

# GLAD: GLOBAL-LOCAL AWARE DYNAMIC MIXTURE-OF-EXPERTS FOR MULTI-TALKER ASR

*Yujie Guo, Jiaming Zhou, Yuhang Jia, Shiwan Zhao and Yong Qin\**

TMCC, College of Computer Science, Nankai University, Tianjin, China  
Email: guoyujie02@mail.nankai.edu.cn

## ABSTRACT

End-to-end multi-talker automatic speech recognition (MTASR) faces significant challenges in accurately transcribing overlapping speech, especially under high-overlap conditions. To address these challenges, we proposed **Global-Local Aware Dynamic Mixture-of-Experts (GLAD)**, which dynamically fuse speaker-aware global information and fine-grained local features to guide expert selection. This mechanism enables speaker-specific routing by leveraging both global context and local acoustic cues. Experiments on LibriSpeechMix show that GLAD outperforms existing MTASR approaches, particularly in challenging multi-talker scenarios. To our best knowledge, this is the first work to apply Mixture-of-Experts to end-to-end MTASR with a global-local fusion strategy. Our code and train dataset can be found at <https://github.com/NKU-HLT/GLAD>.

**Index Terms**— multi-talker ASR, mixture of experts, cocktail party problem, dynamic routing

## 1. INTRODUCTION

Multi-talker automatic speech recognition (MTASR) represents a critical challenge in speech processing, requiring models to simultaneously decode overlapping utterances from multiple speakers in complex acoustic environments. This capability is essential for practical applications including meeting transcription and multi-party dialogue analysis, where speakers frequently interrupt or talk simultaneously.

Current approaches to MTASR can be broadly categorized into two paradigms. Single-input multiple-output (SIMO) methods [1, 2, 3] explicitly separate mixed speech into speaker-specific branches before transcription, typically employing permutation invariant training (PIT) [4, 5], which incurs additional computational cost. To improve efficiency, heuristic error assignment training (HEAT) [6, 7] applies the Hungarian algorithm for optimal permutation assignment. However, SIMO methods assume a fixed number of speakers and require explicit separation, limiting their applicability in real-world scenarios with unknown or variable speaker counts.

Alternatively, single-input single-output (SISO) models [8, 9] adopt serialized output training (SOT) to implicitly handle speaker separation via attention mechanisms, producing unified transcription sequences with speaker-delimiter tokens. This paradigm offers greater flexibility than SIMO, as it does not require predefined speaker counts and can naturally handle varying numbers of speakers. Recent advances have enhanced SOT through integration with large language models [10], combining connectionist temporal classification [11, 12], and incorporating ideas from SIMO [13].

In parallel, the Mixture of Experts (MoE) paradigm [14, 15] enables conditional computation via specialized subnetworks and has shown strong performance in ASR [16, 17]. Recent work extends MoE to LoRA-based expert adaptation [18]. In audio-visual speech recognition, Speaker-number Aware MoE [19] has shown that MoE can transcribe the target speaker’s speech in multi-talker scenarios by leveraging the lip movements of the target speaker.

Despite recent advances, the application of MoE architectures to MTASR remains largely unexplored. MoE is particularly well-suited to this task, as it enables dynamic allocation of specialized experts to handle varying numbers of speakers and different degrees of overlap. In addition, speaker-aware modeling is important: effectively capturing who is speaking can help the model distinguish both the speaker identity and the spoken content. Crucially, raw speech signals inherently carry speaker-specific acoustic characteristics, which are often diluted in deeper network layers.

Motivated by this, we propose **GLAD (Global-Local Aware Dynamic Mixture-of-Experts)**, a novel framework designed to improve expert selection in MTASR. Specifically, we introduce a global router that processes raw speech features to capture speaker-aware context. In parallel, local routers model layer-specific information derived from intermediate encoder layers. A dynamic fusion module then adaptively combines global and local routing signals on a per-frame basis, guiding expert selection with both global speaker cues and local acoustic details. We further apply GLAD to SOT paradigm, resulting in GLAD-SOT. Our main contributions are as follows:

(1) To our best knowledge, we are the first to apply MoE architectures to end-to-end MTASR, showing significant improvements over conventional methods.

