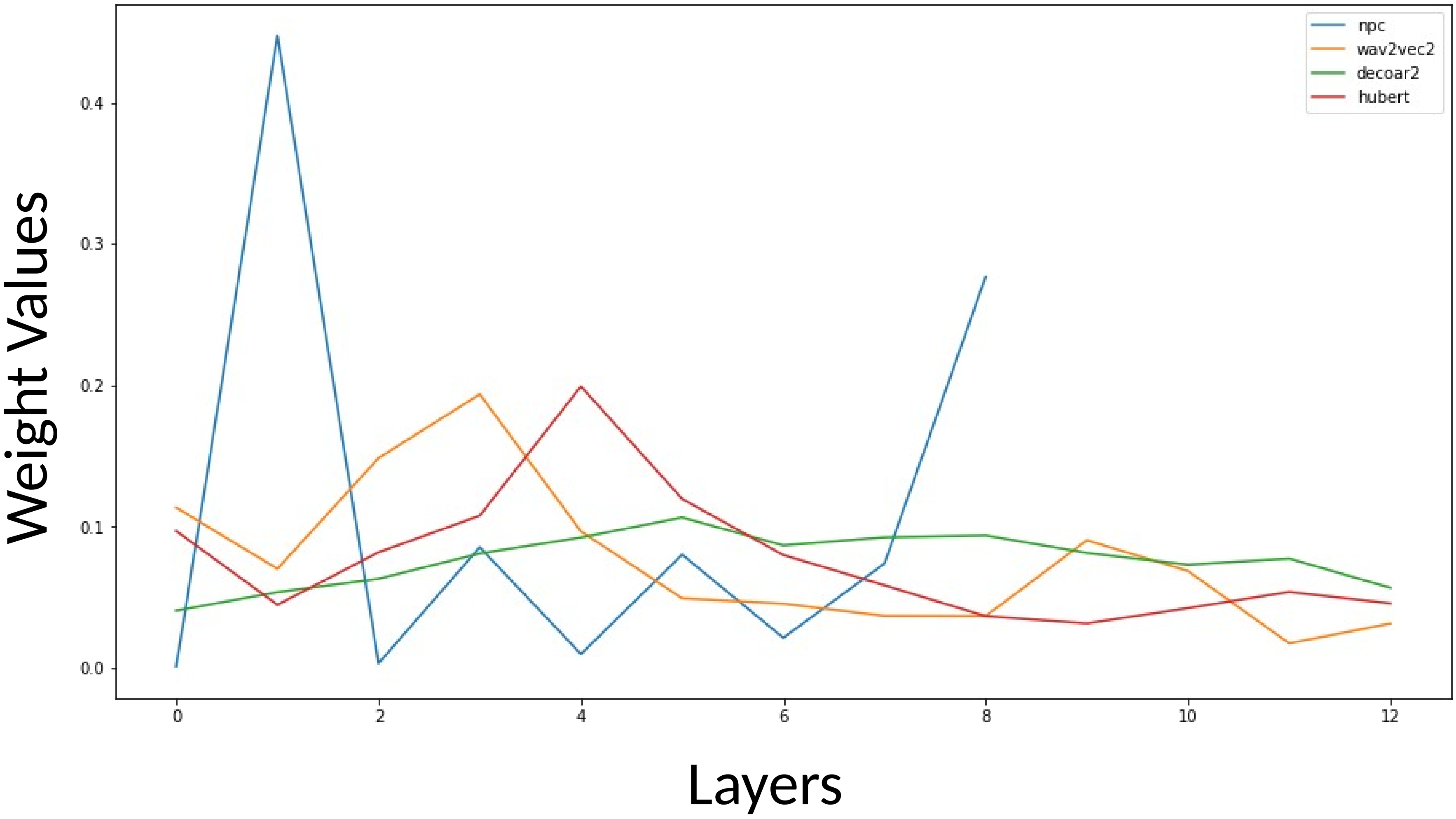
(The weights are normalized by the representation's norms.)



Weight Values

Layers

Layer Weights
– Speaker Verification



Specialist? Universal?

Just name a few …

PASE+ APC NPC Mockingjay DeCoAR Wav2vec HuBERT

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

They are universal!

…. but how can task-agnostic self-supervised learning achieve that?

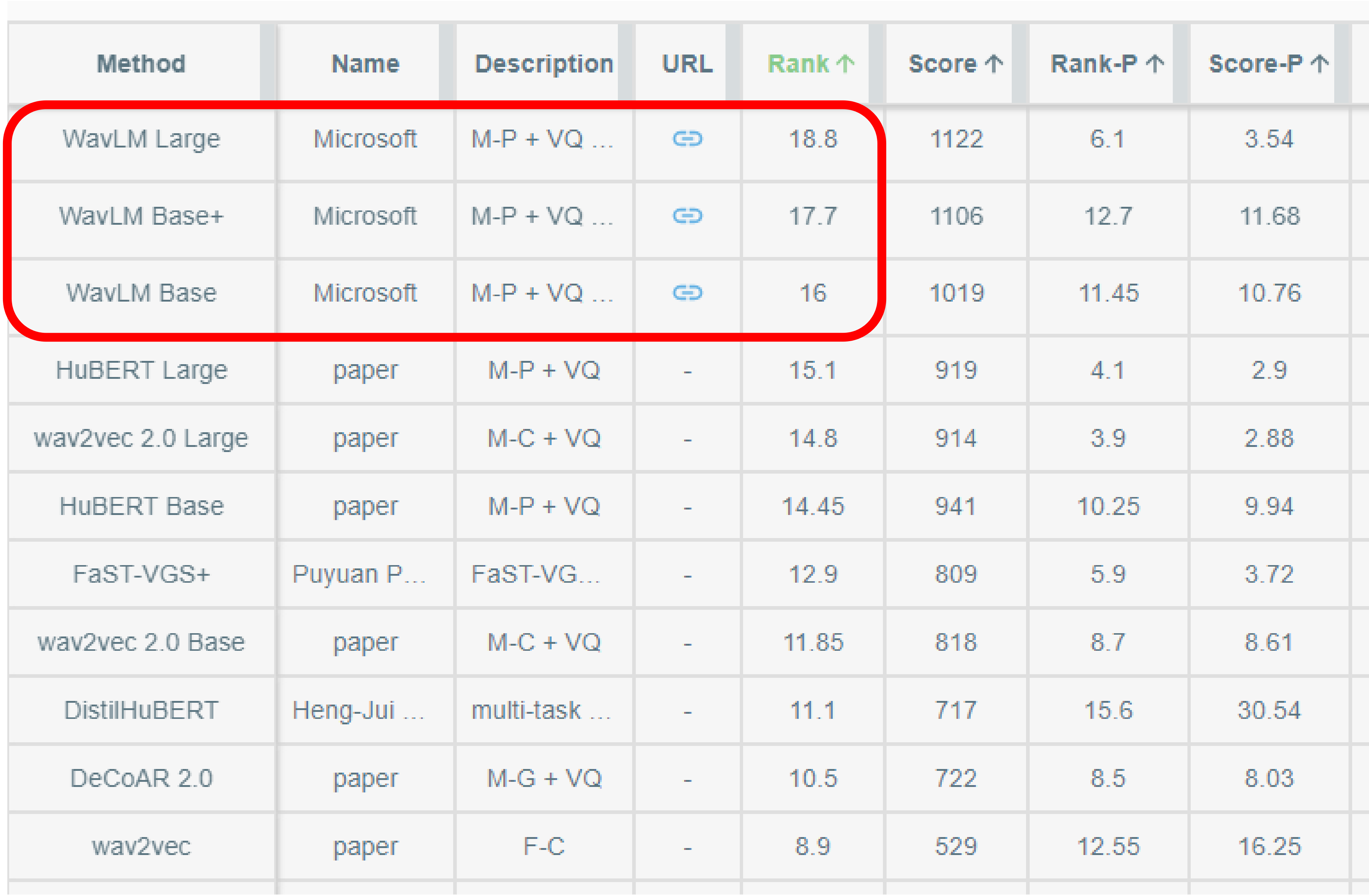
(I don't have the answer now.)

My two cents

(Now)

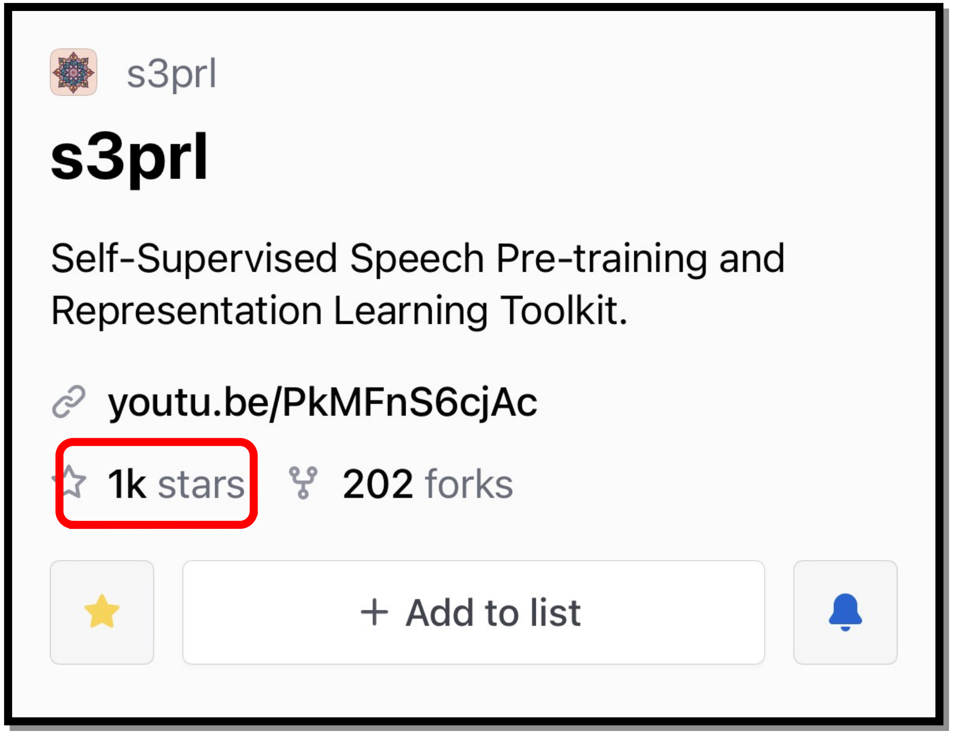
Welcome to Join the Game ⏪

https://superbbenchmark.org/



Method	Name	Description	URL	Rank ↑	Score ↑	Rank-P ↑	Score-P ↑
WavLM Large	Microsoft	M-P + VQ ...	🔗	18.8	1122	6.1	3.54
WavLM Base+	Microsoft	M-P + VQ ...	🔗	17.7	1106	12.7	11.68
WavLM Base	Microsoft	M-P + VQ ...	🔗	16	1019	11.45	10.76
HuBERT Large	paper	M-P + VQ	-	15.1	919	4.1	2.9
wav2vec 2.0 Large	paper	M-C + VQ	-	14.8	914	3.9	2.88
HuBERT Base	paper	M-P + VQ	-	14.45	941	10.25	9.94
FaST-VGS+	Puyuan P...	FaST-VG...	-	12.9	809	5.9	3.72
wav2vec 2.0 Base	paper	M-C + VQ	-	11.85	818	8.7	8.61
DistilHuBERT	Heng-Jui ...	multi-task ...	-	11.1	717	15.6	30.54
DeCoAR 2.0	paper	M-G + VQ	-	10.5	722	8.5	8.03
wav2vec	paper	F-C	-	8.9	529	12.55	16.25

