Multi-Stage Multi-Modal Pre-Training For Automatic Speech Recognition

Yash Jain[1], David Chan[2], Pranav Dheram[3], Aparna Khare[3], Olabanji Shonibare[3], Venkatesh Ravichandran[3], Shalini Ghosh[3]

[1] 

[2] 

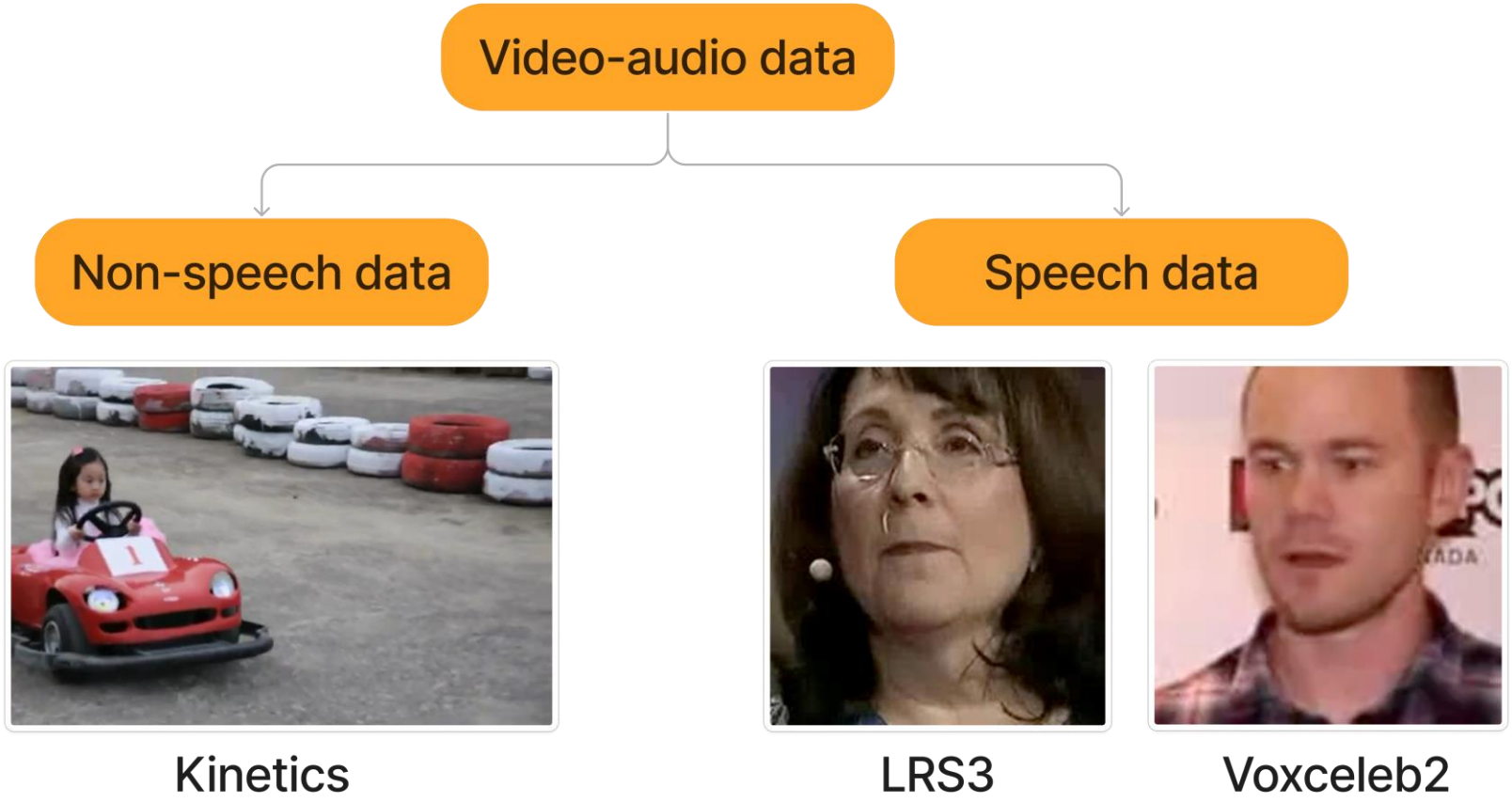
[3] amazon | science

LREC-COLING 2024

Multi-Stage Multi-Modal Pre-Training For Automatic Speech Recognition