(2) We introduce GLAD, which dynamically combines speaker-aware global context with fine-grained local acoustic features. This enables experts to focus on target speakers while capturing fine-grained speech patterns in multi-talker scenarios.

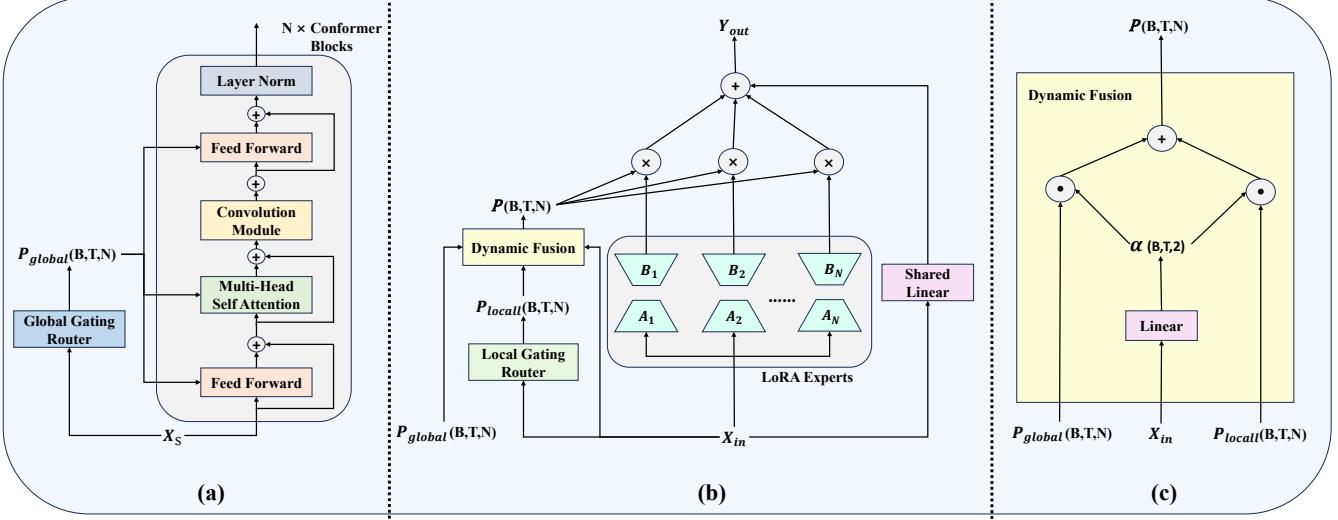
(3) Our ablation studies investigate the role of global acoustic features and their support for speaker-aware expert routing in MTASR.

## 2. OUR METHOD

### 2.1. Serialized Output Training for Multi-talker ASR

Serialized Output Training (SOT) [8] is an efficient and scalable method for MTASR. Unlike conventional approaches that require multiple output branches, SOT uses a single decoder to handle a variable number of speakers. For multi-talker utterances, the SOT model is trained to predict a token sequence that concatenates the transcriptions of all speakers, such as ‘text1 <sc> text2 ...’, where <sc> denotes a speaker change. In our experiments, we define ‘\$’ as the

\*Corresponding author



**Fig. 1:** Overview of the GLAD-SOT architecture, which applies the proposed GLAD to SOT. (a) GLAD-SOT leverages original speech features to generate global expert weights for encoder layers. (b) Each MoLE layer combines global and local weights to coordinate LoRA experts. (c) The global-local aware dynamic fusion module adaptively fuses global and local weights to guide expert collaboration.

speaker-change token, and the transcriptions follow the chronological order of speakers. This serialization strategy enables the model to implicitly separate speakers, while avoiding the combinatorial complexity of permutation-based methods.

