Yuxiao LIN, Tao JIN, Xize CHENG, Zhou ZHAO, Fei WU, 2025. Multi-talker audio–visual speech recognition towards diverse scenarios. *Frontiers of Information Technology & Electronic Engineering*, 26(11):2310-2323. https://doi.org/10.1631/FITEE.2500411

Multi-talker audio–visual speech recognition towards diverse scenarios

Key words: Speech recognition and synthesis; Multi-modal recognition; Curriculum learning; Multi-talker speech recognition

Corresponding author: Fei WU
E-mail: wufei@zju.edu.cn
ORCID: https://orcid.org/0000-0003-2139-8807

Motivation

Audio–visual speech recognition extends traditional automatic speech recognition (ASR) into a multi-modal framework by leveraging visual cues from lip movements. It enables artificial intelligence (AI) systems to perceive speech more similarly as humans, and naturally resolves the permutation ambiguity.

Real-world applications present more complex challenges that current research has yet to fully address. Recording conditions often involve an unknown number of speakers talking simultaneously and have modality misalignment.

Main idea

We propose a speaker-number-aware mixture-of-experts (SA-MoE) mechanism, which explicitly assigns different scenarios to different experts while using speaker counting as an auxiliary task for automatic expert assignment.

We propose a cross-modal realignment (CMR) module that is integrated with pre-trained models to automatically correct input sequence misalignment.

According to the challenges, the scenarios are naturally divided into groups of varying difficulty levels. We apply curriculum learning and propose challenge-based curriculum learning (CBCL) that reduces excessive focus on simpler cases while ensuring sufficient attention to harder scenarios.

SA-MoE encoder

The key components of the SA-MoE encoder include a router and multiple mixture-of-experts (MoE) layers. The router generates frame-wise probabilities for speaker number prediction. The MoE layers consist of a shared linear layer and *N* lightweight expert layers.

