Large Language Model Can Transcribe Speech in Multi-Talker Scenarios with Versatile Instructions

Lingwei Meng, Shujie Hu*, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, Helen Meng*

The Chinese University of Hong Kong, Hong Kong SAR, China

***Abstract*—Recent advancements in large language models (LLMs) have revolutionized various domains, bringing significant progress and new opportunities. Despite progress in speech-related tasks, LLMs have not been sufficiently explored in multi-talker scenarios. In this work, we present a pioneering effort to investigate the capability of LLMs in transcribing speech in multi-talker environments, following versatile instructions related to multi-talker automatic speech recognition (ASR), target talker ASR, and ASR based on specific talker attributes such as sex, occurrence order, language, and keyword spoken. Our approach utilizes WavLM and Whisper encoder to extract multi-faceted speech representations that are sensitive to speaker characteristics and semantic context. These representations are then fed into an LLM fine-tuned using LoRA, enabling the capabilities for speech comprehension and transcription. Comprehensive experiments reveal the promising performance of our proposed system, MT-LLM, in cocktail party scenarios, highlighting the potential of LLM to handle speech-related tasks based on user instructions in such complex settings[1].**

***Index Terms*—large language model, cocktail party problem, multi-talker speech recognition, multi-modal.**

I. Introduction

Large language models (LLMs) have experienced rapid and significant advancements recently, achieving or even surpassing human-level proficiency in numerous natural language processing (NLP) tasks [1]–[3]. These advancements have sparked interest in exploring the capabilities of LLM in multi-modal perception [4], including speech [5]–[7], vision [1], [8], [9], and content generation [10]–[12]. Several studies have investigated speech-related LLMs, which typically involve a fine-tuned text LLM following speech-related instructions and pairing with auxiliary audio encoders [5]–[7]. The audio encoder extracts acoustic representations and adapts them to the input feature space of LLM, allowing LLM to perform various speech tasks such as automatic speech recognition (ASR), speech translation (ST), speaker verification (SV), and speech question answering (SQA), among others. However, despite the progress, the potential of speech LLMs in cocktail party scenarios—where multiple talkers speak simultaneously and overlapping occurs—has not yet been sufficiently exploited.

In recent year, various end-to-end approaches have garnered interest and been developed to tackle multi-talker ASR task, which involves simultaneously transcribing speech from multiple talkers. These studies are based on Permutation Invariant Training (PIT) [13]–[15], Heuristic Error Assignment Training (HEAT) [16], [17], or Serialized Output Training (SOT) [18]–[22] to match predictions with corresponding target labels for loss calculation. However, these approaches typically transcribe speech from all talkers indiscriminately and fail to associate transcriptions with specific talkers, unless an additional external [23], [24] or internal [25]–[27] model is employed to extract speaker information. Although several studies [28], [29] proposed handling multi-talker ASR in conjunction with other tasks within a single model, the addressed tasks remain constrained and lack the flexibility to address various user requirements specifying talker attributes such as *please transcribe the talker who said the word "strawberry"*.

Nevertheless, the rise of large language models illuminates new possibilities for tackling such problems with a unified model. In this work, we leverage the powerful comprehension and instruction-following capabilities of LLM to perform speech recognition based on various instructions in multi-talker scenarios. Specifically, we utilize Llama 2 [3] as our foundational LLM, coupled with the Whisper [30] encoder to extract semantic context, and WavLM [31] multi-layer features to capture acoustic information indicating speaker characteristics, referring to WavLLM [5] and SALMONN [6]. Corresponding adapters are designed to project audio embeddings into the LLM's input space. We denote the proposed model as MT-LLM (Multi-Talker LLM). Versatile instructions are used to prompt MT-LLM to perform tasks including (i) simultaneously transcribing the speech of multiple talkers into text, (ii) transcribing a target talker's speech given a reference audio clip, (iii) transcribing speech based on the talker's specific sex, (iv) transcribing the speech of a specified talker according to their occurrence order, (v) transcribing the speech of the talker where a given keyword appears, and (vi) transcribe the talker who speaks the specific language. The comprehensive experiments demonstrate that MT-LLM can effectively meet user's diverse requirements for transcribing multiple talkers based on instructions specifying talker attributes. Our major contributions are threefold:

- We propose a pioneering effort to explore instruction-based speech recognition in multi-talker scenarios, leveraging the powerful comprehension and generation capabilities of LLM;
- Beyond multi-talker ASR, MT-LLM can transcribe speech from specific talkers according to six versatile instructions, demonstrating promising performances;
- We reveal that speech LLMs can support a more natural and effective human-computer interaction paradigm in complex speech environments, with parameter-efficient training.

