SUPERB

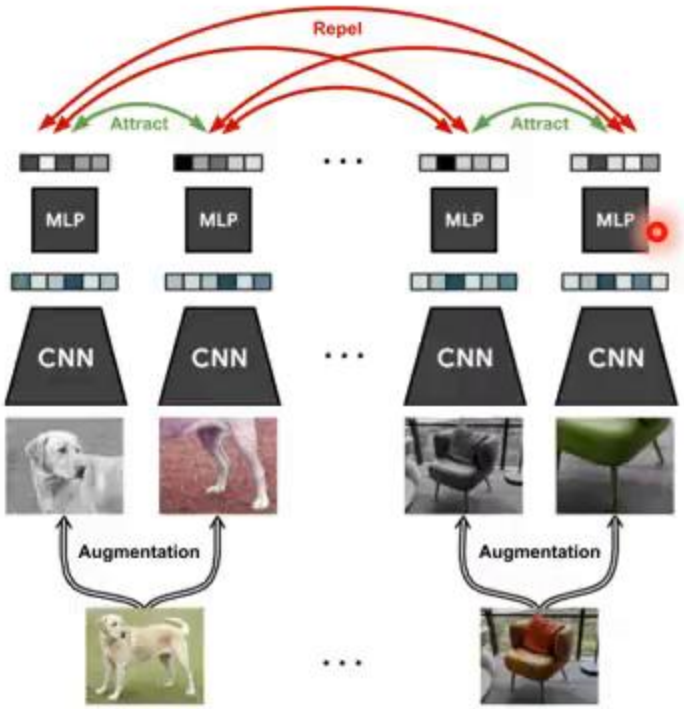
Self-supervised Learning for Speech

SUPERB

Self-supervised Learning for Speech

Self-supervised Learning

BERT (text)



SimCLR (Image)

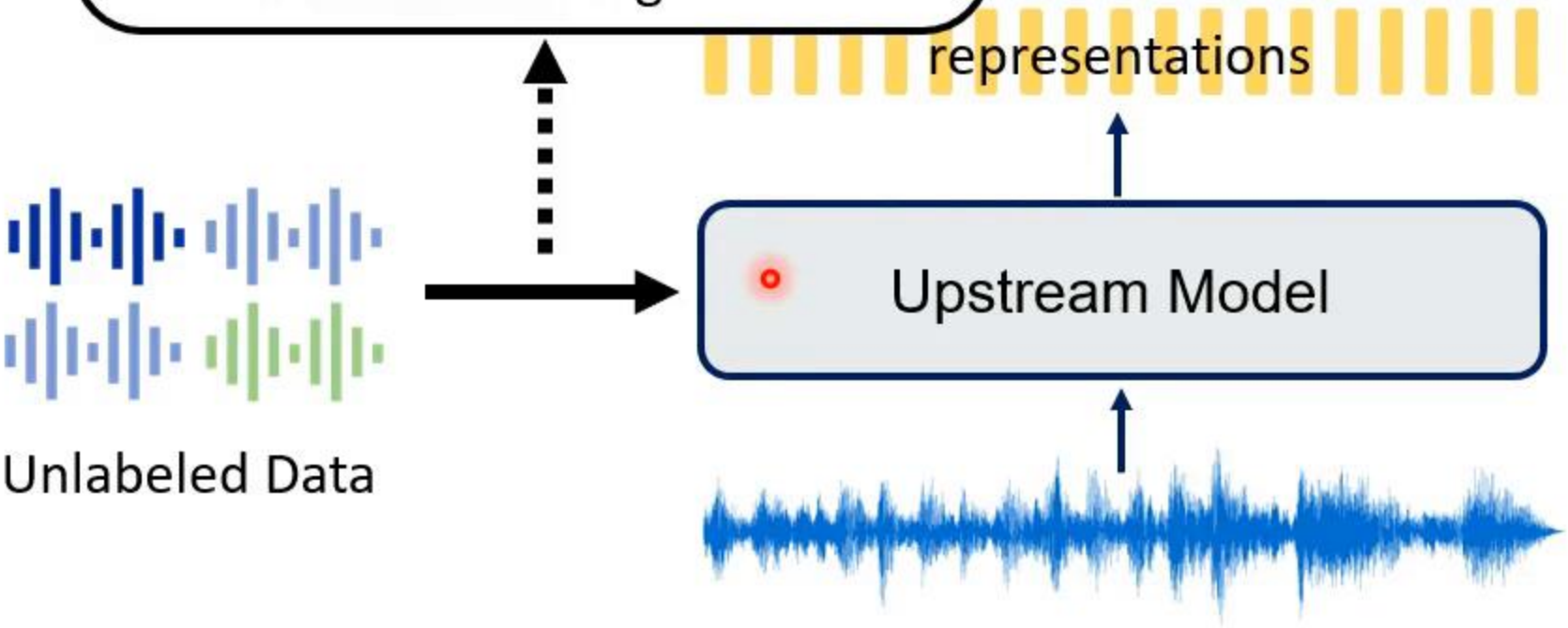
Self-supervised Learning Framework

Phase 1: Pre-train

(not complete survey)

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

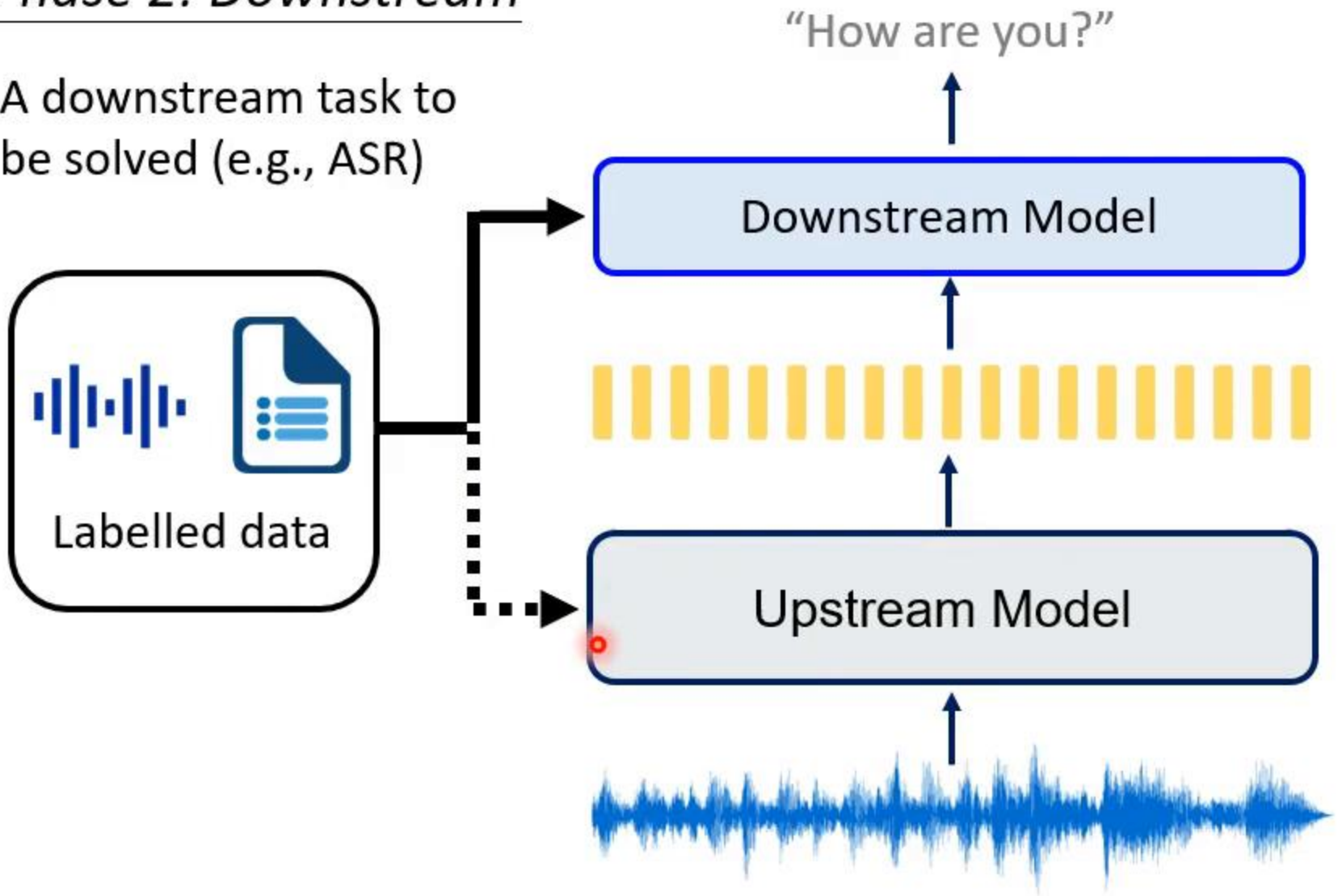
Task-agnostic



Self-supervised Learning Framework

Phase 2: Downstream

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



Specialist? Universal?

Just name a few …

They have shown to achieve good performance on ASR.

Specialist? Universal?