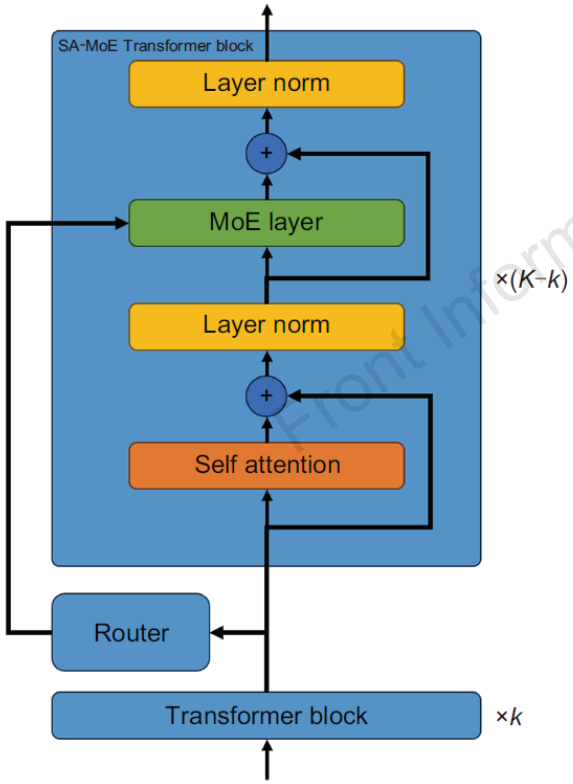


Fig. 2 Structure of SA-MoE

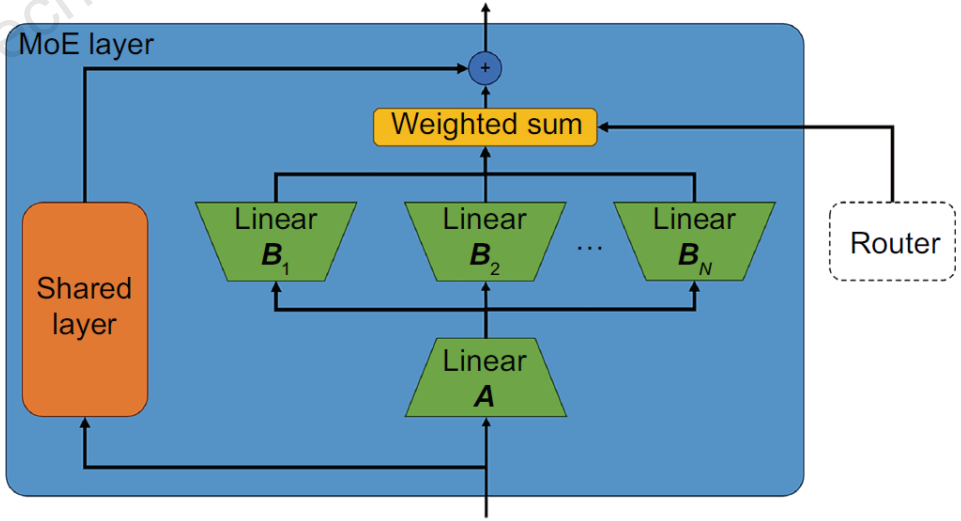


Fig. 3 MoE layer

Cross-modal realignment

Instead of directly concatenating the per-frame input audio features and video features, we use a CMR module to reconstruct an aligned video feature sequence before extracting multi-modal features for speech recognition.

