From Speech to Interaction: Analyzing Multimodal Systems in Cocktail-Party Scenarios

Thai Binh Nguyen
Karlsruhe Institute of Technology
Karlsruhe, Germany
thai-binh.nguyen@kit.edu

Zhaolin Li
Karlsruhe Institute of Technology
Karlsruhe, Germany
zhaolin.li@kit.edu

Jan Niehues
Karlsruhe Institute of Technology
Karlsruhe, Germany
jan.niehues@kit.edu

Alexander Waibel
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
waibel@cs.cmu.edu



Figure 1: Example recording setup for a multi-party conversation session in a cocktail-party scenario.

Abstract

Humans have the remarkable ability to engage in spontaneous informal conversations and selectively attend to individual speakers while filtering out competing speech from nearby conversations. This "cocktail party" scenario still presents severe challenges to speech recognition systems. The CHiME-9 MCoRec task provides a testbed where systems must recognize groups of speakers and transcribe each of their conversations from audio-visual input. In this work, we analyze a diverse set of systems, representing different design directions for addressing the cocktail-party scenario, where the best system achieves up to 57% relative error reduction. We identify three main strategies: (1) explicit or implicit audio-visual target speech separation, (2) improved audio-visual speech recognition for each target speaker, and (3) the use of large language models to group speakers into conversations and enhance conversational consistency. Our analysis shows that these directions address complementary failure modes of the cocktail-party problem, and that high speech overlap alone does not explain performance differences, challenging the common assumption that overlap is the primary source of difficulty in cocktail-party recognition.

This work is licensed under a Creative Commons Attribution 4.0 International License.
ICMI '26, Napoli, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
https://doi.org/10.1145/3776574.3831133

Keywords

multimodal learning, speech recognition, speaker attribution, conversational analysis, cocktail-party problem

ACM Reference Format:
Thai Binh Nguyen, Zhaolin Li, Jan Niehues, and Alexander Waibel. 2026. From Speech to Interaction: Analyzing Multimodal Systems in Cocktail-Party Scenarios. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3776574.3831133

1 Introduction

Recognizing speech in real-world multi-party conversations remains a challenging problem, particularly in so-called cocktail-party scenarios where parallel conversations take place simultaneously in a shared environment. In such settings, systems must handle severe speech overlap, rapid turn-taking, and strong speaker interference, making it difficult to determine not only what is being said, but also who is speaking and with whom they are interacting.

The CHiME-9 Multi-Modal Context-aware Recognition (MCoRec) task provides a realistic benchmark for this problem. Unlike traditional speech recognition tasks that focus on single-speaker or single-conversation settings, MCoRec requires systems to jointly transcribe each speaker's speech and group speakers into their respective conversations from audio-visual recordings. This introduces additional challenges beyond recognition accuracy, as systems must also resolve speaker ambiguity and conversational structure under highly overlapped and fragmented interactions.

Recent submissions to the MCoRec challenge explore diverse design strategies to address these challenges, including improvements in audio-visual representation learning, explicit target-speaker extraction, and the use of large language models (LLMs) for reasoning about conversational context. While these approaches achieve varying levels of success, it remains unclear which design choices are most effective and what factors truly drive performance differences in cocktail-party recognition.

In this work, we present a comparative analysis of multiple CHiME-9 MCoRec systems, with the goal of understanding how different modeling strategies address the core difficulties of multi-talker recognition. Rather than focusing on individual components in isolation, we analyze systems at the pipeline level and organize them into a unified taxonomy based on how they handle target-speaker recognition, speech separation, and conversation clustering. In addition, we perform condition-aware evaluation to examine system behavior under varying conversational factors, such as speaker activity, number of speakers, and conversational structure.

