Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Survey of end-to-end multi-speaker automatic speech recognition for monaural audio

Xinlu He, Jacob Whitehill *

Worcester Polytechnic Institute, 100 Institute Road, Worcester, 01609, MA, USA

ARTICLE INFO

Keywords:
Multi-speaker ASR
End-to-end ASR
Monaural audio
Speech overlap

ABSTRACT

Monaural multi-speaker automatic speech recognition (ASR) remains challenging due to data scarcity and the intrinsic difficulty of recognizing and attributing words to individual speakers, particularly in overlapping speech. Recent advances have driven the shift from cascade systems to end-to-end (E2E) architectures, which reduce error propagation and better exploit the synergy between speech content and speaker identity. Despite rapid progress in E2E multi-speaker ASR, the field lacks a comprehensive review of recent developments. This survey provides a systematic taxonomy of E2E neural approaches for multi-speaker ASR, highlighting recent advances and comparative analysis. Specifically, we analyze: (1) architectural paradigms (single-input-multiple-output (SIMO) vs. single-input-single-output (SISO)) for pre-segmented audio, analyzing their distinct characteristics and trade-offs; (2) recent architectural and algorithmic improvements based on these two paradigms, including multi-modal inputs; (3) extensions to long-form speech, including segmentation strategy and speaker-consistent hypothesis stitching. Further, we (4) evaluate and compare methods across standard benchmarks. We conclude with a discussion of open challenges and future research directions towards building robust and scalable multi-speaker ASR.

Contents

* Corresponding author.
E-mail address: jrwhitehill@wpi.edu (J. Whitehill).

https://doi.org/10.1016/j.csl.2025.101925
Received 16 June 2025; Received in revised form 4 December 2025; Accepted 10 December 2025

Available online 11 December 2025

1. Introduction

Multi-speaker Automatic Speech Recognition (ASR) aims to transcribe speech from audio containing multiple speakers whose speech may overlap. In contrast to single-speaker ASR, which focuses on *what was said*, multi-speaker ASR additionally determines *who says what* (Hershey et al., 2010; Settle et al., 2018; Watanabe et al., 2020). This task is closely related to the well-known *cocktail party problem* (Qian et al., 2018), where humans focus on one speaker in a noisy environment filled with competing talkers. Multi-speaker ASR extends this concept by transcribing all speakers in the mixture, and attributing words to individual speakers. This enables practical applications in real-world scenarios such as meetings, group discussions, and phone call recordings, and supports diverse downstream tasks such as meeting summaries, dialogue analytics, and intelligent conversational assistants.

Compared to single-speaker ASR, multi-speaker ASR poses unique challenges, primarily due to overlapping speech and the need for speaker distinction. These require advanced modeling and are further constrained by the scarcity of large, well-annotated datasets. Additionally, multi-speaker ASR is inherently multifaceted, involving not only recognition and diarization but also overlap detection, turn-taking detection, and target-speaker ASR. Although these tasks are interrelated and can benefit from joint modeling, effectively leveraging their synergy remains challenging.

While there are multiple comprehensive literature surveys for single-speaker ASR (Arora and Singh, 2012; Malik et al., 2021; Prabhavalkar et al., 2024), no recent survey has reviewed end-to-end multi-speaker ASR systems. Our paper seeks to fill this gap. Before exploring end-to-end solutions, we examine the limitations of traditional cascade architectures, which served as the initial attempts to address the multi-speaker ASR task.

1.1. The limits of cascade methods

Early multi-speaker ASR systems often adopted cascade (modular) methods (Kanda et al., 2021c; Yu et al., 2022a), as shown in Fig. 1. One approach is **diarization-segmented cascade system** (Fig. 1(a)): (1) Apply speaker diarization to determine *who spoke when* to obtain time-stamped speaker boundaries. (2) Split the audio into segments by detected speaker boundaries. (3) Use a single-speaker ASR model on each segment to obtain individual transcription. This approach leverages well-established single-speaker ASR and works well under minimal overlap. However, its accuracy degrades with overlapping speech. Moreover, traditional diarization (Anguera et al., 2012; Sell and Garcia-Romero, 2014; Díez et al., 2018; Landini et al., 2020) assumes a single active speaker per frame. Even with later diarization methods that support multi-speaker labeling per frame (Fujita et al., 2019; Li and Whitehill, 2021; Li et al., 2023a), the resulting segments may still contain overlapping speech, posing challenges for single-speaker ASR models.

Another **separation-based cascade system** (Fig. 1(b)) incorporates a speech separation module. (1) A separation model *enhances* and *denoises* the mixed audio into multiple single-speaker streams, addressing overlapping at the signal level. (2) Each stream is then transcribed using a single-speaker ASR model. While this method can separate overlapping speech in the initial stage, its overall accuracy is dependent on the separation models, which typically optimize signal-level objectives rather than directly targeting ASR performance. Consequently, errors introduced during the speech separation phase can propagate to the subsequent ASR process, compounding the overall error rate of the system (Kanda et al., 2021c). These limitations have motivated end-to-end (E2E) multi-speaker ASR approaches that directly map mixed audio to speaker-attributed transcriptions.

1.2. Preview of end-to-end methods

Unlike cascade methods that explicitly separate speakers before transcription, E2E systems model *jointly* optimize the complementary problems “who is speaking” and “what is being said”. This can improve accuracy on both tasks. In E2E frameworks, the input is the raw mixed audio, and the output is the transcriptions partitioned by different speakers.

Initial explorations on E2E multi-speaker ASR (Yu et al., 2017a; Seki et al., 2018; Settle et al., 2018; Chang et al., 2020) primarily adopted a **single-input multiple-output** (SIMO) design, where mixed speech is processed through parallel branches to extract

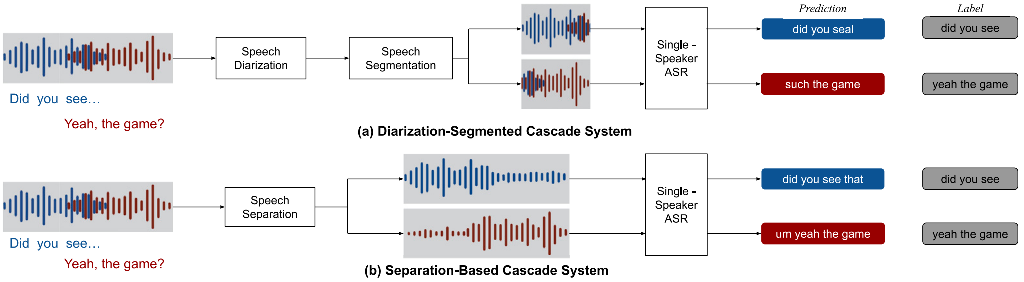

Fig. 1. Two types of cascade multi-speaker ASR systems. (a) Diarization-segmented cascade system: The mixed audio is segmented by speaker and then processed by a single-speaker ASR model, potentially introducing errors in overlapping speech regions. (b) Separation-based cascade system: The mixture is first separated into individual speaker streams using speech separation, followed by single-speaker ASR processing. This method may propagate errors from the separation stage.

speaker-specific representations. These methods typically follow a separation-then-recognition process and are trained end-to-end, often using *permutation invariant training (PIT)* (Yu et al., 2017b; Kolbæk et al., 2017). A key limitation is that they assume a fixed number of speakers. To address this limitation, **single-input single-output** (SISO) methods were proposed (Kanda et al., 2020a, 2021d; Liang et al., 2023; Shi et al., 2024a), notably using *serialized output training* (SOT) (Kanda et al., 2020b), which generates a unified sequence across speakers. Both SIMO and SISO have since been extended with various architectural and training improvements.

In light of recent advances, this review consolidates recent progress in end-to-end multi-speaker ASR by providing a structured taxonomy of representative models. We analyze core designs and improvements, and compare performance across benchmarks. Three key observations are summarized:

1. A key distinction in multi-speaker ASR research lies in whether to separate the speech mixture explicitly. Explicit separation generates multiple outputs (SIMO), offering clearer modularity and easier integration with separation and ASR. In contrast, direct mixture processing produces a single output (SISO), preserving contextual information across stages and enabling multi-task learning with mixture-based tasks.
2. There is a growing trend to adapt foundation speech models (Radford et al., 2023; Baevski et al., 2020; Hsu et al., 2021) to multi-speaker scenarios via lightweight structure modifications and fine-tuning, aiming to mitigate data scarcity.
3. Model comparison is hindered by limited open-source implementations and inconsistent settings. We provide setting-wise comparisons on standard datasets (AMI, LibriSpeechMix, LibriMix), and analyze models on their design focus.

1.3. Review scope and organization

In this paper, we review the recent progress on end-to-end multi-speaker ASR where multiple speakers may talk simultaneously. In particular, we focus on monaural (single-channel) audio recordings and offline (not real-time/streaming) speech analysis. To structure the comparison of recent models, we identify four key dimensions: **(1) Model architecture:** whether the system follows a SIMO or SISO framework. **(2) Multi-modal inputs**: whether the model incorporates additional modalities beyond audio; this has become an emerging design choice in recent systems. **(3) Input granularity:** whether the system is designed and evaluated on pre-segmented clips or long-form continuous audio. **(4) Speaker enrollment:** whether the system incorporates speaker enrollment to improve the system.

The rest of the paper is organized as follows:

- Section 2 reviews background techniques for multi-speaker ASR: end-to-end ASR and speech separation techniques.
- Section 3 categorizes and reviews recent advances in SISO and SIMO models for pre-segmented audio.
- Section 4 discusses methods that use extra modalities, such as visual information of the speakers' faces, or textual context to make ASR predictions.
- Section 5 focuses on long-form audio, discussing segmentation techniques and hypothesis stitching.
- Section 6 reviews datasets and metrics for multi-speaker ASR, and presents performance comparisons across different experimental settings.

2. Background techniques

This section reviews two background techniques for multi-speaker ASR. First, we discuss E2E ASR architectures, detailing prevalent model designs and objectives. Second, we outline the speech separation techniques that process mixed audio into isolated streams, which inspire architecture or serve as pre-processing modules for multi-speaker ASR.

2.1. End-to-end ASR

Recent advances in deep learning have enabled E2E network approaches to achieve state-of-the-art performance in ASR. These systems utilize sequence-to-sequence models to directly map speech signals to text outputs. The three primary end-to-end ASR architectures are: (1) Connectionist Temporal Classification (CTC) which relies solely on the previous input; (2) Recurrent Neural Network Transducer (RNN-T), which depends on both previous input and previous output; and (3) Attention-based Encoder-Decoder (AED) architectures, which consider all inputs and previous outputs.

Among these, AED has gained wide appeal through the development of state-of-the-art systems such as Whisper (Radford et al., 2023). In AED, the encoder processes the input acoustic features into a sequence of hidden states, and the decoder predicts output tokens conditioned on past outputs via an attention mechanism over encoder representations. Early AED implementations relied on RNNs for both the encoder and decoder, but recent models increasingly adopt Transformer (Vaswani et al., 2023) and Conformer (Gulati et al., 2020) architectures for their ability to model long-range dependencies with global attention. Recently, large speech foundation models (e.g., Whisper (Radford et al., 2023), Wav2Vec (Baevski et al., 2020), Hubert (Hsu et al., 2021)) leverage self-supervised learning or multi-task training on massive datasets. After fine-tuning for ASR, these models achieve state-of-the-art performance.

The training objective typically employs a cross-entropy loss to maximize the probability of transcription. Additionally, a joint CTC loss is often applied to the encoder outputs to enforce monotonic alignment, leveraging its dependence only on previous inputs. This also enhances the encoder's acoustic modeling. However, due to CTC's strict monotonic constraint, it struggles with overlapping speech, requiring careful adaptation in multi-speaker ASR.

2.2. Speech separation techniques

In end-to-end multi-speaker ASR, it sometimes includes a speech separation module and is trained together with the ASR part, inspired by the cascade model. Here we look back at the deep-learning-based speech separation techniques.