## 2.2. Mixture of LoRA Experts Integration in SOT

The Mixture of Experts (MoE) paradigm has proven effective in scaling neural networks by distributing computation across specialized modules. Building upon this concept, the Mixture of LoRA Experts (MoLE) framework substitutes traditional dense expert layers with parameter-efficient low-rank adaptation (LoRA) [20] modules, offering improved scalability and domain adaptability.

In our approach, We integrate MoLE with SOT by replacing all linear layers in the Transformer encoder with MoLE modules. Each layer contains  $N$  LoRA-based experts operating in parallel with a shared linear transformation. Given an input  $X \in \mathbb{R}^{d_{in}}$ , the MoLE layer output is formulated as:

$$Y_{out} = W_L X_{in} + \frac{\alpha}{r} \sum_{i=1}^N P_i B_i A_i X + b, \quad (1)$$

where  $W_L$  and  $b$  denote the shared linear layer weight and bias. Each LoRA is parameterized by low-rank matrices  $A_i \in \mathbb{R}^{r \times d_{in}}$  and  $B_i \in \mathbb{R}^{d_{out} \times r}$  with rank  $r \ll \min(d_{in}, d_{out})$ , scaled by a factor  $\alpha$ . The expert weight  $P_i$  is computed via a gating mechanism combining global and local routing (see Sections 2.3 and 2.4).

## 2.3. Gating Router Design

In MTASR, speaker-related information is critical for distinguishing overlapping speech. As the raw speech signal contains rich speaker characteristics, we propose a global router that operates on original speech  $X_S \in \mathbb{R}^{T \times d_h}$ , where  $T$  is the number of frames and  $d_h$  is the feature dimension. Unlike local routers that rely on intermediate representations, the global router directly uses the original speech signal to compute expert weights, enabling speaker-aware routing based on global context and guiding experts to focus on speaker-specific content across time.

Specifically, the global router computes the global experts distribution  $P_{global} \in \mathbb{R}^{T \times N}$  using a linear transformation:

$$P_{global} = \text{softmax}(X_S W_{global}), \quad (2)$$

where  $W_{global} \in \mathbb{R}^{d_h \times N}$  is a trainable weight matrix and  $N$  denotes the number of experts. This enables the global router to compute speaker-aware expert weights from the original speech features, providing informed guidance based on global contextual information.

The global router computes speaker-aware expert probabilities  $P_{global}$  directly from the raw speech features  $X_S$ . In parallel, the local router computes fine-grained expert probabilities  $P_{local} \in \mathbb{R}^{T \times N}$  from the intermediate layer input  $X_{in} \in \mathbb{R}^{T \times d_h}$ :

$$P_{local} = \text{softmax}(X_{in} W_{local}), \quad (3)$$

where  $W_{local} \in \mathbb{R}^{d_h \times N}$  is the trainable weight matrix of the local router.

## 2.4. Global-Local Aware Dynamic Information Fusion

While the global router captures speaker-aware context from raw speech features to guide expert selection, the local router provides fine-grained information critical for accurate speech modeling. To harness their complementary strengths, we introduce a global-local aware Mixture-of-Experts that adaptively balances global and local routing signals at each frame. This enables expert activation to be guided by both global speaker cues and local acoustic details.

To integrate these complementary signals, we use a dynamic fusion module that generates per-frame fusion weights based on  $X_{in}$ :

$$\alpha = \text{softmax}(X_{in} W_{fusion}), \quad (4)$$

where  $W_{fusion} \in \mathbb{R}^{d_h \times 2}$  is a trainable fusion matrix, and  $\alpha \in \mathbb{R}^{T \times 2}$  provides per-frame fusion weights, with  $\alpha_0$  and  $\alpha_1$  as contributions of global and local expert distributions, respectively. The final expert probabilities are then computed as follows:

$$P_i = \alpha_0 \odot P_{global,i} + \alpha_1 \odot P_{local,i}, \quad (5)$$

where  $\odot$  denotes element-wise multiplication, and  $P_i$  represents the final probability for the  $i$ -th expert. This dynamic fusion allows each expert to adaptively leverage global speaker cues and local acoustic details for more effective routing.