Our analysis reveals several key findings. First, different system designs address distinct failure modes of the cocktail-party problem, and no single strategy dominates across all conditions. Second, conversation grouping appears to be a comparatively easier problem, as it remains robust to imperfect transcriptions, suggesting that coarse semantic and temporal cues are often sufficient for reliable clustering. Finally, contrary to common assumptions, high speech overlap alone does not fully explain performance differences, suggesting that other factors such as target-speaker representation and conversational modeling play an equally important role. These findings provide a clearer understanding of the challenges in multi-talker recognition and offer guidance for designing more effective multi-modal systems in complex conversational environments.

2 Related Work

Audio-visual speech recognition (AVSR) has been extensively studied to improve speech recognition in noisy and multi-speaker environments by leveraging visual cues such as lip movements to complement the acoustic signal [7, 16, 17, 22, 42, 60, 61, 64, 70, 71]. In overlapping multi-speaker settings, prior work has explored incorporating target-speaker cues or visual speech information directly into the recognition process [10, 67]. More recent AVSR systems further improve robustness through self-supervised learning, automatic labeling, and stronger cross-modal attention [35, 41, 55, 57]. However, most of these methods are developed for single-speaker or controlled conditions and are primarily evaluated based on recognition accuracy, leaving their behavior in realistic cocktail-party scenarios less well understood.

The renewed interest in audio-visual speech recognition has been accompanied by the availability of large-scale benchmarks. LRS introduced sentence level lip reading in the wild and established a benchmark based on broadcast video [14]. LRS3 further scaled this setting with TED talks and became one of the common benchmarks for audio visual speech recognition in unconstrained conditions [2]. MuAViC extended this line to multilingual and noise robust learning, supporting both audio visual speech recognition and speech to text translation across multiple languages [4]. These benchmarks have progressively increased the realism of AVSR through greater speaker diversity, language coverage, and conversational speech. However, they remain centered on a single visible speaker, moderate acoustic conditions, or prepared speech, and therefore do not capture the concurrent conversational dynamics found in environments such as restaurants, bars, and offices.

Several multimodal corpora have introduced more realistic conversational settings. The CHIL Audiovisual Corpus was one of the first multimodal meeting datasets, introducing synchronized audio, video, and interaction annotations for smart-room meeting and lecture analysis [43]. The AMI Meeting Corpus provided a benchmark for multimodal meeting analysis with synchronized audio, video, and interaction annotations [8]. The MISP challenges moved toward home environments and multi participant speech processing [11, 65]. The MMCSG dataset further considered multimodal conversations recorded from smart glasses [72]. These datasets improve realism, but they still simplify the cocktail party setting as most sessions remain centered on a single shared conversation, and the task mainly focuses on recognition and diarization. As a result, they provide limited support for studying how systems jointly model who speaks, what is said, when speech occurs, and how parallel conversational streams are structured.

Beyond direct audio-visual speech recognition, another line of work uses visual information for target-speaker extraction or speech separation. These methods use a face track, target identity, or cross-modal alignment to recover one speaker from a mixture [18, 20, 34, 50]. They provide effective front-end components for noisy and overlapped conditions, but they usually assume a known target speaker and do not address the full ASR task.

Another direction works on audio-visual diarization partially to address the cocktail party problem by studying who spoke when. Recent work has progressed from probabilistic fusion of tracking and sound localization to end-to-end neural activity modeling, late fusion for in-the-wild videos, and multimodal clustering with semantic cues [13, 21, 28, 48]. Although these methods improve diarization in challenging multi-party conditions, they still focus on speaker activity estimation rather than jointly modeling transcription, speaker attribution, and conversation-level interaction structure.

Beyond recognition, conversational clustering is also central to the cocktail-party problem because it helps determine who is speaking to whom. Prior work has used acoustic interaction patterns to infer conversational structure [6, 15, 44], while visual cues such as gaze and head orientation have been leveraged to model addressee relationships and interaction flow [9, 31, 63]. More recently, semantic information from dialogue context has been incorporated to infer reply structure and addressee relations in multi-party conversations [23, 29, 33]. Despite this progress, these approaches mainly focus on interaction structure alone and do not consider speech content, speaker identity, and conversational flow together in cocktail-party settings.