The deep-learning-based speech separation can be divided into frequency-domain methods and time-domain methods. The frequency-domain methods process the frequency feature of mixed speech and estimate time-frequency masks or spectral magnitudes for each speaker. The deep clustering framework (DPCL) (Hershey et al., 2016) first maps each time-frequency spectral magnitude into a speaker-discriminative embedding, and then the clustering algorithm is used to get the speaker label. Alternatively, a mask output for speech separation can be directly estimated by a deep neural network without embedding. Chimera (Wang et al., 2018) combines the two, outputting both speaker embedding and mask. The time domain methods, such as TasNet (Luo and Mesgarani, 2017), directly consume the waveforms using the encoder-separator-decoder framework. The encoder decomposes the mixture into learnable basis functions, the separator estimates speaker-specific weights, and the decoder reconstructs clean waveforms.

In training, label ambiguity arises when multiple outputs must be matched to multiple labels without a predefined order. Permutation Invariant Training (PIT) (Yu et al., 2017b) addresses this by dynamically aligning model outputs with reference signals. It evaluates all possible permutations and selects the one that minimizes the loss. Consider a mixture of speech from S speakers, and a model produces a set of estimated signals $\{\hat{Y}^s\}_{s=1}^S$. The corresponding reference signals $\{Y^s\}_{s=1}^S$ have no inherent correspondence to the outputs due to the unknown order of speakers. The objective function is defined as:

$$\mathcal{L}_{\text{PIT}} = \min_{\pi \in \mathcal{P}(S)} \sum_{s=1}^{S} \mathcal{L}(\hat{Y}^s, Y^{\pi(s)}), \tag{1}$$

where $\mathcal{P}(S)$ represents all permutations of $\{1, \ldots, S\}$, and $\mathcal{L}(\hat{Y}^s, Y^{\pi(s)})$ computes loss between $\hat{Y}^s$ and the permuted reference $Y^{\pi(s)}$.

Despite its effectiveness, PIT faces computational challenges due to the need to evaluate all possible permutations. Additionally, alternative approaches like HEAT (Lu et al., 2021) have been developed to reduce computational cost by using the Hungarian algorithm to efficiently find the optimal permutation.

3. End-to-end multispeaker ASR

In this section, we explore end-to-end multi-speaker ASR systems on two prominent frameworks: single-input single-output (SISO) and single-input multi-output (SIMO) (as shown in Fig. 2). While SIMO frameworks generate separate transcriptions for each speaker through parallel branches, SISO frameworks process mixed audio to produce a single transcription output. We review their architectural designs, training methodologies, and key improvements of both frameworks, highlighting their strengths and limitations. This section focuses on processing predefined audio clips, typically segmented from continuous audio using silence points and heuristic rules (e.g., duration thresholds). In single-speaker ASR, such segments are commonly known as *utterances*. For multi-speaker ASR, the concept of *utterance group* (Kanda et al., 2021) has been introduced as multiple utterances linked through overlapping regions (as shown in Fig. 3).

3.1. Single-input multiple-output (SIMO)

SIMO frameworks process mixture audio with exactly S speakers and produce S transcriptions, one per speaker through parallel branches, as illustrated in Fig. 2(a). Here, S is a fixed parameter that denotes the number of speakers. From the standard architecture, we discuss advancements in separation enhancement, speaker scalability, and leveraging pre-trained models.

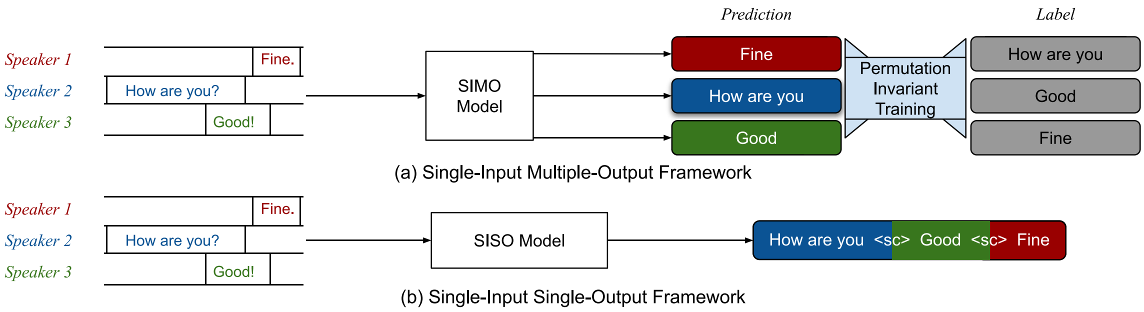

Fig. 2. Single-input multiple-output (SIMO) and single-input single-output (SISO) processes for multi-speaker ASR.

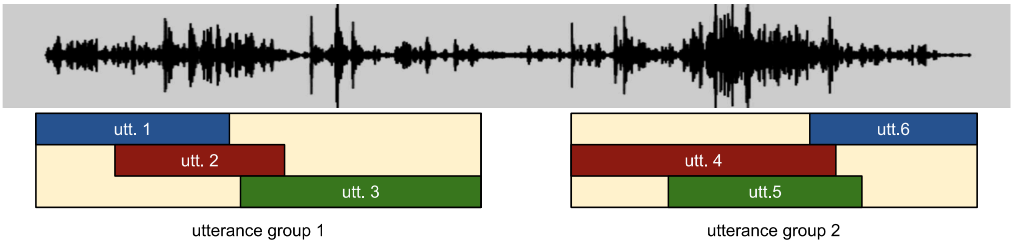

Fig. 3. Example of utterance groups consisting of overlapping speech segments from multiple speakers.

3.1.1. Model architecture

The SIMO framework generates a separate transcription for each speaker through distinct output branches. A simple implementation involves a shared network followed by S individual networks, each dedicated to a speaker's transcription. For instance, Yu et al. (2017a) employs a shared RNN with S linear heads to produce S transcriptions. Alternatively, inspired by separation-based cascade methods, a class of SIMO frameworks (e.g., Seki et al., 2018; Lin et al., 2022) was proposed to integrate speech separation and ASR in a unified structure.

Fig. 4a illustrates a typical SIMO model architecture (Seki et al., 2018; Chang et al., 2020; Zhang et al., 2019, 2020) that integrates the speaker separation process and ASR into a single framework, enabling end-to-end training from scratch. The architecture adopts a stacked design comprising shared and unshared modules across branches. The process begins with a mixture encoder, $\text{Encoder}_{\text{Mix}}$, which extracts an intermediate feature sequence $\mathbf{Z}$ from the mixed audio input $\mathbf{X}$, serving as the input for subsequent separation and ASR. This feature sequence is then fed into S parallel branches, each dedicated to one of the S speakers. Within branch s, a speaker differentiating encoder, $\text{Encoder}_{\text{SD}^s}$, disentangles the speech content $\mathbf{Z}^s$ of the corresponding speaker from the mixture feature $\mathbf{Z}$. Finally, a shared ASR model, such as an attention-based encoder–decoder (AED) or transformer, generates the transcription hypothesis $\mathbf{H}^s$ for speaker s. The process can be formally described as:

$$\mathbf{Z} = \text{Encoder}_{Mix}(\mathbf{X}), \tag{2}$$

$$\mathbf{Z}^s = \text{Encoder}_{SD^s}(\mathbf{Z}), \quad s = 1, \dots, S \tag{3}$$

$$\mathbf{H}^s = \text{ASR}(\mathbf{Z}^s), \quad s = 1, \dots, S \tag{4}$$

The training objective of each branch follows that of single-speaker ASR: cross-entropy loss $\mathcal{L}_{\text{CE}}$ to maximize the transcription accuracy, and an optional CTC loss $\mathcal{L}_{\text{CTC}}$ for monotonic alignment. Similar to speech separation, SIMO models with multiple output branches introduce label ambiguity, where the correspondence between hypotheses and references is unclear. Permutation Invariant Training (PIT) is also utilized to align hypotheses $\mathbf{H}^s = (h_1^s, \dots, h_{N_s}^s)$ with reference transcriptions $\mathbf{R}^s = (r_1^s, \dots, r_{N_{ss}}^s)$ through permutations (Yu et al., 2017a). To reduce the permutation computational cost, label matching is often determined solely based on the CTC loss, while both CTC and cross-entropy losses are used (Seki et al., 2018; Chang et al., 2019; Zhang et al., 2019, 2020).

SIMO provides a natural framework for integrating traditional separation and ASR methods into an end-to-end system. This design facilitates the incorporation of state-of-the-art separation techniques to enhance performance. However, the fixed number of branches limits the model's ability to handle a variable number of speakers, which remains a significant constraint in real-world scenarios. Also, limited separation performance still causes redundancy and omissions in the final transcription output.

3.1.2. SIMO improvements

To address SIMO's limitations, recent research has focused on three key goals: (1) enhancing separation performance, (2) enabling a flexible number of speakers, and (3) mitigating data scarcity. The following sections detail these categories. Fig. 5 summarizes the corresponding goals and methods.

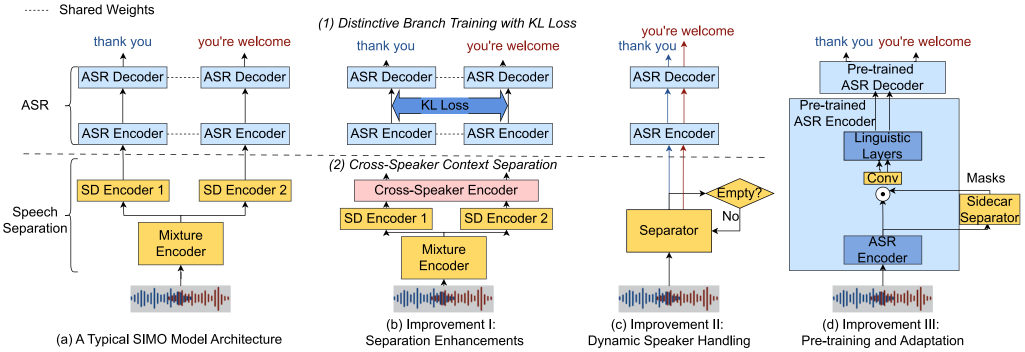

Fig. 4. A typical SIMO architecture (a) and three types of key improvements (b–d). (b) Enhance separation by (1) Introducing an auxiliary KL Loss to promote distinctive separation between speakers. (2) Incorporating a cross-speaker encoder to provide cross-speaker contextual cues for compensating omission and reducing repetitions between branches; (c) Support dynamic speaker counts using iterative separation. (d) Adapt pre-trained large speech foundation model with a Sidecar separator.

3.1.2.1. Separation enhancement. While end-to-end models jointly train separation and recognition modules, independent branches can propagate early separation errors, resulting in repeated or omitted transcriptions. Recent work addresses this by increasing inter-branch distinctiveness and leveraging context to refine separation.

Distinctive branch training: To encourage distinct transcriptions across output branches, auxiliary loss functions can be applied between ASR hidden state vectors. In AED-based models, Seki et al. (2018) proposed a contrastive loss that maximizes Kullback–Leibler (KL) divergence between the ASR encoder states from different branches, as shown in Fig. 4(b.1).

This loss penalizes similarity between streams, reducing the likelihood of redundant transcriptions.

Context-aware separation: More recently, the separation has been further improved by introducing cross-speaker context to enhance the separation quality. Traditional SIMO systems process each speaker independently in parallel branches, restricting the model's capacity to capture cross-speaker dependencies and perform mutual correction. To address this, Kang et al. (2024) proposes a *Cross-Speaker Encoder* ($\text{Encoder}_{\text{CSE}}$), positioned between the Encoder_{SD} and ASR model to enable information sharing across branches (Fig. 4(b.2)). It consumes the concatenation of mixture encoding from $\text{Encoder}_{\text{Mix}}$ and different speaker's individual encodings from $\text{Encoder}_{\text{SD}}$, and employs a conformer block to enable information sharing across branches. The conformer output is then partitioned into branch-specific representations and passed to the respective ASR encoders. This context-aware mechanism leverages the local context of the mixed speech to refine the single-speaker features, thereby improving separation accuracy.

3.1.2.2. Dynamic speaker handling. A key limitation of traditional SIMO frameworks is their reliance on a fixed number of speaker streams, which constrains their applicability to real-world scenarios. Recent work addresses this challenge by introducing iterative and adaptive systems to determine dynamically the number of speakers. As illustrated in Fig. 5(c), von Neumann et al. (2020) propose an iterative approach where the system separates one speaker's audio at a time with a Dual-Path RNN TasNet (Luo et al., 2020) separator. The residual mixture is then passed to subsequent iterations until no speech remains. Each separated feature is then fed into a shared ASR model, which can be jointly fine-tuned with the separator. This design supports end-to-end training while supporting varying numbers of speakers.