Just name a few …

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

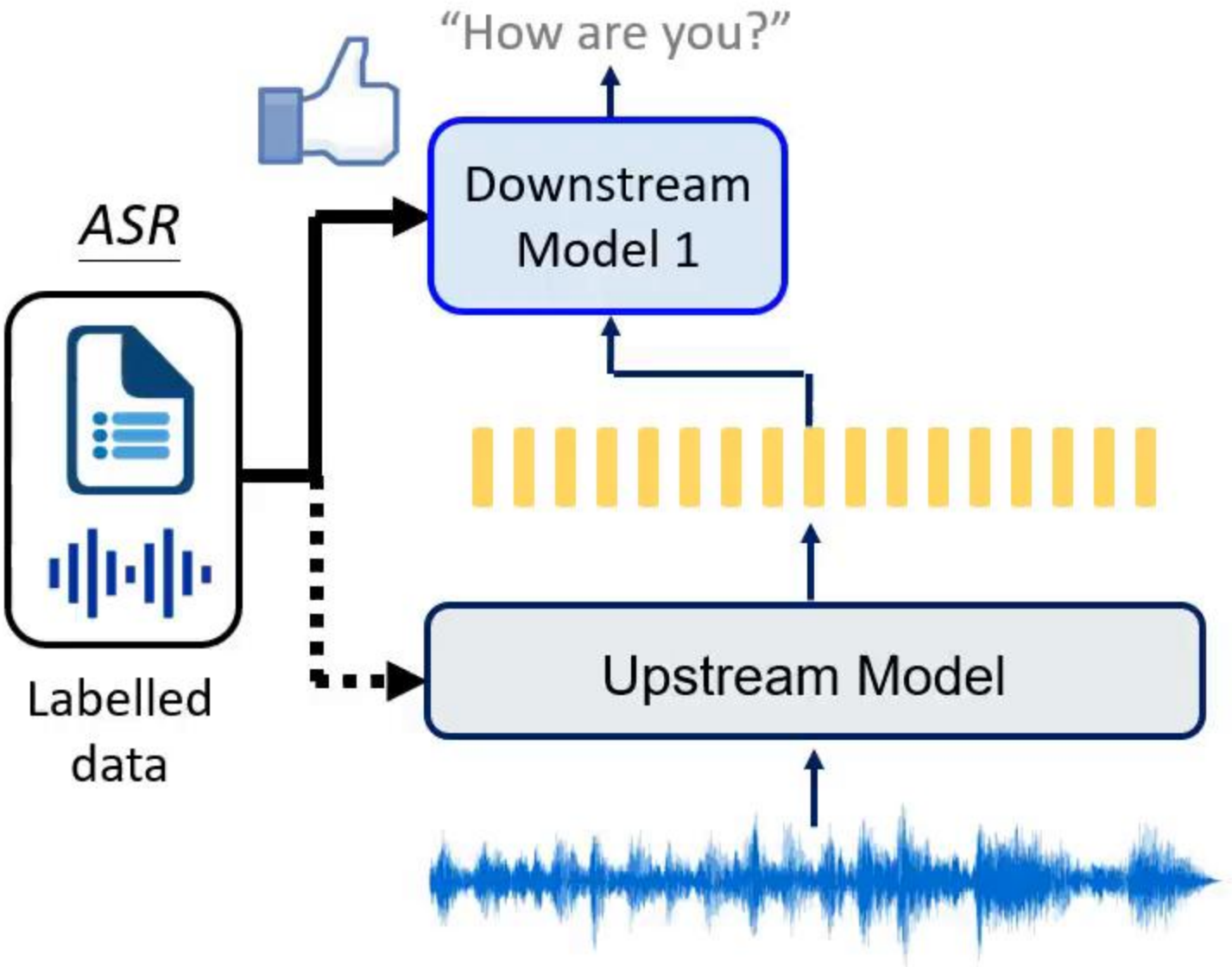
Specialist? Universal?

Just name a few …

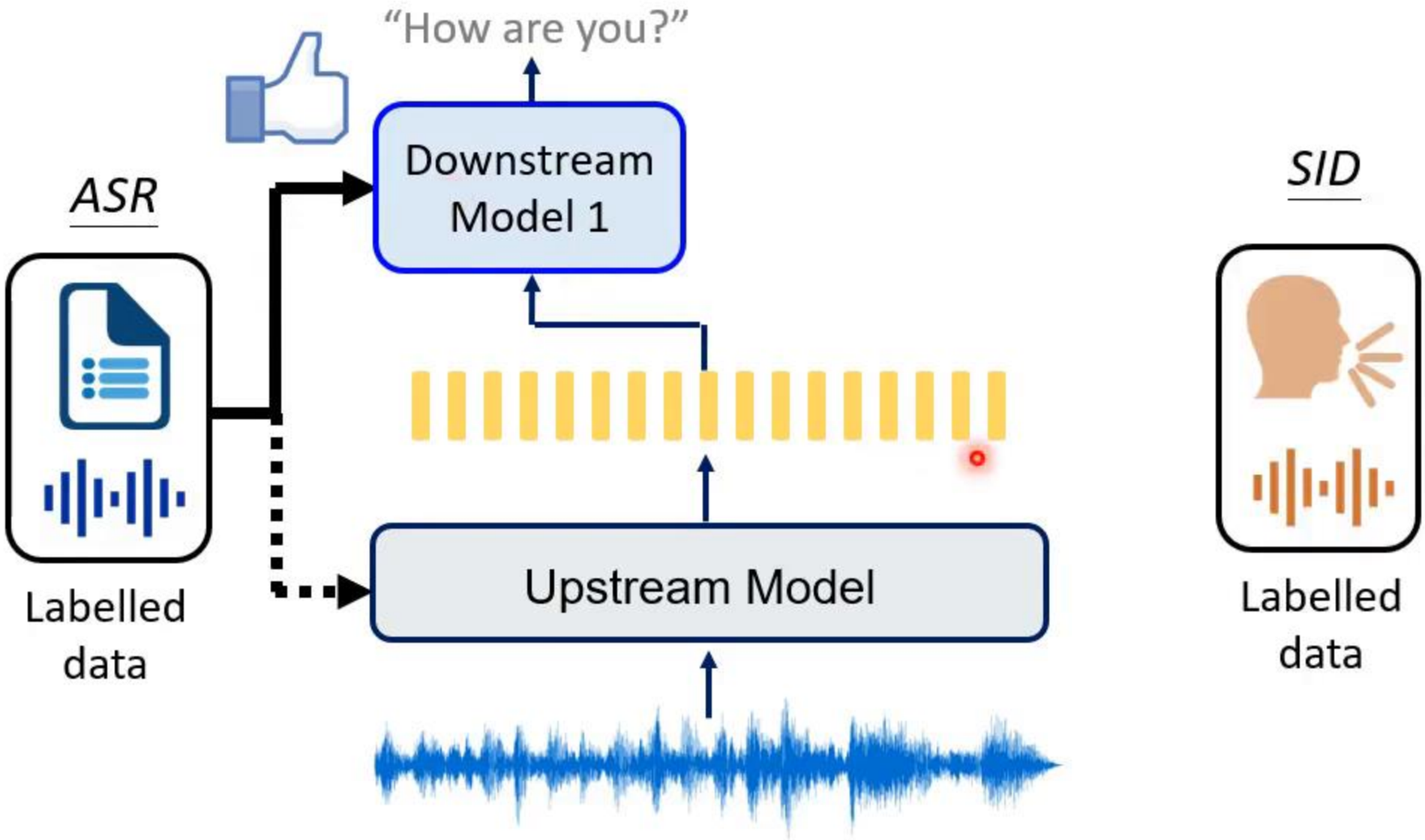
They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