Toolkit – S3PRL



s3prl

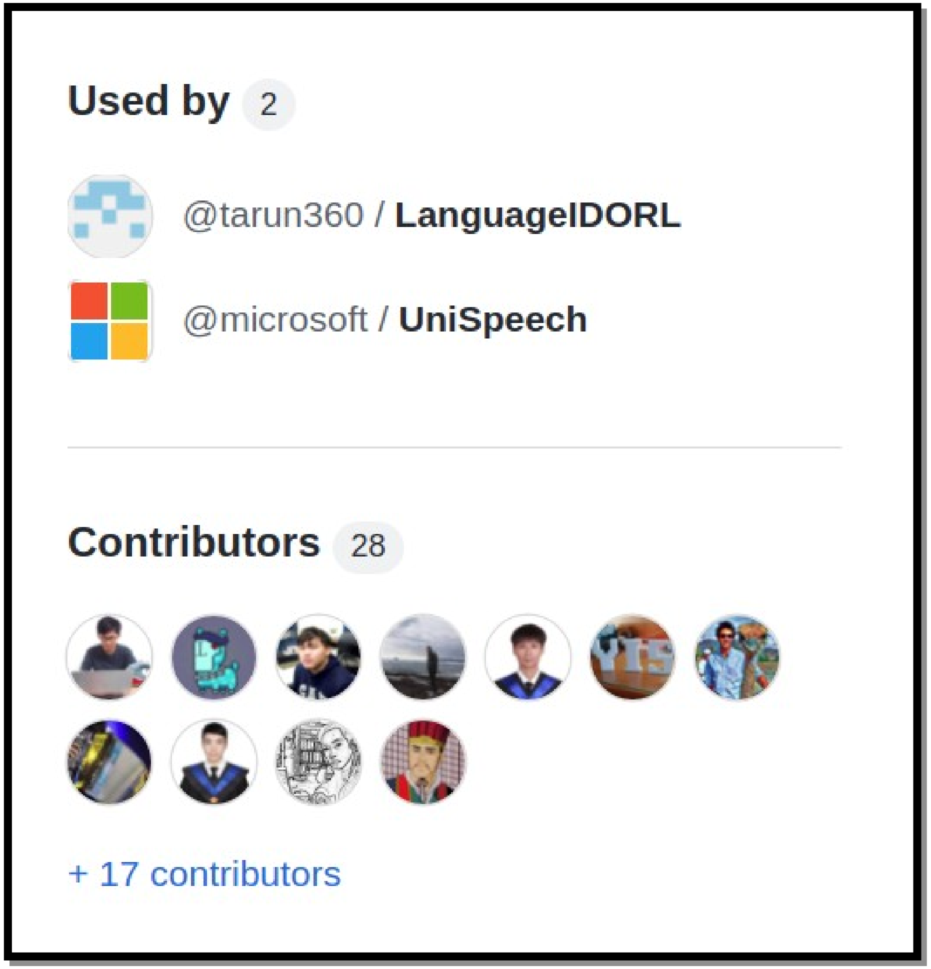
s3prl

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

youtu.be/PkMFnS6cjAc

1k stars 202 forks

https://github.com/s3prl/s3prl



Used by 2

@tarun360 / **LanguageIDORL**

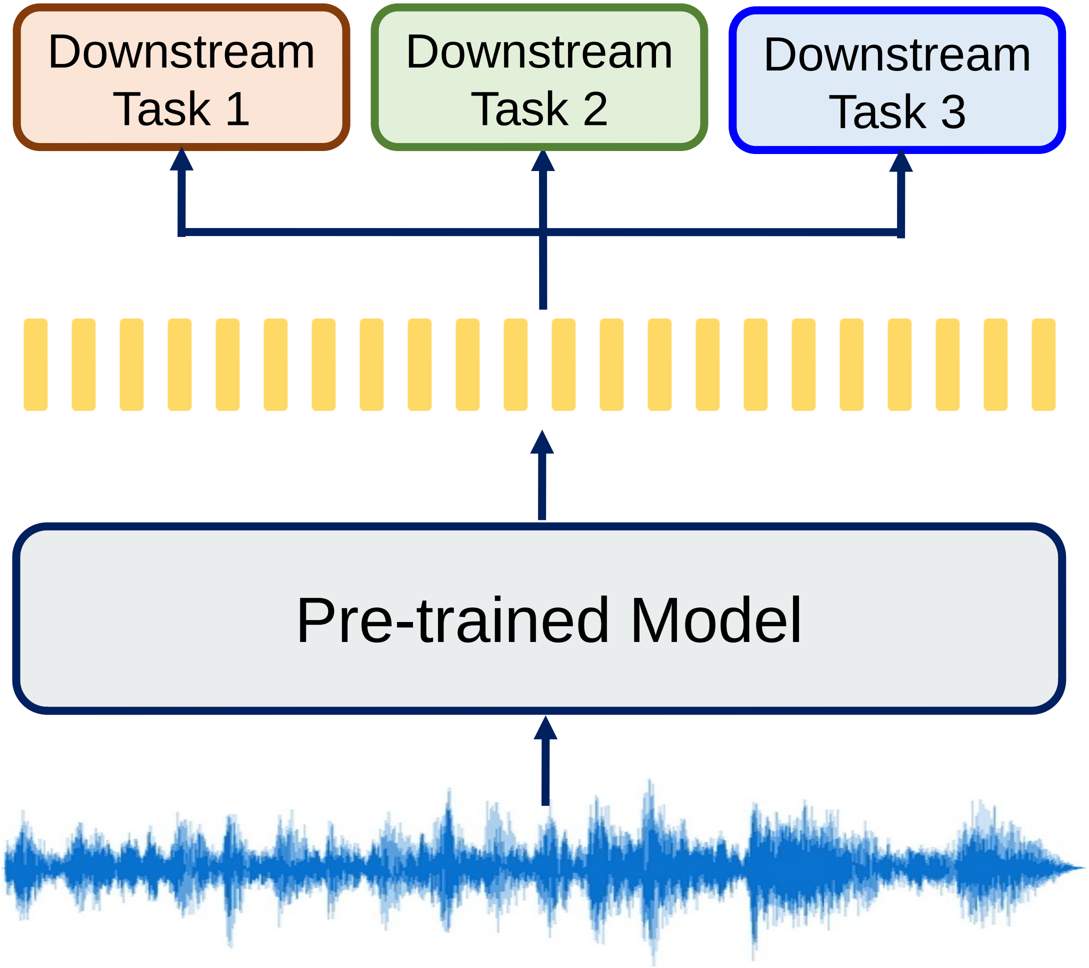
@microsoft / **UniSpeech**

Contributors 28

+ 17 contributors

Applications

Operating Systems



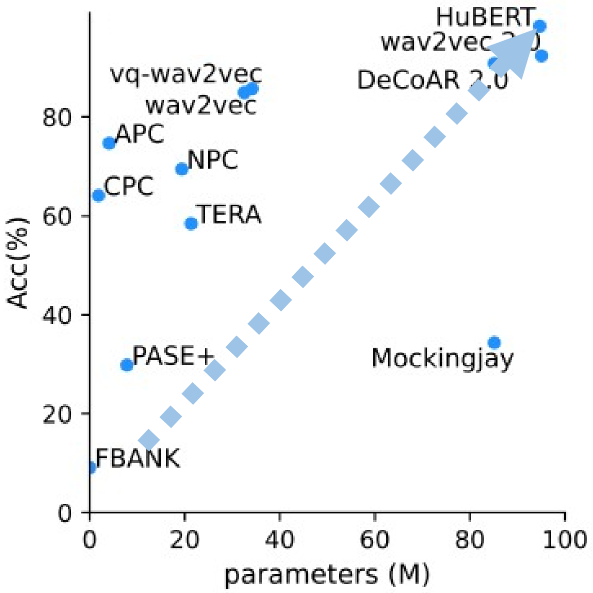
Let’s welcome the era of Pre-training.

Research in Progress based on Self-supervised Learning

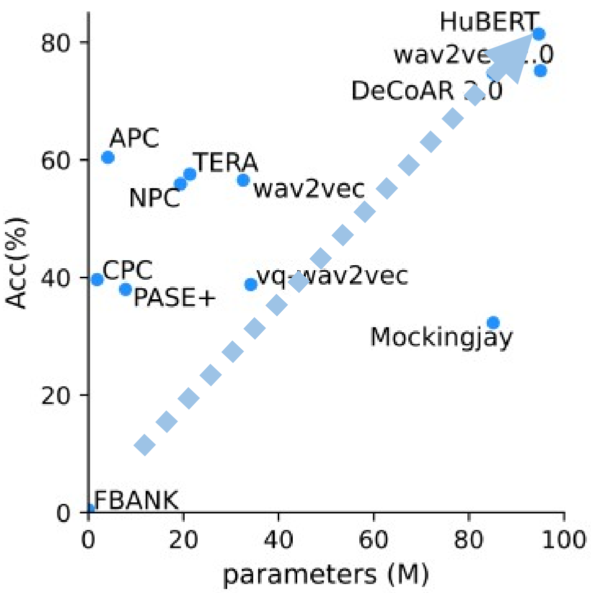
More ……

- 1. Make Pre-trained Model Smaller
- 2. Attacking Pre-trained Model
- 3. Privacy Issue of Pre-trained Model
- 4. Data Bias vs. Pre-training
- 5. Unsupervised Speech Recognition
- 6. Spoken Question Answering