3.1.2.3. Pretraining and adaptation. Leveraging pre-trained models mitigates the data scarcity challenge in mult-speaker ASR, as shown in Fig. 4(d). In the SIMO paradigm, a common approach is to adopt a dual-module framework consisting of a pre-trained speech separation module and a pre-trained ASR module, with joint fine-tuning to align their objectives. For example, Settle et al. (2018) includes Chimera++, a pre-trained separation module, to generate speaker masks over mixture features, producing S separate streams. These streams are then fed into a shared ASR model for transcription, and the two modules are subsequently fine-tuned together to enable end-to-end optimization.

More recent advances introduce large speech foundation models like Whisper (Radford et al., 2023) and Wav2Vec (Baevski et al., 2020) into the SIMO pipeline. Since these models are not inherently designed for multi-speaker inputs, a separation module must be inserted to enable SIMO-style processing. To this end, Meng et al. (2023, 2024b) propose a lightweight convolutional network Sidecar separation module inserted between layers of a frozen Wav2Vec2.0 or Whisper model. This design splits the mixed audio into multiple streams, each processed independently by subsequent Wav2Vec layers to produce distinct transcriptions. Instead of building an external separation module, it operates directly on the internal feature representation of the pre-trained model, enabling an efficient and seamless adaptation of pre-trained single-speaker ASR models to multi-speaker scenarios.

3.2. Single-input single-output (SISO)

SISO is another multi-speaker ASR framework that only generates a single output sequence containing all speakers' transcriptions. Unlike SIMO, which produces separate outputs per speaker, SISO relies on serialized output training (SOT) (Kanda et al., 2020b) to serialize transcriptions into a unified stream. The transcriptions of all speakers are concatenated into one output.

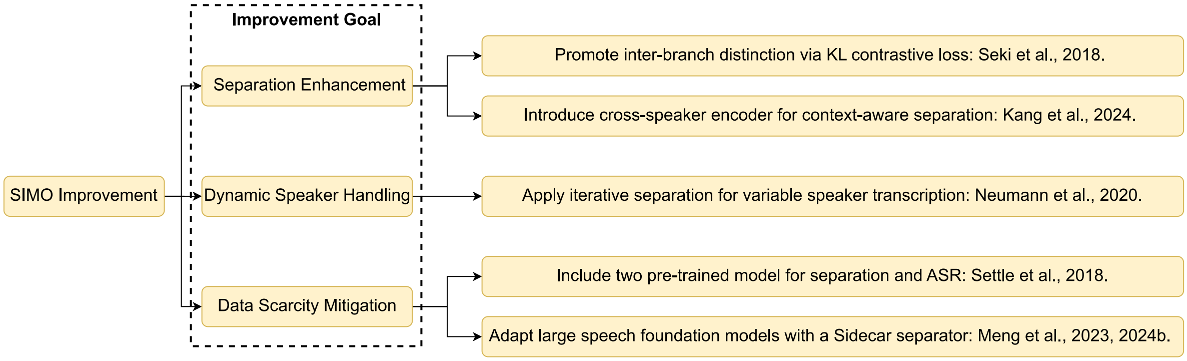
Improvement Goal
SIMO Improvement
Separation Enhancement
Dynamic Speaker Handling
Data Scarcity Mitigation
Promote inter-branch distinction via KL contrastive loss: Seki et al., 2018.
Introduce cross-speaker encoder for context-aware separation: Kang et al., 2024.
Apply iterative separation for variable speaker transcription: Neumann et al., 2020.
Include two pre-trained model for separation and ASR: Settle et al., 2018.
Adapt large speech foundation models with a Sidecar separator: Meng et al., 2023, 2024b.

Fig. 5. SIMO improvements by goal. Targeting better separation performance, variable speaker handling, and data scarcity mitigation by module refinement, iterative processing, and pre-trained model integration.

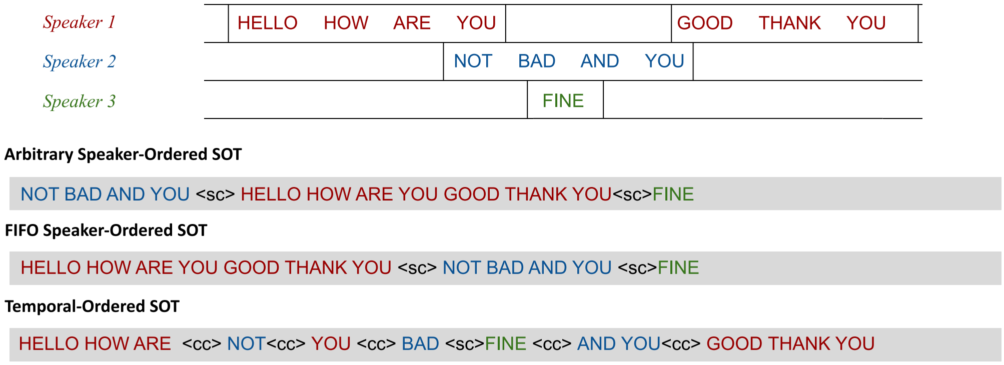
Speaker 1
HELLO HOW ARE YOU
GOOD THANK YOU
Speaker 2
NOT BAD AND YOU
Speaker 3
FINE
Arbitrary Speaker-Ordered SOT
NOT BAD AND YOU <sc> HELLO HOW ARE YOU GOOD THANK YOU<sc>FINE
FIFO Speaker-Ordered SOT
HELLO HOW ARE YOU GOOD THANK YOU <sc> NOT BAD AND YOU <sc>FINE
Temporal-Ordered SOT
HELLO HOW ARE <cc> NOT<cc> YOU <cc> BAD <sc>FINE <cc> AND YOU<cc> GOOD THANK YOU

Fig. 6. An example of different SOT texts, containing speaker-change symbol <sc> and <cc>.

SISO offers several advantages. First, it naturally handles varying numbers of speakers, making it well-suited to real-world scenarios with unknown speaker counts. Second, by generating a single output sequence, SISO captures inter-speaker dependencies, improving both coherence and overall transcription accuracy. Third, it reduces computational cost by enforcing a fixed output order, thereby simplifying training.

This section first introduces two forms of SOT output: speaker-ordered sentence-based SOT and temporal-ordered token-based SOT. We then discuss improvements to the SISO architecture.

3.2.1. Serialized output

SOT can be implemented in two primary forms: sentence-based SOT and token-based SOT (Fig. 6). In sentence-based SOT, all sentences of each speaker are concatenated into a sequence. Special token <sc> is inserted between the transcriptions to indicate changes in the speaker. Consecutive sentences from the same speaker are simply concatenated in the transcript, even if those sentences may be partially or even completely interrupted by another speaker's sentence. The order of speakers can be arbitrary, with PIT used in loss. Alternatively, sentence-based SOT can follow a first-in, first-out (FIFO) order, where transcriptions are arranged by each speaker's start time, from the earliest to the latest.

In contrast, token-based SOT serializes transcriptions based on token timestamps. The channel change symbol <cc> is utilized in token-based SOT. This approach provides finer granularity in capturing overlapping speech by preserving the timing order, while remaining compatible with CTC loss, which is well-suited for mostly time-monotonic sequences (Li et al., 2024; Fan et al., 2024). However, when the same speaker's sentence is interleaved with others, a more advanced decoder is needed to effectively model long-range context and maintain coherence within that speaker's speech.

3.2.2. SISO model and improvements

Since SISO models produce a single output stream, they typically build on single-speaker architectures without requiring explicit speaker separation. Fig. 7(a) shows an encoder–decoder model that has been successfully adapted to SISO-based multi-speaker ASR in Kanda et al. (2020b). In this setting, the ASR encoder generates hidden representations $\mathbf{Z}^{\mathrm{asr}}$, which are then passed to ASR decoder to obtain serialized transcriptions.

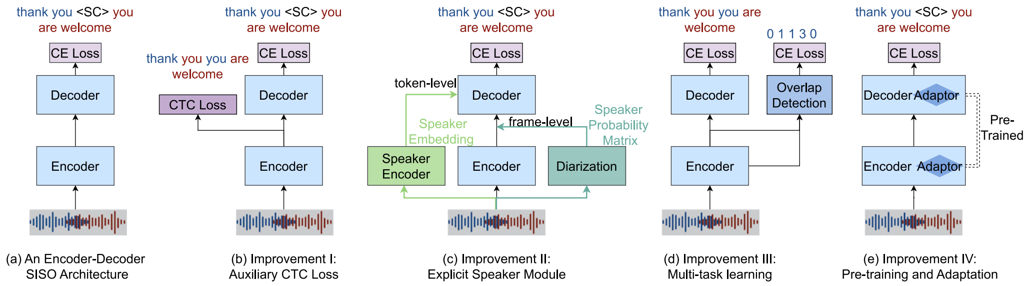
thank you <SC> you are welcome
CE Loss
Decoder
Encoder
(a) An Encoder-Decoder SISO Architecture
thank you you are welcome
CTC Loss
thank you <SC> you are welcome
CE Loss
Decoder
Encoder
(b) Improvement I: Auxiliary CTC Loss
thank you <SC> you are welcome
CE Loss
token-level
Decoder
Speaker Embedding
frame-level
Speaker Probability Matrix
Speaker Encoder
Encoder
Diarization
(c) Improvement II: Explicit Speaker Module
thank you you are welcome
CE Loss
0 1 1 3 0
CE Loss
Decoder
Overlap Detection
Encoder
(d) Improvement III: Multi-task learning
thank you <SC> you are welcome
CE Loss
Decoder Adaptor
Pre-Trained
Encoder Adaptor
(e) Improvement IV: Pre-training and Adaptation

Fig. 7. An encoder–decoder SISO architecture (a) and four representative improvements (b–e). (b) Adding an auxiliary CTC loss to enhance acoustic modeling and speaker-awareness capability; (c) Incorporating explicit speaker modules to inject speaker information at the frame or token level; (d) Applying multi-task learning with auxiliary tasks such as overlap detection; (e) Integrating pre-trained models and fine-tuning them with mechanisms such as adapters.

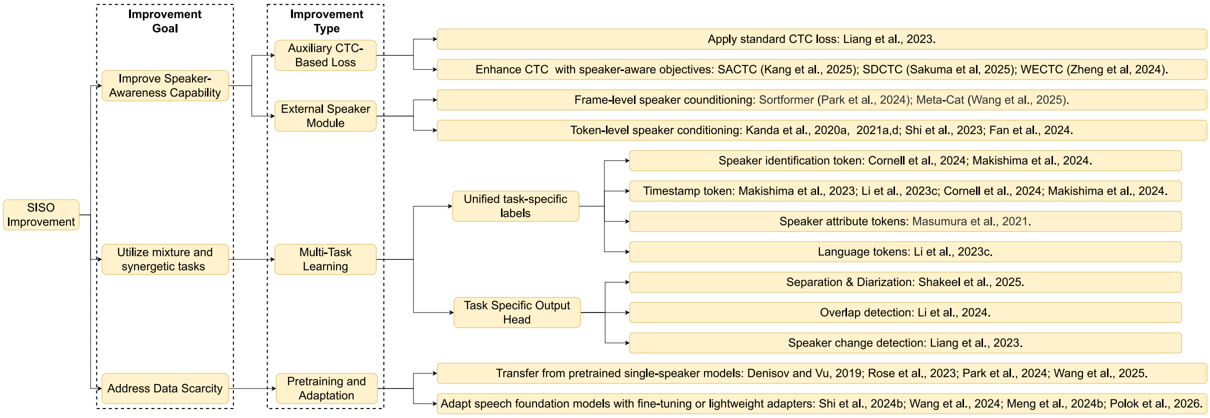
Improvement Goal
Improvement Type
SISO Improvement
Improve Speaker-Awareness Capability
Utilize mixture and synergetic tasks
Address Data Scarcity
Auxiliary CTC-Based Loss
External Speaker Module
Multi-Task Learning
Pretraining and Adaptation
Apply standard CTC loss: Liang et al., 2023.
Enhance CTC with speaker-aware objectives: SACTC (Kang et al., 2025); SDCTC (Sakuma et al, 2025); WECTC (Zheng et al, 2024).
Frame-level speaker counditioning: Sortformer (Park et al., 2024); Meta-Cat (Wang et al., 2025).
Token-level speaker conditioning: Kanda et al., 2020a, 2021a,d; Shi et al., 2023; Fan et al., 2024.
Unified task-specific labels
Speaker identification token: Cornell et al., 2024; Makishima et al., 2024.
Timestamp token: Makishima et al., 2023; Li et al., 2023c; Cornell et al., 2024; Makishima et al., 2024.
Speaker attribute tokens: Masumura et al., 2021.
Language tokens: Li et al., 2023c.
Task Specific Output Head
Separation & Diarization: Shakeel et al., 2025.
Overlap detection: Li et al., 2024.
Speaker change detection: Liang et al., 2023.
Transfer from pretrained single-speaker models: Denisov and Vu, 2019; Rose et al., 2023; Park et al., 2024; Wang et al., 2025.
Adapt speech foundation models with fine-tuning or lightweight adapters: Shi et al., 2024b; Wang et al., 2024; Meng et al., 2024b; Polok et al., 2026.

