Unified Speech Recognition: A Single Model for Auditory, Visual and Audiovisual Inputs

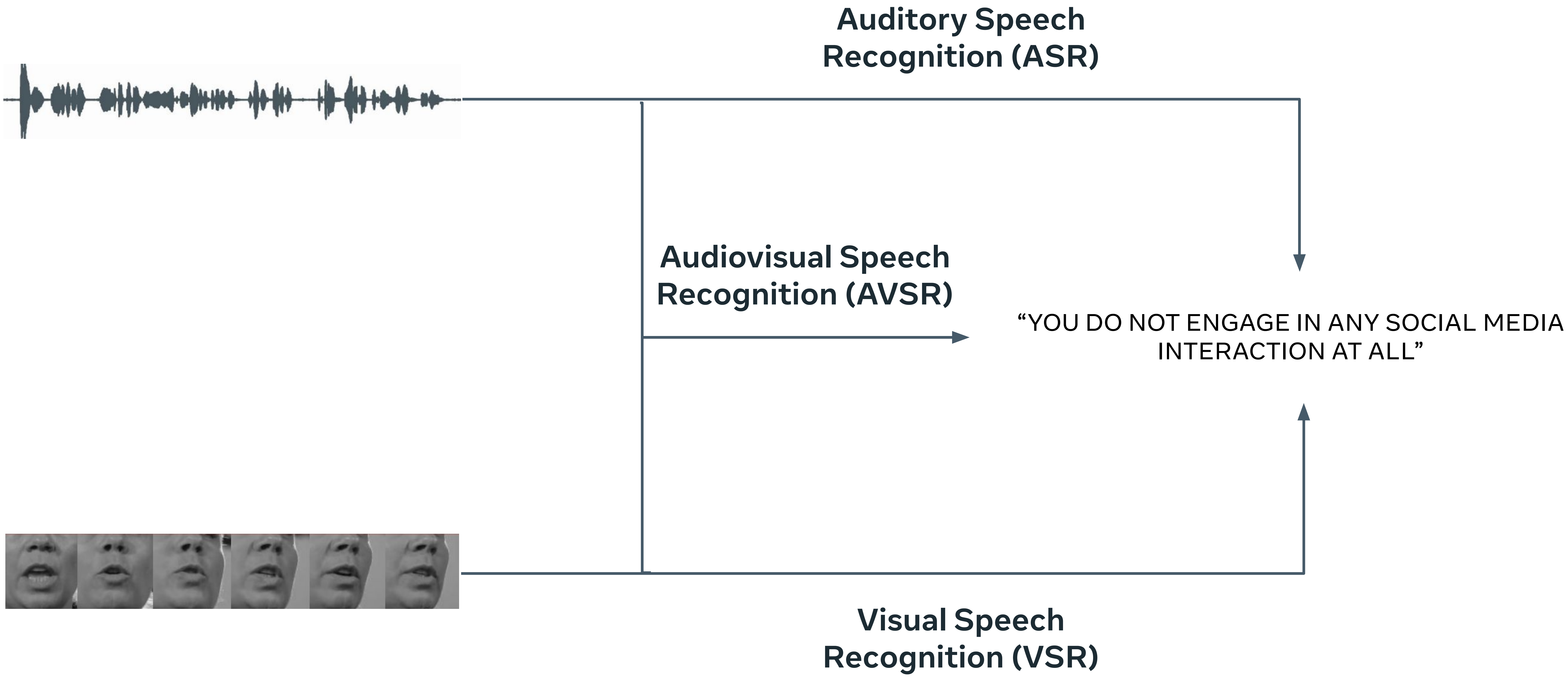
Alexandros Haliassos, Rodrigo Mira, Honglie Chen, Zoe Landgraf, Stavros Petridis, Maja Pantic