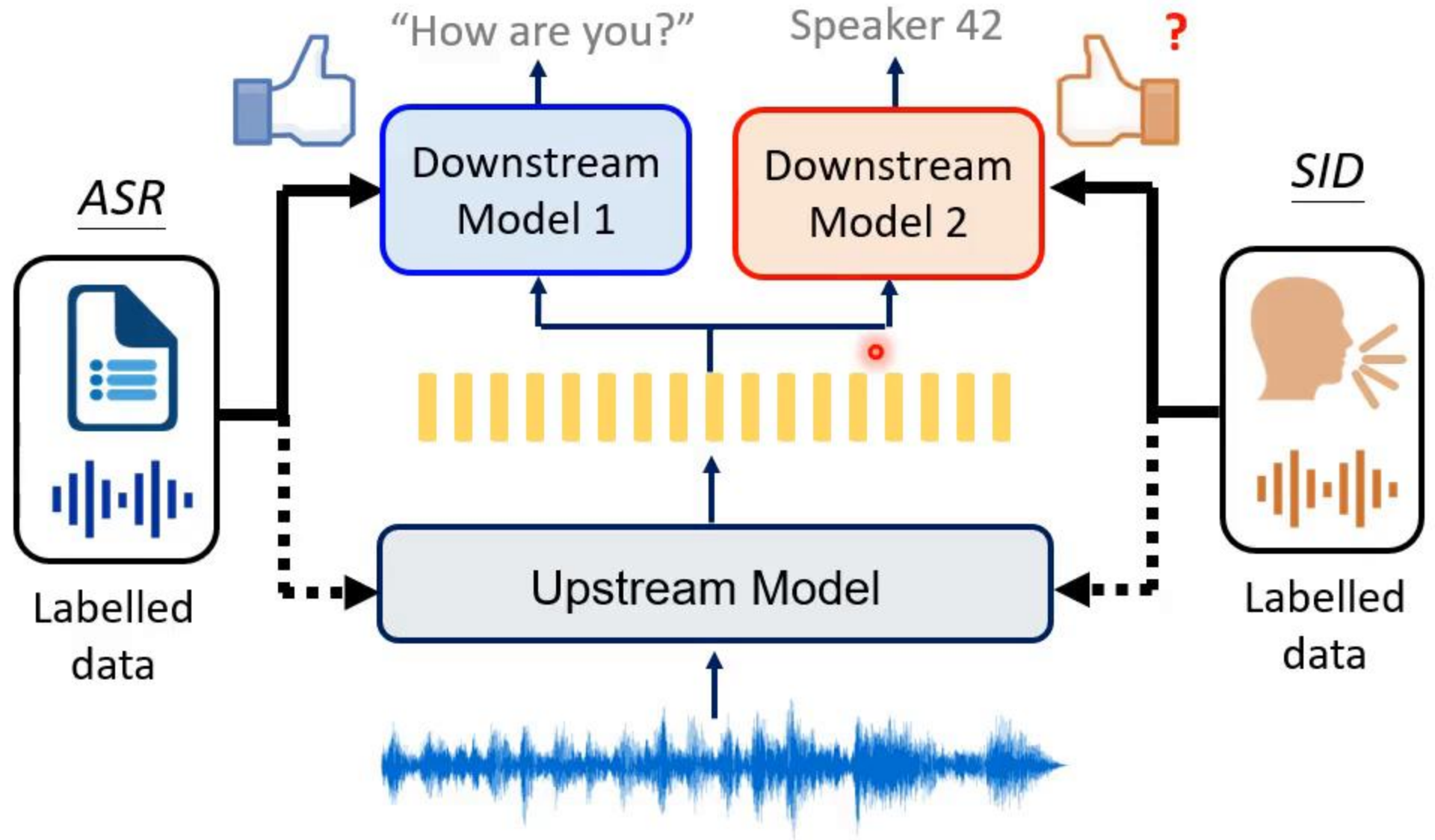
Specialist? Universal?



Specialist? Universal?



Specialist? Universal?



Specialist? Universal?

Just name a few …

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

My two cents
(one year ago)

Specialist? Universal?

Just name a few …

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

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

SUPERB

Ref: https://arxiv.org/abs/2105.01051


PASE+


APC


NPC


Mockingjay


DeCoAR

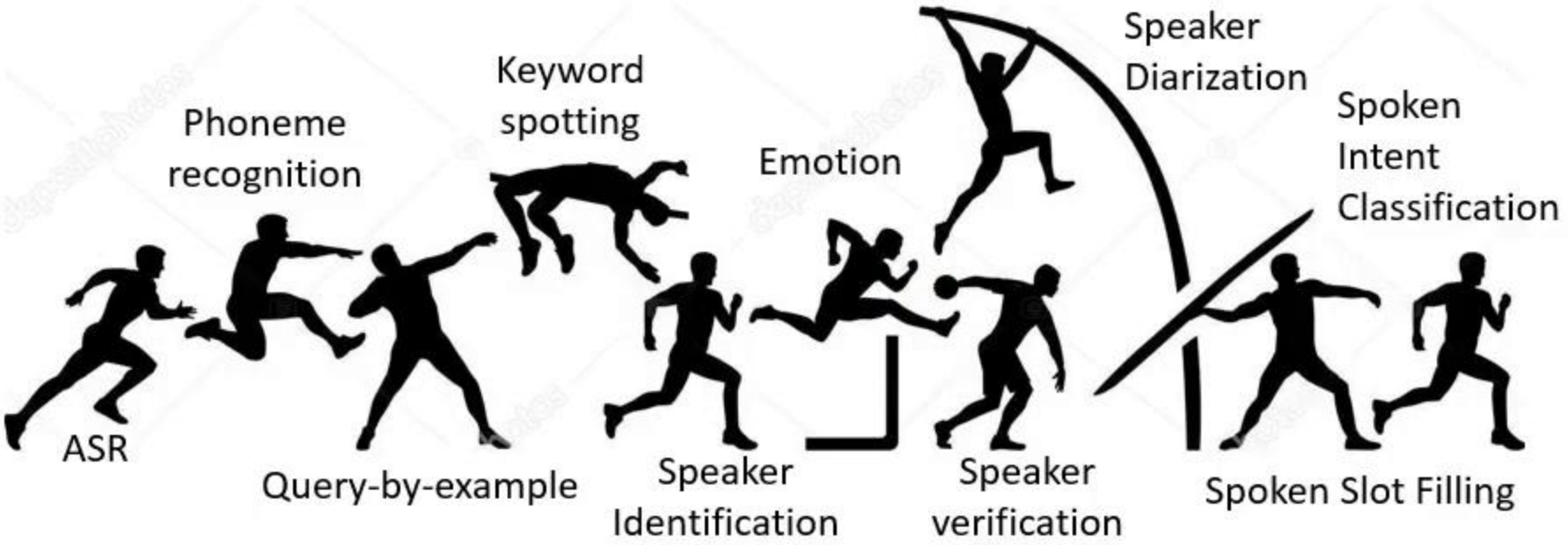

Wav2vec


HuBERT


ASR

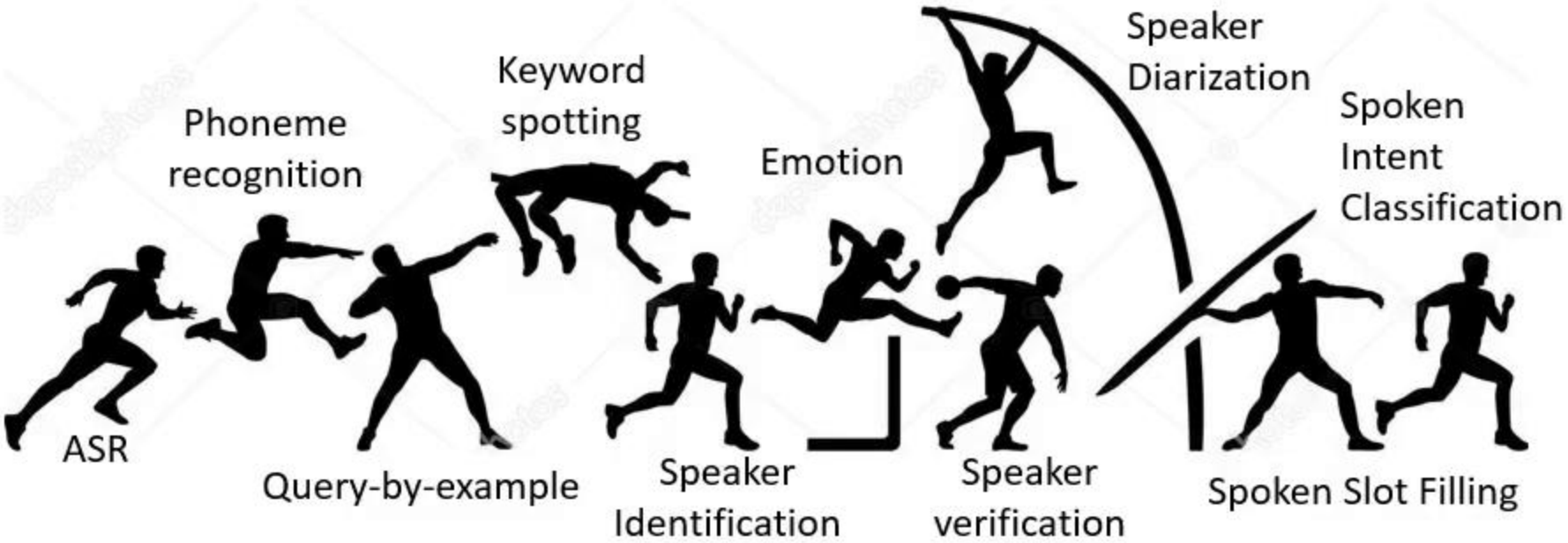
SUPERB

Ref: https://arxiv.org/abs/2105.01051



SUPERB

Ref: https://arxiv.org/abs/2105.01051



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

Tasks – Speaker

Speaker Identification

Speaker Verification

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 ...**

Ref: https://arxiv.org/abs/2105.01051

Game Start!