Fig. 8. SISO improvements by goal and type. Aiming to enhance speaker-aware capabilities, utilize shared representation and alleviate data scarcity via auxiliary losses, external speaker modules, multi-task learning, and pre-trained models adaptation.

Although using a single processing path enables cross-speaker context modeling, the lack of explicit separation mechanisms in SISO systems makes accurate transcription of overlapping speech challenging. To address this and compensate for limited multi-speaker training data, researchers have proposed four main strategies: (1) auxiliary CTC-related losses, (2) integration of external speaker modules, (3) multi-task learning, and (4) pre-training and adaptation. Fig. 8 shows the improvement type.

3.2.2.1. Auxiliary CTC-related loss. As described in Section 2.1, CTC loss commonly serves as an auxiliary objective to cross-entropy loss in AED to enhance the encoder's acoustic modeling. In the SISO framework, it can be applied similarly, through a parallel branch connected to the encoder output, independent of the attention-based decoder, as illustrated in Fig. 7. In addition, modified CTC variants can incorporate speaker information to improve speaker differentiation in multi-speaker transcription, as described below.

CTC in SISO: To leverage CTC for improving acoustic alignment in SISO systems, the CTC loss objective must maintain temporal consistency with the original mixed acoustic features. Typically, token-based SOT serves as the CTC target in this setting. If the model's final output follows sentence-based SOT, the remaining components (not supervised by CTC) need to handle the utterance-level reordering. For instance, Liang et al. (2023) proposed a two-stage CTC approach: the first CTC loss is applied after the initial encoder layers, optimized for token-based SOT to enhance acoustic modeling, while the second CTC loss operates on the full encoder output with sentence-based SOT supervision.

Speaker-enhanced CTC: The second approach modifies CTC to explicitly integrate speaker information. A noticeable example is Speaker-Aware CTC (SACTC) (Kang et al., 2025) proposed by Kang et al. which formulates a Bayes-risk-based CTC that constrains the encoder to distinguish speaker-specific features at specific time frames, explicitly modeling speaker separation. More recently, Speaker-Disguishable CTC (SD-CTC) (Sakuma et al., 2025) extends CTC by jointly assigning a token and its corresponding speaker label to each frame. Another advancement comes from Zheng et al. (2024) with Weakening and Enhancing CTC (WECTC) loss, which enhances speaker change token (<sc>) prediction by adjusting pseudo-label posteriors.

Despite its widespread use in SISO frameworks and extensions with speaker-enhanced modules, CTC remains limited in overlapping speech scenarios, as shown in the results of Kang et al. (2025). It assumes conditional independence between output tokens and enforces a one-token-per-frame constraint, producing a single, linear output sequence. This makes it fundamentally incompatible with simultaneous speech from multiple speakers, often resulting in entangled or incomplete transcriptions. Future research may investigate approaches that relax this constraint while preserving CTC's alignment ability.

3.2.2.2. External speaker module. One direction for improving the SISO framework is the integration of explicit speaker information into the model architecture. By incorporating speaker-specific cues, the model can enhance its ability to differentiate between speakers and better handle overlapping speech. Current approaches can be broadly categorized into two methodologies: (1) Frame-level Speaker Conditioning (FSC) and (2) Token-level Speaker Conditioning (TSC).

Frame-level speaker conditioning (FSC) incorporates speaker information by aligning speaker and speech timestamps at the frame level. Modular FSC (Yu et al., 2022a) directly aligns diarization results with SOT transcriptions in a post-processing stage, which is not an end-to-end approach and may lead to error propagation. Later end-to-end FSC methods (right side of Fig. 7(c)) integrate speaker information before the decoder, aligning the diarization output with ASR encoding representation at the frame level. Specifically, the diarization module generates an S-speaker assignment probability matrix for T frames, denoted as $\mathbf{P} \in \mathbb{R}^{S\times T}$. This matrix is used to incorporate speaker information into the ASR encoder representations $\mathbf{Z} \in \mathbb{R}^{M\times T}$ before passing the fused representation to the decoder through different integration strategies. One approach (Park et al., 2024) incorporates a sinusoidal matrix, similar to the sinusoidal positional encoding used in Transformers (Vaswani et al., 2023). Each speaker is associated with a unique sinusoidal pattern, which is then weighted by the probability matrix $\mathbf{P}$ and added to the original encoder representation $\mathbf{Z}$. This mechanism can differentiate speakers in a structured, deterministic way, and can be disabled to fall back to a standard ASR pipeline. Alternatively, Wang et al. proposed Meta-Cat (Wang et al., 2025), which first computes a speaker-specific representation $\mathbf{Z}_s$ by element-wise multiplying the encoder output $\mathbf{Z}$ with the frame-level probability vector P_s for each speaker s. These representations are then concatenated across all speakers to form a supervector, which is passed to the decoder. This method expands the ASR embedding multiple times according to the number of speakers. Both approaches ensure the entire model remains differentiable, enabling joint training of the whole system.

Token-level speaker conditioning (TSC) introduces speaker embeddings as auxiliary input to the decoder during token generation. Unlike FSC, which aligns frame-level probabilities, TSC typically uses a speaker encoder to extract speaker embeddings for speaker-aware decoding (left side of Fig. 7(c)). A typical TSC speaker-attributed ASR system (Kanda et al., 2020a, 2021a,d) consists of four modules: ASR encoder, speaker encoder, speaker decoder, and ASR decoder. The input mixture is processed in parallel by the ASR and speaker encoders to generate the speech representation $\mathbf{Z}^{\text{asr}}$ and speaker embeddings $\mathbf{Z}^{\text{spk}}$, respectively. These are fed into the decoder, along with the previous token sequence, to produce the context-aware speaker representation. The ASR decoder then takes this speaker representation as extra input with ASR encoder states and previous output to predict the next token.

Here, the speaker representation is related to a pre-defined speaker inventory $\mathcal{D} = \{\mathbf{d}_1, \dots, \mathbf{d}_K\}$, where $\mathbf{d}_k$ represents a speaker profile in the inventory. The output of the speaker decoder is used as a query to compute attention over this speaker inventory $\mathcal{D}$, and generates a weighted speaker profile $\bar{\mathbf{d}}_n$ as final speaker representation given to the speaker decoder. Later approaches proposed by Shi et al. (2023) considered contextual information for speaker representation generation. A separate contextual text encoder is deployed before the speaker decoder to aggregate the semantic information of the whole output utterance. In addition, when calculating $\bar{\mathbf{d}}_n$, an extra context-dependent scorer is employed to model the local speaker discriminability by contrasting with speakers in the context. Fan et al. (2024) also enhance the speaker contextural relationship by replacing the weighted profile $\bar{\mathbf{d}}_n$ as a similarity matrix and passing this matrix to the ASR decoder. In this way, the system can incorporate the speaker information without the speaker inventory $\mathcal{D}$.

3.2.2.3. Multi-task learning. Multi-task learning enables the joint optimization of synergistic tasks via a fully or partially shared network, which can in turn improve the performance of individual tasks (e.g., multi-speaker ASR and overlapping detection). Compared to its application in SIMO models, multi-task learning in SISO can directly leverage the shared representation and jointly train the ASR model with mixture-related tasks such as diarization and overlap detection by sharing network components.

Unified labeling: Task-specific labels can be inserted into the serialized output as special tokens. This unified labeling enables joint modeling of multiple tasks, such as speaker identification and timestamp predictions, without modifying the original auto-regressive multi-speaker ASR architecture. First, specific speaker tokens (e.g., `<spk0>` `<spk1>`) can replace general speaker changing tokens (`<sc>` and `<cc>`) to differentiate speakers, as in Cornell et al. (2024) and Makishima et al. (2024). This reduces the speaker ambiguity in SOT, especially token-level SOT. Second, quantized timestamp tokens (e.g., 20 ms resolution) can be inserted as special tokens to indicate the start and end time of an utterance (Makishima et al., 2023; Li et al., 2023c; Cornell et al., 2024; Makishima et al., 2024). Timestamp tokens provide extra information about speaker turn-taking, temporal order, and overlapping speech. In some systems, additional timestamp tokens are inserted based on heuristic rules, such as when the gap between consecutive tokens exceeds two seconds, to indicate pauses or silence (Li et al., 2023c). Additionally, Masumura et al. (2021) and Li et al. (2023c) independently explored the use of speaker attribute tokens (e.g., gender, age) and language tokens, respectively, to enrich the output with contextual information. All tasks under this unified labeling scheme are handled by a single model with a shared output head.

Task specific output head: Separate labels and output heads for each auxiliary task can serve as another implementation of multi-task learning in SISO. In this case, tasks share parts of the network, typically lower-layer acoustic or speaker-related representations, while maintaining task-specific output branches in the later stages. For example, Shakeel et al. (2025) introduce a unified multi-speaker encoder (UME) that jointly addresses the speech separation, speech diarization, and multi-speaker ASR tasks within a single architecture. Their approach uses the Open Whisper-style Speech Model (OWSMv3.1) (Peng et al., 2024) to extract representations from multiple layers, and fuses them using learnable weighting parameters. The resulting shared representation is then fed into task-specific heads. For the separation module, the model also directly incorporates the original speech mixture as an additional input. Overall, UME demonstrates substantial gains over strong single-task SOTA baselines across three tasks. Li et al. (2024) jointly optimize the overlapping prediction task and the ASR task to enhance multi-speaker ASR (Fig. 7(d)). Specifically, the overlap-aware task predicts two binary states for each token: whether it overlaps with other tokens and whether it is a boundary token. Both the ASR and overlapping prediction models adopt an encoder–decoder structure, sharing lower layers of the encoder to leverage acoustic features while diverging to predict overlap-aware labels. The total loss combines the ASR loss and the overlap-aware loss, enabling joint training to improve performance in overlapping speech scenarios.

Speaker change detection has also been integrated as an auxiliary task in multi-task learning frameworks. For example, Liang et al. (2023) proposed Boundary-Aware SOT (BA-SOT). Specifically, the model adds a speaker change detection (SCD) module as the output head for the speaker change task after several decoder blocks. A binary speaker change label is used for the auxiliary task, while the ASR task adopts FIFO-style SOT labels without speaker change tokens. The total loss consists of ASR loss, SCD loss, and a newly introduced boundary constraint loss. Studies (Liang et al., 2023; Li et al., 2024) have shown that joint modeling of ASR with auxiliary tasks such as speaker change detection and overlap detection can lead to improved recognition accuracy.

3.2.2.4. Pretraining and adaptation. Similar to SIMO, leveraging pre-trained modules in SISO architectures helps mitigate the data scarcity in multi-speaker ASR. Because SISO models share the same model structure and single-output format with single-speaker ASR, SISO models can be directly initialized from pre-trained single-speaker models without architectural modification (Denisov and Vu, 2019; Rose et al., 2023). This facilitates transfer learning from large-scale simulated mixtures, which are artificially created multi-speaker audio samples (see Section 6). These simulated datasets can be used for pre-training and subsequently fine-tuned on real-world corpora to achieve competitive performance (Kanda et al., 2021). Moreover, speaker-specific modules in SISO models – such as speaker encoder and diarization model – can also be initialized from pre-trained modules (Park et al., 2024; Wang et al., 2025).