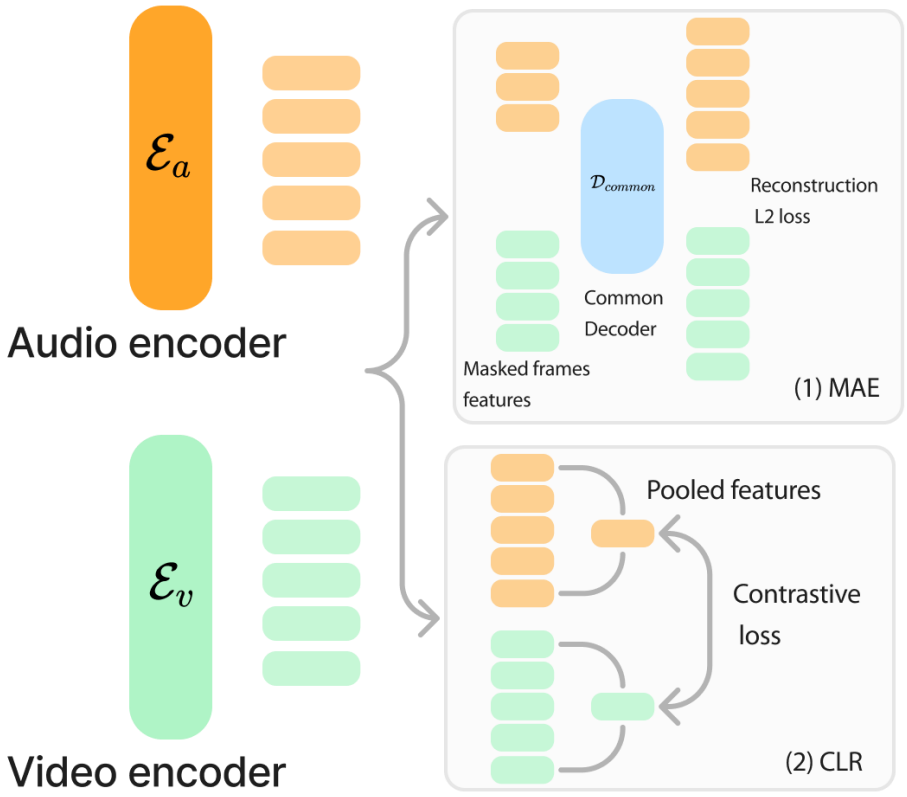
- Large scale evaluation of audio-visual pre-training methods with varying characteristics.
- Evaluate them on Speech recognition task and SUPERB benchmark.
- Introduce a novel mid-training stage between pre-training and fine-tuning for better feature alignment.