### 3. EXPERIMENTAL SETUP

#### 3.1. Dataset

We conduct our experiments using the LibriSpeechMix (LSM) dataset [8], a standard benchmark for MTASR derived from LibriSpeech [21] by simulating overlapping speech. LSM includes both two-speaker (LSM-2mix) and three-speaker (LSM-3mix) mixtures.

As the original LSM provides only development and test sets, we construct our training set following the same protocol as prior work [8, 13, 11]. Two-speaker mixtures are generated by randomly pairing utterances from the 960-hour LibriSpeech training set with random temporal offsets. To enable a unified model for both single- and multi-talker scenarios, we combine a randomly sampled subset of these two-speaker mixtures with a subset of original single-speaker utterances, resulting in a final training set of approximately 1.35k hours<sup>1</sup>.

Following previous studies [13, 11], we categorize our evaluation scenarios into three overlap levels—low, medium, and high—to systematically analyze model performance across varying degrees of speaker overlap. The overlap ratio is defined as the proportion of overlapping speech duration relative to the total duration of the mixed utterance. Specifically, we define low overlap as (0, 0.2], medium overlap as (0.2, 0.5], and high overlap as (0.5, 1.0]. This provides a structured framework to assess multi-talker ASR robustness under different overlap conditions. The details of our training data is shown as table 1.

**Table 1:** Training Dataset Composition

	1mix	2mix			Total
		Low	Mid	High	
Utt.	202472	39413	59983	45423	347291
Dur.(hrs)	692.1	181.5	275.5	202.5	1351.6

#### 3.2. Model settings

We build our models on the Conformer [22] architecture using ES-Pnet2 [23]. Both SOT and GLAD-SOT adopt a Conformer encoder with 12 blocks and a Transformer decoder comprising 6 blocks. Each block contains 4-head self-attention with 256 hidden units. The encoder uses macaron-style Conformer blocks with two 1024-dimensional feed-forward layers, while the decoder blocks contain two 2048-dimensional feed-forward layers.

In GLAD-SOT, all encoder linear layers are replaced with MoLE containing 3 experts, with LoRA rank and scaling factor  $\alpha$  set to 8.

To control for model size, we also train an extended SOT model, denoted SOT-14, which comprises 14 Conformer encoder blocks and 6 Transformer decoder blocks. With 36.01M parameters, SOT-14 closely matches the size of GLAD-SOT (35.18M).

#### 3.3. Training settings and Metrics

In our experiments, all models are trained for 50 epochs. The final models are obtained by averaging the top 10 checkpoints selected

based on performance on the dev-clean 2mix set. We use the Adam optimizer with a learning rate of 5e-4 and apply 25,000 warm-up steps. All experiments are conducted on 8 NVIDIA GeForce RTX 3090 GPUs.

For single-talker evaluation, we use the standard Word Error Rate (WER). For multi-talker scenarios, we adopt Permutation-Invariant WER [8], a common metric for SOT that resolves speaker permutation ambiguity. Since the test set has an uneven distribution of overlap levels, Permutation-Invariant WER may not accurately reflect performance in challenging overlapping conditions. Therefore, following [13, 11], we also use Overlap-Aware WER (OA-WER), which provides a more balanced evaluation by averaging WERs across different overlap levels.

### 4. RESULTS AND DISCUSSIONS

Table 2 shows the results of GLAD method and other methods. All methods are trained on single-talker and two-talker data. We use SOT+SACTC results in [11] and CSE-SOT results in [13].

#### 4.1. Performance of GLAD-SOT

Compared to the baselines (S1 and S2), our method S5 (GLAD-SOT) achieves consistent improvements in training scenarios (LibriSpeech and LSM-2mix) and generalization scenarios (LSM-3mix). In particular, S5 outperforms S2 despite having a smaller model size, indicating that the performance improvements are attributed to the GLAD architecture rather than simply scaling up the model.