II. Methods

We propose empowering a text-based LLM to act as a versatile instruction follower for speech recognition in multi-talker speech scenarios. The proposed method consists of three major components: a large language model as the foundational model fine-tuned with low-rank adaptation (LoRA) [32], dual speech encoders with corresponding adapters, and training data construction. We denote the proposed model as MT-LLM in the subsequent sections.

*Equal Contribution

[1]The code, model, and samples are available at https://github.com/cuhealthybrains/MT-LLM

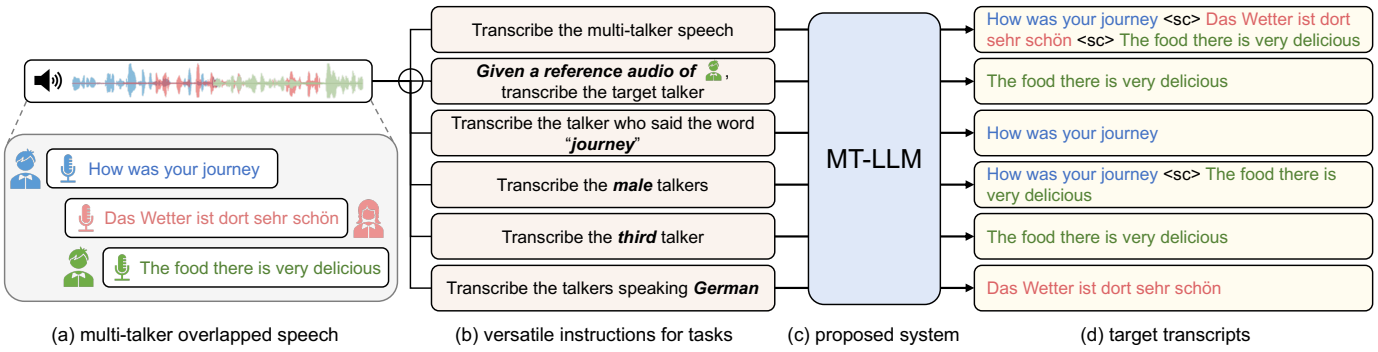

Fig. 1. MT-LLM supports versatile ASR instructions in multi-talker scenarios. Given a multi-talker overlapped speech input (a) and a text instruction prompt (b), the proposed MT-LLM (c) is expected to autoregressively generate corresponding target transcripts (d). For tasks that involve multiple talkers, MT-LLM follows the SOT-style output, transcribing the utterances of multiple talkers in the order of their start times, separated by "<sc>" indicating "speaker change".

A. Problem Formulation

This study regards the speech recognition in multi-talker scenarios as the next-token-prediction language modeling task. Conditioned on the input speech waveform $\boldsymbol{X}$ and text instruction $\boldsymbol{I}$, MT-LLM is optimized to autoregressively generate the target text output $\boldsymbol{Y} = [\boldsymbol{y}_0, \boldsymbol{y}_1, ..., \boldsymbol{y}_{N-1},]$, by maximizing the following distribution:

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{I}; \boldsymbol{\theta}) = \prod_{t=0}^{T-1} p(\boldsymbol{y}_t \mid \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{Y}_{<t}; \boldsymbol{\theta}) \quad (1)$$

where $\boldsymbol{\theta}$ represents the parameters of MT-LLM.

Among $\boldsymbol{Y}$, The transcripts for multiple talkers require a permutation assignment to determine the talker order, thereby addressing the label ambiguity issue [33]. Drawing on previous experiences [18], [20], we employed the straightforward Serialized Output Training (SOT) method to address this issue. As illustrated in Fig. 1 (d), SOT arranges the transcripts based on the speaking order of the talkers, with plain text "<sc>" inserted between them to signify speaker changes.