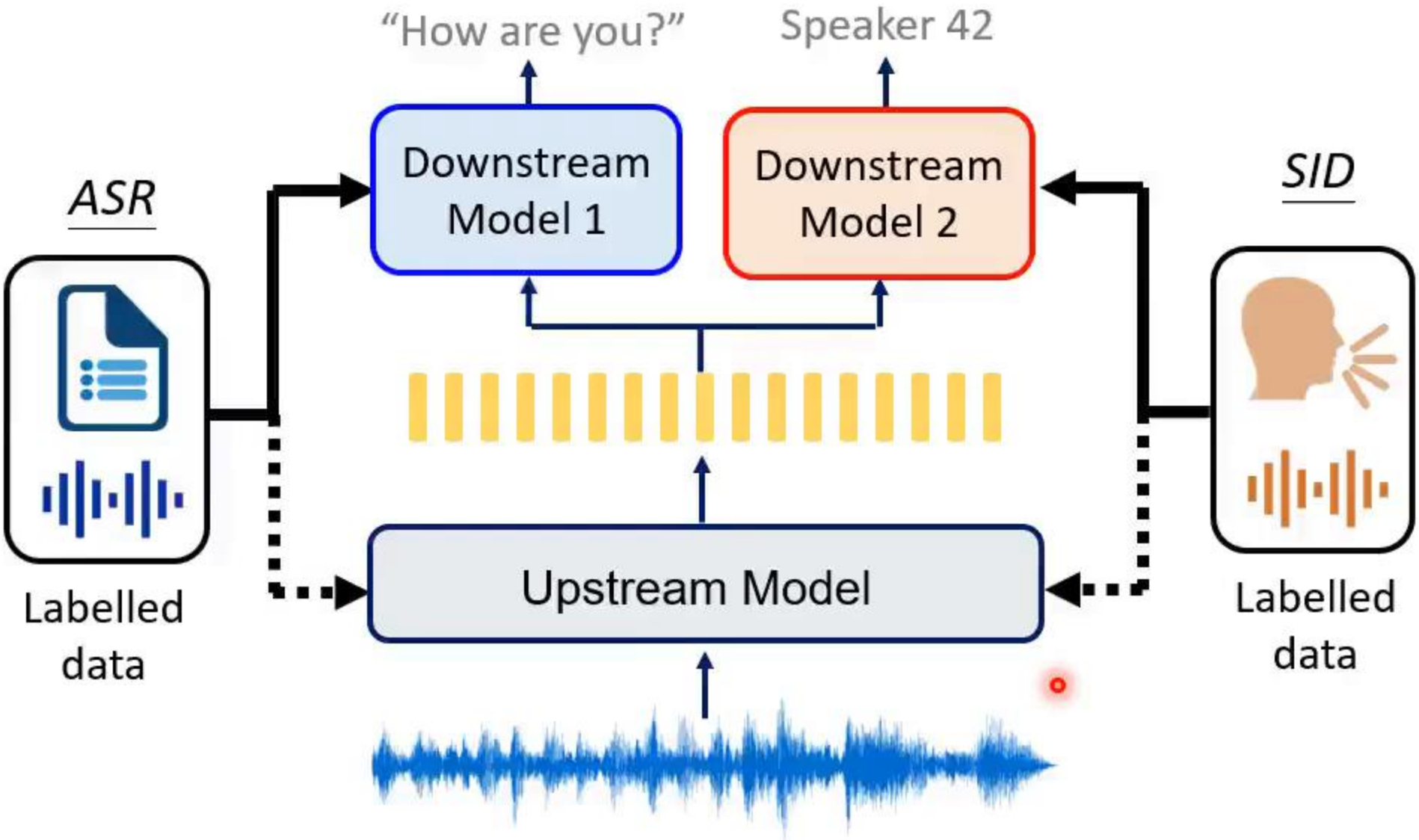
Round 1

Game Start!

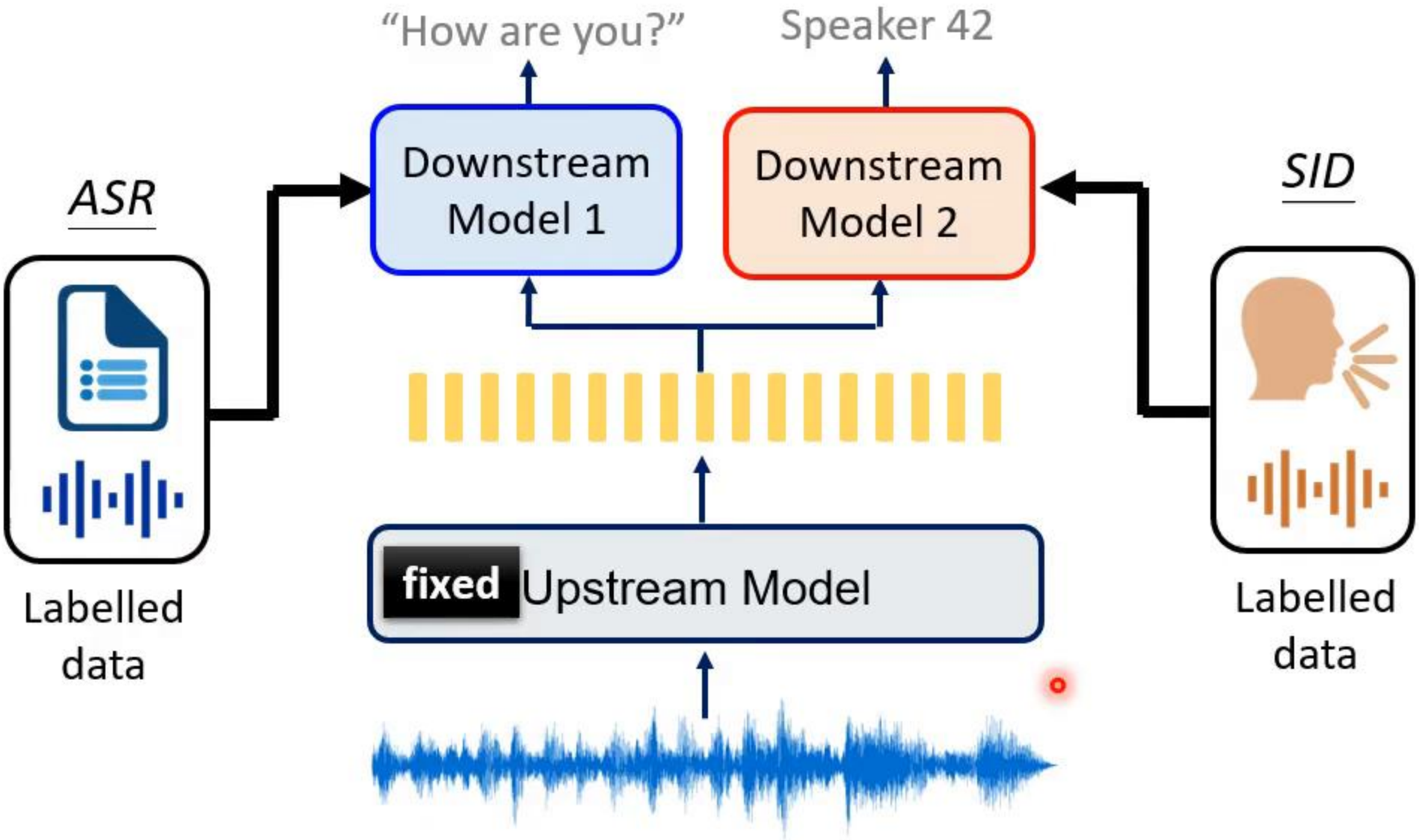
Round 1

Rules in Round 1

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



Rules in Round 1



Rules in Round 1

The network architecture of a downstream model is predefined.

Keep it simple

"How are you?"