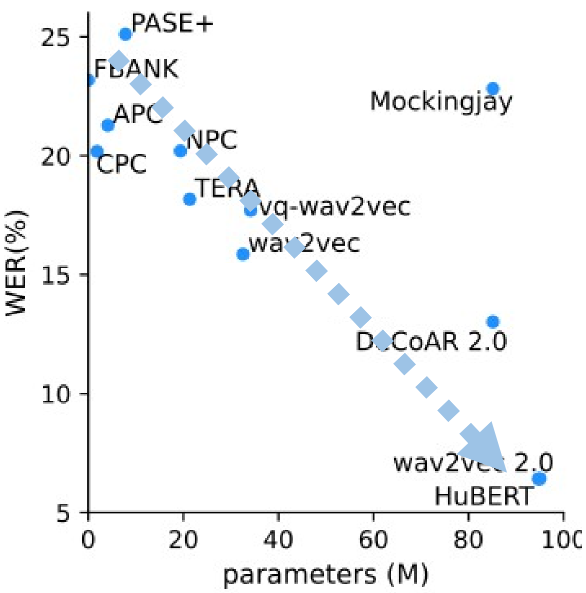
1. Make Pre-trained Model Smaller



Intent Classification



Speaker Identification

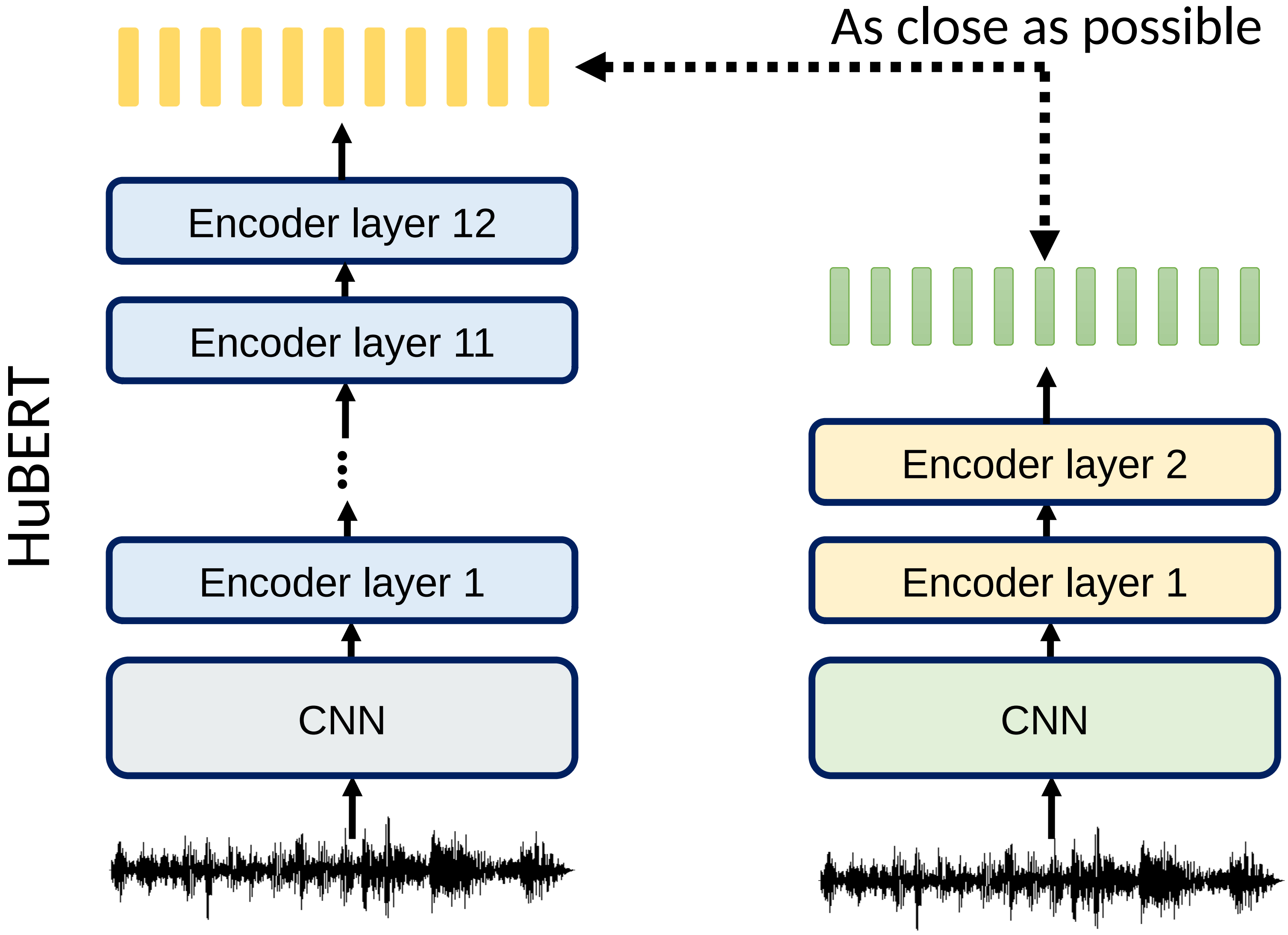


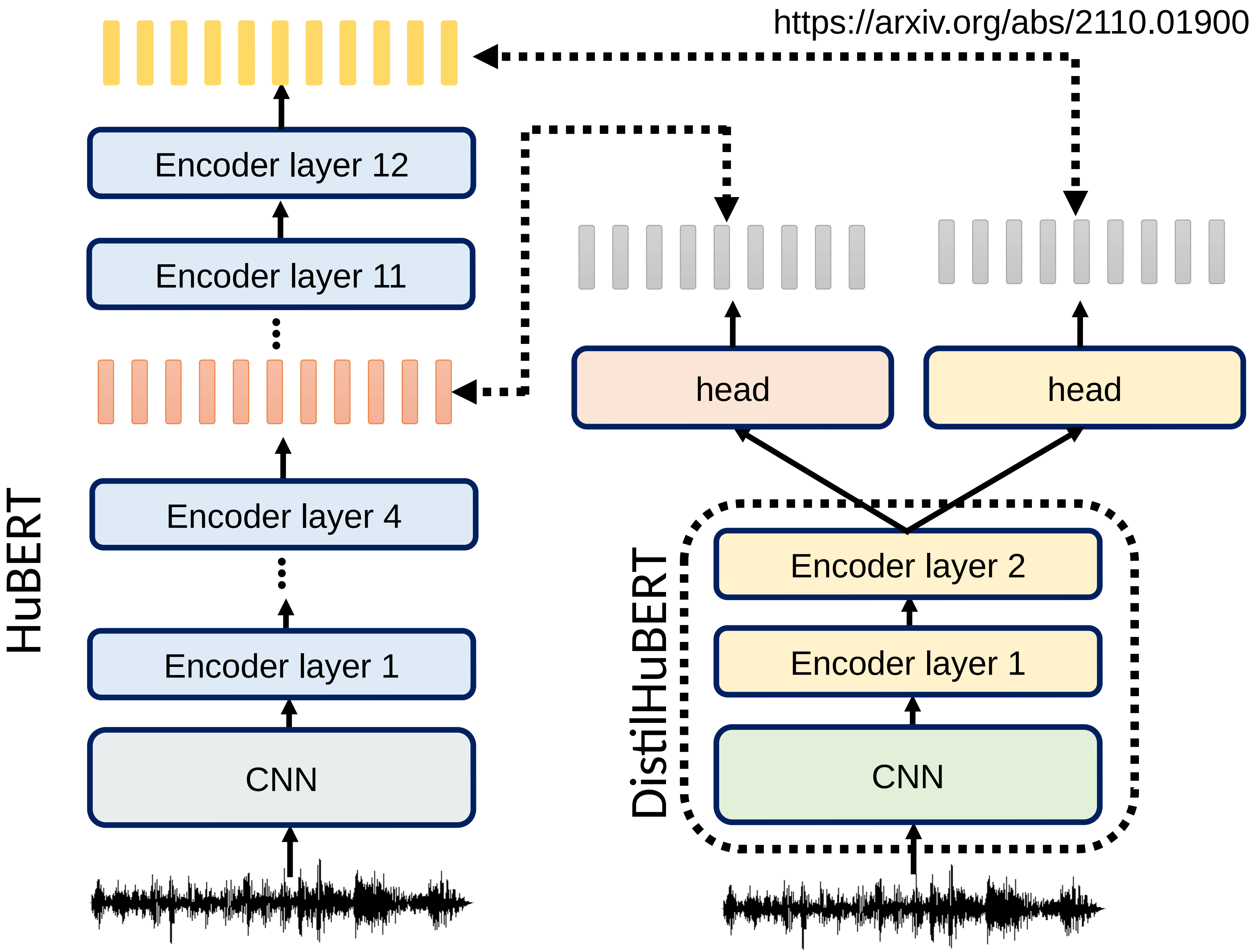
ASR

Larger models usually lead to **better** performance.

Typical Knowledge Distillation

Each layer contains different information. Learning from the last layer is not sufficient.

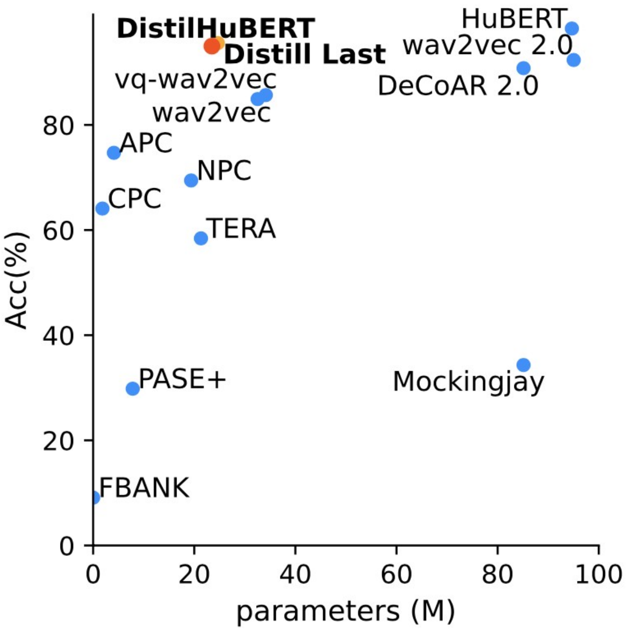




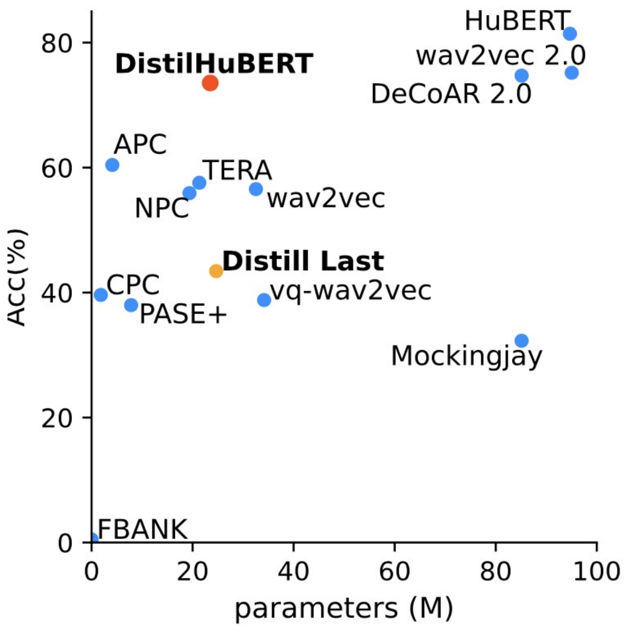
https://arxiv.org/abs/2110.01900

https://arxiv.org/abs/2110.01900

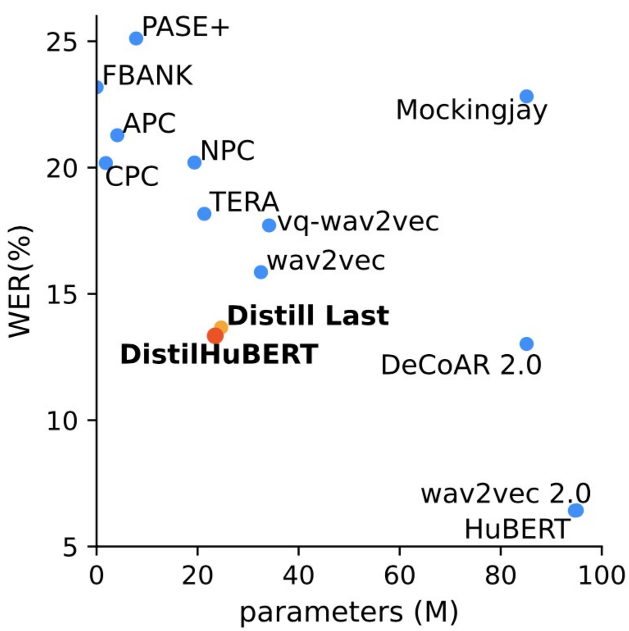
1. Make Upstream Model Smaller



Intent Classification



Speaker Identification



ASR

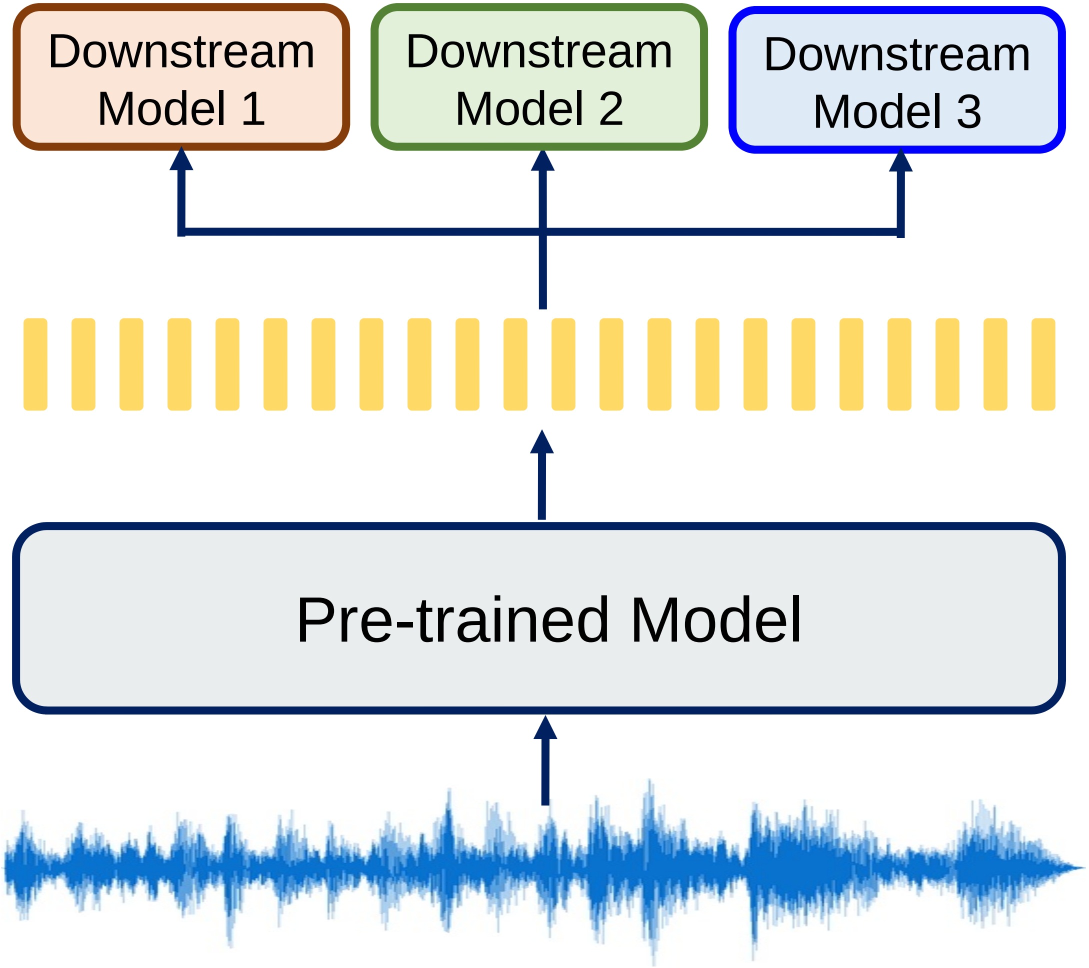
DistilHuBERT is better than the models with the same size.