B. Model Architecture

As shown in Fig. 2, the speech representations synthesized by the dual speech encoders are fused and subsequently projected into the feature space of the backbone LLM. The LLM then leverages both the speech input and text instructions to predict the target transcripts. The backbone LLM and speech encoders remain frozen, while the system is fine-tuned using parameter-efficient adapters, equipping it with the capability for speech processing and comprehension.

Dual Speech Encoders and Corresponding Adapters Referring to [5], [6], we utilize two pre-trained speech encoders, Whisper encoder and WavLM, to capture multi-faceted speech information. Whisper [30] is a speech recognition and translation model trained on web-scale weakly supervised data, with its encoder being sensitive to speech semantic context [34]. We exploit its last-layer output embedding to capture rich semantic information. In contrast, WavLM [31] is a self-supervised learning model trained using a masked speech prediction approach, leading to different layers encoding acoustic information sensitive to various downstream speech tasks. To better leverage the multi-scale acoustic information extracted by WavLM, we aggregate its multi-layer hidden states by summing them with learnable weights for each layer. The modality adapters for the two speech encoders, along with a linear projector, are designed to better align speech representations with the LLM feature space.

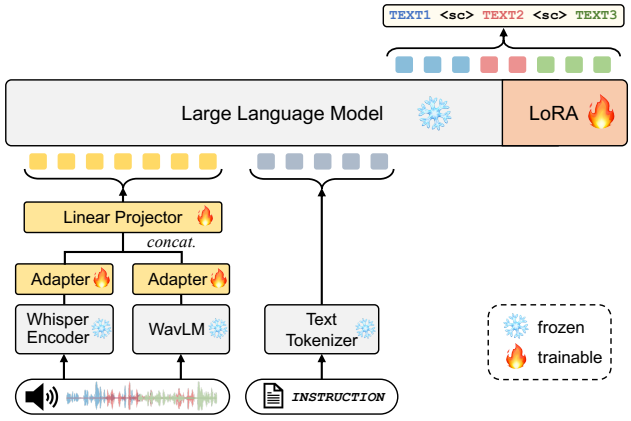

Fig. 2. Model architecture of MT-LLM. Multi-faceted speech representations are extracted using dual speech encoders and projected into the LLM feature space. Fine-tuned with LoRA, the LLM acquires the capability to comprehend and transcribe speech in multi-talker scenarios based on text instructions.

Backbone LLM and LoRA We exploit Llama-2-7b-chat[2], developed by Meta AI [3] on extensive and diverse training data, as our foundation backbone. Llama 2 excels across a broad spectrum of NLP tasks, showcasing superior capabilities in context understanding and text generation. To incorporate the speech modality into the LLM, we employ a parameter-efficient fine-tuning technique, LoRA [32]. LoRA is applied to the key, query, value, and output weight matrices within Llama's attention modules, enabling the model's capability to process and comprehend speech modality inputs.

C. Task Descriptions

To validate the performance of MT-LLM in executing speech recognition based on instructions in multi-talker scenarios, we simulated multi-talker overlapped speech audios from single-talker speech corpora, and designed six tasks along with corresponding instructions and generated the respective target text label. As illustrated in Fig. 1 (b), the text instructions pertain to the following tasks:

Multi-Talker (MT) ASR Simultaneously transcribing the speech of multiple talkers into text. This basic task challenges the model's ability to handle overlaps and distinguish between different talkers' voices within a single audio stream.

[2]https://llama.meta.com/llama2

Target-Talker (TT) ASR A random talker is selected as the target, and a 3-second audio clip of the target talker, along with 3 seconds of silence, is concatenated with the input multi-talker speech. The model is instructed to transcribe only the target talker's speech from the overlapping audio. This task tests the model's capability to differentiate and isolate the speech of a specified individual given a reference audio clip as the clue.

Keyword-Tracing (KT) ASR For each multi-talker sample, we first collect a set of unique words that appear only once across all talkers' speech content, each with a minimum length of six characters. From this set, we randomly select a keyword and instruct the model to transcribe the speech of the talker who said that word. This task assesses the model's proficiency in tracking specific lexical items and attributing them to the correct speaker.