This structural compatibility also holds for speech foundation models: single-speaker models, such as Whisper and WavLM-based ASR, can be directly fine-tuned under the SISO framework for multi-speaker ASR, unlike SIMO models that require additional separation modules. These models can either be fully fine-tuned, or adapted using lightweight modules like LoRA (Hu et al., 2022), enabling resource-efficient training while keeping most pre-trained parameters frozen (Shi et al., 2024b; Wang et al., 2024; Meng et al., 2024b; Polok et al., 2026). In addition, Shi et al. (2024b) explored maintaining the multilingual property of foundation models by leveraging adapters.

3.3. SIMO vs. SISO: Summary

By performing speaker separation and speech recognition separately, SIMO provides more modularity than SISO, which performs both tasks in an integrated way. However, SIMO is less flexible than SISO in three key aspects: (1) SISO accommodates variable speaker counts, whereas SIMO requires knowing the number of speakers *a priori*; (2) SIMO's branch-specific separation lacks cross-speaker validation for redundancy and complementarity and further underutilizes cross-speaker information in ASR. Very few (e.g., Kang et al. (2024)) SIMO approaches attempt to address this shortcoming. (3) SISO's retention of mixed features enables multi-task learning with mixture-based tasks, such as overlapping detection. Additionally, SISO's fixed speaker order in transcription (e.g., FIFO) reduces computational overhead from label matching in SIMO training.

These structural differences lead to distinct improvement priorities. SIMO's overall recognition performance is fundamentally limited by its separation module: poor separation introduces residual noise or truncated phonemes, degrading ASR accuracy. Thus, SIMO requires enhanced separation for optimal performance. In contrast, SISO lacks explicit separation, and thus harnessing speaker information to disambiguate overlapping speech is a focus for SISO improvement. Some advances are largely incremental, such as variations of the CTC loss or other light-weight architectural refinements. In contrast, other developments, such as the incorporation of large foundation models, represent more transformative progress in multi-speaker ASR.

Both SIMO and SISO address data scarcity by utilizing pre-trained models, particularly recent large foundation models. For SIMO, this requires adding a separation process to the foundation model. SISO directly fine-tune the foundation ASR model on multi-speaker data. Both frameworks can achieve strong performance by tuning only 8%–10% of all parameters.

3.4. Future directions

One promising direction is **SIMO-SISO hybridization**, which aims to combine SIMO's separation precision with SISO's comprehensive modeling. In such framework, the model can first separate the input audio into multiple speaker-specific branches, whose representations are then concatenated and passed to additional network layers for integrated decoding. Recent work has explored this hybridization at different representation levels. For example, Cross-speaker encoder (Kang et al., 2024) combined acoustic features from different branches before ASR, to enhance separation by providing cross-speaker context. Also, Huang et al. (2023) adapt WavLM to extract target-speaker transcriptions from mixture audio using speaker embeddings. The resulting utterances are concatenated and passed through a joint speaker module to generate the final serialized transcript. Future studies may further explore multi-level integration with advanced model architecture to enhance SIMO's contextual modeling and improve SISO's ability to handle overlapping speech.

Another direction is the **adapted foundation model enhancement**. Foundation-model adaptation represents a transformative direction with the potential to reshape multi-speaker ASR. There are lots of questions that remain underexplored. For instance, can a foundation-adapted SIMO model also leverage cross-speaker context? How can foundation-adapted SISO models benefit from multi-task learning? Beyond architecture-specific adaptations, future work may further investigate the integration of multi-speaker ASR into multi-modal LLM-based unified frameworks. This can extend current instruction-based approaches (Shi et al., 2024c) to more practical, real-world applications.

4. Multi-modal multispeaker ASR

The previous section categorized end-to-end multi-speaker ASR systems primarily based on their output format. However, from the input perspective, a growing body of work explores feeding multi-speaker ASR models with multi-modal inputs. When the acoustic mixture is excessively noisy, is highly overlapped, or contains rare or long-tail words that challenge audio-only systems, additional modalities can provide complementary cues to better resolve both speaker identity and linguistic content.

Most existing multi-modal approaches follow the SISO architectures, as this design retains the complete acoustic mixture throughout the encoder and decoder, thus allowing high-level visual or textual context to be fused directly into the joint modeling of speakers and speech. In contrast, SIMO pipelines must decide where to incorporate cross-modal signals: either before separation or after separation.

The auxiliary context usually falls into two categories: (1) visual information, such as video and images; (2) textual information that leverages the text-understanding capabilities of multi-modal LLMs.

4.1. Audio-visual multi-speaker ASR

Visual information in audio-visual speech recognition typically comes from face tracks and mouth movements. Prior single-speaker studies (Ma et al., 2023; Li et al., 2023b; Seo et al., 2023) have consistently shown that visual cues can improve ASR robustness, especially under adverse acoustic conditions, because the visual modality is unaffected by background noise. In most systems, visual features are incorporated by concatenating them with audio features, which is then given to audio-visual ASR model.

In multi-speaker scenarios, visual cues also help improve transcription accuracy. Early multi-person audio-visual ASR studies (Braga et al., 2020; Braga and Siohan, 2022b) explored how to incorporate face-track features into a standard ASR model without explicitly predicting speaker labels. Importantly, these approaches do not require any training labels regarding which face corresponds to which voice; rather, this correspondence is inferred dynamically by the model. Let $\mathbf{Z}^a$ denote the audio encoder representation and $\mathbf{Z}^{v_s}$ the visual representation for speaker $s \in \{1, \dots, S\}$. A cross-attention module takes $\mathbf{Z}^a$ as the query and $\{\mathbf{Z}^{v_s}\}_{s=1}^{S}$ as key and values. This allows the model to produce a weighted visual summary $\tilde{\mathbf{Z}}^v$ that emphasizes the most relevant speaker-specific visual cues. The fused representation $\mathbf{Z} = \text{Concat}(\mathbf{Z}^a, \tilde{\mathbf{Z}}^v)$ is then fed into the audio-visual ASR model. Subsequent work (Braga and Siohan, 2022a) extends this by adding an auxiliary active-speaker detection task, using the same fused representation $\mathbf{Z}$ to jointly predict the speaking face.

For multi-speaker ASR that distinguishes speakers in a sentence, where multiple speakers must be differentiated within a single mixture, both SIMO and SISO systems benefit from visual conditioning, although in different ways. In the SIMO architecture, Wu et al. (2021) injects visual representations $\mathbf{Z}^v$ into both the encoder and decoder using the same cross-modal attention module. Their experiments show that it significantly outperforms simply concatenating all face-track features before the mixture encoder $\text{Encoder}_{\text{Mix}}$. Visual information also helps resolve output-stream permutation. In SISO settings, where all speakers are decoded in a single transcript, Makishima et al. (2025) concatenates all available visual features with the audio representation prior to decoding, similar to single-speaker audio-visual ASR, and reports consistent performance gains.

4.2. LLM-based text conditioning

Recently, the emergence of multi-modal large language models (MLLMs) (Latif et al., 2023) has provided a unified framework for a wide range of speech and language tasks. These models typically consist of a speech encoder and a projector that projects acoustic features into the LLM embedding space, followed by an LLM that performs sequence generation. Like single-speaker ASR with multimodal LLM (Trinh et al., 2024), multi-speaker ASR can also be cast into this unified architecture (Li et al., 2023c; Shi et al., 2025). Because the unified output format of SISO models naturally aligns with the generation paradigm of MLLMs, current LLM-based multi-speaker systems almost exclusively adopt the SISO output format.

MLLM-based systems benefit from instruction-based multi-task learning, which enhances speech understanding and enables flexible, user-defined interaction scenarios. The recent work of Meng et al. (2024a) exemplifies this interactive paradigm: users can specify multi-speaker ASR tasks through natural language instructions, such as transcribing only the first speaker, a female speaker, or utterances containing specific keywords. The system integrates speech representations from the Whisper encoder and WavLM with an LLM. Compared to training with only a generic multi-speaker ASR instruction, this instruction-based training with additional contextual prompts improves the model's speech understanding and overall multi-speaker ASR performance.

In addition, MLLMs can ingest auxiliary textual context that benefits both speech and speaker recognition. Long-tail or rare words can be supplied as external context; prior work in single-speaker ASR (Trinh et al., 2025) demonstrates the effectiveness of providing such terms to a multimodal LLM, and He et al. (2025) extends this idea to the multi-speaker setting with similar gains. Diarization information can also be integrated into an MLLM to assist multi-speaker decoding, as shown in Lin et al. (2025). In this framework, the LLM consumes not only projected speech tokens but also structured diarization triplets (speaker embeddings, start indices, and end indices), each aligned to the LLM embedding space through learnable linear layers.

5. Long-form multispeaker ASR

The previous section focuses on pre-segmented audio. We turn to continuous long-form audio in this section. Unlike pre-segmented data, long-form audio must be partitioned before being processed by either SIMO or SISO ASR models, and the results must be integrated into a coherent global transcription with consistent speaker identities. This section addresses these two key challenges. An ideal segmentation algorithm should be computationally efficient while preserving linguistic coherence. The segmentation methods are introduced in Section 5.1. Section 5.2 discusses how to merge local hypotheses into a unified global transcription, with consistent speaker identities across all segments.

5.1. Segmentation methods

Segmenting long-form multi-speaker audio is crucial for enabling accurate and efficient ASR. Strategies need to balance computational efficiency and preserve linguistic coherence. Existing methods typically rely on acoustic or semantic cues.

Voice activity detection (VAD) segments speech by detecting silent intervals based on acoustic features such as energy levels and spectral patterns. Its simplicity leads to widespread use in both cascade and end-to-end systems (Kanda et al., 2021c,a; Yu et al., 2022a). However, VAD-based segments may still be excessively long, necessitating further clipping to meet the input requirements of downstream ASR models. Also, since VAD operates purely on acoustic cues, it may inadvertently split sentences at unnatural boundaries, particularly when speakers pause or hesitate mid-sentence, which can disrupt linguistic coherence and degrade the performance of subsequent ASR tasks.

Sliding window segmentation is another strategy, which processes audio using fixed-length windows and a pre-defined stride. Unlike VAD, the sliding window method generates uniformly sized segments ready for ASR models. However, it still risks disrupting semantic coherence by splitting sentences at arbitrary boundaries, a problem exacerbated by rigid window constraints. Additionally, shorter strides can inflate computational costs due to extensive overlap.

To address these limitations, recent studies incorporate semantic information into segmentation. For example, Cornell et al. (2024) introduce an adaptive sliding window strategy, which inserts the special token `<trunc>` to mark truncation, and resumes from the last silence point to preserve linguistic continuity and reduce redundancy. Huang et al. (2022) predict segment boundaries in a streaming manner, based on both acoustic and text-level cues, enabling dynamic segmentation with minimal overhead.

5.2. Hypothesis stitching methods

Generating the final global hypothesis requires concatenating and aligning local transcriptions from segmented audio. A straightforward approach is to concatenate local transcriptions directly with VAD-segmented clips (Kanda et al., 2022, 2021a; Cornell et al., 2024). When long audio is segmented with overlaps, the final global transcription cannot be obtained by simple concatenation due to redundancy and potential mismatch. High-confidence word selection can be employed in overlapping parts (Chiu et al., 2019), while a neural-network-based hypothesis stitcher (Chang et al., 2021) fuses segment outputs into a coherent transcription without requiring alignment.