https://arxiv.org/abs/2111.04330

2. Adversarial Attack

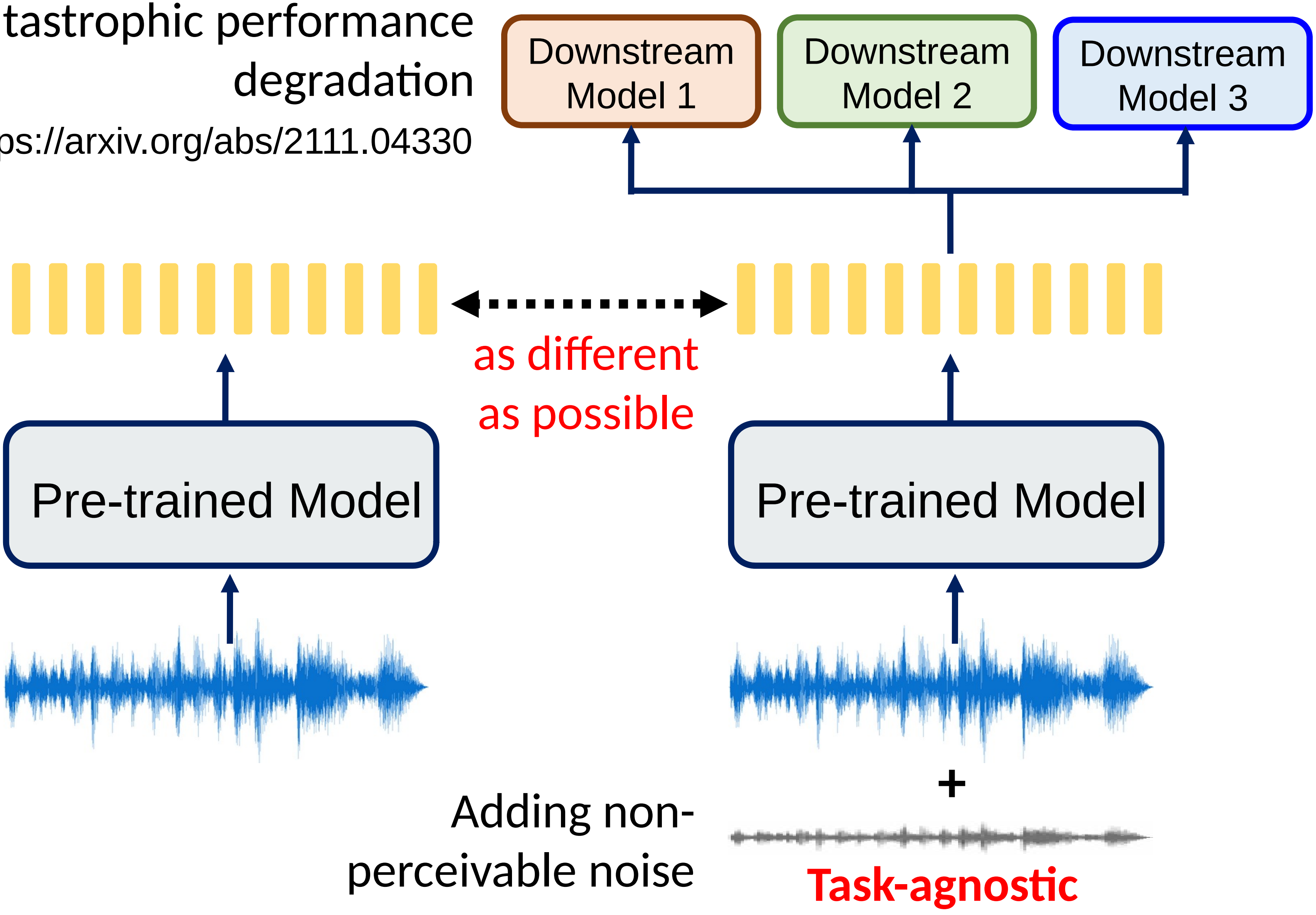
For all tasks

single point of failure?



Catastrophic performance degradation

https://arxiv.org/abs/2111.04330



Downstream Model 1

Downstream Model 2

Downstream Model 3

as different as possible

Pre-trained Model

Pre-trained Model

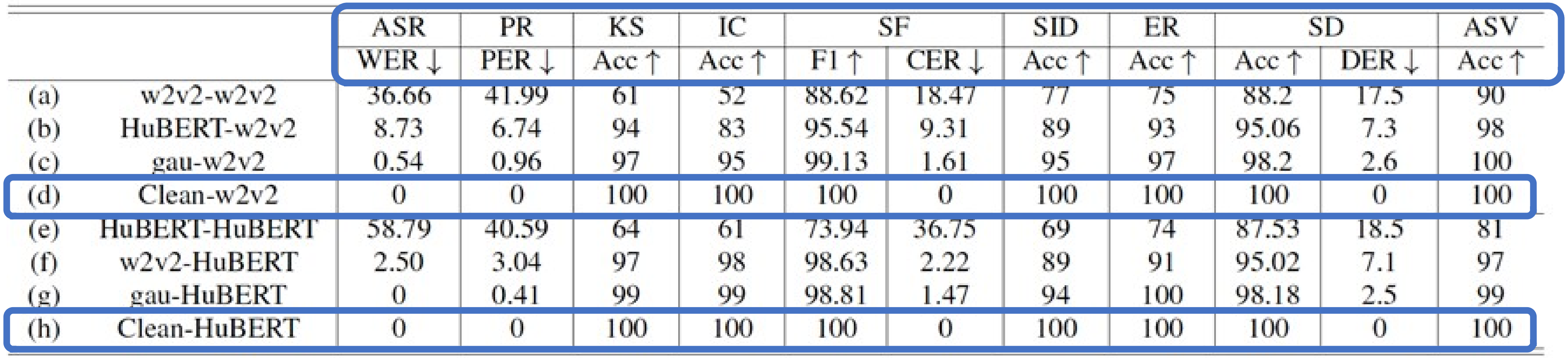
+

Adding non-perceivable noise

Task-agnostic

2. Adversarial Attack

The directions of the arrows denote the directions towards the better performance of the task.



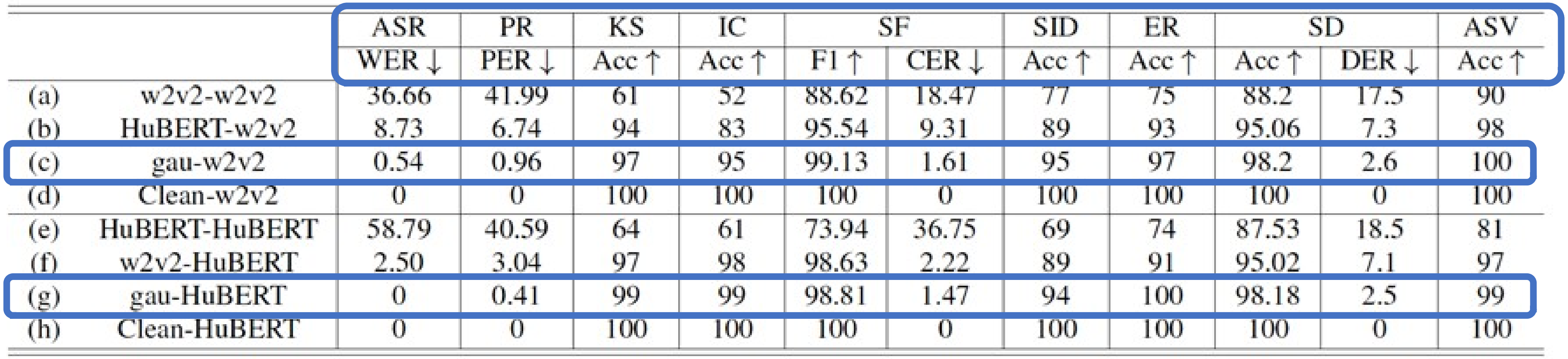
		ASR	PR	KS	IC	SF		SID	ER	SD		ASV
		WER ↓	PER ↓	Acc ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Acc ↑	Acc ↑	DER ↓	Acc ↑
(a)	w2v2-w2v2	36.66	41.99	61	52	88.62	18.47	77	75	88.2	17.5	90
(b)	HuBERT-w2v2	8.73	6.74	94	83	95.54	9.31	89	93	95.06	7.3	98
(c)	gau-w2v2	0.54	0.96	97	95	99.13	1.61	95	97	98.2	2.6	100
(d)	Clean-w2v2	0	0	100	100	100	0	100	100	100	0	100
(e)	HuBERT-HuBERT	58.79	40.59	64	61	73.94	36.75	69	74	87.53	18.5	81
(f)	w2v2-HuBERT	2.50	3.04	97	98	98.63	2.22	89	91	95.02	7.1	97
(g)	gau-HuBERT	0	0.41	99	99	98.81	1.47	94	100	98.18	2.5	99
(h)	Clean-HuBERT	0	0	100	100	100	0	100	100	100	0	100

w2v2 and HuBERT are self-supervised models.

Without attack: Only select the samples with the correct predictions (e.g., 0% WER for ASR, 0% PER for PR, etc.)

2. Adversarial Attack

The directions of the arrows denote the directions towards the better performance of the task.



		ASR	PR	KS	IC	SF		SID	ER	SD		ASV
		WER ↓	PER ↓	Acc ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Acc ↑	Acc ↑	DER ↓	Acc ↑
(a)	w2v2-w2v2	36.66	41.99	61	52	88.62	18.47	77	75	88.2	17.5	90
(b)	HuBERT-w2v2	8.73	6.74	94	83	95.54	9.31	89	93	95.06	7.3	98
(c)	gau-w2v2	0.54	0.96	97	95	99.13	1.61	95	97	98.2	2.6	100
(d)	Clean-w2v2	0	0	100	100	100	0	100	100	100	0	100
(e)	HuBERT-HuBERT	58.79	40.59	64	61	73.94	36.75	69	74	87.53	18.5	81
(f)	w2v2-HuBERT	2.50	3.04	97	98	98.63	2.22	89	91	95.02	7.1	97
(g)	gau-HuBERT	0	0.41	99	99	98.81	1.47	94	100	98.18	2.5	99
(h)	Clean-HuBERT	0	0	100	100	100	0	100	100	100	0	100

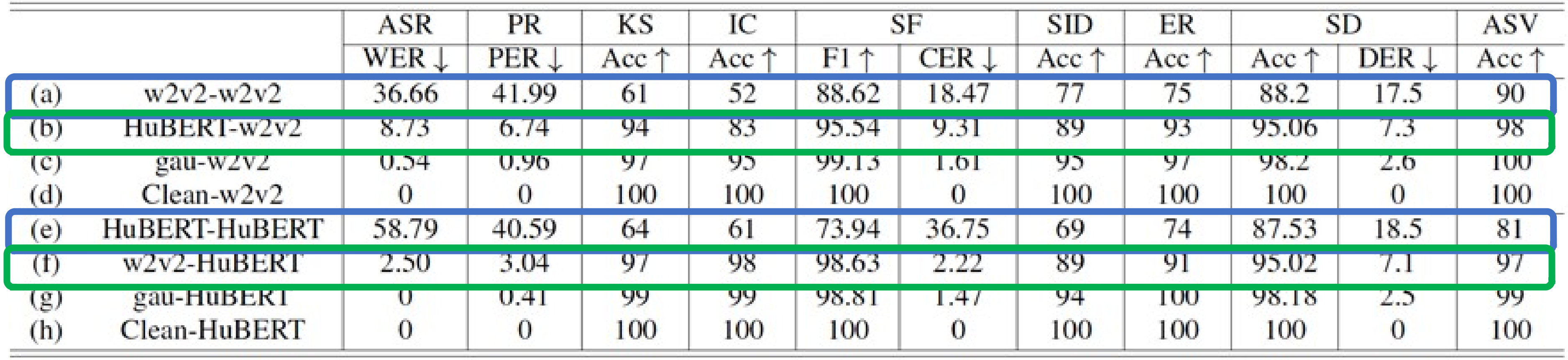
w2v2 and HuBERT are self-supervised models.