3 Dataset and Tasks

3.1 Task Overview

The MCoRec task [46] targets speech recognition in realistic multi-party conversational settings, where multiple speakers may speak simultaneously and form parallel conversations. Unlike conventional ASR tasks that focus solely on transcription, MCoRec requires systems to jointly address three challenges: (i) recognizing speech

content, (ii) attributing speech to the correct speaker, and (iii) modeling conversational interactions.

Formally, given a 360° video V and a set of target speakers $S = \{s_1, \ldots, s_N\}$, each represented by a sequence of face bounding boxes, the system predicts a set of speaker-dependent transcriptions $\{\hat{Y}_i\}_{i=1}^{N}$ and a clustering $\hat{C}$ that assigns each speaker to a conversation group. This formulation captures the goal of determining *who speaks what and with whom* in complex multi-speaker environments.

3.2 Dataset and Recording Setup

The MCoRec dataset captures natural, unscripted multi-party conversations under realistic conditions. Each session consists of up to eight participants engaged in up to four simultaneous conversations, recorded using a single 360° camera and a single-channel microphone. Figure 1 shows an example of a recording session with 6 participants engaged in 3 separate conversations, each consisting of 2 speakers.

The recording setup reflects real-world scenarios using consumer devices. A 360° camera is placed at the center of a table, providing a panoramic view of all participants, while the integrated microphone records the entire acoustic scene. The distance between the camera and participants ranges from approximately 0.5 m to 2 m. Each session lasts around six minutes and contains spontaneous discussions on everyday topics such as work, hobbies, and personal experiences.

Recordings are collected across diverse indoor environments, including living rooms, meeting rooms, and lecture halls, resulting in varied acoustic and visual conditions. Within each session, speakers are organized into fixed conversation groups, while allowing natural and unconstrained interactions.

A key characteristic of the dataset is the presence of extreme conversational conditions. Due to natural interactions, speech overlap can reach up to 100%, and conversational turns are highly fragmented. These properties make the task significantly more challenging than conventional multi-speaker benchmarks, requiring systems to handle both acoustic interference and complex interaction dynamics. The MCoRec dataset [1] consists of 150 sessions, divided into training, development, and evaluation splits, with no speaker overlap across splits. A summary of the dataset is shown in Table 1.

3.3 Baseline Pipeline

The baseline system follows a cascaded architecture consisting of three main components: Active Speaker Detection (ASD), Audio-Visual Speech Recognition (AVSR), and conversation clustering [46].

First, ASD identifies time segments during which each speaker is active, reducing the search space for downstream recognition. These segments are then processed by an AVSR model, which integrates audio and visual cues (e.g., lip movements) to generate speaker-dependent transcriptions under noisy and overlapping conditions. Finally, a clustering module groups speakers into conversations.

[1]The MCoRec dataset was developed externally by Interactive-AI LLC and is made available for the CHiME Challenge for research purposes. For commercial use, please contact authors. Dataset and baseline systems can be accessed via https://github.com/MCoRec/mcorec_baseline.

Table 1: Overview of the MCoRec dataset.

Aspect	Description
Train / Dev / Eval	56 / 25 / 67 sessions
Duration	5.6h / 2.5h / 6.9h
Speakers per session	2 – 8
Conversations per session	2 – 4
Recording	360° video + single-channel audio
Extra modality	Ego-view videos (train only)
Annotations	Speaker's transcription + groups

The clustering module is based on a temporal overlap assumption: speakers within the same conversation tend to exhibit turn-taking behavior, while speakers from different conversations are more likely to speak simultaneously. Pairwise overlap ratios between speakers are computed and converted into a distance matrix, which is used by an agglomerative hierarchical clustering algorithm to form conversation groups.

3.4 Evaluation Metrics

MCoRec evaluates systems using three complementary metrics that capture both transcription accuracy and interaction modeling.

Speaker-dependent WER. Word error rate (WER) is computed for each speaker based on substitutions, deletions, and insertions, and then averaged across all speakers.