Multi-Stage Multi-Modal Pre-Training For Automatic Speech Recognition



Multi-Stage Multi-Modal Pre-Training For Automatic Speech Recognition

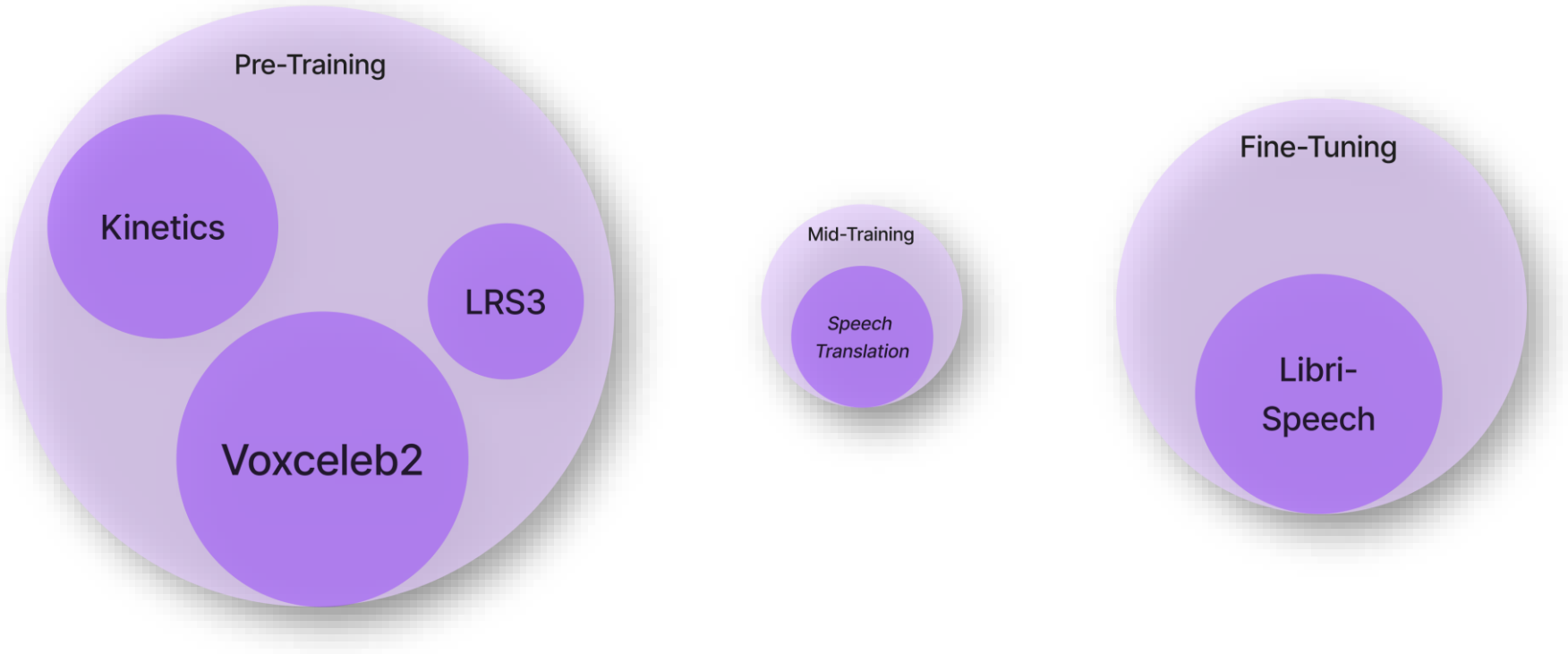
1. Masked Autoencoder (MAE)
2. Contrastive Learning (CLR)
3. MAE + CLR