Without attack: Only select the samples with the correct predictions (e.g., 0% WER for ASR, 0% PER for PR, etc.)

Adding Gaussian noises: Only a small impact on performance

2. Adversarial Attack

The directions of the arrows denote the directions towards the better performance of the task.

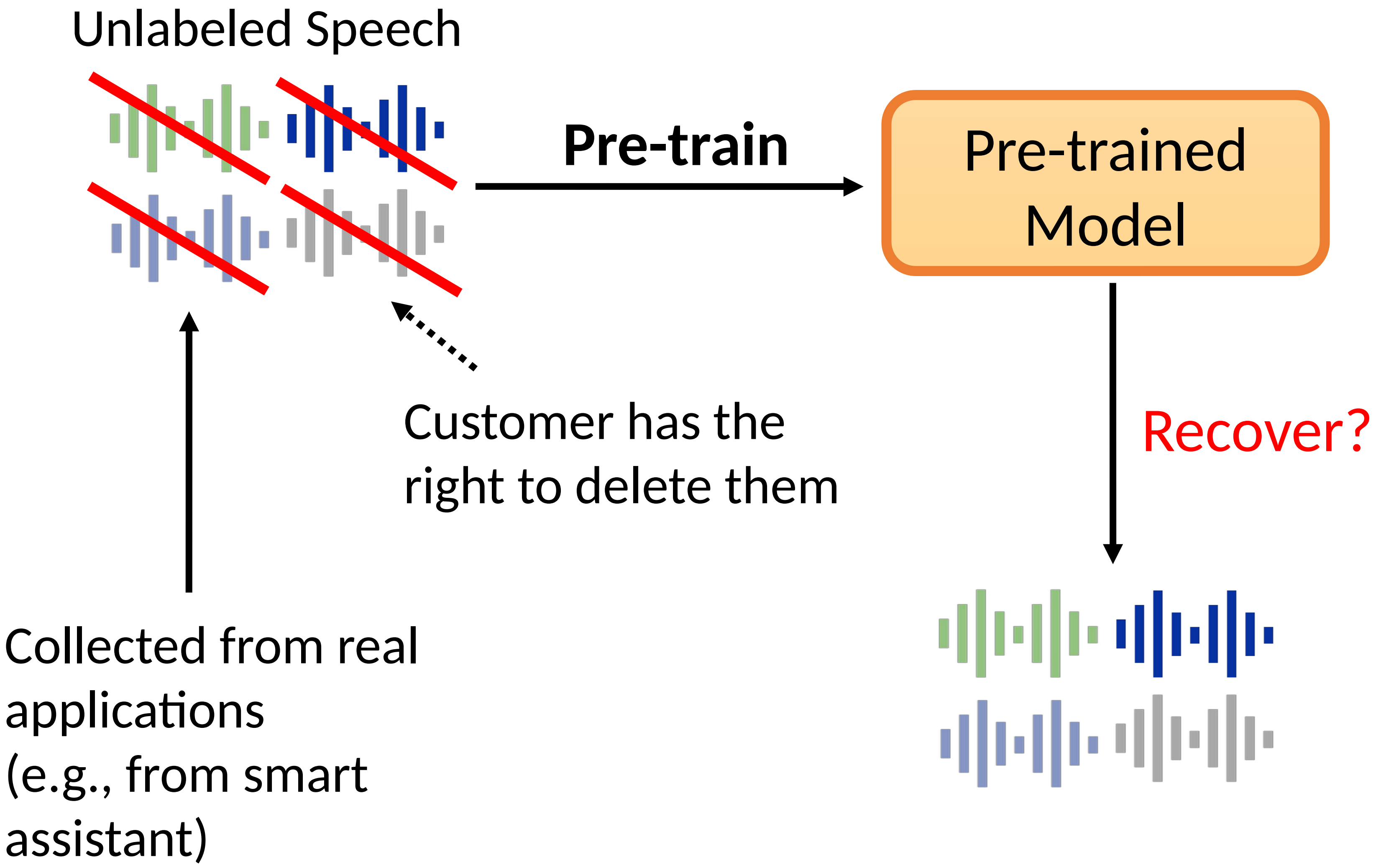


		ASR	PR	KS	IC	SF		SID	ER	SD		ASV
		WER ↓	PER ↓	Acc ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Acc ↑	Acc ↑	DER ↓	Acc ↑
(a)	w2v2-w2v2	36.66	41.99	61	52	88.62	18.47	77	75	88.2	17.5	90
(b)	HuBERT-w2v2	8.73	6.74	94	83	95.54	9.31	89	93	95.06	7.3	98
(c)	gau-w2v2	0.54	0.96	97	95	99.13	1.61	95	97	98.2	2.6	100
(d)	Clean-w2v2	0	0	100	100	100	0	100	100	100	0	100
(e)	HuBERT-HuBERT	58.79	40.59	64	61	73.94	36.75	69	74	87.53	18.5	81
(f)	w2v2-HuBERT	2.50	3.04	97	98	98.63	2.22	89	91	95.02	7.1	97
(g)	gau-HuBERT	0	0.41	99	99	98.81	1.47	94	100	98.18	2.5	99
(h)	Clean-HuBERT	0	0	100	100	100	0	100	100	100	0	100

White-box attack: the attack is very effective.

Black-box attack: not as effective as while-box attack

3. Privacy Issue



Unlabeled Speech

Pre-train

Pre-trained Model

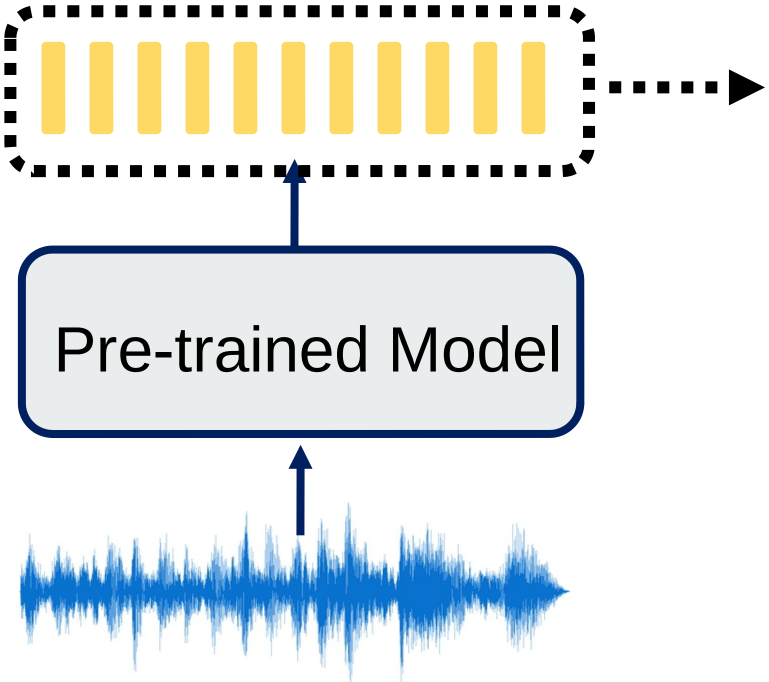
Customer has the right to delete them

Recover?

Collected from real applications (e.g., from smart assistant)

3. Privacy Issue

- Membership Inference Attack



Pre-trained Model

Given an utterance

The input utterance is in training set or not?

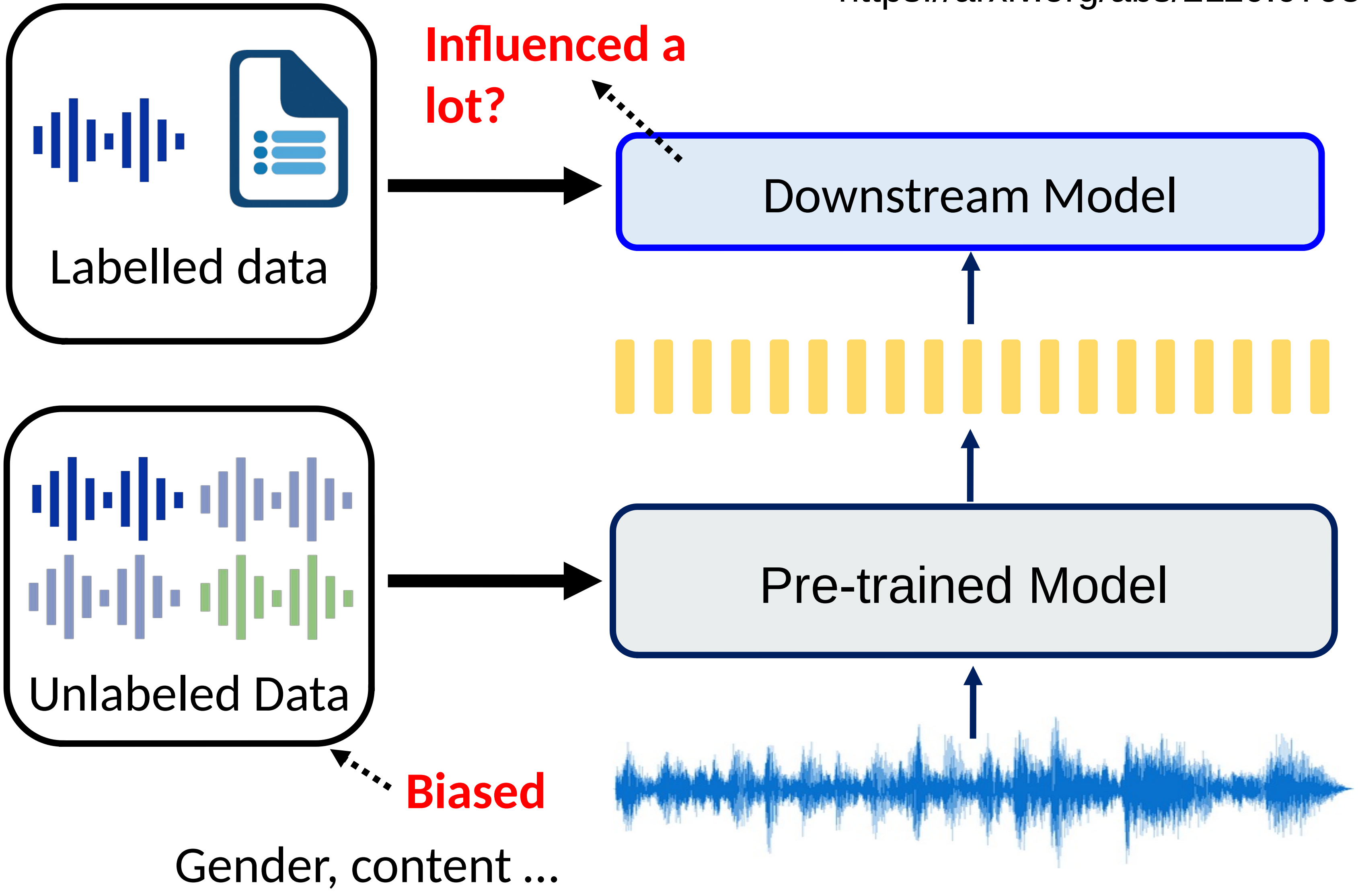
With a native approach, we can get very high accuracy.

https://arxiv.org/abs/2111.05113

4. Would Biased Unlabeled Data become an Issue?

Don't speak too fast: The impact of data bias on self-supervised speech models

https://arxiv.org/abs/2110.07957



Labelled data

Influenced a lot?