Imperial College London

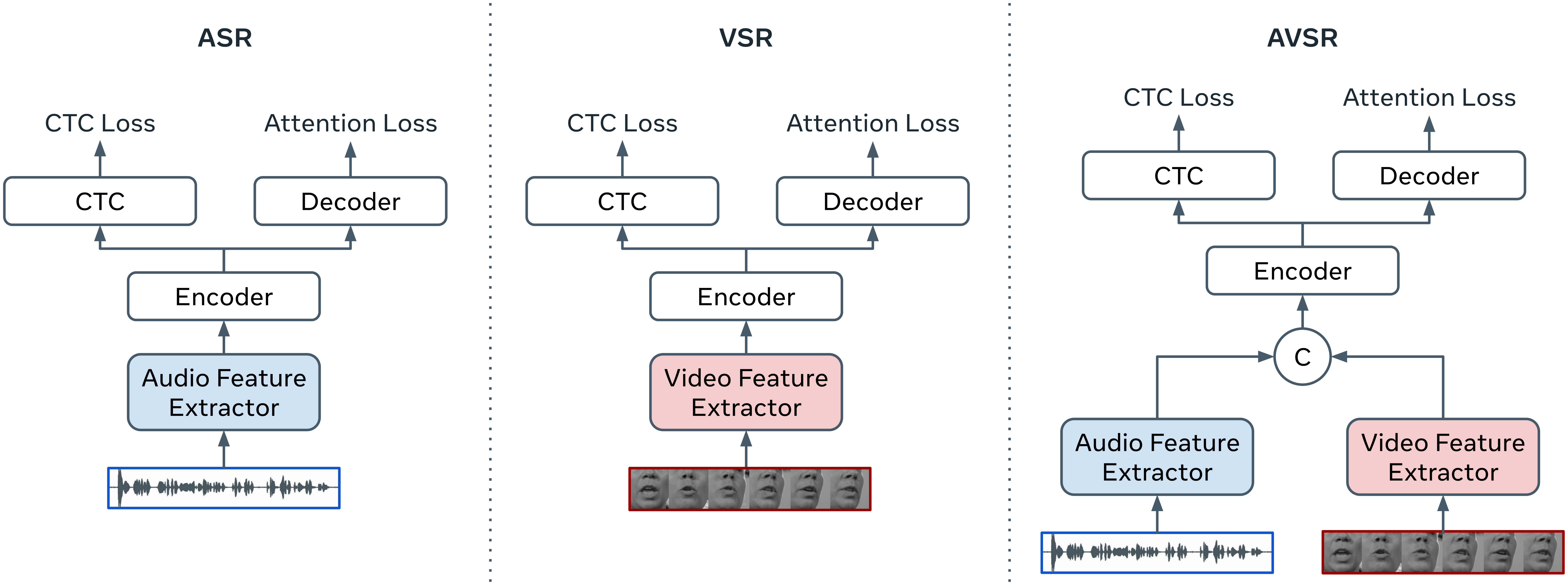
Meta

NEURAL INFORMATION PROCESSING SYSTEMS

Continuous speech recognition



Separate Model per Task

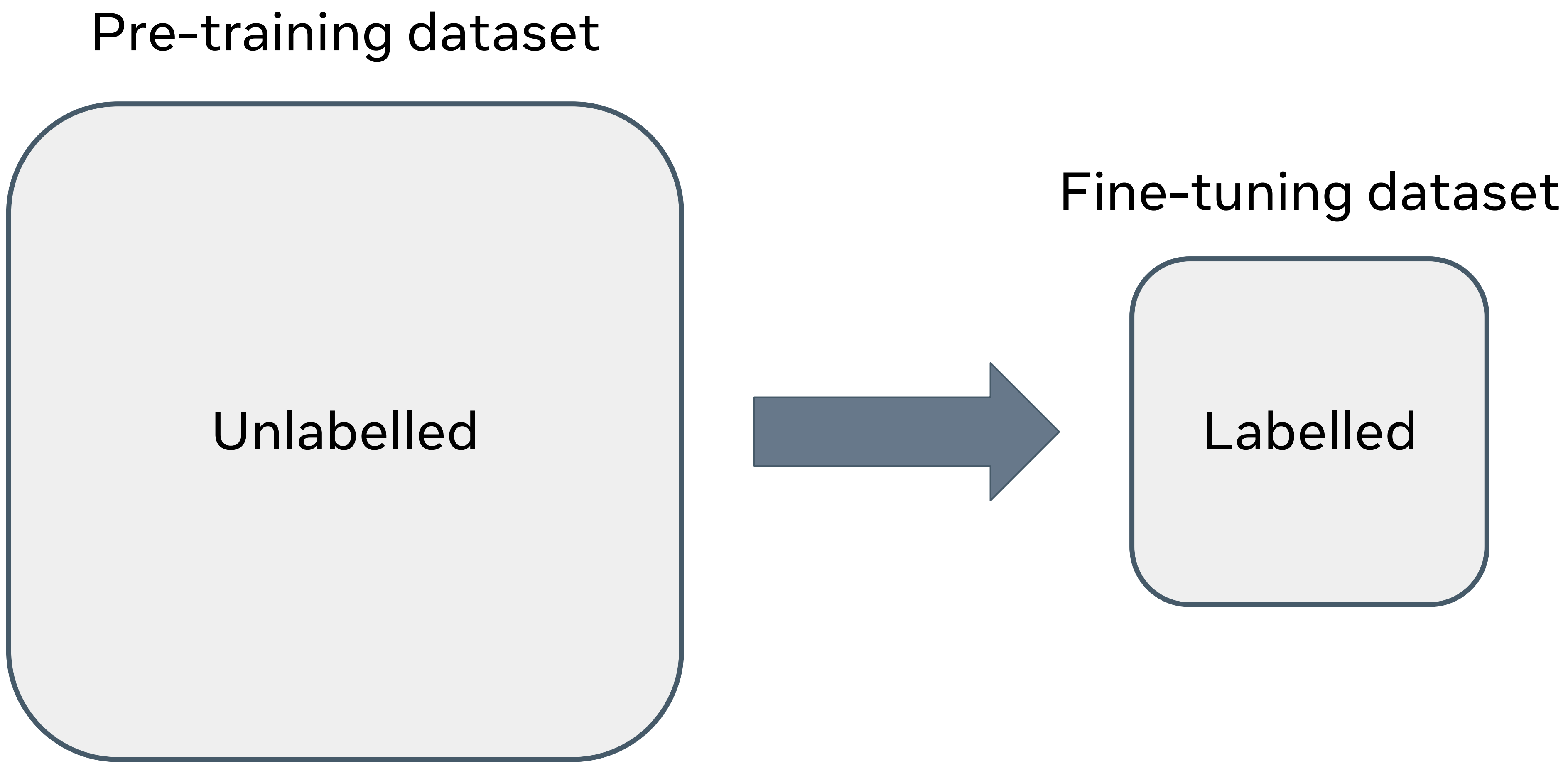


Self-Supervised Learning

- Leverage unlabelled data by training the encoder on a pretext task
- Popular pretext tasks use ideas from cross-modal learning and masked prediction
- Pre-trained network is then fine-tuned on a typically smaller labelled dataset