Against stronger baselines such as CSE-SOT (S3) and SACTC (S4), GLAD-SOT achieves competitive or superior results. On LSM-2mix, S5 attains the best overall performance, while on LSM-3mix, it matches S4 in overall WER. More importantly, in challenging high-overlap conditions (mid/high cases in both LSM-2mix and LSM-3mix), S5 consistently achieves state-of-the-art performance. As mid and high overlap situations are the most common in real-world scenarios, the results demonstrate the practical value of our approach in MTASR field.

These findings align with our design motivation. In mid-overlap and high-overlap conditions, speaker information is often entangled across frames, making global context essential for correct speaker identification. The GLAD architecture dynamically integrates global speaker-aware routing and local fine-grained information, enhancing the model’s ability to determine both “who is speaking” and “what is being said.” In contrast, in low-overlap conditions where speech streams are largely separable, local routing alone is often sufficient, and global routing may introduce redundancy or noise, explaining minor performance drops. As low-overlap scenarios are relatively easy and well-handled by existing methods, GLAD is particularly effective for challenging multi-talker ASR cases.

#### 4.2. Ablations study and analysis

**Global-Local Aware Dynamic Information Fusion:** We compare S5, S6, and S7 to study the impact of global and local routing information. In S6, expert weights are set as the sum of global and local weights ( $P = P_{local} + P_{global}$ ). S7 uses only local routing, while S5 employs our proposed GLAD mechanism.

Results show that S7 performs well on simpler tasks (LSM-2mix-low and LSM-3mix-low), where local fine-grained information is sufficient. However, in more challenging cases (LSM-2mix-mid/high and LSM-3mix-mid/high), S5 significantly outperforms S7, demonstrating the benefit of incorporating global information.

<sup>1</sup><https://github.com/NKU-HLT/GLAD>

**Table 2:** WER(%) Results on Librispeech, LibrispeechMix-2mix and LibrispeechMix-3mix. All models are trained on single-talker and two-talker data. In the table, the best results are highlighted in bold with an underline, and the second-best results are indicated with an underline.

System	Method	Parm.(M)	Librispeech		LibrispeechMix-2mix						LibrispeechMix-3mix					
			Dev	Test	Overall		Test(Conditional)				Dev	Test	Test(Conditional)			
					Dev	Test	low	mid	high	OA-WER			low	mid	high	OA-WER
S1	SOT	32.90	4.1	4.6	10.0	10.7	10.4	10.2	12.7	11.1	26.6	26.7	26.1	25.7	29.7	27.2
S2	SOT-14	36.01	4.0	<u>4.2</u>	8.3	8.4	7.3	8.4	11.5	9.1	23.6	24.4	21.8	22.8	30.4	25.0
S3	CSE-SOT [13]	45.24	4.5	5.5	8.1	8.4	<u>7.2</u>	8.3	12.0	9.2	24.2	24.5	<u>18.1</u>	24.1	31.8	24.7
S4	SOT+SACTC [11]	34.18	<u>3.9</u>	<b>4.1</b>	8.2	<b>8.0</b>	<b>6.0</b>	8.4	12.8	9.1	<b>22.6</b>	<b>22.6</b>	<b>15.9</b>	<u>22.7</u>	29.1	<b>22.6</b>
S5	GLAD-SOT (G-S)	35.18	<u>3.9</u>	4.3	<b>7.5</b>	<b>8.0</b>	7.8	<b>7.5</b>	<u>10.1</u>	<b>8.5</b>	22.7	<u>23.0</u>	23.8	<b>21.5</b>	<b>25.5</b>	23.6
S6	G-S w/o dynamic fusion	35.09	<b>3.8</b>	4.2	8.1	8.8	9.2	<u>7.6</u>	<b>10.3</b>	9.0	24.0	25.6	26.9	23.6	<b>28.6</b>	26.4
S7	G-S w/o global router	35.09	4.0	4.6	<u>7.6</u>	<u>8.3</u>	7.7	8.0	10.7	<u>8.8</u>	23.4	24.1	23.0	<u>22.7</u>	<b>28.6</b>	24.8
S8	G-S w/o GLAD in att	33.55	<b>3.8</b>	<b>4.1</b>	8.4	<u>8.3</u>	7.5	8.0	11.1	8.9	24.8	24.8	23.4	23.5	29.2	25.4
S9	G-S w/o GLAD in ffd	34.52	4.9	5.8	12.6	12.8	12.3	11.8	16.7	13.6	30.5	30.8	29.6	29.0	35.9	31.5

This improvement stems from global routing’s ability to capture speaker-aware information from original speech signal and directs experts to focus on the appropriate speaker.