Downstream Model

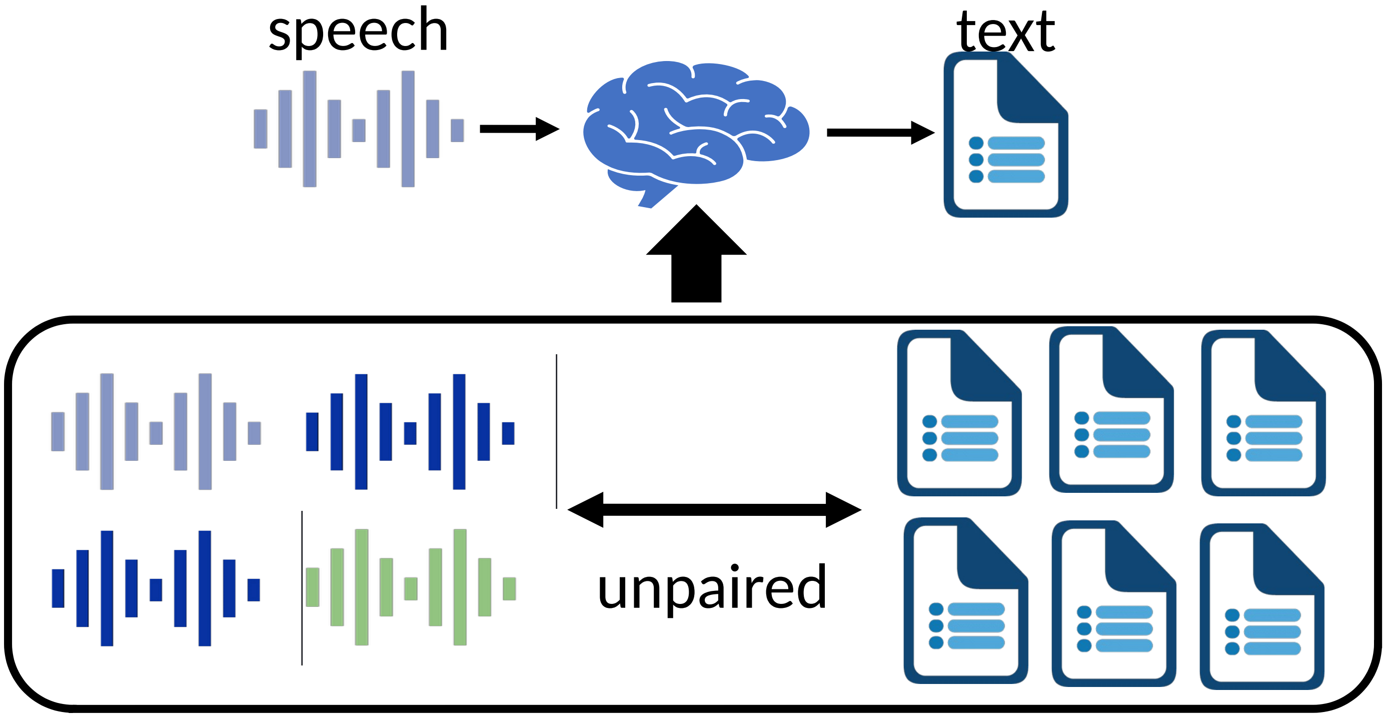
Pre-trained Model

Unlabeled Data

Biased

Gender, content …

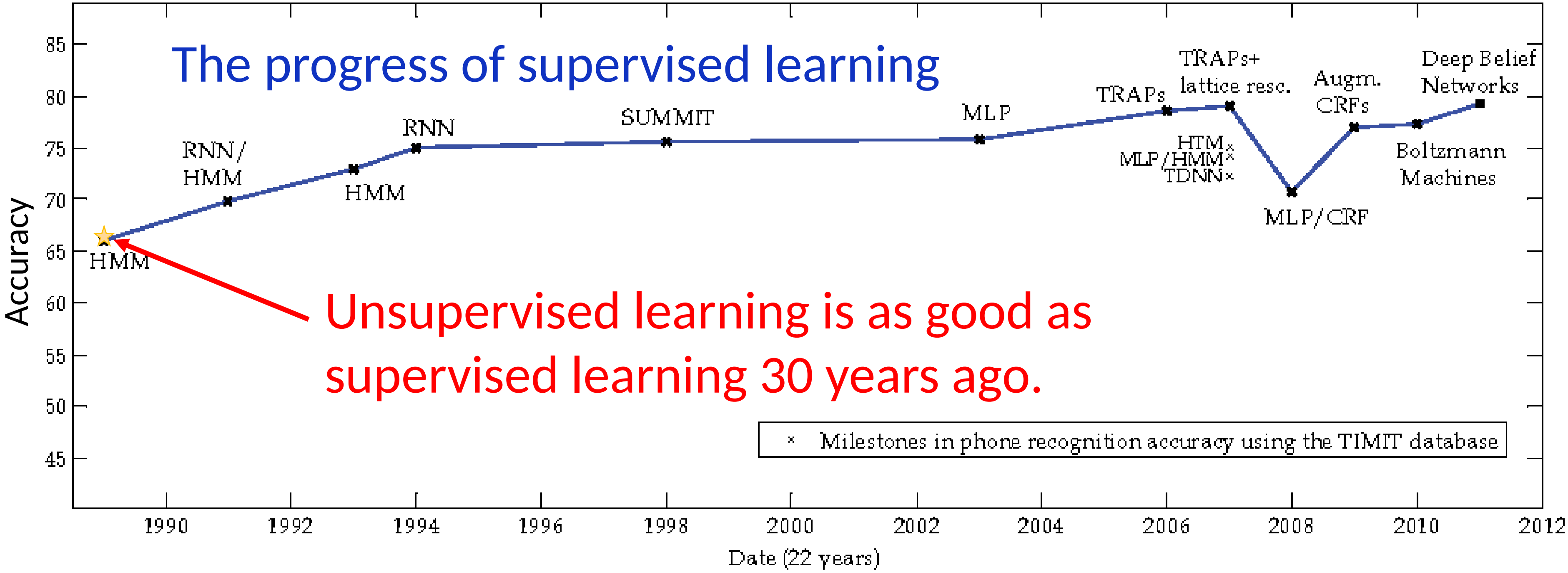
5. Unsupervised Speech Recognition



This can be achieved by Generative Adversarial Network (GAN).

How is the results?

- Unsupervised setting on TIMIT (text and audio are unpair, text is not the transcription of audio)
  - 63.6% PER (oracle boundaries) [Liu, et al., INTERSPEECH 2018]
  - 41.6% PER (automatic segmentation) [Yeh, et al., ICLR 2019]
  - 33.1% PER (automatic segmentation)[Chen, et al., INTERSPEECH 2019]

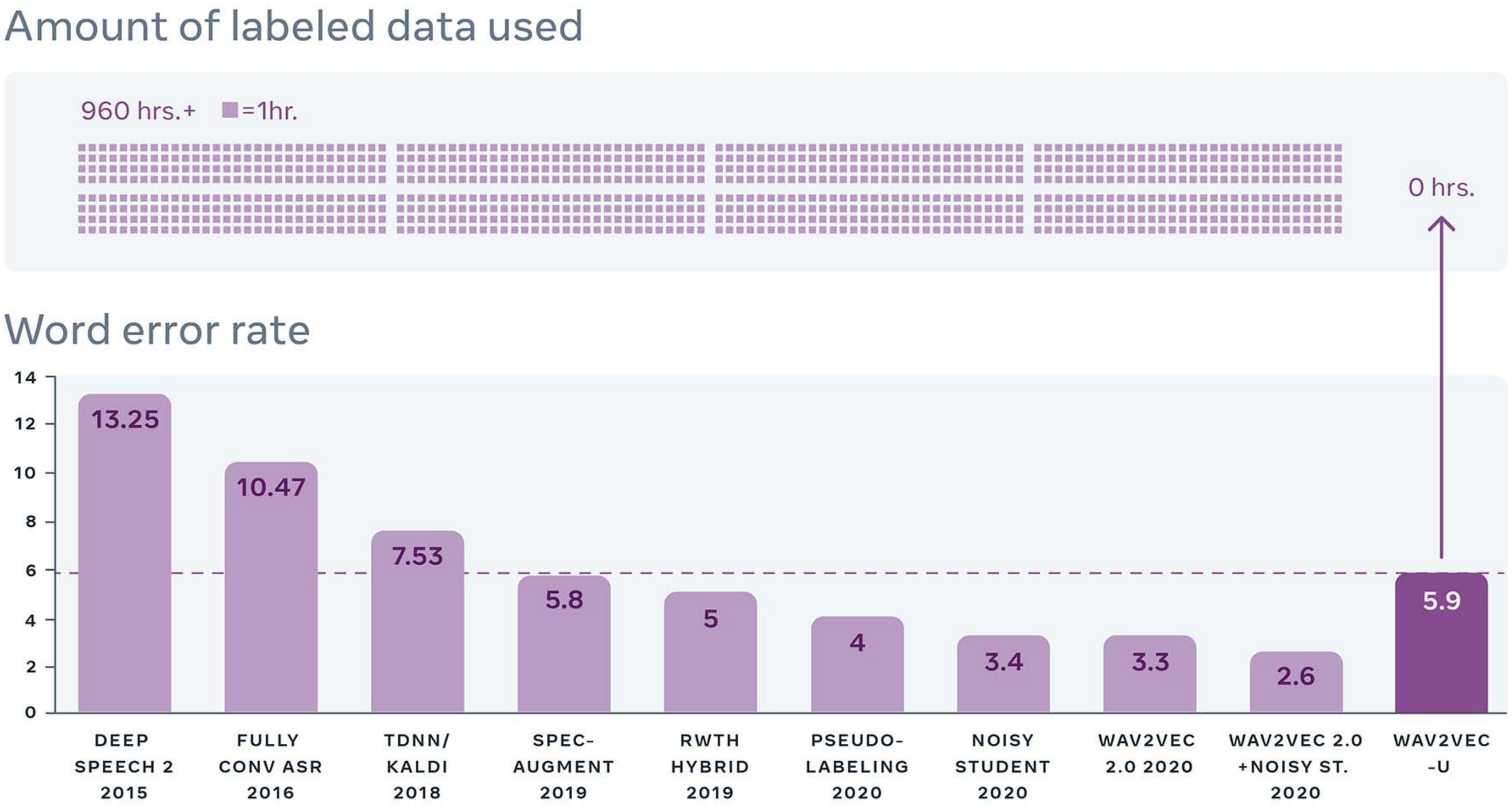


The progress of supervised learning

Unsupervised learning is as good as supervised learning 30 years ago.

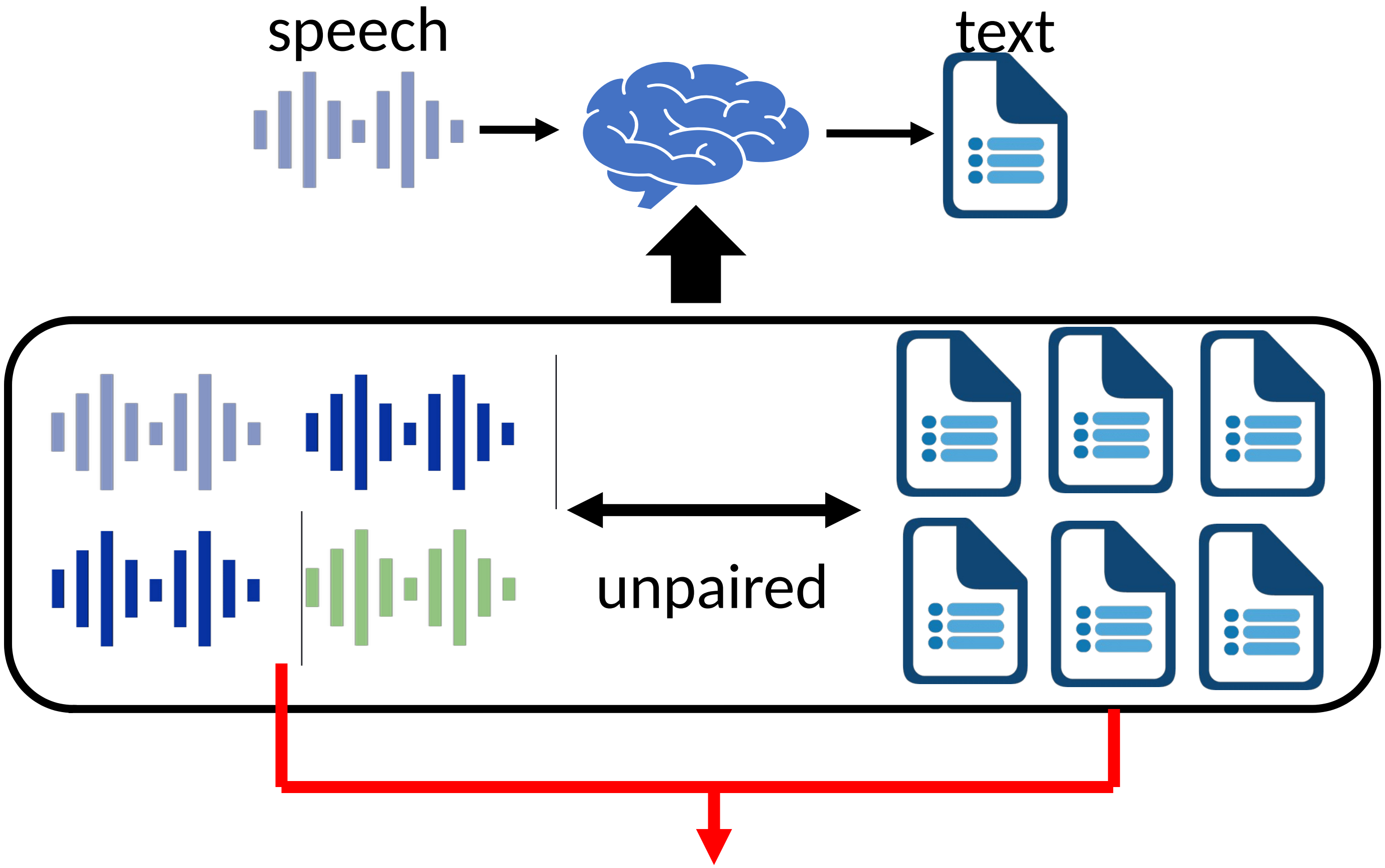
The image is modified from: Phone recognition on the TIMIT database Lopes, C. and Perdigão, F., 2011. Speech Technologies, Vol 1, pp. 285--302.