Speaker 42

e.g., 2-layer LSTM

e.g., Linear layer

ASR

SID

Downstream Model 1

Downstream Model 2

Last Layer Output

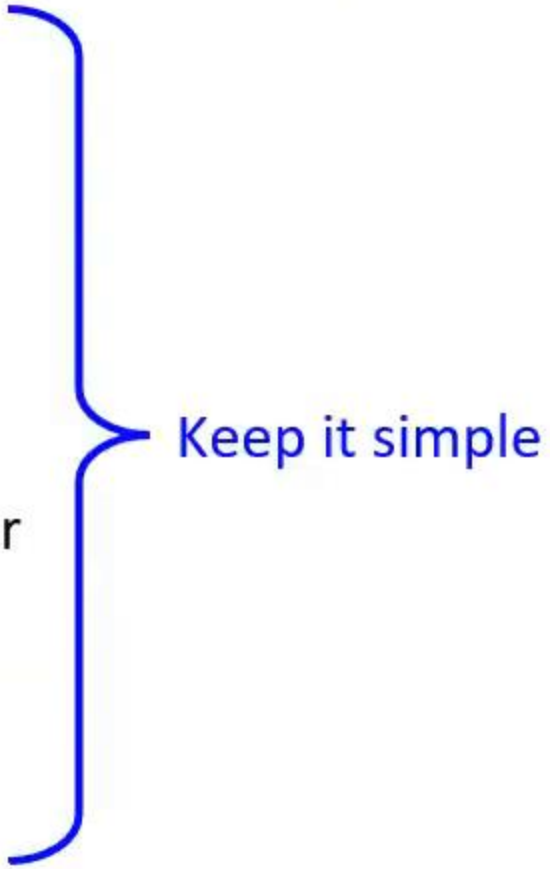
fixed Upstream Model

Labelled data

Labelled data

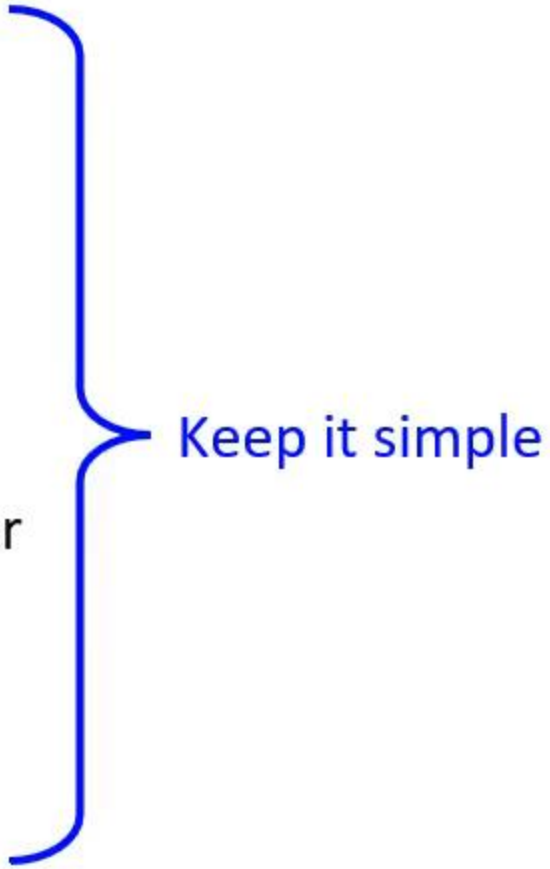
Rules in Round 1 – Downstream

- 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

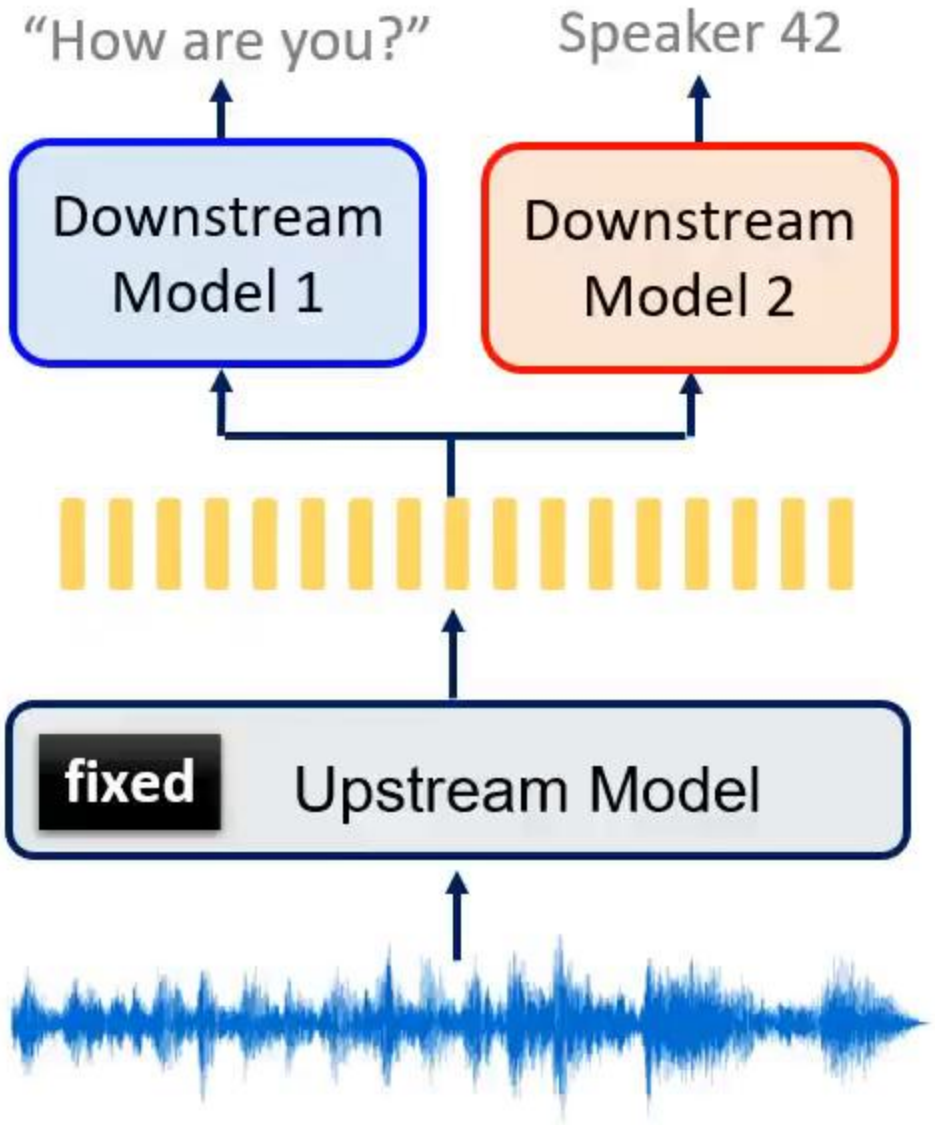


Rules in Round 1 – Downstream

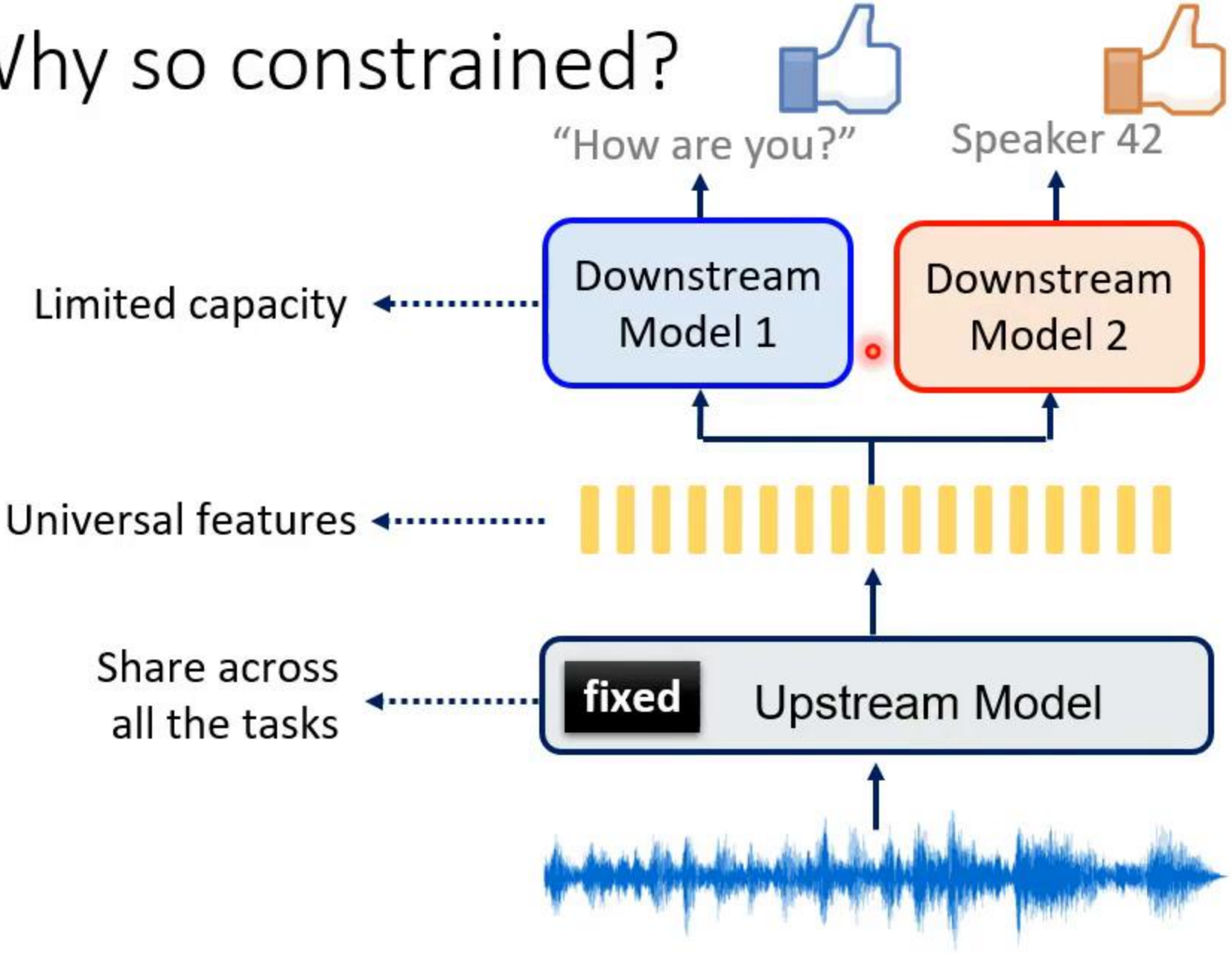
- Phoneme Recognition: linear layer
- Keyword Spotting: linear layer
- Speech Recognition: 2-layer LSTM
- Query-by-example: none
- Speaker Identification: linear layer
- Speaker Verification: the same as x-vector
- Speaker Diarization: 1-layer LSTM
- Intent Classification: linear layer
- Slot Filling: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



Why so constrained?