The fixed combination approach (S6) shows inferior performance, especially in LSM-3mix-low, due to its inability to adaptively balance global and local contributions. In contrast, GLAD (S5) dynamically adjusts weighting based on speaker-aware global information and fine-grained local information, achieving strong performance across varying conditions.

**The position where GLAD is used:** We compare S5, S8, and S9 to investigate the impact of GLAD placement within the model. S8 applies GLAD mechanism in feed forward modules. S5 applies it in attention modules, and S5 applies GLAD both in feed forward modules and attention modules.

Results show that S8 significantly outperforms S9, while S5 achieves the best overall performance. This can be attributed to the functional differences between feed-forward and attention layers. Feed-forward layers primarily handle feature transformation and non-linear mapping, making them more suitable for dynamic expert selection guided by speaker-aware global routing. In contrast, attention layers focus on capturing temporal dependencies and benefit from more stable representations. These observations align with the use of MoE in large language models, where experts are applied to feed-forward layers. The superior performance of S5 indicates that GLAD in feed-forward provides the main improvement, while GLAD in attention offers complementary fine-grained adjustments that further enhance modeling capability.

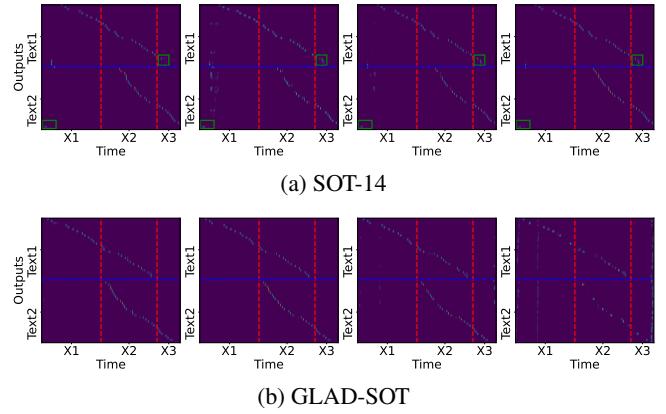
### 4.3. Case Study

To better understand the behavior of GLAD-SOT, we visualize the cross-attention matrices from the last decoder block on a representative sample (LSM-test-clean-2mix-1014). Figure 2 compares 4 attention heads for SOT-14 and GLAD-SOT. The cross-attention matrices indicate which parts of the input each output token should attend to, revealing the speech segments the model relies on to generate that token. This visualization illustrates the differences in feature modeling between SOT-14 and GLAD-SOT.

In Figure 2 (a). SOT-14 attention frequently crosses speaker boundaries. Specifically, certain tokens associated with Speaker 1 exhibit high attention weights toward temporal segments where only Speaker 2 is active, while tokens of Speaker 2 similarly attend to Speaker 1 segments (particularly in the region highlighted by the

green solid-line box), indicating limited awareness of speaker segmentation and difficulty handling overlapping speech.

In contrast, figure 2 (b) shows that GLAD-SOT exhibits more consistent, well-localized attention aligned with the corresponding speaker segments. This indicates that GLAD enables each output token to effectively listen to the relevant portion of the input, improving the model’s ability to capture speaker-specific information. By leveraging both global speaker-aware and local fine-grained cues, GLAD enhances the modeling of speaker-switching patterns, leading to more accurate transcription in multi-talker scenarios.



**Fig. 2:** Cross-attention matrices from the last decoder block. Text1/2 denote transcriptions of Speaker 1/2. X1, X2 and X3 indicate segments for Speaker 1 only, overlapping speech, and Speaker 2 only, respectively.