Sex-Specific (SS) ASR We randomly instruct the model to transcribe all male or all female talkers from the multi-talker speech input. This task evaluates the model's ability to distinguish voices based on sex-related characteristics and filter the transcription accordingly.

Order-Specific (OS) ASR We randomly instruct the model to transcribe a talker based on their appearance order. This requires the model to keep track of the sequence of speakers and accurately extract the speech of a designated individual in the order they spoke.

Target-Lingual (TL) ASR We randomly instruct the model to transcribe the speech of talkers speaking either English or German. This task evaluates the model's ability to discriminate languages and transcribe spoken content based on the specified language.

III. Experimental Setup

A. Training Data Construction

We simulated multi-talker speech audios from single-talker speech corpora to accommodate the versatile task instructions, following the protocol outlined in [18]. Specifically, the utterances are simulated primarily from the 960-hour LibriSpeech [35] training set, comprising mixtures of two or three talkers. The start time for each talker is randomly sampled, resulting in overlapped mixtures. Additionally, the 180-hour German subset of CoVoST 2[3] [36], along with its corresponding German target text, is employed to mix with LibriSpeech utterances to support the Target-Lingual ASR task.

Combined with original single-talker corpora, the total training speech data amounts to $\sim$6.3K hours, with $\sim$10% containing German.

B. Evaluation and Metrics

We evaluate the performance of MT-LLM on versatile instruction-based tasks, using the official 2- and 3-speaker LibriSpeechMix test set and an additional home-made En-De-Mixed test set for Target-Lingual ASR task, ensuring that none of the speakers from evaluation sets are included in the training data. We also test single-talker ASR performance on LibriSpeech test-clean set and CoVoST 2 German (De) test set.

The word error rate (WER) is calculated between the predicted text and the target labels. For speech recognition tasks involving multiple talkers, permutations with minimum errors are used to compute WER, following previous studies [20], [37], [38].

C. Model Settings and Training details

The proposed MT-LLM employs the encoder of Whisper-large-v2[4] and WavLM-base[5] to extract speech representations, with Llama-2-chat-7b as the backbone LLM. All parameters of above models are frozen. The LLM is fine-tuned using LoRA with a rank of 32. The adapters for both speech encoders consist of two convolution layers to down-sample and align the representations within the temporal domain, followed by a bottleneck adapter [39] and a linear layer. The outputs of both adapters have a time stride of 80 ms and a dimension of 2,048. The total number of parameters in MT-LLM is 7.55 billion, of which 1% (76.6 million) are trainable.

MT-LLM is trained on 32 NVIDIA A100-40G GPUs with a batch size equivalent to 60 seconds per GPU for 150K updates. We optimize the model using AdamW optimizer, warming up the learning rate to a peak of 1e-4 over the first 10% updates, followed by a linear decay. All tasks are mixed together for training to develop a unified model.

TABLE I
Single-talker and multi-talker ASR performance on English and German test set. Evaluated by WER (%).

System	LibriSpeech	LibriSpeechMix		De	En-De-Mix	
	1-spk	2-spk	3-spk	1-spk	2-spk	3-spk
D2V-Sidecar-DB [28]	-	7.5	11.9	-	-	-
SOT-Conformer [25]	3.6	4.9	6.2	-	-	-
SALMONN-7B [6]	2.4	32.9	45.9	-	-	-
XLSR-Large-De [40]	-	-	-	12.1	-	-
MT-LLM (ours)	2.3	5.2	10.2	11.7	21.0	22.8