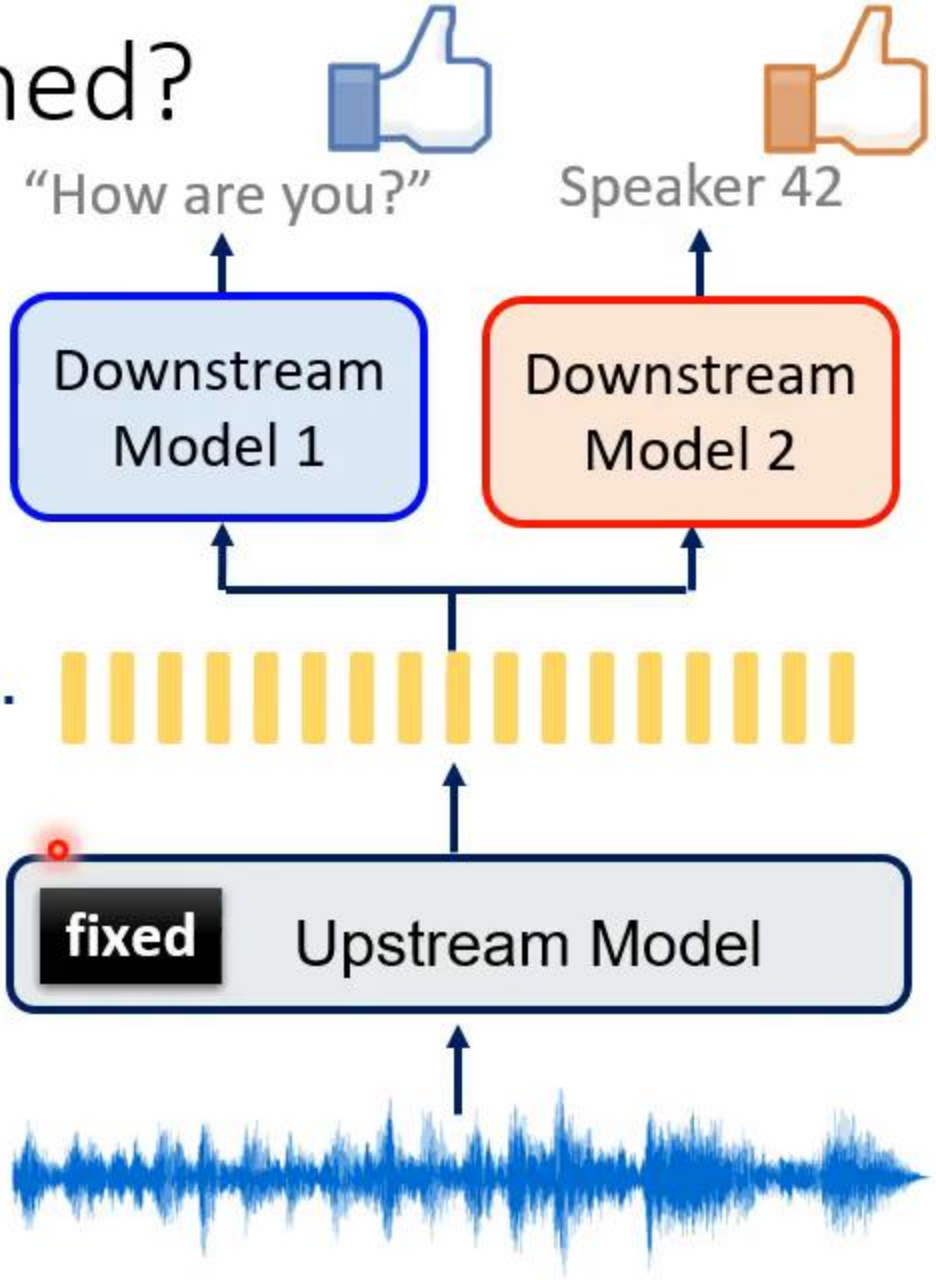
Why so constrained?



Why so constrained?

Easy to build new applications!

This sounds too good to be true ……



Results of Round 1

	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

Results of Round 1

Emotion

	Content				Speaker			Semantic		
	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

Upstream Models

Results of Round 1

Emotion

	Content				Speaker			Semantic		
	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

Upstream Models

Results of Round 1

Emotion

	Content				Speaker			Semantic		
	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

Results of Round 1

Emotion

	Content				Speaker			Semantic		
	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

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

Results of Round 1

Emotion

	Content				Speaker			Semantic		
	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 self-supervised models.

Game Start!

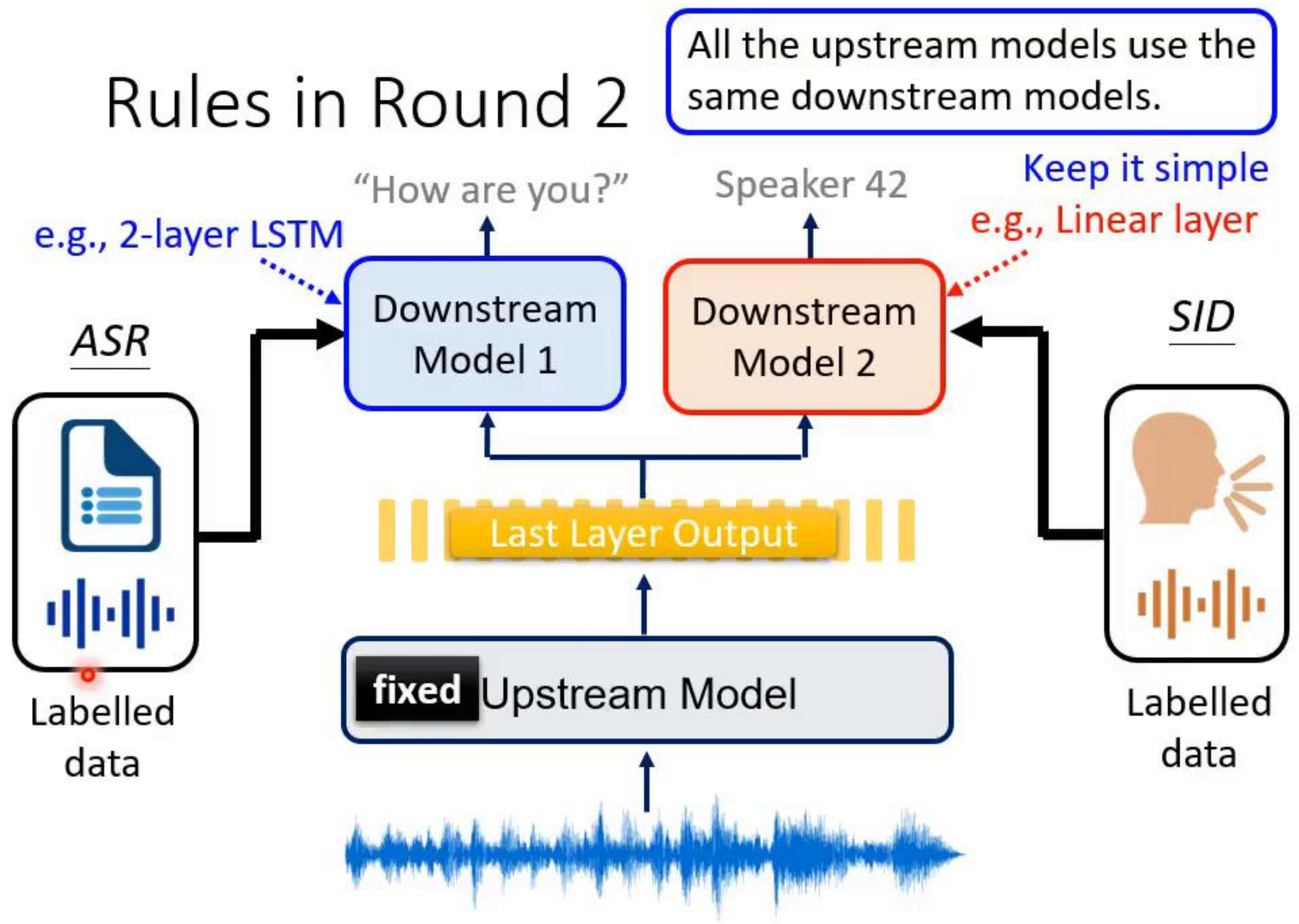
Round 2

Game Start!

Round 2

Rules in Round 2

All the upstream models use the same downstream models.



Rules in Round 2

All the upstream models use the same downstream models.

"How are you?"