Unsupervised ASR + Self-supervised Pre-training



https://ai.facebook.com/blog/wav2vec-unsupervised-speech-recognition-without-supervision/

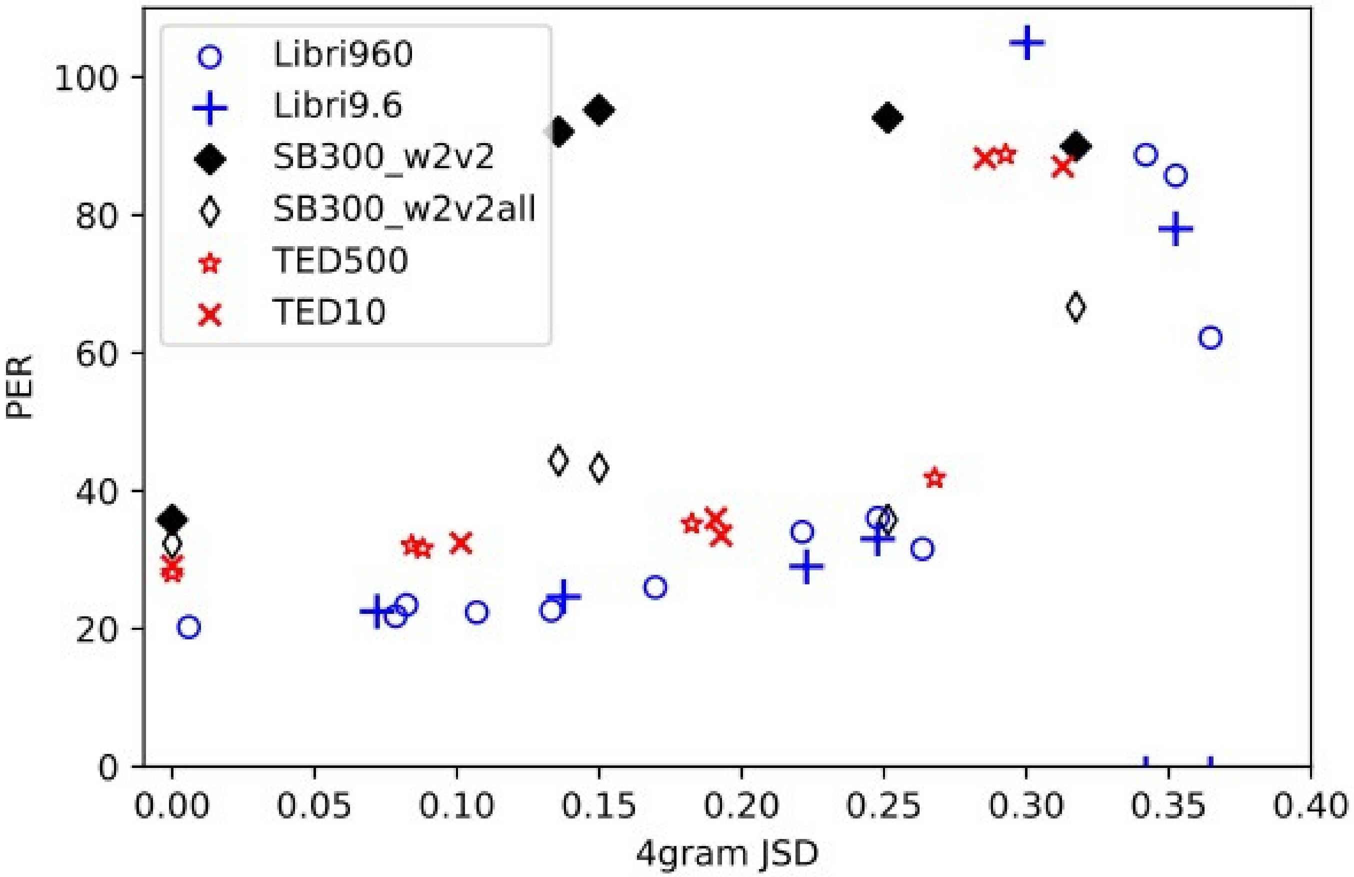
5. Unsupervised Speech Recognition



Can they come from different domains?

5. Unsupervised Speech Recognition

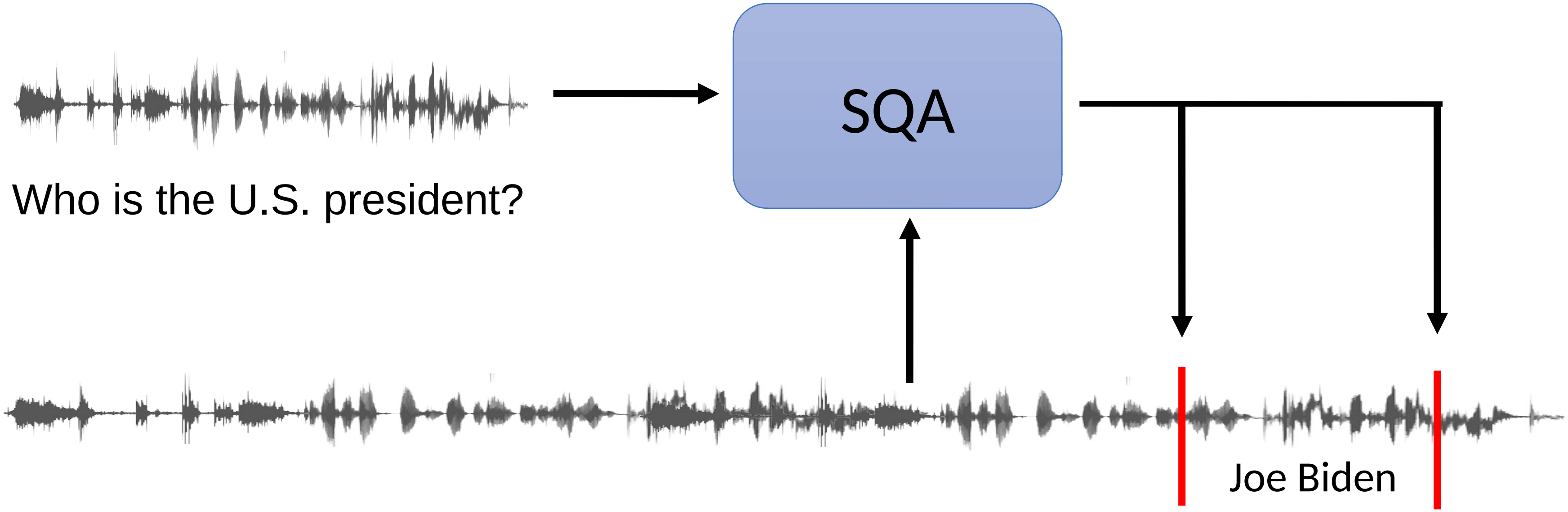
https://arxiv.org/abs/2110.03509



6. Spoken Question Answering

- Spoken Question Answering (SQA)