IV. Results and Discussion

A. Results on Multi-Talker ASR

We evaluate the single-talker ASR performance on the LibriSpeech and CoVoST 2 German test set, and multi-talker ASR performance on the 2- and 3-speaker LibriSpeechMix and the home-made En-De-Mixed test set. As shown in Table I, the proposed MT-LLM demonstrates promising results across all datasets, outperforming D2V-Sidecar-DB [28], a recent representative work employing PIT. SOT-Conformer [25] is a state-of-the-art model specifically designed and extensively trained for multi-talker ASR on the LibriSpeechMix dataset. On the 2-speaker LibriSpeechMix, MT-LLM achieves results comparable to SOT-Conformer, while lags behind on the 3-speaker set. Given that MT-LLM serves as an exploratory work aimed at exploring the potential of LLM for versatile instruction-based ASR in complex environments, rather than solely pushing the limits of multi-talker ASR, we argue that the gap is understandable. Despite being trained on both single-talker and multi-talker data, we note that SOT-Conformer falls short in single-talker scenarios, indicating its specialization in multi-talker ASR impairs its single-talker performance. In contrast, MT-LLM performs consistently good in single-talker ASR, demonstrating that its instruction-understanding capabilities help maintain high performance in simpler, single-talker settings. SALMONN [6] is a speech LLM trained on various speech-related tasks including overlapped speech recognition. However, using its official prompt, we observed significant inferior performance compared to our approach as shown in Table I.

We list wav2vec2-Large-XLSR-53-German model[6] performance on De for reference. Given that the audio samples containing German is relatively limited and noisier than LibriSpeech, MT-LLM still demonstrates satisfactory results on De and En-De-Mix test sets, showcasing the model's ability to simultaneously distinguish and transcribe multiple languages from multi-talker speech.

[3]Note that CoVoST 2 is originally designed for speech translation, while we use its German subset for ASR in our study.

[4]https://huggingface.co/openai/whisper-large-v2

[5]https://huggingface.co/microsoft/wavlm-base

[6]https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german

B. Results on Versatile Tasks Based on Instructions

We investigate the performance of MT-LLM on various ASR tasks described in Section II-C. The evaluations are conducted using the 2- and 3-speaker LibriSpeechMix and En-De-Mix test sets, with the results presented in Table II and Table III.

To validate MT-LLM's ability to accurately capture and transcribe a specified talker in multi-talker scenarios, we evaluate a *best-matching* result as a reference. Specifically, for each sample, we select the target text of the correct talker based on instructions specifying talker attributes (such as sex or keywords), and calculate WER against each ASR transcript of multiple talkers produced by MT-LLM in Section IV-A. The lowest WER obtained is reported as the best-matching result, eliminating the impact of speaker confusion. This result is compared with the model's performance when directly executing instruction-based ASR tasks, thereby reflecting MT-LLM's capability to follow instructions and correctly identify the intended talker.

The results in Table II and Table III highlight the effectiveness of the MT-LLM model across different instruction-based ASR tasks. In both the 2-speaker and 3-speaker mixed scenarios. MT-LLM achieves impressive performance for Target-Talker (TT), Keyword Tracing (KT), and Sex-Specific (SS) ASR tasks, which are within a reasonable range of the best-matching results. This suggests that MT-LLM is proficient at isolating and accurately transcribing speech from specified talkers based on instructions focusing on different speaker attributes. However, for Order-Specific (OS) ASR tasks, there is a gap compared to the best-matching result, indicating that MT-LLM slightly falls short in determining the order in which the talker appears. We anticipate significant improvements by using additional positional embedding for the speech part, employing larger speech modality adapters, or fully fine-tuning the speech encoder with heavier training as in [25]. Target-Lingual (TL) ASR task presents a more pronounced challenge for the model, due to that the mixed audios containing German are relatively limited and noisier compared to English set. We observe that performance of the TL task in both English and German is close, resulting in similar WER to best-matching situation. This indicates room for improvement through language-specific embeddings or higher-quality multi-lingual data.

For the more complex 3-speaker scenarios, MT-LLM consistently manages to handle the added intricacies of increased overlapping of multiple talkers, despite a rise in WER compared to the 2-speaker cases. This discrepancy illustrates the challenge posed by denser overlapping in speech, yet MT-LLM still demonstrates a competent ability to follow instructions and transcribe the target speaker accordingly.

C. Ablation Study

First, we conduct an ablation study to examine the impact of using dual speech encoders. We train two models specifically on the multi-talker (MT) ASR task: one using dual speech encoders and the other using only the Whisper encoder. As shown in Table IV, incorporating the WavLM encoder into the architecture leads to notable performance improvements, underscoring the importance of multi-layer acoustic information captured by WavLM in improving speech recognition accuracy, particularly in three-talker scenarios.