Maintaining globally consistent speaker labels is essential for long-form multi-speaker ASR. Existing approaches rely on speaker profiles or embeddings learned jointly within the ASR model. In profile-based methods, speaker identities can be resolved during speaker-attributed decoding when speaker enrollment is available. When enrollment is not possible, speaker profiles can be approximated through unsupervised clustering of embeddings from a pretrained speaker encoder (Kanda et al., 2022). Even supplying dummy profiles, which do not appear in the input audio, can improve performance (Kanda et al., 2021a). In the joint modeling approach, the multi-speaker ASR system outputs both transcriptions and speaker embeddings, typically by a multi-task learning framework. Global labels can be obtained by clustering the embeddings. In Cornell et al. (2024), each window of the E2E DAST model provides local speaker transcription and diarization with time-averaged speaker embeddings. Meanwhile, Mao et al. (2020) demonstrates the advantages of jointly learning speaker embeddings and transcriptions for hour-long multi-speaker podcasts, using lexical cues to improve speaker label assignment. After processing all windows, final diarization and speaker embeddings can be obtained by clustering these time-averaged embeddings. This method eliminates the need for additional speaker embedding models and improves computational efficiency, but its performance is heavily dependent on embedding quality.

6. Evaluation

Multi-speaker ASR research relies on both real-world and simulated datasets. Real-world datasets are collected from natural conversation scenarios such as meetings or phone calls; they provide authentic acoustic conditions, spontaneous speech patterns, and natural overlaps. However, they are often small, noisy and domain-specific, posing challenges for early-stage model training and broader applicability. Simulated datasets are created by overlapping single-speaker recordings and introducing noise at the configurable overlapping rate and noise level, enabling scalable as well as controlled training and evaluation of overlap and noise robustness. To improve realism, Yang et al. (2023) leverages statistical language models to guide overlap patterns. Additionally, Moriya et al. (2024) introduce on-the-fly data generation to support dynamic parameter adjustment and improve memory efficiency. Table 1 highlights commonly used multi-speaker ASR datasets.

Evaluating multi-speaker ASR systems requires measuring both transcription accuracy and the system's ability to assign speaker labels. **Word error rate (WER)** evaluates overall transcription accuracy by measuring insertions, deletions, and substitutions.

Table 1
Common real-world and simulated datasets. # **Speakers** refers to the number of speakers per recording in real datasets, and the exact mixed speaker number per sample in simulated datasets.

	Dataset	Scenario	Hours	Language	# Speakers
Real	AMI	Meetings	100	English	3–5
	AliMeeting (Yu et al., 2022b)	Meetings	120	Chinese	2–4
	CallHome	Phone calls	60	Multilingual	2
	LibriCSS (Chen et al., 2020)	Meetings	10	English	8
Sim	WSJ0-2mix (Hershey et al., 2016)	Read speech	45	English	2
	LibriMix (Cosentino et al., 2020)	Read speech	500	English	2–3
	LibriHeavyMix (Jin et al., 2024)	Read speech	20,000	English	2–4

Character Error Rate (CER) is adopted for languages with non-alphabetic scripts, while **Sentence Error Rate (SER)** captures the proportion of sentences containing any error. In multi-speaker settings, these metrics can be extended with special tokens for speaker changes, overlapping speech, timestamps, and speaker identities. This extension enables evaluation not only for content accuracy but also of the system's capability to handle speaker turns and overlaps.

In addition to standard WER, several evaluation metrics have been proposed to account for speaker distinctions in multi-speaker ASR. **Concatenated minimum-permutation WER (cpWER)** (Watanabe et al., 2020) concatenates utterances per speaker and computes the WER across all permutations of hypotheses and references, selecting the lowest:

$$\text{cpWER} = \min_{\pi \in \mathcal{P}(S)} \frac{\sum_{s=1}^{S} \text{WER}(H^s, R^{\pi(s)})}{S}, \tag{5}$$

where $\mathcal{P}(S)$ denotes all permutations of $\{1, \dots, S\}$, and H^s, $R^{\pi(s)}$ denote the hypothesis and reference for speaker s, respectively. cpWER jointly reflects transcription and speaker assignment accuracy. **Speaker-attributed WER (SA-WER)** further enforces speaker correspondence by requiring that hypotheses be matched with the reference of the correct speaker label. It penalizes speaker assignment errors and provides a stricter evaluation of speaker-aware performance. Among these, WER, cpWER, and SA-WER are increasingly strict in evaluation criteria. To incorporate temporal alignment, von Neumann et al. (2023) propose **time-constrained cpWER (tcpWER)**, which restricts word matching to a fixed time window in addition to speaker permutation, which requires real or estimated token timestamps.

6.1. Performance comparison of E2E approaches

This section compares various techniques and their accuracy across standard multi-speaker benchmarks. We report results on both real-world data (AMI) and simulated mixtures (LibriSpeechMix and LibriMix), as shown in Table 2. LibrispeechMix (Kanda et al., 2020b) provides the standard dev/test set for evaluation. To facilitate comparison between methods, we adopt cpWER as the primary evaluation metric to include speaker determination in WER. Note that some methods only report results on standard WER and SA-WER (see the results marked with [a] and [b]); those accuracies cannot be directly compared with those of other methods. All results are reported directly from the original papers. Our analysis examines how different scenarios and training approaches impact results.

The results demonstrate no consistent superiority between SISO and SIMO approaches in multi-speaker ASR. For instance, SISO methods (Kanda et al., 2021; Shi et al., 2024d) outperform SIMO (Huang et al., 2023; Meng et al., 2023) on AMI and LibriMix respectively, while SIMO method (Meng et al., 2024b) surpasses SISO (Kanda et al., 2021d) on LibriSeechMix. Moreover, cpWER has not shown consistent improvement throughout the six-year development period of end-to-end multi-speaker ASR approaches. The currently best performance on AMI comes from a relatively small 50M model trained on extensive 900k hours of simulated mixture data (Kanda et al., 2021) in 2021. This suggests a stagnation in real-world benchmark progress, despite many recent methodological proposals.

In general, contemporary research on multi-speaker ASR is less about pursuing marginal WER improvements on specific benchmark datasets. Instead, it tends to follow three trends: (1) **Understanding scenario-dependent factors**: For example, LibriMix methods demonstrate this evolution: while early works (Kanda et al., 2020b) used no speaker enrollment, subsequent studies (Kanda et al., 2020a, 2021d) introduced speaker-attributed (SA) methods with enrollment. This was further advanced by CIF SA-SOT (Fan et al., 2024), which achieved speaker attribution without enrollment. Meanwhile, Transcribe-to-Diarize (Kanda et al., 2022) extends the approach from Kanda et al. (2020a) to long-form scenarios. (2) **Exploring novel architectures and information fusion methods**, such as cross-speaker encoding for SIMO (Kang et al., 2024), enhancing CTC loss with speaker attribute (Kang et al., 2025). (3) **Efficiently adapting single-speaker foundation ASR systems to multi-speaker settings with minimal training**, such as Adapted USM (Li et al., 2023c), WavLM/wTSE&JSM (Huang et al., 2023), Whisper-SS-TTI (Meng et al., 2024b), W2V-Sidecar (Meng et al., 2023). The # Parameters (Train/Total) in tables partially reflect the reduced training effort, though they do not fully capture cases like META-CAT (Wang et al., 2025), which achieves efficiency through fewer training epochs despite its larger trained size.

Table 2
cpWER of multi-speaker ASR models on **AMI**, **LibrispeechMix**, and **LibriMix** corpus. **SDM** denotes single distant microphone; **IHM** denotes mixture of independent headset microphones. **Mix Pre. Hrs** indicates the hours of simulated mixture data used for pretraining. **Params Tr/To** indicates trainable/total model parameters in millions. **# Tr. Spk** refers to the speaker configuration used during training. Methods are grouped by input granularity and listed chronologically.

Dataset	Model	Input Gran.	Framework	Spk. Enrl.	Mix Pre. Hrs	Params Tr/To	SDM dev	SDM eval	IHM dev	IHM eval
AMI	Conformer AED (Kanda et al., 2021)	Utterance group	SISO	×	900k	50/50	18.4	21.2	13.5	14.9
	WavLM/wTSE&JSM (Huang et al., 2023)	Utterance group	Hybrid	✓	58	13/108	–	–	–	28.4[b]
	Adapted USM (Li et al., 2023c)	Utterance group	SISO	×	–	84/1630	–	21.4	–	–
	META-CAT (Wang et al., 2025)	Utterance Group	SISO	×	–	600/723	–	–	–	22.8
	CMT-LLM (He et al., 2025)	Utterance Group	SISO	×	–	24/7300	30.0	32.9	21.9	22.8
	Transcribe-to-Diarize (Kanda et al., 2022)	Long-Term	SISO	×	–	146/146	22.6	24.9	15.9	16.4
	SLIDAR (Cornell et al., 2024)	Long-Term	SISO	×	5k	655/655	21.8	24.5	14.2	15.6
Dataset	Model	Input Gran.	Framework	Spk. Enrl.	# Tr. Spk	Params Tr/To	2spk dev	2spk eval	3spk dev	3spk eval
LibrispeechMix	LSTM SOT (Kanda et al., 2020b)	Utterance Group	SISO	×	1, 2, 3	136/136	–	11.2[a]	–	24.0[a]
	Transformer SOT (Kanda et al., 2021d)	Utterance Group	SISO	×	1, 2, 3	129/129	–	4.9[a]	–	6.2[a]
	LSTM SA-ASR (Kanda et al., 2020a)	Utterance Group	SISO	✓	1, 2, 3	146/146	–	9.9[b]	–	23.1[b]
	SA-MBR (Kanda et al., 2021b)	Utterance Group	SISO	✓	1, 2, 3	146/146	–	9.5[b]	–	20.7[b]
	Transformer SA-ASR (Kanda et al., 2021d)	Utterance Group	SISO	✓	1, 2, 3	142/142	–	6.4[b]	–	8.5[b]
	W2V-Sidecar (Meng et al., 2023)	Utterance Group	SIMO	×	2	9/104	6.0	5.7	–	–
	CSE Network (Kang et al., 2024)	Utterance Group	SIMO	×	1, 2	33/33	11.8	10.7	24.2	24.3
	CIF SA-SOT (Fan et al., 2024)	Utterance Group	SISO	×	2	136/136	–	3.4	–	–
	SA-CTC SOT (Kang et al., 2025)	Utterance Group	SISO	×	2	56/56	3.9	4.1	22.6	22.6
	SD-CTC SOT (Sakuma et al., 2025)	Utterance Group	SISO	×	1,2	114/114	–	3.5	–	–
	Whisper-SS-TTI (Meng et al., 2024b)	Utterance Group	SIMO	×	2, 3	9/250	–	5.2	–	8.6
	Whisper-SS-TTI (Meng et al., 2024b)	Utterance Group	SIMO	×	2, 3	13/779	–	4.0	–	7.5
	Whisper-SS-TTI (Meng et al., 2024b)	Utterance Group	SIMO	×	2, 3	18/1950	–	3.4	–	6.8
	MT-LLM (Meng et al., 2024a)	Utterance Group	SISO	×	2, 3	76.6/7550	–	5.2	–	10.2
Dataset	Model	Input Gran.	Framework	Spk. Enrl.	# Tr. Spk	Params Tr/To	2spk dev	2spk eval	3spk dev	3spk eval
LibrMix	WavLM/wTSE&JSM (Huang et al., 2023)	Utterance Group	SIMO	✓	2	13/108	–	10.7[b]	–	–
	Whisper-SS-TTI (Meng et al., 2024b)	Utterance Group	SIMO	×	2, 3	9/250	–	9.4	–	26.8
	Whisper-SS-TTI (Meng et al., 2024b)	Utterance Group	SIMO	×	2, 3	13/779	–	6.6	–	21.5
	Whisper-SS-TTI (Meng et al., 2024b)	Utterance Group	SIMO	×	2, 3	18/1950	–	4.7	–	16.8
	W2V-Sidecar (Meng et al., 2023)	Utterance Group	SIMO	×	2	9/104	7.7	8.1	–	–

(*continued on next page*)

Table 2 (*continued*).

	GEncSep (Shi et al., 2024d)	Utterance Group	SISO	×	2, 3	–	6.4	6.6	13.3	13.1
	CMT-LLM (He et al., 2025)	Utterance Group	SISO	×	2	24/7300	7.3	8.1	–	–
	Hypothesis stitcher (Chang et al., 2021)	Long-term	SISO	×	1 - 6	–	–	11.5	–	13.4
	Hypothesis clustering (Kashiwagi et al., 2024)	Long-term	SISO	✓	1, 2, 3	–	–	8.2[b]	–	21.5[b]

[a] Signifies the paper reported results as standard WER.
[b] Signifies the paper reported results as SA-WER.