Conversation Clustering F1. Clustering performance is evaluated using a pairwise F1 score. For each pair of speakers, the system predicts whether they belong to the same conversation. Precision and recall are computed over all speaker pairs and averaged across sessions.

Joint ASR-Clustering Error. The primary evaluation metric combines ASR and clustering into a single score:

$$\text{JointError}(s) = 0.5 \cdot \text{WER}(s) + 0.5 \cdot (1 - F1(s)), \tag{1}$$

where $F1(s)$ is computed using a one-vs-rest formulation for each speaker. The final score is averaged over all speakers and sessions.

This metric enforces a balance between transcription accuracy and interaction modeling. In particular, improvements in clustering can significantly impact the overall score, even when WER remains unchanged, highlighting the importance of jointly optimizing both components.

4 Taxonomy and Design Space

We refer to individual systems using anonymized identifiers[2] (S1–S7) and analyze them through their architectural design choices rather than team identities. Our goal is not only to compare systems by performance, but also to identify the main design axes along which current multimodal cocktail-party systems differ.

We organize the design space around the two components that are directly reflected in the task definition and evaluation: *target-speaker transcription* and *conversation clustering*. This decomposition aligns with the two main sources of error in the benchmark: speaker-dependent word error rate (WER), which captures how well a system transcribes each target speaker, and clustering F1, which captures how well it models interaction structure. Although we

[2]Each system is described in a separate system paper to be released soon.

Table 2: Component-level comparison between systems

System	Target-Speaker Transcription Pipeline			Conversation Clustering	
	ASD	**AVTSE (Extraction)**	**AVSR (Recognition)**	**Modality / Cues**	**Algorithm / Model**
Baseline[46]	Light-ASD	-	AV-HuBERT CTC/Attention	Temporal (Overlap / Turn-taking)	Agglomerative Hierarchical Clustering (AHC)
S1 [30]	ConvNeXt + Conformer with Binary Classifier	Ensemble: BRAVEn, ResNet-18	Ensemble: ConvNeXt + Conformer fuse (Whisper/WavLM) + Transformer decoder	Semantic (Text) + Timestamps	Qwen-2.5 & DeepSeek R1 (Zero-shot)
S2.1 [59]	Baseline	-	AV-HuBERT	Semantic (Text) + Timestamps	Qwen3-8B (Deep reasoning)
S2.2 [59]	Baseline	-	AV-HuBERT + Qwen3-8B (post correction)	Semantic (Text) + Timestamps	Qwen3-8B (Deep reasoning)
S2.3 [59]	Baseline	-	Whisper-Flamingo	Semantic (Text) + Timestamps	Qwen3-8B (Deep reasoning)
S3 [1]	Baseline (Tuned thresholds)	-	Baseline	Baseline	AHC (optimized distance/linkage)
S4 [32]	-	-	Parakeet FastConformer + AV-HuBERT visual features	Semantic (Text) + Temporal fallback	Qwen3 similarity + AHC (+ fallback)
S5 [27]	Baseline	- (Auxiliary AVTSE loss only)	AV-HuBERT (multitask learning)	Baseline	Baseline
S6.1 [39]	Baseline	AV-TFGridNet (pseudo-supervised)	AV-HuBERT	Baseline	Baseline
S6.2 [39]	Baseline	AV-TFGridNet (pseudo-supervised)	Muavic-EN	Baseline	Baseline
S7.1 [37]	Baseline	-	AV-HuBERT + Whisper (dual-use fallback policy)	Visual (Mutual Gaze)	UniGaze + MGaze scoring + AHC
S7.2 [37]	Baseline	-	AV-HuBERT + Whisper (dual-use fallback policy)	Audio + Visual (Overlap + Gaze)	AVFuse (gaze + overlap fusion) + AHC

discuss these components separately for clarity, they are strongly coupled in practice. Segmentation errors propagate into recognition, recognition quality affects semantic clustering, and failures in either component directly affect the joint evaluation metric.