We also investigate the effect of multi-task training on improving the basic MT ASR task. As illustrated in Table IV, training on a variety of tasks not only endows the system with the versatility to handle diverse ASR tasks but also improves performance on the basic MT ASR task. This indicates that the different tasks are interdependent and complementary, helping to supervise the model and enhance its overall speech comprehension capabilities.

TABLE II
RESULTS FOR VERSATILE ASR TASKS ON 2-SPEAKER LIBRISPEECHMIX AND EN-DE-MIX TEST SETS. EVALUATED BY WER (%).

System	LibriSpeechMix 2-spk				En-De-Mix 2-spk
	TT	KT	SS	OS	TL
best-matching	5.5	4.8	4.9	5.7	21.0
MT-LLM	6.7	5.0	5.5	9.0	21.8

TABLE III
RESULTS FOR VERSATILE ASR TASKS ON 3-SPEAKER LIBRISPEECHMIX AND EN-DE-MIX TEST SETS. EVALUATED BY WER (%).

System	LibriSpeechMix 3-spk				En-De-Mix 3-spk
	TT	KT	SS	OS	TL
best-matching	12.0	9.7	12.4	11.7	24.0
MT-LLM	16.2	12.6	15.0	15.4	24.1

TABLE IV
ABLATION STUDY ON THE USAGE OF DUAL SPEECH ENCODERS AND MULTI-TASK TRAINING. EVALUATED BY WER (%).

With WavLM	Multi-Task Training	LibriSpeech	LibriSpeechMix	
		1-spk	2-spk	3-spk
✗	✗	4.3	6.6	16.1
✓	✗	2.4	5.5	10.7
✓	✓	2.3	5.2	10.2

D. Limitations and Future Work

Despite MT-LLM's promising performance on various ASR tasks in multi-talker scenarios, we acknowledge several limitations. First, the primary intention of this study is to explore the ability of LLMs to capture specific talkers according to instructions, and transcribe their speech in multi-talker scenarios. Therefore, MT-LLM is not designed to be a universal speech LLM for a broader spectrum of tasks. Second, the experiments are conducted on limited simulated datasets for insightful observation, rather than on real-world datasets, due to resource constraints. In the future, we anticipate the development of a more comprehensive speech LLM for cocktail party environments with meticulously crafted training data and schemes.

V. CONCLUSION

In this work, we present a pioneering exploration into the use of large language models (LLMs) for instruction-based speech recognition in multi-talker scenarios. We utilize the Whisper encoder to extract semantic context information and WavLM to capture multi-layer acoustic information indicating speaker characteristics, thereby enabling the foundational LLM to effectively handle speech modality input. With parameter-efficient fine-tuning, the proposed MT-LLM demonstrates remarkable capabilities in comprehending and transcribing speech based on versatile instructions related to multi-talker ASR, target talker ASR, and ASR based on specific talker attributes such as sex, occurrence order, language, and keyword spoken. Comprehensive experiments reveal promising performance in complex multi-talker environments, highlighting the potential of LLMs to enhance speech-related tasks and improve human-computer interaction in challenging settings.

VI. ACKNOWLEDGMENTS

This research is partially supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N) and by the CUHK Stanley Ho Big Data Decision Analytics Research Centre.

REFERENCES

[1] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, et al., "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[4] Liang Chen, Zekun Wang, Shuhuai Ren, et al., "Next token prediction towards multimodal intelligence: A comprehensive survey," *arXiv preprint arXiv:2412.18619*, 2024.

[5] Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei, "WavLLM: Towards robust and adaptive speech large language model," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Nov. 2024, pp. 4552–4572.

[6] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang, "SALMONN: Towards generic hearing abilities for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[7] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[8] Shaohan Huang, Li Dong, Wenhui Wang, et al., "Language is not all you need: Aligning perception with language models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei, "Grounding multimodal large language models to the world," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei, "Kosmos-g: Generating images in context with multimodal large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[11] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh, "Video generation models as world simulators," 2024.