Currently, limited open-source availability makes fair comparisons difficult and slows research progress. Lots of methods show their improvements by comparing with their own baseline systems that do not include the new architectures. This highlights the importance of developing standardized benchmarks and sharing reproducible models as a community. A coordinated effort towards standardized open-source benchmark suites and consistent experiment protocols would substantially improve comparability and accelerate progress in multi-speaker ASR.

In addition to standardized benchmarks, several open-source toolkits provide practical starting points for building and evaluating multi-speaker ASR systems. ESPnet (Watanabe et al., 2018), SpeechBrain Ravanelli et al. (2021), NVIDIA NeMo (Kuchaiev et al., 2019), and Pyannote (Bredin et al., 2020) respectively provide SOT recipes, audio-visual ASR pipelines, integrated diarization–ASR modules, and state-of-the-art diarization components. These resources further support reproducibility and practical usage.

7. Conclusion and future directions

Recent research on multi-speaker ASR has increasingly focused on end-to-end (E2E) methods, aiming to produce more accurate speaker-distinguishable transcriptions in scenarios such as phone calls, meetings, and group discussions. E2E methods can overcome the limitations of traditional modular systems, such as error propagation and failure to leverage cross-task synergies. Furthermore, progress in single-speaker ASR, speech separation, and diarization has provided the architectural foundation and pretrained model for initialization that facilitate the training of a multi-speaker ASR system, although with the scarcity of large-scale multi-speaker data.

This review provided an overview of end-to-end multi-speaker ASR approaches, from segment-level methods to processing continuous long-form recordings. Various architectural designs (Section 3) were described for how to disentangle different speakers, and how to effectively leverage contextual cues from mixed speech, such as inter-speaker, overlapping, and temporal dependencies. Beyond acoustic-only designs, Section 4 surveyed emerging multi-modal extensions, ranging from audio-visual fusion to LLM-based text conditioning, which provide complementary cues for speaker attribution and linguistic disambiguation, especially under heavy overlap or noise. Recent work on long-form processing (Section 5), which has further improved applicability to real-world cases, was also presented. Our accuracy comparisons indicate that, while end-to-end models often outperform modular approaches, no single architecture consistently outperforms others across end-to-end designs. Ongoing research continues to explore more sophisticated designs leveraging complex interaction patterns, while integrating advances in large-scale pretrained ASR models. Also, practical deployment factors, such as latency, streaming constraints, and computational cost, are becoming increasingly important and warrant further attention.

Overall, the future of the multi-speaker ASR lies in developing more robust, adaptive, and scalable systems for various scenarios. While recent designs have made significant progress, key challenges remain in robustly handling overlapping speech, and designing effective SISO/SIMO hybrids (Section 3.4). Moreover, future research may also explore adaptive modeling, such as transforming single-speaker models into multi-speaker systems without performance degradation, and optional speaker enrollment. In addition, advancing multi-speaker understanding — through joint training with downstream objectives, cross-modal fusion (Section 4), and injecting multi-speaker ASR into a unified multi-task foundation model.

CRediT authorship contribution statement

Xinlu He: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Jacob Whitehill:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jacob Whitehill reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805, and also from grant #2046505. The opinions expressed are those of the authors and do not represent the views of the NSF. Finally, we are grateful for the feedback from the anonymous reviewers.

Data availability

No data was used for the research described in the article.

References

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O., 2012. Speaker diarization: A review of recent research. IEEE Trans. Audio Speech Lang. Process. http://dx.doi.org/10.1109/TASL.2011.2125954.

Arora, S., Singh, R., 2012. Automatic speech recognition: A review. Int. J. Comput. Appl. http://dx.doi.org/10.5120/9722-4190.

Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv:2006.11477.

Braga, O., Makino, T., Siohan, O., Liao, H., 2020. End-to-end multi-person audio/visual automatic speech recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6994–6998. http://dx.doi.org/10.1109/ICASSP40776.2020.9053974.

Braga, O., Siohan, O., 2022a. Best of both worlds: Multi-task audio-visual automatic speech recognition and active speaker detection. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6047–6051. http://dx.doi.org/10.1109/ICASSP43922.2022.9746036.

Braga, O., Siohan, O., 2022b. A closer look at audio-visual multi-person speech recognition and active speaker selection. arXiv:2205.05684. URL https://arxiv.org/abs/2205.05684.

Bredin, H., et al., 2020. Pyannote.audio: neural building blocks for speaker diarization. In: Proc. ICASSP.

Chang, X., Kanda, N., Gaur, Y., Wang, X., Meng, Z., Yoshioka, T., 2021. Hypothesis stitcher for end-to-end speaker-attributed ASR on long-form multi-talker recordings. In: ICASSP. http://dx.doi.org/10.1109/ICASSP39728.2021.9414432.

Chang, X., Qian, Y., Yu, K., Watanabe, S., 2019. End-to-end monaural multi-speaker ASR system without pretraining. In: ICASSP. http://dx.doi.org/10.1109/ICASSP.2019.8682822.

Chang, X., Zhang, W., Qian, Y., Roux, J.L., Watanabe, S., 2020. End-to-end multi-speaker speech recognition with transformer. In: ICASSP. http://dx.doi.org/10.1109/ICASSP40776.2020.9054029.

Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., Wu, J., Li, J., 2020. Continuous speech separation: dataset and analysis. arXiv:2001.11482.

Chiu, C.-C., Han, W., Zhang, Y., Pang, R., Kishchenko, S., Nguyen, P., Narayanan, A., Liao, H., Zhang, S., Kannan, A., Prabhavalkar, R., Chen, Z., Sainath, T.N., Wu, Y., 2019. A comparison of end-to-end models for long-form speech recognition. ASRU.

Cornell, S., Jung, J.-W., Watanabe, S., Squartini, S., 2024. One model to rule them all? Towards end-to-end joint speaker diarization and speech recognition. In: ICASSPs.

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., Vincent, E., 2020. LibriMix: An open-source dataset for generalizable speech separation. arXiv:2005.11262.

Denisov, P., Vu, N.T., 2019. End-to-end multi-speaker speech recognition using speaker embeddings and transfer learning. ArXiv.

Díez, M., Burget, L., Matejka, P., 2018. Speaker diarization based on Bayesian HMM with eigenvoice priors. In: The Speaker and Language Recognition Workshop.

Fan, Z., Dong, L., Zhang, J., Lu, L., Ma, Z., 2024. SA-SOT: Speaker-aware serialized output training for multi-talker ASR. In: ICASSP.

Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., Watanabe, S., 2019. End-to-end neural speaker diarization with permutation-free objectives. In: Interspeech.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R., 2020. Conformer: Convolution-augmented transformer for speech recognition. arXiv:2005.08100.

He, J., Sawada, N., Miyazaki, K., Toda, T., 2025. CMT-LLM: Contextual multi-talker ASR utilizing large language models. pp. 2575–2579. http://dx.doi.org/10.21437/Interspeech.2025-943.

Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S., 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In: ICASSP. http://dx.doi.org/10.1109/ICASSP.2016.7471631.

Hershey, J., Rennie, S., Olsen, P., Kristjansson, T., 2010. Superhuman multi-talker speech recognition: A graphical modeling approach. Comput. Speech Lang. http://dx.doi.org/10.1016/j.csl.2008.11.001.

Hsu, W., Bolte, B., Tsai, Y.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. arXiv:2106.07447.

Hu, E.J., Shen, y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations.

Huang, W.R., Chang, S.-y., Rybach, D., Prabhavalkar, R., Sainath, T.N., Allauzen, C., Peyser, C., Lu, Z., 2022. E2E segmenter: Joint segmenting and decoding for long-form ASR. In: Interspeech.

Huang, Z., Raj, D., García, P., Khudanpur, S., 2023. Adapting self-supervised models to multi-talker speech recognition using speaker embeddings. In: ICASSP. http://dx.doi.org/10.1109/ICASSP49357.2023.10097139.

Jin, Z., Yang, Y., Shi, M., Kang, W., Yang, X., Yao, Z., Kuang, F., Guo, L., Meng, L., Lin, L., Xu, Y., Zhang, S.-X., 2024. LibriheavyMix: A 20,000-hour dataset for single-channel reverberant multi-talker speech separation, ASR and speaker diarization. http://dx.doi.org/10.21437/Interspeech.2024-90.

Kanda, N., Chang, X., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2021a. Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings. In: SLT. http://dx.doi.org/10.1109/SLT48900.2021.9383600.

Kanda, N., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Zhou, T., Yoshioka, T., 2020a. Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. ArXiv.

Kanda, N., Gaur, Y., Wang, X., Meng, Z., Yoshioka, T., 2020b. Serialized output training for end-to-end overlapped speech recognition. In: Interspeech.

Kanda, N., Meng, Z., Lu, L., Gaur, Y., Wang, X., Chen, Z., Yoshioka, T., 2021b. Minimum Bayes risk training for end-to-end speaker-attributed ASR. In: ICASSP. http://dx.doi.org/10.1109/ICASSP39728.2021.9415062.

Kanda, N., Xiao, X., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2022. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR. In: ICASSP.

Kanda, N., Xiao, X., Wu, J., Zhou, T., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2021c. A comparative study of modular and joint approaches for speaker-attributed ASR on monaural long-form audio. In: ASRU. http://dx.doi.org/10.1109/ASRU51503.2021.9687974.

Kanda, N., Ye, G., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2021d. End-to-end speaker-attributed ASR with transformer. http://dx.doi.org/10.21437/Interspeech.2021-101.

Kanda, N., Ye, G., Wu, Y., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2021. Large-scale pre-training of end-to-end multi-talker ASR for meeting transcription with single distant microphone.

Kang, J., Meng, L., Cui, M., Guo, H., Wu, X., Liu, X., Meng, H., 2024. Cross-speaker encoding network for multi-talker speech recognition. ICASSP.

Kang, J., Meng, L., Cui, M., Wang, Y., Wu, X., Liu, X., Meng, H., 2025. Disentangling speakers in multi-talker speech recognition with speaker-aware CTC. In: ICASSP. http://dx.doi.org/10.1109/ICASSP49660.2025.10888841.

Kashiwagi, Y., Futami, H., Tsunoo, E., Arora, S., Watanabe, S., 2024. Hypothesis clustering and merging: Novel MultiTalker speech recognition with speaker tokens. http://dx.doi.org/10.48550/arXiv.2409.15732.

Kolbæk, M., Yu, D., Tan, Z.-H., Jensen, J.H., 2017. Multi-talker speech separation and tracing with permutation invariant training of deep recurrent neural networks. ArXiv.

Kuchaiev, O., et al., 2019. Nemo: a toolkit for building AI applications using neural modules. In: Proc. ICASSP.

Landini, F., Profant, J., Diez, M., Burget, L., 2020. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks. http://dx.doi.org/10.48550/arXiv.2012.14952.

Latif, S., Shoukat, M., Shamshad, F., Usama, M., Cuayáhuitl, H., Schuller, B., 2023. Sparks of large audio models: A survey and outlook. http://dx.doi.org/10.48550/arXiv.2308.12792.

Li, Z., He, X., Whitehill, J., 2023a. Compositional clustering: Applications to multi-label object recognition and speaker identification. Pattern Recognit. http://dx.doi.org/10.1016/j.patcog.2023.109829.

Li, J., Li, C., Wu, Y., Qian, Y., 2023b. Robust audio-visual ASR with unified cross-modal attention. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10096893.

Li, C., Qian, Y., Chen, Z., Kanda, N., Wang, D., Yoshioka, T., Qian, Y., Zeng, M., 2023c. Adapting multi-lingual ASR models for handling multiple talkers. ArXiv.

Li, T., Wang, F., Guan, W., Huang, L., Hong, Q., Li, L., 2024. Improving multi-speaker ASR with overlap-aware encoding and monotonic attention. In: ICASSP.

Li, Z., Whitehill, J., 2021. Compositional embedding models for speaker identification and diarization with simultaneous speech from 2+ speakers. In: ICASSP. http://dx.doi.org/10.1109/ICASSP39728.2021.9413752.