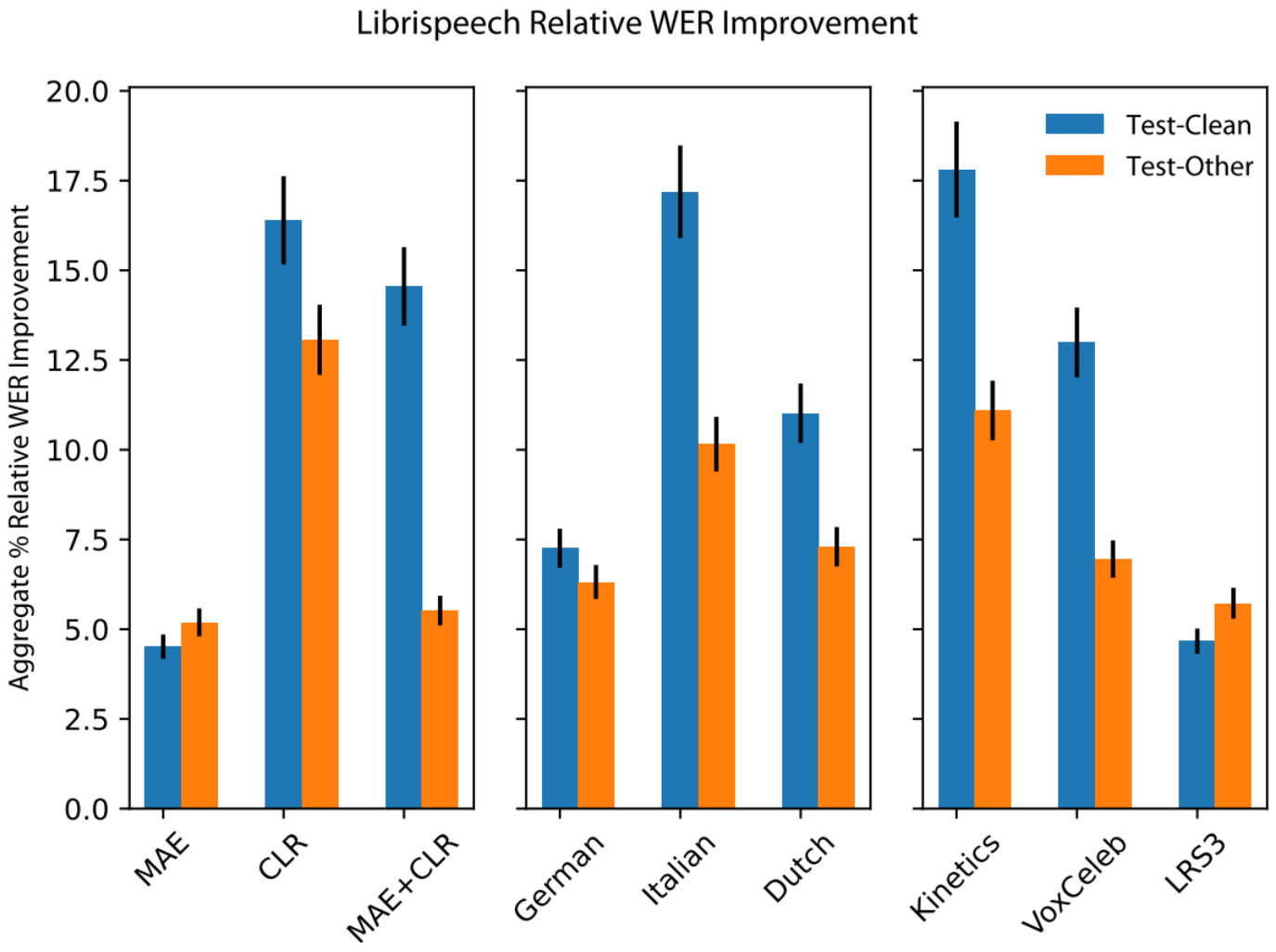
Multi-Stage Multi-Modal Pre-Training For Automatic Speech Recognition

- Automatic Speech Recognition
  - Librispeech dataset
- SUPERB benchmark
  - Keyword spotting
  - Intent Classification
  - Phoneme Recognition
  - Speaker Diarization