Table 2 summarizes the systems at the component level. Compared with the baseline, the submitted systems span a considerably richer design space. They differ in their target-speaker transcription pipelines, including how speech segments are defined (ASD), whether audio-visual target-speaker extraction (AVTSE) is performed before recognition, and which AVSR backbone is used. They also differ in conversation clustering, depending on whether it is based on temporal overlap, semantic reasoning, or visual interaction cues.

4.1 Target-Speaker Transcription Pipelines

The target-speaker transcription problem in MCoRec can be viewed as a sequence of three subproblems: identifying when a speaker is active, isolating or emphasizing the target speaker under overlap, and mapping the resulting signal to text. Across the systems, we observe three major design dimensions: segmentation strategy, target-speaker extraction strategy and recognition backbone.

Segmentation: *explicit ASD* versus *long-context recognition.* Most systems retain the baseline assumption that recognition should operate on short speaker-dependent segments obtained from an active speaker detection module. The baseline itself uses Light-ASD[38] to identify active regions and segment long video into smaller chunks before AVSR. S2, S3, S5, S6, and S7 largely follow this paradigm, though with different levels of dependence on the baseline segmentation. S2, S5, S6 and S7 keep the baseline Light-ASD unchanged and focuses improvements on downstream AVSR. S3 keeps the baseline pipeline but tunes onset/offset and chunking-related parameters to improve segment quality for both recognition and clustering. Two systems (S1 and S4) depart more clearly from this standard setup. S1 replaces the lightweight baseline ASD with a stronger audio-visual detector based on the ConvNeXt encoder [40] combined with Conformer[24] for frame binary classifiers (this ConvNeXt + Conformer encoder uses both in ASD and AVSR). At the other extreme, S4 avoids explicit short-segment ASD-based recognition and instead performs long-context target-speaker AVSR by concatenating speaker-conditioned visual tracks (lip crop frames) into a continuous stream and decoding substantially longer contexts in

a single pass. This shifts part of the burden from segmentation to the recognizer itself.

***Target-speaker extraction**: explicit extraction* versus *implicit conditioning*. A major axis of variation concerns whether systems explicitly separate or extract target speech before recognition. The baseline does not include a dedicated extraction stage, it relies on visual conditioning inside AVSR to bias recognition toward the visible target speaker. Several submissions continue this implicit strategy like S2, S3, S4, S7. In contrast, S1 and S6 adopt explicit audio-visual target speech extraction (AVTSE). S1 develops four AVTSE variants and uses an ensemble strategy, including BRAVEn-based[26] semantic-phonetic encoders and a ResNet-18-based dual-tower design[12], to disentangle the target stream before AVSR. S6 follows a two-stage pseudo-supervised paradigm: close-talk recordings are first enhanced to reduce cross-talk, then used as pseudo-labels to fine-tune an AV-TFGridNet[49] extraction model that operates on far-field audio plus target visual input. Recognition is subsequently performed on the extracted signal, and ASD is applied after extraction for segmentation. These systems reflect the view that target-speaker isolation should be solved explicitly before transcription.

S5 occupies an intermediate position. It does not perform explicit extraction at inference time, but it introduces an auxiliary AVTSE objective during training. Concretely, it adds a SEANet-style[62] reconstruction branch on top of intermediate AVSR representations so that the shared encoder is encouraged to preserve target-speaker acoustic information, while the test-time path remains a standard AVSR decoder.

Recognition backbones: AV-HuBERT[45, 58] dominance, but with different extensions. The most common recognition backbone is AV-HuBERT with CTC/Attention decoding[66]. This is the baseline model and remains the reference point for most submissions. S2 uses it as one of two main AVSR options. S3 keeps it and improves results mainly through better segmentation and decoding parameters. S5 builds directly on AV-HuBERT and strengthens it via multitask training. S6 evaluates several AVSR backends on top of extracted speech, including baseline AV-HuBERT, Muavic-EN[5]. S7 combines baseline AV-HuBERT with a Whisper-based system in a dual-model[36] inference policy: whenever the AV-HuBERT model produces an empty output, whisper-based replace it by the dual-use hypothesis to potentially fill the gap.