Speaker 42

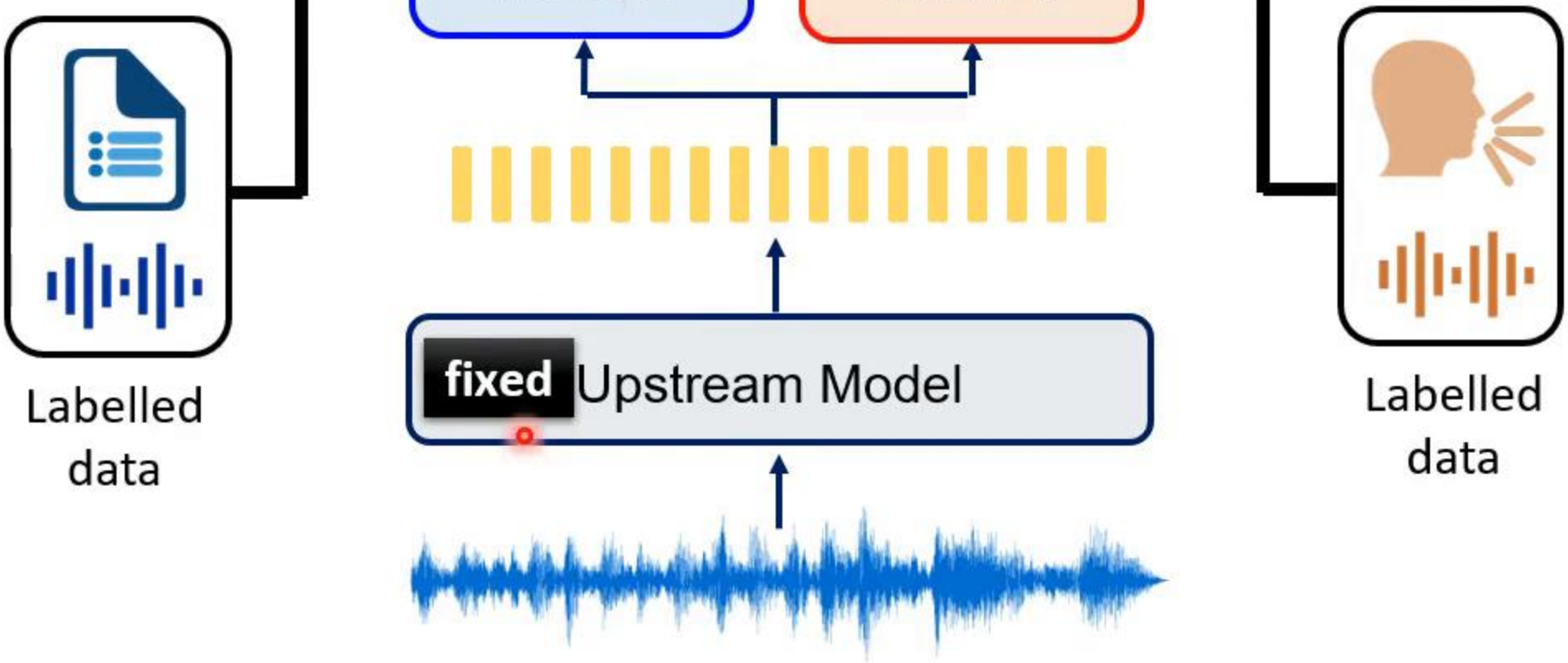
Keep it simple

e.g., 2-layer LSTM

e.g., Linear layer

ASR

SID



Rules in Round 2

All the upstream models use the same downstream models.

"How are you?"