Without speech recognition

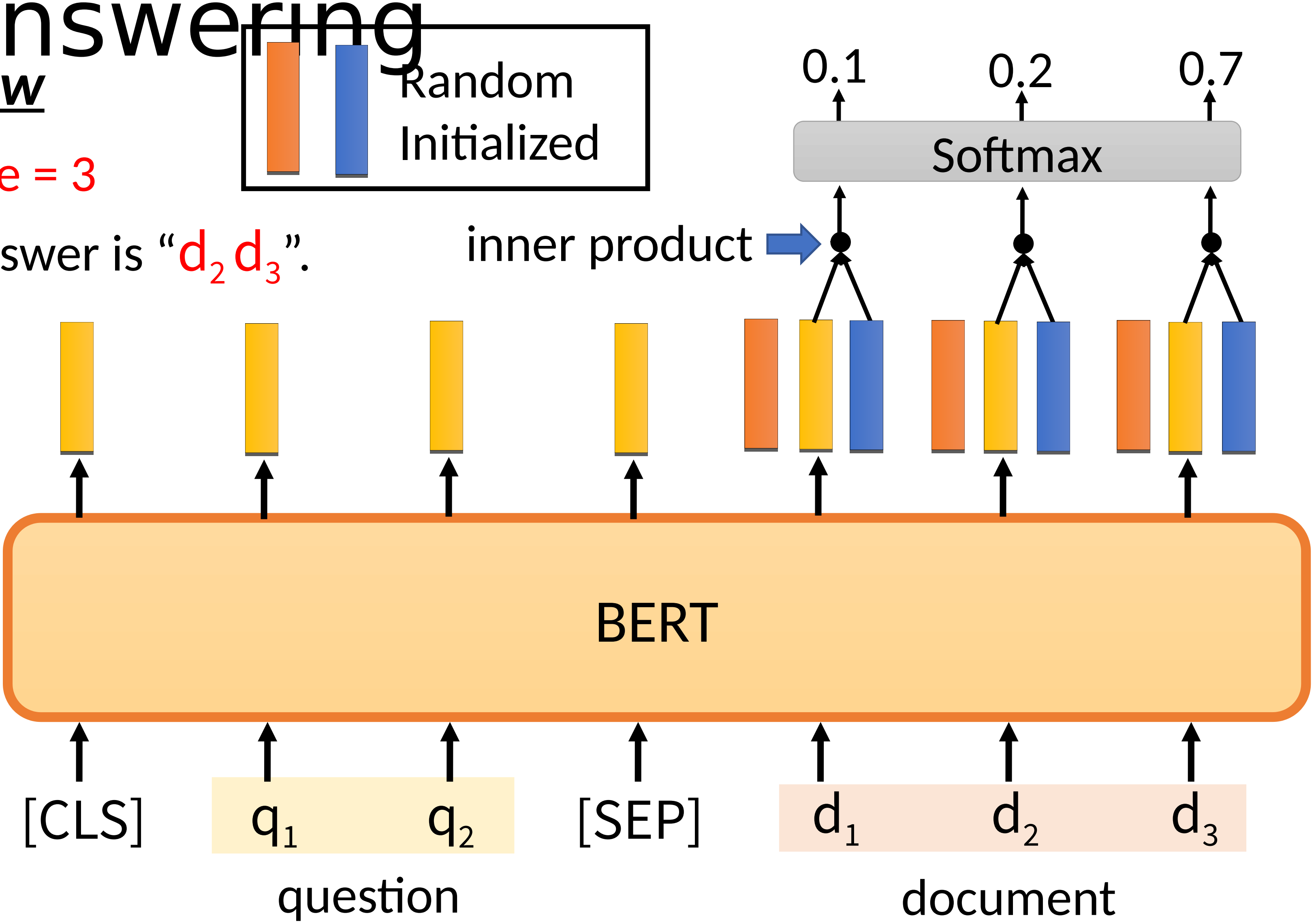


6. Spoken Question Answering

Review

$s = 2 \quad e = 3$

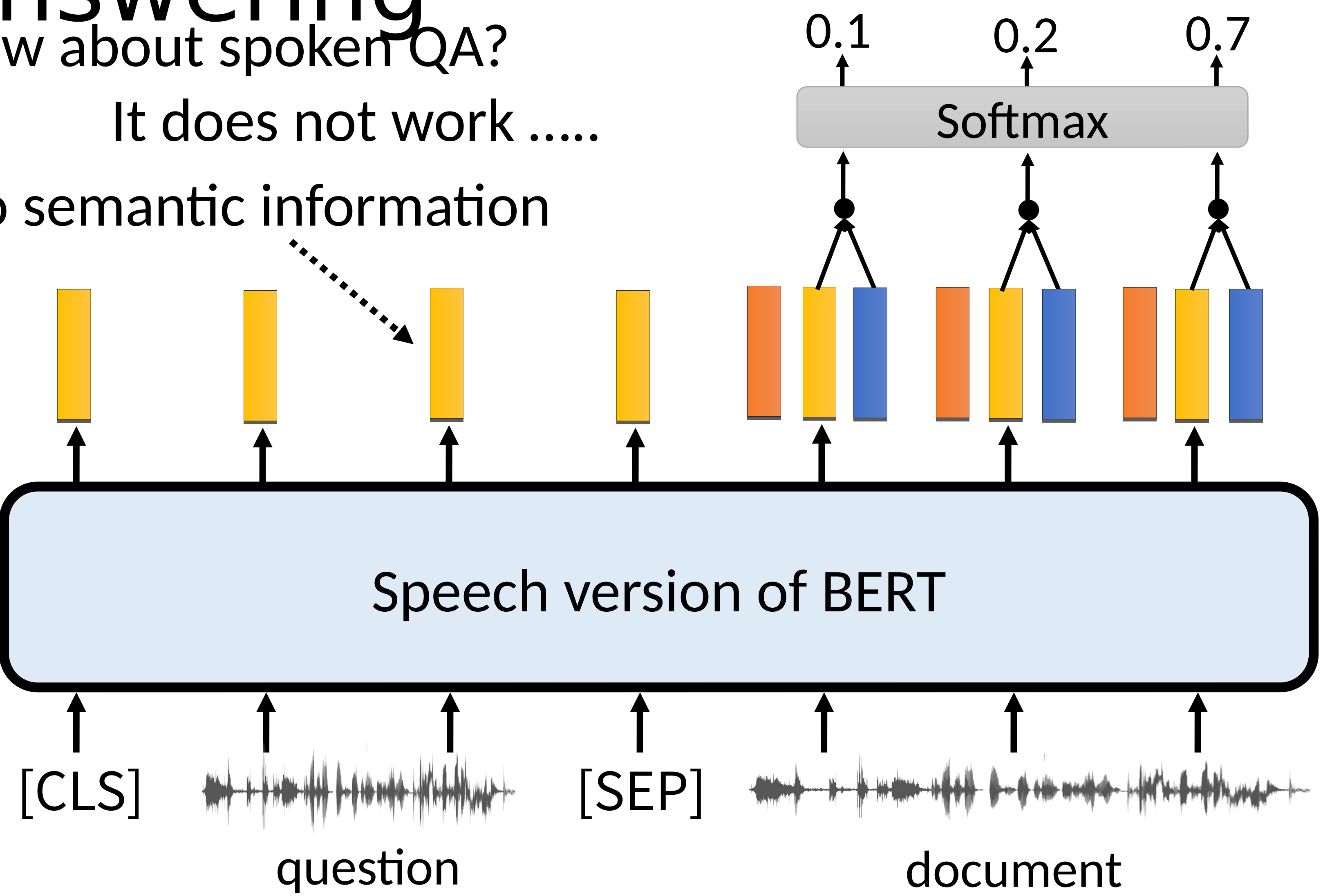
The answer is “$d_2\, d_3$”.



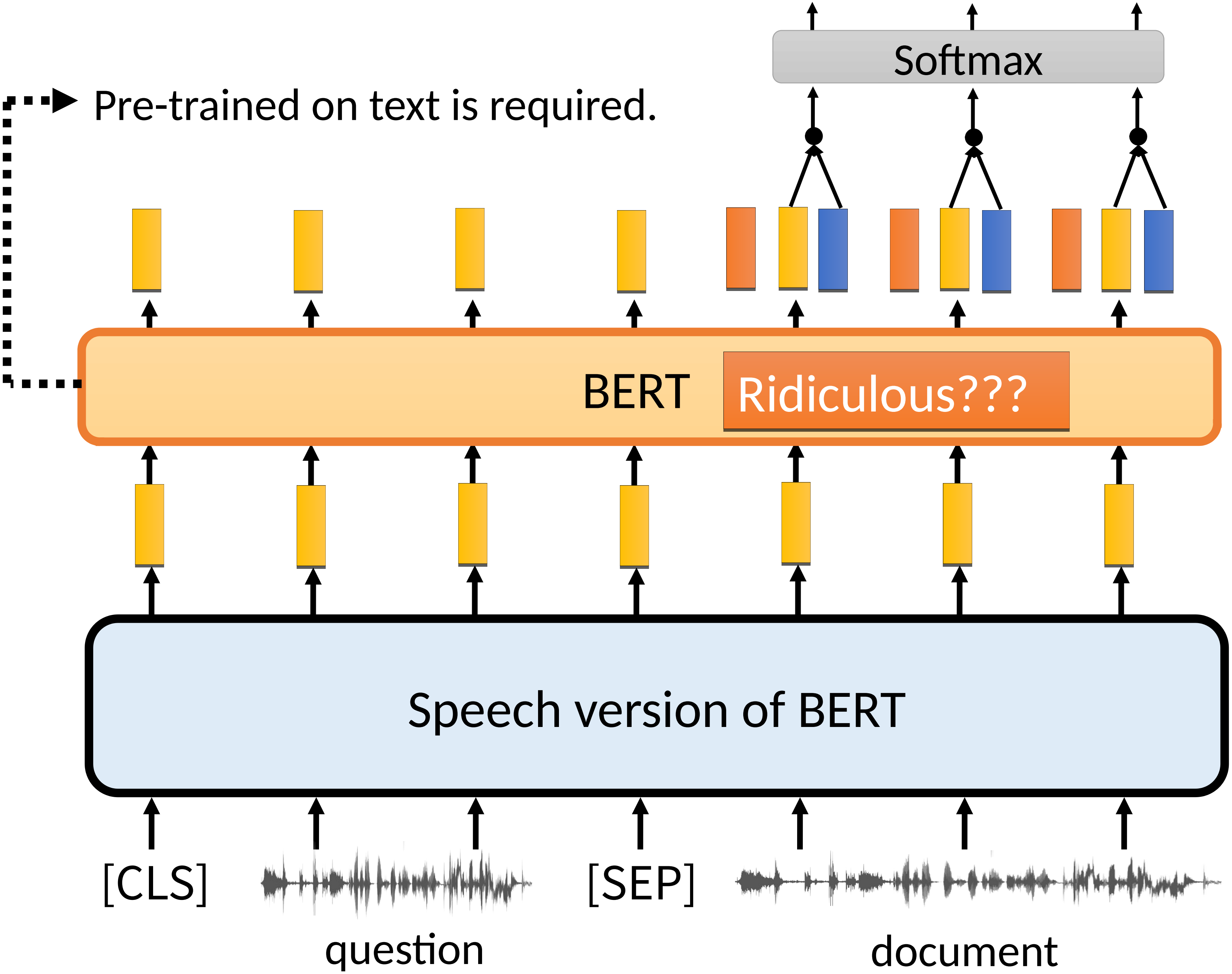
6. Spoken Question Answering

How about spoken QA?

It does not work .....



No semantic information



Softmax

Pre-trained on text is required.

BERT

Ridiculous???

Speech version of BERT

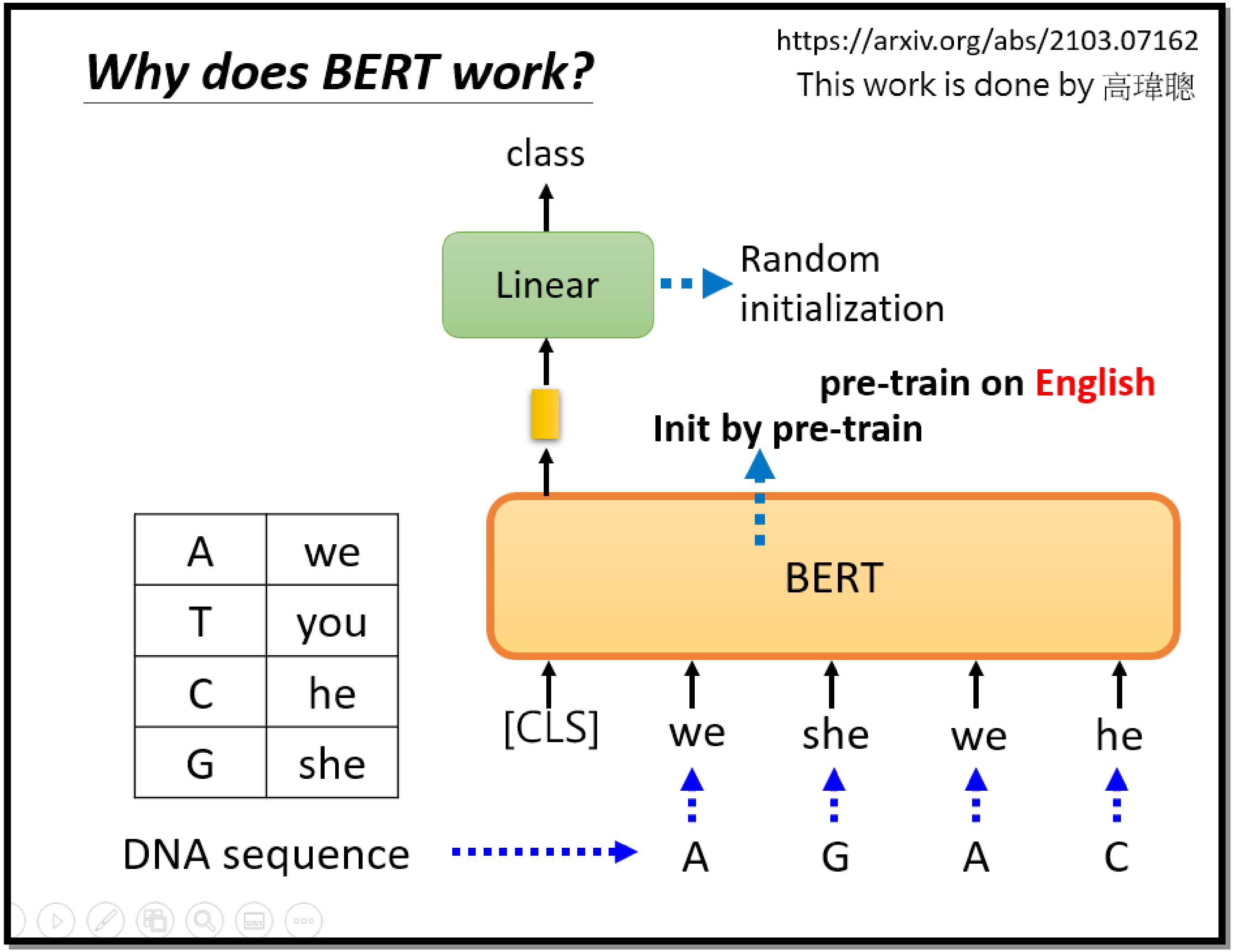
[CLS]

question

[SEP]

document

Recall these experiments ….



Why does BERT work?

https://arxiv.org/abs/2103.07162

This work is done by 高瑋聰

class

Linear

Random initialization

pre-train on English

Init by pre-train

BERT

[CLS] we she we he

DNA sequence

A G A C

A	we
T	you
C	he
G	she

No pre-train: **6.12** F1 score Pre-training on text: **54.22** F1 score

More ……

- 1. Unsupervised Speech Recognition
- 2. Make Pre-trained Model Smaller
- 3. Attacking Pre-trained Model
- 4. Privacy Issue of Pre-trained Model
- 5. Data Bias vs. Pre-training
- 6. Spoken Question Answering