## 5. CONCLUSIONS

In this paper, we proposed GLAD, a novel approach that leverages Mixture-of-LoRA-Experts to address the multi-talker ASR task by dynamically integrating global and local information. Extensive experiments on LibriSpeechMix demonstrate that our GLAD mechanism significantly outperforms existing MTASR methods, especially in challenging multi-talker scenarios, demonstrating its potential for real-world applications.

## 6. REFERENCES

- [1] Dong Yu, Xuankai Chang, and Yanmin Qian, “Recognizing multi-talker speech with permutation invariant training,” *arXiv preprint arXiv:1704.01985*, 2017.
- [2] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe, “End-to-end multi-speaker speech recognition with transformer,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6134–6138.
- [3] Shane Settle, Jonathan Le Roux, Takaaki Hori, Shinji Watanabe, and John R Hershey, “End-to-end multi-speaker speech recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4819–4823.
- [4] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.
- [5] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.
- [6] Anshuman Tripathi, Han Lu, and Hasim Sak, “End-to-end multi-talker overlapping speech recognition,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6129–6133.
- [7] Liang Lu, Naoyuki Kanda, Jinyu Li, and Yifan Gong, “Streaming end-to-end multi-talker speech recognition,” *IEEE Signal Processing Letters*, vol. 28, pp. 803–807, 2021.
- [8] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, “Serialized output training for end-to-end overlapped speech recognition,” *arXiv preprint arXiv:2003.12687*, 2020.
- [9] Zhiyun Fan, Linhao Dong, Jun Zhang, Lu Lu, and Zejun Ma, “Sa-sot: Speaker-aware serialized output training for multi-talker asr,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 9986–9990.
- [10] Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng, “Large language model can transcribe speech in multi-talker scenarios with versatile instructions,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [11] Jiawen Kang, Lingwei Meng, Mingyu Cui, Yuejiao Wang, Xixin Wu, Xunying Liu, and Helen Meng, “Disentangling speakers in multi-talker speech recognition with speaker-aware ctc,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [12] Asahi Sakuma, Hiroaki Sato, Ryuga Sugano, Tadashi Kumanou, Yoshihiko Kawai, and Tetsuji Ogawa, “Speaker-distinguishable ctc: Learning speaker distinction using ctc for multi-talker speech recognition,” *arXiv preprint arXiv:2506.07515*, 2025.
- [13] Jiawen Kang, Lingwei Meng, Mingyu Cui, Haohan Guo, Xixin Wu, Xunying Liu, and Helen Meng, “Cross-speaker encoding network for multi-talker speech recognition,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11986–11990.
- [14] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” *arXiv preprint arXiv:1701.06538*, 2017.
- [15] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng, “Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications,” *CoRR*, 2023.
- [16] He Wang, Xucheng Wan, Naijun Zheng, Kai Liu, Huan Zhou, Guojian Li, and Lei Xie, “Camel: Cross-attention enhanced mixture-of-experts and language bias for code-switching speech recognition,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [17] Fengrun Zhang, Wang Geng, Hukai Huang, Yahui Shan, Cheng Yi, and He Qu, “Boosting code-switching asr with mixture of experts enhanced speech-conditioned llm,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [18] Bingshen Mu, Kun Wei, Qijie Shao, Yong Xu, and Lei Xie, “Hdmole: Mixture of lora experts with hierarchical routing and dynamic thresholds for fine-tuning llm-based asr models,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [19] Yuxiao Lin, Tao Jin, Xize Cheng, Zhou Zhao, and Fei Wu, “Curriculum learning aided audio-visual speech recognition with arbitrary speaker number,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., “Lora: Low-rank adaptation of large language models.,” *ICLR*, vol. 1, no. 2, pp. 3, 2022.
- [21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [23] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, “ESPnet: End-to-end speech processing toolkit,” in *Proceedings of Interspeech*, 2018, pp. 2207–2211.