[12] Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, et al., "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.

[13] Wangyou Zhang, Xuankai Chang, Yanmin Qian, and Shinji Watanabe, "Improving end-to-end single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1385–1394, 2020.

[14] 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)*, 2020.

[15] Lingwei Meng, Jiawen Kang, Mingyu Cui, Yuejiao Wang, Xixin Wu, and Helen Meng, "A Sidecar separator can convert a single-talker speech recognition system to a multi-talker one," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[16] Liang Lu, Naoyuki Kanda, Jinyu Li, and Yifan Gong, "Streaming multi-talker speech recognition with joint speaker identification," in *Interspeech 2021*, 2021, pp. 1782–1786.

[17] Desh Raj, Daniel Povey, and Sanjeev Khudanpur, "SURT 2.0: Advances in transducer-based multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3800–3813, 2023.

[18] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *Interspeech 2020*, 2020, pp. 2797–2801.

[19] Naoyuki Kanda, Jian Wu, Yu Wu, Xiong Xiao, Zhong Meng, Xiaofei Wang, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Takuya Yoshioka, "Streaming Multi-Talker ASR with Token-Level Serialized Output Training," in *Interspeech 2022*, 2022, pp. 3774–3778.

[20] Ying Shi, Lantian Li, et al., "Serialized output training by learned dominance," in *Interspeech 2024*, 2024, pp. 712–716.

[21] 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," *arXiv preprint arXiv:2409.12388*, 2024.

[22] Lin Zheng, Han Zhu, Sanli Tian, et al., "Unsupervised domain adaptation on end-to-end multi-talker overlapped speech recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 3119–3123, 2024.

[23] Zili Huang, Desh Raj, Paola García, and Sanjeev Khudanpur, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[24] Ryo Masumura, Naoki Makishima, Taiga Yamane, Yoshihiko Yamazaki, et al., "End-to-End Joint Target and Non-Target Speakers ASR," in *Interspeech 2023*, 2023, pp. 2903–2907.

[25] Naoyuki Kanda, Guoli Ye, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka, "End-to-End Speaker-Attributed ASR with Transformer," in *Interspeech 2021*, 2021, pp. 4413–4417.

[26] Yang Zhang, Krishna C. Puvvada, Vitaly Lavrukhin, and Boris Ginsburg, "Conformer-based target-speaker automatic speech recognition for single-channel audio," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[27] Ryo Masumura, Naoki Makishima, Tomohiro Tanaka, Mana Ihori, et al., "Unified multi-talker ASR with and without target-speaker enrollment," in *Interspeech 2024*, 2024, pp. 727–731.

[28] Lingwei Meng, Jiawen Kang, Mingyu Cui, Haibin Wu, Xixin Wu, and Helen Meng, "Unified modeling of multi-talker overlapped speech recognition and diarization with a Sidecar separator," in *Interspeech 2023*, 2023, pp. 3467–3471.

[29] Lingwei Meng, Jiawen Kang, Yuejiao Wang, Zengrui Jin, Xixin Wu, Xunying Liu, and Helen Meng, "Empowering Whisper as a joint multi-talker and target-talker speech recognition system," in *Interspeech 2024*, 2024, pp. 4653–4657.

[30] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, vol. 202, pp. 28492–28518.

[31] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[33] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[34] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass, "Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *Interspeech 2023*, 2023, pp. 2798–2802.

[35] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[36] Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino, "CoVoST 2 and massively multilingual speech translation," in *Interspeech 2021*, 2021, pp. 2247–2251.

[37] Jiawen Kang, Lingwei Meng, et al., "Cross-speaker encoding network for multi-talker speech recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[38] Samuele Cornell, Taejin Park, Steve Huang, Christoph Boeddeker, Xuankai Chang, Matthew Maciejewski, Matthew Wiesner, Paola Garcia, and Shinji Watanabe, "The CHiME-8 DASR challenge for generalizable and array agnostic distant automatic speech recognition and diarization," *arXiv preprint arXiv:2407.16447*, 2024.

[39] Neil Houlsby, Andrei Giurgiu, et al., "Parameter-efficient transfer learning for NLP," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[40] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.