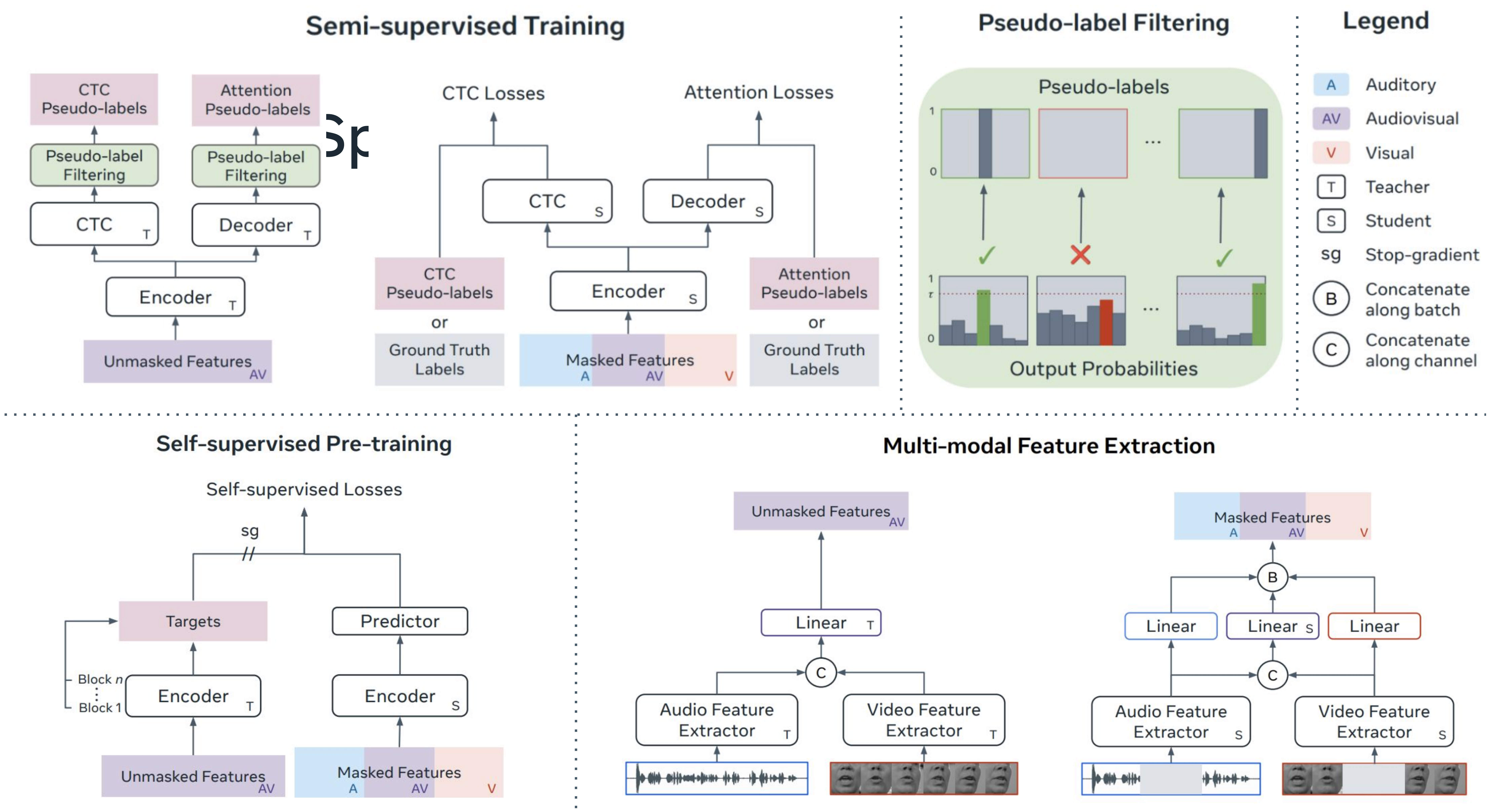
SSL Methods

AV-HuBERT, RAVEn, BRAVEn, AV-data2vec,...