Speaker 42

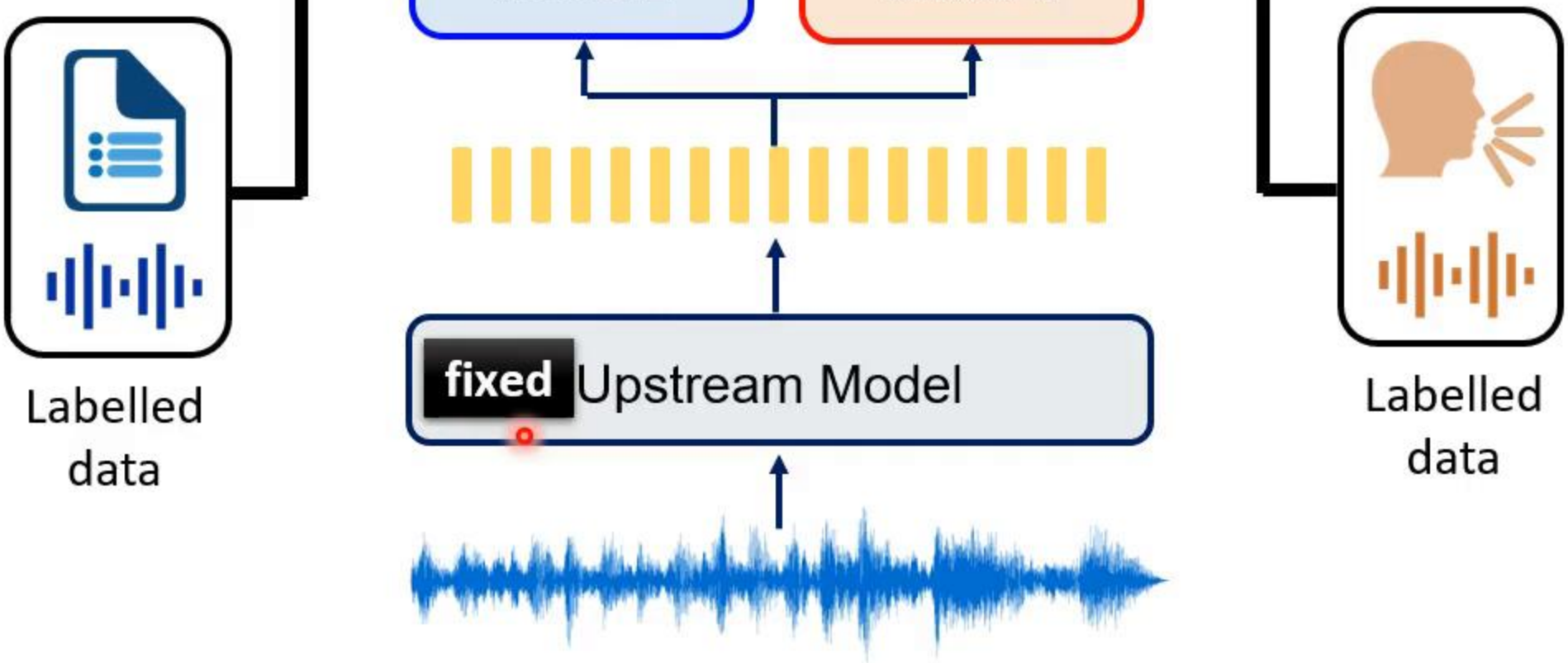
Keep it simple

e.g., 2-layer LSTM

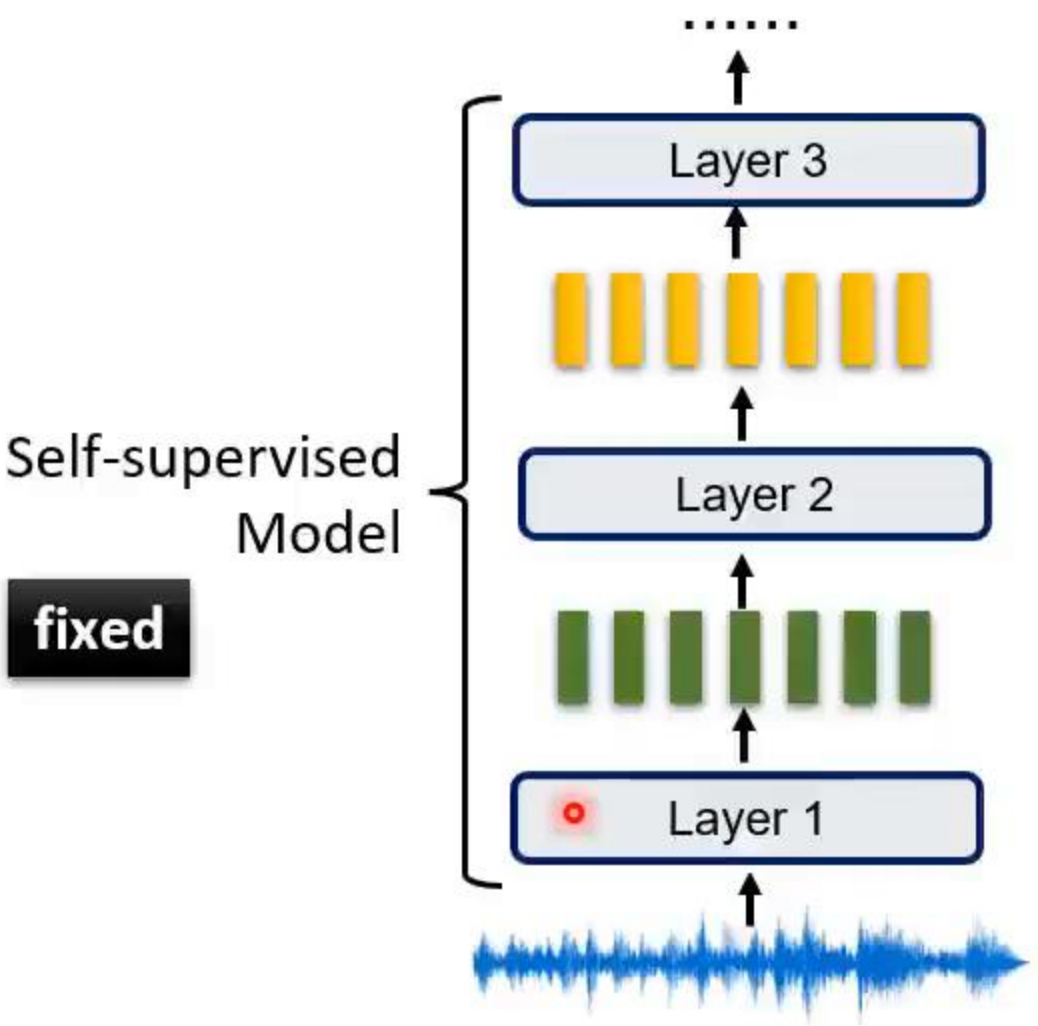
e.g., Linear layer

ASR

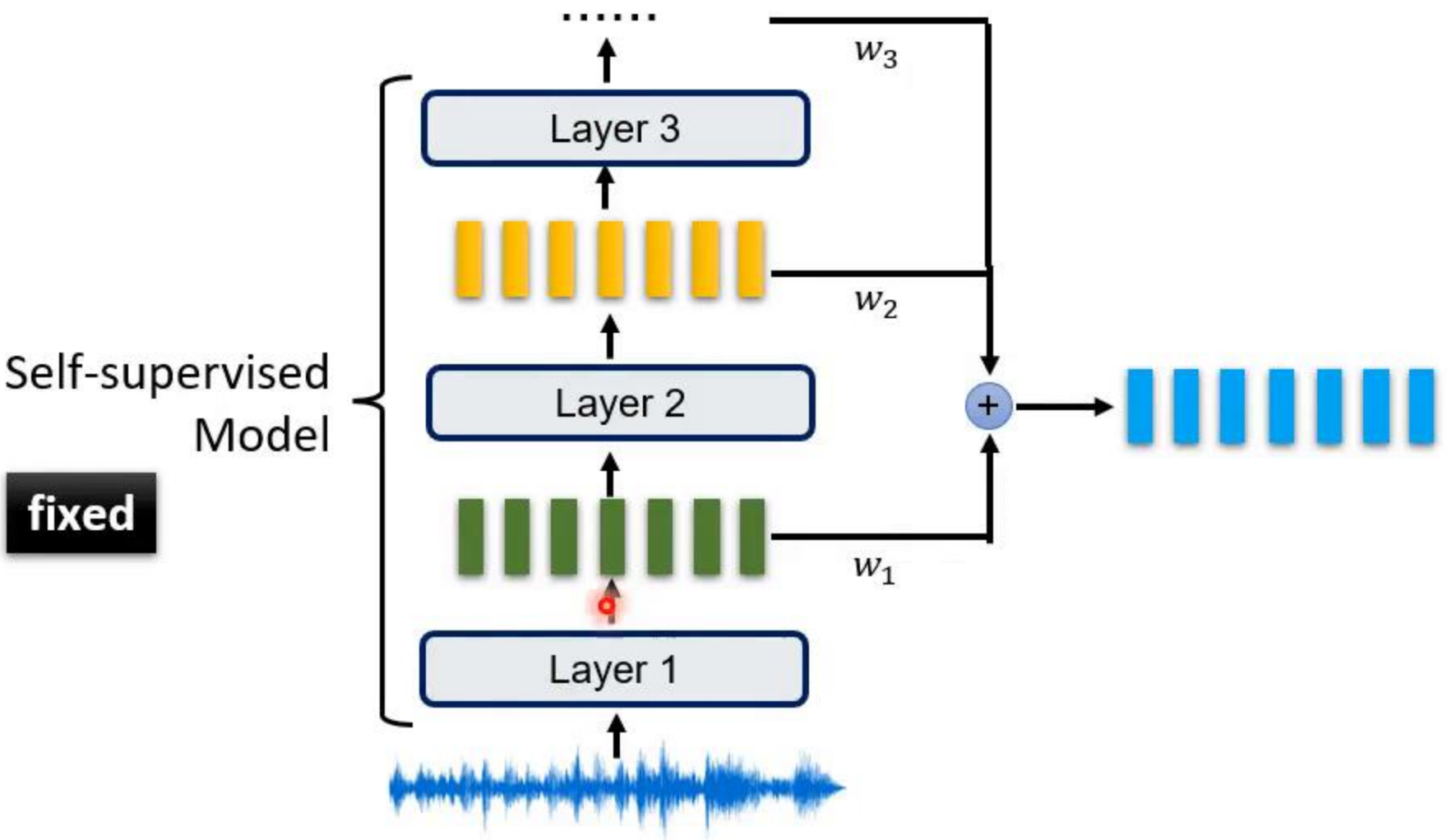
SID



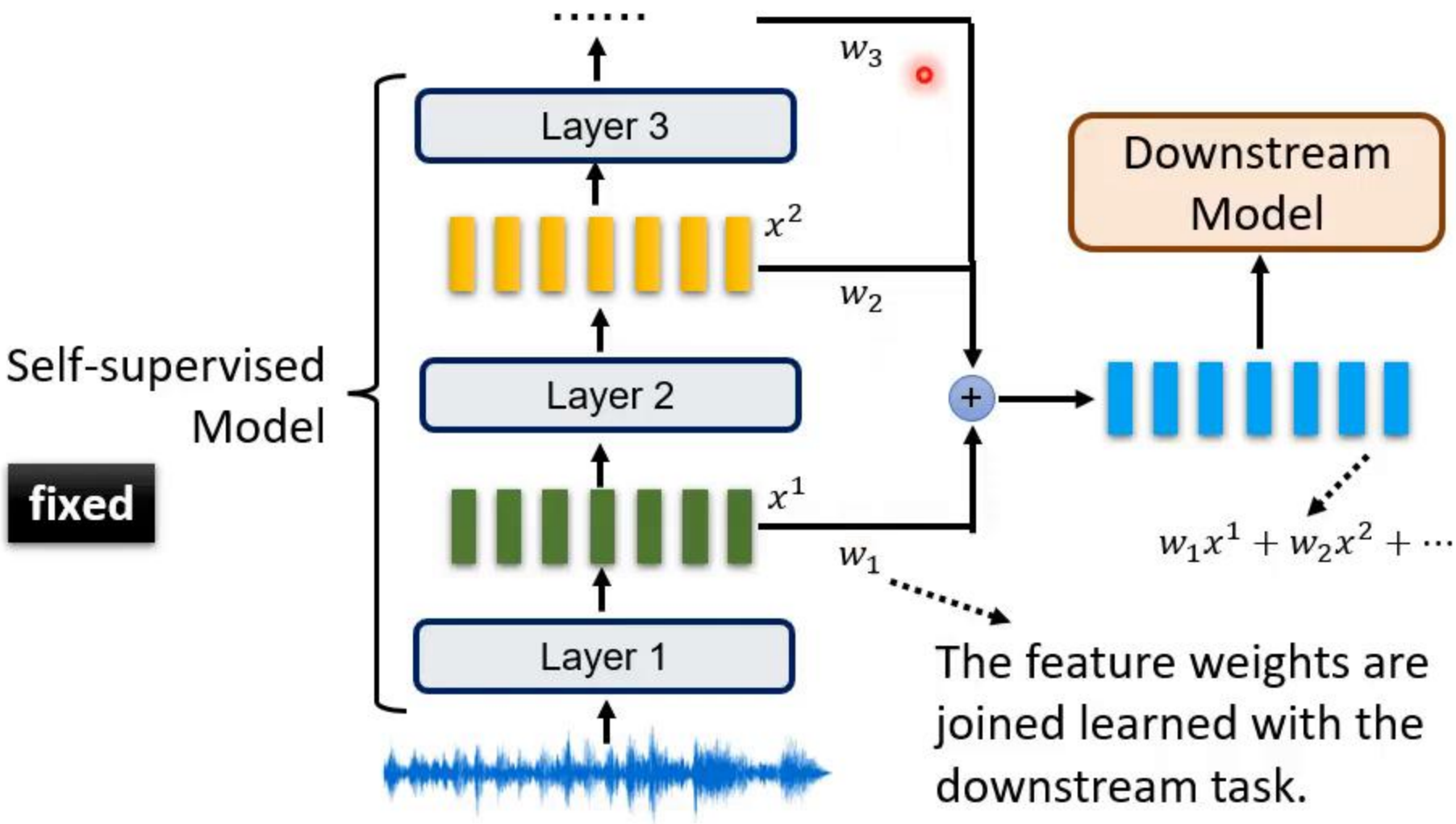
Rules in Round 2



Rules in Round 2



Rules in Round 2



Results of Round 2

Emotion

	Content				Speaker			Semantic		
	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

Emotion

	Content				Speaker			Semantic		
	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

Emotion

	Content				Speaker			Semantic		
	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

Results of Round 2

Emotion

	Content				Speaker			Semantic		
	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 self-supervised models are all-around.

Results of Round 2

Emotion

	Content				Speaker			Semantic		
	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 self-supervised models are all-around.

Specialist? Universal?

Just name a few ...

They have shown to achieve good performance on ASR.

Are they specialist for ASR? Or are they universal?

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)

Specialist? Universal?

Just name a few …

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 ☺

Welcome to Join the Game ☺

SUPERB Challenge

- Stay tuned to the SUPERB's webpage: https://superbbenchmark.org/
- You can submit to public leaderboard now.
- Oct 15, 2021: Hidden-set leaderboard will be online and accepts submissions.

SUPERB Challenge

- Stay tuned to the SUPERB's webpage: https://superbbenchmark.org/
- You can submit to public leaderboard now.
- Oct 15, 2021: Hidden-set leaderboard will be online and accepts submissions.

SUPERB Challenge

- Stay tuned to the SUPERB's webpage: https://superbbenchmark.org/
- You can submit to public leaderboard now.
- Oct 15, 2021: Hidden-set leaderboard will be online and accepts submissions.

AAAI 2022 Workshop

- **Self-Supervised Learning for Speech and Audio Processing**
- Webpage: https://aaai-sas-2022.github.io/
- Deadline: November 12, 2021

Special Issue

- IEEE JSTSP Special Issue on **Self-Supervised Learning for Speech and Audio Processing**
- Deadline: December 31, 2021
- Link: https://signalprocessingsociety.org/blog/ieee-jstsp-special-issue-self-supervised-learning-speech-and-audio-processing

Concluding Remarks

- Participate in the SUPERB challenge
- Self-supervised Learning for Audio and Speech Processing @ AAAI 2022
  - Deadline: November 12, 2021
- IEEE JSTSP Special Issue on Self-Supervised Learning for Speech and Audio Processing
  - Deadline: December 31, 2021