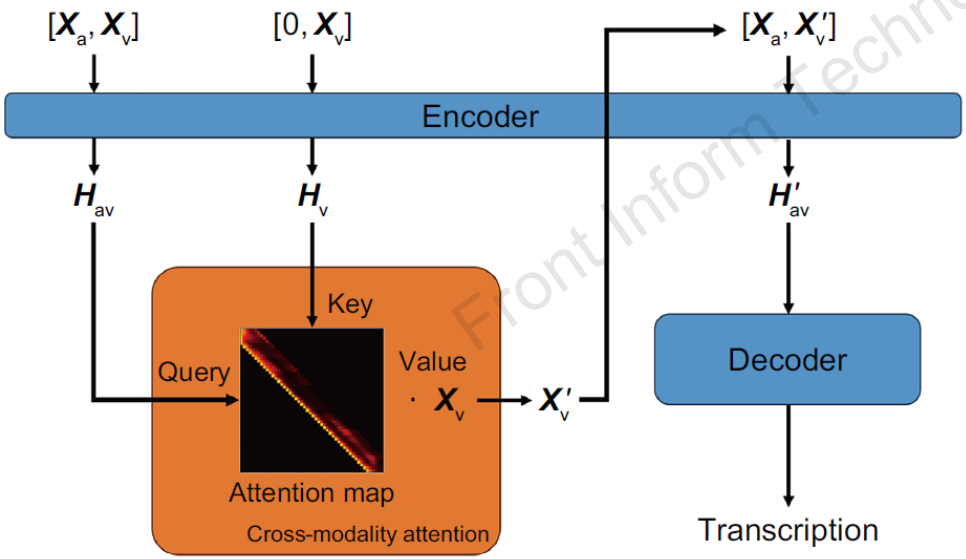


Fig. 4 Workflow of CMR

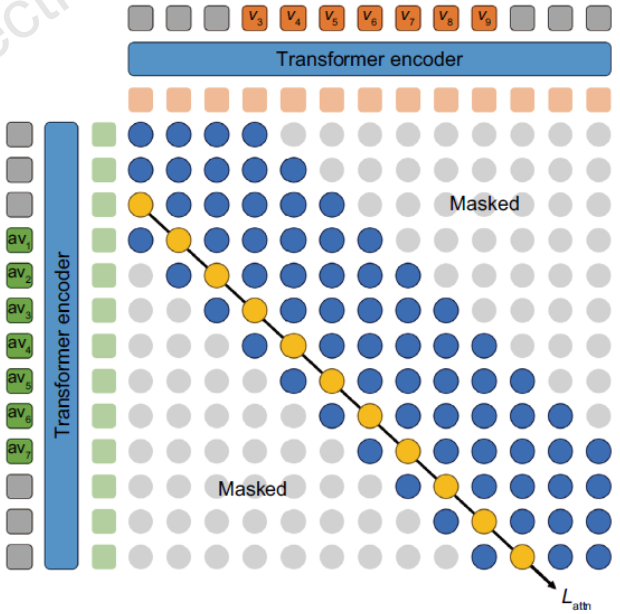


Fig. 5 An example of computing the cross-modal attention with $F = 3$ and $f = -2$. Gray squares denote dummy frames padded to the input sequence, and gray circles denote masked attention. The attention loss is computed along the yellow circles. References to color refer to the online version of this figure

Challenge-based curriculum learning

We propose a challenge-based curriculum learning (CBCL) strategy, which always focuses on the hardest data in the current subset. At each stage, there is a chance of 50% that a data point is sampled from the current highest difficulty.

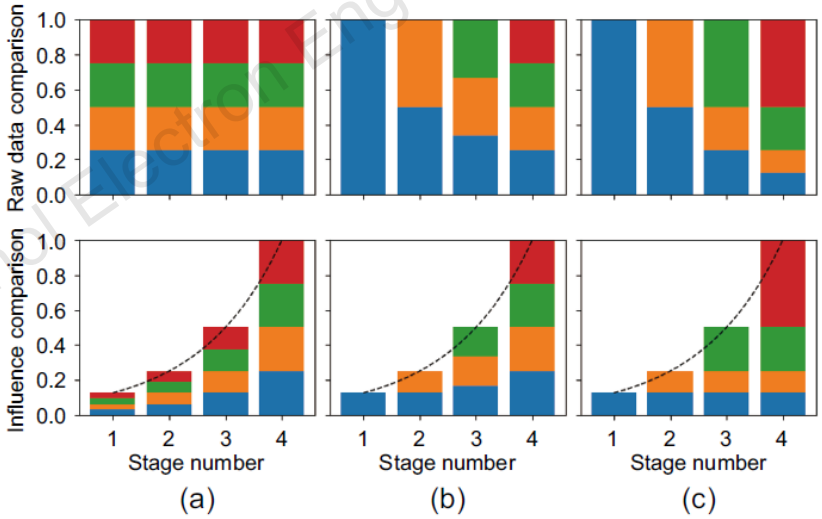


Fig. 6 Comparison of data ratio using different training strategies: (a) mix training; (b) standard CL; (c) CBCL. Different colors indicate data added at different stages of the curriculum. Figures on the top row show the ratio of raw data at each stage, and the bottom row illustrates the effect of the data on the model, assuming the effect decreases exponentially over time

Major results

Scenarios with different speaker numbers

Table 2 WER combining different structures and training strategies

Training strategy	SA-MoE	Performance					Overall
		N_{sp}=1	2	3	4	5	
Mix	×	2.15±0.04	4.36±0.03	**8.03±0.13**	13.54±0.07	17.81±0.36	9.18±0.18
	✓	**2.08±0.12**	4.37±0.04	8.46±0.06	**12.81±0.07**	17.48±0.25	9.04±0.13
Std. CL	×	2.19±0.06	4.44±0.07	8.22±0.01	14.04±0.09	17.93±0.15	9.36±0.09
	✓	2.19±0.12	4.74±0.06	8.30±0.05	13.36±0.04	17.70±0.20	9.26±0.11
CBCL	×	2.23±0.04	4.83±0.08	8.22±0.07	13.20±0.05	17.46±0.32	9.19±0.15
	✓	2.15±0.07	**4.27±0.09**	8.18±0.08	12.98±0.07	**16.68±0.40**	**8.85±0.18**

N_{sp} means the speaker number. Best results are in bold. The underline means the second-best result