Problems with Self-Supervised Learning

- Limited alignment between pretext and fine-tuning tasks
- Catastrophic forgetting of encoder, necessitating various training tricks
- Decoder is randomly initialised, prone to overfitting to small labelled dataset


Semi-supervised Training
CTC Pseudo-labels
Attention Pseudo-labels
Pseudo-label Filtering
Pseudo-label Filtering
CTC T
Decoder T
Encoder T
Unmasked Features AV
CTC Losses
Attention Losses
CTC S
Decoder S
CTC Pseudo-labels
or
Ground Truth Labels
Encoder S
Attention Pseudo-labels
or
Ground Truth Labels
Masked Features A AV V
Pseudo-label Filtering
Pseudo-labels
1
0
...
1
τ
0
Output Probabilities
Legend
A Auditory
AV Audiovisual
V Visual
T Teacher
S Student
sg Stop-gradient
B Concatenate along batch
C Concatenate along channel
Self-supervised Pre-training
Self-supervised Losses
sg
Targets
Predictor
Block n
Block 1
Encoder T
Encoder S
Unmasked Features AV
Masked Features A AV V
Multi-modal Feature Extraction
Unmasked Features AV
Linear T
C
Audio Feature Extractor T
Video Feature Extractor T
Masked Features A AV V
B
Linear
Linear S
Linear
C
Audio Feature Extractor S
Video Feature Extractor S

Unified Semi-Supervised Training: Main Properties

Setting: Semi-supervised training with 30-hour LRS3 subset as labelled data and full 433-hour LRS3 as unlabelled data (low-resource setting)

Confidence threshold

τ	WER (%)		
	V	A	AV
0.0	40.7	4.9	4.7
0.8	**37.8**	**4.0**	**3.9**
1.0	61.8	8.9	8.4

A threshold between 0 and 1 is important for performance

Relative labelled weights

γ_a	γ_v	WER (%)		
		V	A	AV
0.5	0.5	42.3	4.1	4.0
0.2	0.2	38.0	4.2	4.1
0.5	0.2	**37.8**	**4.0**	**3.9**

VSR benefits from abundance, ASR from quality

CTC vs. CTC-attention

Loss type	WER (%)		
	V	A	AV
CTC	45.6	5.2	5.0
CTC-att	**37.8**	**4.0**	**3.9**

A CTC-attention hybrid framework improves performance

Unified Self-Supervised Training: Main Properties

Setting: Self-supervised pre-training + semi-supervised fine-tuning with low-resource setting

Target type

Target	WER (%)		
	V	A	AV
Scratch	37.8	4.0	3.9
V	36.2	3.7	3.4
A	37.3	**3.2**	3.1
AV	**36.0**	**3.2**	**3.0**

Pre-training with AV targets yields best performance

Averaging blocks

Target	WER (%)		
	V	A	AV
Last block	37.2	3.4	3.1
Avg blocks	**36.0**	**3.2**	**3.0**

Using the average of encoder blocks as targets outperforms using only the last block

Predictor depth

Depth	WER (%)		
	V	A	AV
1	37.0	3.2	3.0
2	**36.0**	3.2	3.0
4	36.9	**3.1**	**2.9**

A predictor depth of 2 works best

Impact of Semi- and Self-supervised Training

LRS3 low-resource setting

Setting	Self-supervised pre-training	Fine-tuning	WER (%)		
			V	A	AV
Only labelled data	✗	Supervised	61.8	8.9	8.4
Self-supervised	✓	Supervised	43.9	4.8	4.6
Semi-supervised	✗	Semi-supervised	37.8	4.0	3.9
Self- + semi-supervised	✓	Semi-supervised	**36.0**	**3.2**	**3.0**

Comparisons with Self-Supervised Methods

- State-of-the-art results on LRS3 low-resource (30h labelled data) and high-resource (433h) settings
- Increasing data/model size improves results
- Results achieved with a *single model* for all tasks