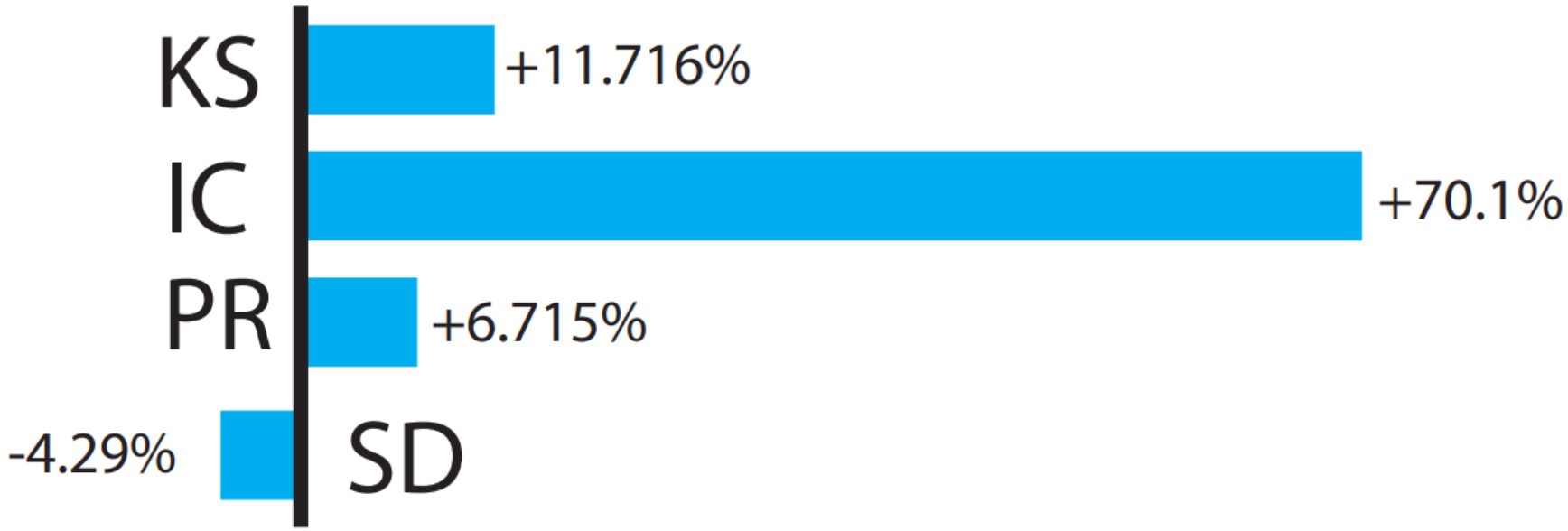
Multi-Stage Multi-Modal Pre-Training For Automatic Speech Recognition



Results: Automatic Speech Recognition



Results: SUPERB benchmark



More Results in the paper!

Method	PT	MT	en-de ↓		en-it ↓		en-nl ↓	
			Test-clean	Test-other	Test-clean	Test-other	Test-clean	Test-other
No Pre-training	None	-	6.84 ± 0.22	12.91 ± 0.47	6.84	12.91	6.84	12.91
MAE	K600	-	7.54	13.88	7.54	13.88	7.54	13.88
	K600	✓	5.69	11.34	5.95	12.55	5.73	12.04
	VC2	-	5.28	11.51	5.28	11.51	5.28	11.51
	VC2	✓	5.11	11.12	5.56	10.42	5.64	12.46
	LRS3	-	4.73	10.27	4.73	10.27	4.73	10.27
	LRS3	✓	5.61	10.85	**4.21**	**9.53**	5.32	10.33
CLR	K600	-	6.85	12.92	6.85	12.92	6.85	12.92
	K600	✓	5.02	10.85	4.72	10.62	4.65	10.41
	VC2	-	6.47	12.42	6.47	12.42	6.47	12.42
	VC2	✓	6.43	12.31	5.1	10.61	4.62	10.77
	LRS3	-	6.35	12.12	6.35	12.12	6.35	12.12
	LRS3	✓	6.74	10.59	5.84	11.33	6.01	10.13
MAE + CLR	K600	-	5.56	11.91	5.56	11.91	5.56	11.91
	K600	✓	5.02	11.68	5.23	11.37	6.39	12.03
	VC2	-	6.75	12.11	6.75	12.11	6.75	12.11
	VC2	✓	5.36	11.22	4.77	10.84	5.03	10.73
	LRS3	-	7.51	12.54	7.51	12.54	7.51	12.54
	LRS3	✓	7.16	12.29	5.08	11.13	6.17	12.32

Thank you!

Meet us at the poster!