Liang, Y., Yu, F., Li, Y., Guo, P., Zhang, S., Chen, Q., Xie, L., 2023. BA-SOT: Boundary-aware serialized output training for multi-talker ASR. In: Interspeech.

Lin, Y., Cheng, M., Li, Z., Tang, B., Li, M., 2025. Diarization-aware multi-speaker automatic speech recognition via large language models. URL https://api.semanticscholar.org/CorpusID:279244049.

Lin, Y., Du, Z., Zhang, S., Yu, F., Zhao, Z., Wu, F., 2022. Separate-to-recognize: Joint multi-target speech separation and speech recognition for speaker-attributed ASR. In: ISCSLP. http://dx.doi.org/10.1109/ISCSLP57327.2022.10037902.

Lu, L., Kanda, N., Li, J., Gong, Y., 2021. Streaming end-to-end multi-talker speech recognition. IEEE Signal Process. Lett. 28.

Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. arXiv:1910.06379.

Luo, Y., Mesgarani, N., 2017. TasNet: time-domain audio separation network for real-time, single-channel speech separation. arXiv:1711.00541.

Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M., 2023. Auto-AVSR: Audio-visual speech recognition with automatic labels. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10096889.

Makishima, N., Kawata, N., Ihori, M., Tanaka, T., Orihashi, S., Ando, A., Masumura, R., 2024. SOMSRED: Sequential output modeling for joint multi-talker overlapped speech recognition and speaker diarization. In: Interspeech 2024.

Makishima, N., Kawata, N., Yamane, T., Ihori, M., Tanaka, T., Suzuki, S., Orihashi, S., Masumura, R., 2025. Unified audio-visual modeling for recognizing which face spoke when and what in multi-talker overlapped speech and video. In: Interspeech 2025. pp. 1838–1842. http://dx.doi.org/10.21437/Interspeech.2025-451.

Makishima, N., Suzuki, K., Suzuki, S., Ando, A., Masumura, R., 2023. Joint autoregressive modeling of end-to-end multi-talker overlapped speech recognition and utterance-level timestamp prediction. In: Interspeech.

Malik, M., Malik, M., Mehmood, K., Makhdoom, I., 2021. Automatic speech recognition: a survey. Multimedia Tools Appl. http://dx.doi.org/10.1007/s11042-020-10073-7.

Mao, H.H., Li, S., McAuley, J., Cottrell, G., 2020. Speech recognition and multi-speaker diarization of long conversations. In: Interspeech.

Masumura, R., Okamura, D., Makishima, N., Ihori, M., Takashima, A., Tanaka, T., Orihashi, S., 2021. Unified autoregressive modeling for joint end-to-end multi-talker overlapped speech recognition and speaker attribute estimation. In: Interspeech.

Meng, L., Hu, S., Kang, J., Li, Z., Wang, Y., Wu, W., Wu, X., Liu, X., Meng, H., 2024a. Large language model can transcribe speech in multi-talker scenarios with versatile instructions. arXiv:2409.08596.

Meng, L., Kang, J., Cui, M., Wang, Y., Wu, X., Meng, H., 2023. A sidecar separator can convert a single-talker speech recognition system to a multi-talker one. In: ICASSP. http://dx.doi.org/10.1109/ICASSP49357.2023.10095295.

Meng, L., Kang, J., Wang, Y., Jin, Z., Wu, X., Liu, X., Meng, H., 2024b. Empowering whisper as a joint multi-talker and target-talker speech recognition system. http://dx.doi.org/10.21437/Interspeech.2024-971.

Moriya, T., Horiguchi, S., Delcroix, M., Masumura, R., Ashihara, T., Sato, H., Matsuura, K., Mimura, M., 2024. Alignment-free training for transducer-based multi-talker ASR. http://dx.doi.org/10.48550/arXiv.2409.20301.

Park, T.J., Medennikov, I., Dhawan, K., Wang, W., Huang, H., Koluguri, N.R., Puvvada, K.C., Balam, J., Ginsburg, B., 2024. Sortformer: Seamless integration of speaker diarization and ASR by bridging timestamps and tokens. ArXiv.

Peng, Y., Tian, J., Jung, J.-W., Watanabe, S., et al., 2024. OWSM v3.1: Better and faster open whisper-style speech models based on E-branchformer. In: Proc. Interspeech.

Polok, A., Klement, D., Kocour, M., Han, J., Landini, F., Yusuf, B., Wiesner, M., Khudanpur, S., Černocký, J., Burget, L., 2026. DiCoW: Diarization-conditioned whisper for target speaker automatic speech recognition. Comput. Speech Lang. 95, 101841. http://dx.doi.org/10.1016/j.csl.2025.101841, URL https://www.sciencedirect.com/science/article/pii/S088523082500066X.

Prabhavalkar, R., Hori, T., Sainath, T.N., Schlüter, R., Watanabe, S., 2024. End-to-end speech recognition: A survey. IEEE/ACM Trans. Audio Speech Lang. Process. http://dx.doi.org/10.1109/TASLP.2023.3328283.

Qian, Y.-m., Weng, C., Chang, X.-k., Wang, S., Yu, D., 2018. Past review, current progress, and challenges ahead on the cocktail party problem. Front. Info. Technol. Electron. Eng. http://dx.doi.org/10.1631/FITEE.1700814.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision. In: ICML. JMLR.org.

Ravanelli, M., et al., 2021. SpeechBrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624.

Rose, R., Chang, O., Siohan, O., 2023. Cascaded encoders for fine-tuning ASR models on overlapped speech. arXiv:2306.16398.

Sakuma, A., Sato, H., Sugano, R., Kumano, T., Kawai, Y., Ogawa, T., 2025. Speaker-distinguishable CTC: Learning speaker distinction using CTC for multi-talker speech recognition. In: Proc. Interspeech.

Seki, H., Hori, T., Watanabe, S., Le Roux, J., Hershey, J.R., 2018. A purely end-to-end system for multi-speaker speech recognition. In: Proceedings of the Annual Meeting of the Assoc. for Comp. Linguistics. http://dx.doi.org/10.18653/v1/P18-1244.

Sell, G., Garcia-Romero, D., 2014. Speaker diarization with plda i-vector scoring and unsupervised calibration. SLT.

Seo, P.H., Nagrani, A., Schmid, C., 2023. Avformer: Injecting vision into frozen speech models for zero-shot AV-ASR. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 22922–22931.

Settle, S., Roux, J.L., Hori, T., Watanabe, S., Hershey, J.R., 2018. End-to-end multi-speaker speech recognition. In: ICASSP. http://dx.doi.org/10.1109/ICASSP.2018.8461893.

Shakeel, M., Sudo, Y., Peng, Y., Lin, C.-J., Watanabe, S., 2025. Unifying diarization, separation, and ASR with multi-speaker encoder. URL https://openreview.net/forum?id=5oaUMZEjWe.

Shi, M., Du, Z., Chen, Q., Yu, F., Li, Y., Zhang, S., Zhang, J., Dai, L.-R., 2023. CASA-ASR: Context-aware speaker-attributed ASR. In: Interspeech. http://dx.doi.org/10.21437/Interspeech.2023-601.

Shi, H., Fujita, Y., Mizumoto, T., Liu, L., Kojima, A., Sudo, Y., 2025. Serialized output prompting for large language model-based multi-talker speech recognition. arXiv:2509.04488. URL https://arxiv.org/abs/2509.04488.

Shi, H., Gao, Y., Ni, Z., Kawahara, T., 2024a. Serialized speech information guidance with overlapped encoding separation for multi-speaker automatic speech recognition. http://dx.doi.org/10.48550/arXiv.2409.00815.

Shi, M., Jin, Z., Xu, Y., Xu, Y., Zhang, S.-X., Wei, K., Shao, Y., Zhang, C., Yu, D., 2024b. Advancing multi-talker ASR performance with large language models. arXiv:2408.17431.

Shi, M., Jin, Z., Xu, Y., Xu, Y., Zhang, S.-X., Wei, K., Shao, Y., Zhang, C., Yu, D., 2024c. Advancing multi-talker ASR performance with large language models. SLT.

Shi, Y., Li, L., Yin, S., Wang, D., Han, J., 2024d. Serialized output training by learned dominance. ArXiv.

Trinh, V.A., He, X., Whitehill, J., 2025. Improving named entity transcription with contextual LLM-based revision. arXiv:2506.10779. URL https://arxiv.org/abs/2506.10779.

Trinh, V.A., Southwell, R., Guan, Y., He, X., Wang, Z., Whitehill, J., 2024. Discrete multimodal transformers with a pretrained large language model for mixed-supervision speech processing. arXiv:2406.06582. URL https://arxiv.org/abs/2406.06582.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention is all you need. arXiv:1706.03762.

von Neumann, T., Boeddeker, C., Delcroix, M., Haeb-Umbach, R., 2023. MeetEval: A toolkit for computation of word error rates for meeting transcription systems. http://dx.doi.org/10.48550/arXiv.2307.11394.

von Neumann, T., Boeddeker, C., Drude, L., Kinoshita, K., Delcroix, M., Nakatani, T., Haeb-Umbach, R., 2020. Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR. ArXiv.

Wang, W., Dhawan, K., Park, T., Puvvada, K., Medennikov, I., Majumdar, S., Huang, H., Balam, J., Ginsburg, B., 2024. Resource-efficient adaptation of speech foundation models for multi-speaker ASR. http://dx.doi.org/10.1109/SLT61566.2024.10832215.

Wang, Z.-Q., Roux, J.L., Hershey, J.R., 2018. Alternative objective functions for deep clustering. In: ICASSP. http://dx.doi.org/10.1109/ICASSP.2018.8462507.

Wang, J., Wang, W., Dhawan, K., Park, T., Kim, M., Medennikov, I., Huang, H., Koluguri, N., Balam, J., Ginsburg, B., 2025. META-CAT: Speaker-informed speech embeddings via meta information concatenation for multi-talker ASR. In: ICASSP 2025. http://dx.doi.org/10.1109/ICASSP49660.2025.10889841.

Watanabe, S., Mandel, M.I., Barker, J., Vincent, E., 2020. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. arXiv:2004.09249.

Watanabe, S., et al., 2018. Espnet: End-to-end speech processing toolkit. In: Proc. Interspeech.

Wu, Y., Li, C., Yang, S., Wu, Z., Qian, Y., 2021. Audio-visual multi-talker speech recognition in a cocktail party. In: Interspeech 2021. pp. 3021–3025. http://dx.doi.org/10.21437/Interspeech.2021-2128.

Yang, M., Kanda, N., Wang, X., Wu, J., Sivasankaran, S., Chen, Z., Li, J., Yoshioka, T., 2023. Simulating realistic speech overlaps improves multi-talker ASR. In: ICASSP.

Yu, D., Chang, X., Qian, Y., 2017a. Recognizing multi-talker speech with permutation invariant training. In: Interspeech.

Yu, F., Du, Z., Zhang, S., Lin, Y., Xie, L., 2022a. A comparative study on speaker-attributed automatic speech recognition in multi-party meetings. In: Interspeech.

Yu, D., Kolbæk, M., Tan, Z.-H., Jensen, J., 2017b. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: ICASSP. http://dx.doi.org/10.1109/ICASSP.2017.7952154.

Yu, F., Zhang, S., Fu, Y., Xie, L., Zheng, S., Du, Z., Huang, W., Guo, P., Yan, Z., Ma, B., Xu, X., Bu, H., 2022b. M2Met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In: ICASSP. http://dx.doi.org/10.1109/ICASSP43922.2022.9746465.

Zhang, W., Chang, X., Qian, Y., 2019. Knowledge distillation for end-to-end monaural multi-talker ASR system. In: Interspeech 2019. http://dx.doi.org/10.21437/Interspeech.2019-3192.

Zhang, W., Chang, X., Qian, Y., Watanabe, S., 2020. Improving end-to-end single-channel multi-talker speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. http://dx.doi.org/10.1109/TASLP.2020.2988423.

Zheng, L., Zhu, H., Tian, S., Zhao, Q., Li, T., 2024. Unsupervised domain adaptation on end-to-end multi-talker overlapped speech recognition. IEEE Signal Process. Lett. http://dx.doi.org/10.1109/LSP.2024.3487795.