Method	Pre-train data	Shared params	WER (%) LR			WER (%) HR		
			V	A	AV	V	A	AV
Base(+) models								
AV-HuBERT [13]	LRS3	✗	51.8	4.9	4.7	44.0	3.0	2.8
VATLM [14]	LRS3	✗	48.0	-	3.6	-	-	-
RAVEn [17]	LRS3	✗	47.0	4.7	-	39.1	2.2	-
AV-data2vec [15]	LRS3	✗	45.2	4.4	4.2	39.0	2.0	1.8
Lip2Vec [20]	LRS3	✗	49.5	-	-	42.0	-	-
BRAVEn [18]	LRS3	✗	43.4	4.0	4.0	36.0	**1.9**	-
USR	LRS3	✓	**36.0**	**3.2**	**3.0**	**34.3**	**1.9**	**1.6**
Base(+) models								
AV-HuBERT [13]	LRS3+Vox2	✗	46.1	4.6	4.0	34.8	2.0	1.8
VATLM [14]	LRS3+Vox2	✗	42.6	-	3.4	34.2	-	1.7
RAVEn [17]	LRS3+Vox2	✗	40.2	3.8	-	33.1	1.9	-
AV-data2vec [15]	LRS3+Vox2	✗	37.8	3.7	3.3	32.9	1.7	1.4
Lip2Vec [20]	LRS3+Vox2	✗	40.6	-	-	34.1	-	-
BRAVEn [18]	LRS3+Vox2	✗	35.1	3.0	-	28.8	**1.4**	-
USR	LRS3+Vox2	✓	**28.4**	**2.6**	**2.5**	**26.5**	1.6	**1.3**
Large models								
AV-HuBERT [13]	LRS3+Vox2	✗	32.5	2.9	3.3	28.6	1.3	1.4
VATLM [14]	LRS3+Vox2	✗	31.6	-	2.7	28.4	-	1.2
RAVEn [17]	LRS3+Vox2	✗	32.5	2.7	-	28.2	1.4	-
AV-data2vec [15]	LRS3+Vox2	✗	30.8	2.7	2.7	28.5	1.3	1.3
Lip2Vec [20]	LRS3+Vox2	✗	31.2	-	-	26.0	-	-
BRAVEn [18]	LRS3+Vox2	✗	30.8	**2.3**	-	26.6	**1.2**	-
u-HuBERT [16]	LRS3+Vox2	✓	-	-	-	29.1	1.5	1.3
USR	LRS3+Vox2	✓	**26.9**	2.4	**2.4**	**22.3**	**1.2**	**1.1**

Comparisons with the State-of-the-Art on LRS3

Method	Labelled hours	Unlabelled hours	Language model	Shared params	WER (%)		
					V	A	AV
Supervised*							
V2P [50]	3,886	-	✗	✗	55.1	-	-
RNN-T [38]	31,000	-	✗	✓	33.6	4.8	4.5
VTP [51]	2,676	-	✓	✗	30.7	-	-
Auto-AVSR [27]	1,902	-	✓	✗	23.5	**1.0**	1.0
Auto-AVSR [27]	3,448	-	✓	✗	19.1	**1.0**	**0.9**
ViT3D-CM [52]	90,000	-	✗	✗	17.0	-	1.6
SynthVSR [53]	6,720	-	✓	✗	16.9	-	-
LP Conf [54]	100,000	-	✗	✗	**12.8**	-	**0.9**
Self/semi-supervised							
AV-HuBERT w/ ST [13]	433	1,326	✗	✗	28.6	-	-
RAVEn w/ ST [17]	433	1,326	✓	✗	23.1	1.4	-
USR	433	1,326	✓	✓	**21.5**	**1.2**	**1.1**

- USR surpasses multiple methods which use significantly more labelled data
- USR outperforms self-supervised methods that use self-training strategy

Conclusion / Future Work

- Proposed a single model for VSR, ASR, and AVSR tasks
- Combined self-supervised learning with a semi-supervised method to achieve state-of-the-art performance
- Future work: improve pseudo-label quality and incorporate extra audio-only data

Code: https://github.com/ahaliassos/usr

Self-Supervised Works (AV-HuBERT, AV-data2vec,...)