Major results

Scenarios with temporal misalignment

Table 7 WER of different methods under different frame shift numbers

Method	Performance											Overall
	$f=-5$	-4	-3	-2	-1	0	1	2	3	4	5	
Unseen	76.83	69.90	52.47	22.33	6.65	**4.67**	5.89	16.17	46.02	65.86	73.48	40.02
Misaligned	7.01	6.44	5.74	5.17	4.98	5.02	5.04	**4.93**	5.16	5.92	6.16	5.60
GT-Aligned	**5.52**	5.48	5.19	5.15	**4.85**	4.84	5.04	5.02	**5.08**	5.32	**5.24**	5.16
VoiceFormer	6.86	6.27	5.88	5.55	5.18	5.19	5.15	5.50	5.61	5.96	6.08	5.75
CMR	5.63	**5.17**	**5.13**	**4.91**	4.94	4.76	**4.99**	5.08	5.20	**5.20**	5.30	**5.12**

Best results are in bold. The underline means the second-best result

Major results

Diverse multi-talker scenarios

Table 10 WER of models under different datasets

Method	Performance		
	FULL	SPEAKER	MISALIGNMENT
AV-HuBERT (Shi et al., 2022b)	12.37±0.07	11.51±0.18	6.05±0.09
VoiceFormer (Rahimi et al., 2022)	12.56±0.18	11.70±0.12	5.64±0.32
GILA (Hu YC et al., 2023)	15.15±0.12	13.87±0.07	7.10±0.23
Ours	**11.33±0.12**	**11.12±0.14**	**4.90±0.17**

Best results are in bold

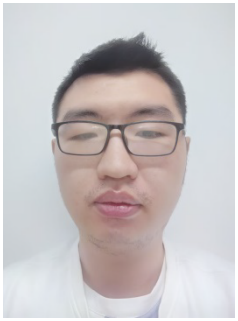
Conclusions

1. To explicitly model scenario differences across varying speaker counts, we design the SA-MoE encoder that dynamically assigns processing pathways according to the speaker number.

2. For handling temporal misalignment between audio–visual streams, we introduce a novel CMR module that automatically corrects input asynchrony.

3. Recognizing the inherent difficulty hierarchy across speaker counts, we propose the CBCL training strategy, where the proportion of simpler scenarios decreases exponentially during training.

4. All proposed methods are compatible with pre-trained audio–visual encoders and can use knowledge learned from unlabeled data.



Xiaoyu LIN is currently pursuing his PhD degree at Zhejiang University. His core research focuses on speech separation and recognition tasks, particularly in scenarios involving overlapping speech from multiple speakers. By fully leveraging information from different modalities, the research aims to assist in extracting the speech content of target speakers, achieving accurate separation and recognition.

Fei WU is a Qiusi Distinguished Professor at Zhejiang University. He serves as dean of the Undergraduate College of Zhejiang University and previously as director of the Artificial Intelligence Research Institute, Zhejiang University. He was a visiting scholar at the Department of Statistics, University of California, Berkeley (Oct. 2009–Aug. 2010). He was the recipient of the National Science Fund for Distinguished Young Scholars (2016), member of the Academic Degree Committee of the State Council for the discipline of Intelligent Science and Technology, head of the Expert Working Group on Artificial Intelligence Innovation and Technology under the Ministry of Education (Aug. 2018–Dec. 2020), expert for the Ministry of Science and Technology's 2030 Innovation and Technology Major Project Guide on "Next-Generation Artificial Intelligence," and principal investigator of Key R&D Program Projects under the Ministry of Science and Technology, principal investigator of two National Natural Science Foundation key projects. He serves as an executive editor of *Engineering*, an executive associate editor-in-chief of *Frontiers of Information Technology & Electronic Engineering*, and a fellow of the Chinese Association for Artificial Intelligence. His primary research areas encompass artificial intelligence, multimedia analysis and retrieval, and statistical learning theory.