Other systems push beyond the standard AV-HuBERT recipe in more radical ways. S1 adopts a heterogeneous ensemble of AVSR models spanning both self-supervised and large-scale encoder–decoder paradigms. The ensemble includes a ConvNeXt Conformer encoder trained with masked audio-visual pretraining and discrete unit prediction, as well as Whisper-based[53] encoder–decoder models augmented with Flamingo-style[3] gated cross-attention for visual conditioning. Additional variants leverage pretrained speech representations such as WavLM to improve acoustic robustness. Some branches further incorporate LLM-conditioned decoding (e.g., Qwen-based refinement) to enhance linguistic coherence and error correction. Final hypotheses are combined via posterior-level fusion and ROVER[19] system combination to exploit complementarity across models. S2 also explores Whisper-Flamingo[56] as an alternative to AV-HuBERT, investigating how a large single-speaker ASR backbone can be adapted to the multi-speaker audiovisual condition. S4 takes a notably different route by conditioning a Parakeet FastConformer/TDT[54, 68] recognizer on baseline encoder AV-HuBERT's visual features, allowing long-context decoding with a strong audio backbone rather than using a conventional AV-HuBERT encoder-decoder stack.

4.2 Conversation Clustering Strategies

The second component is to group speakers into conversation-level interactions. Here the design space is even more diverse than for transcription. We observe three broad families of approaches: temporal-overlap heuristics, semantic clustering based on transcripts, and multimodal interaction modeling using visual cues. The main distinction is the level of abstraction at which the system reasons about interaction: low-level activity patterns, utterance content, or speaker attention cues.

Temporal clustering from speech activity: The baseline clustering module derives pairwise speaker similarity from overlap ratios between active speaking regions and then applies agglomerative hierarchical clustering (AHC)[47]. The core assumption is that speakers in the same conversation tend to alternate, whereas speakers from different conversations overlap more often. S3 remains closest to this philosophy, retaining the same basic temporal clustering approach but tuning the distance and linkage hyperparameters to better match the MCoRec development set. S5 and S6 also keep the baseline clustering unchanged. This approach is attractive because it's simple, efficient, and does not depend on linguistic quality. However, it is also tightly constrained by the quality of ASD or post-extraction segmentation. If activity estimates are noisy, temporal clustering becomes unreliable. This dependence is especially important in MCoRec, where overlaps are frequent and utterances are fragmented, so even small segmentation shifts can alter pairwise overlap statistics substantially.

Semantic clustering with large language models: A second cluster of systems treats conversation clustering as a semantic reasoning problem rather than a temporal one. S1 uses large language models, including Qwen 2.5[52](70B) and DeepSeek R1[25](671B), in a multi-stage zero-shot prompting pipeline that selects speaker-to-conversation assignments from transcripts and timestamps. S2 uses Qwen3-8B[69] with deep reasoning mode, explicitly prompting the model to infer conversational grouping from recognized text and utterance timing. S4 uses Qwen3 to derive semantic relations between speakers: it first classifies speakers as topic-bearing or passive based on their transcripts, then estimates pairwise topic similarity scores between active speakers. These scores form a similarity matrix, on which AHC is applied to obtain core conversation groups. Passive speakers are finally assigned to these groups using a temporal overlap-based fallback.

Compared with temporal clustering, these systems reason at the level of discourse coherence rather than local turn-taking. This is potentially much more powerful in MCoRec because simultaneous conversations are often separable by topic even when speaking patterns are ambiguous. However, semantic clustering is only as reliable as the transcripts it receives. Recognition errors, short backchannels, and semantically sparse utterances can all reduce the usefulness of LLM-based grouping, which is why S4 explicitly

Table 3: Datasets used by systems.

System	Trained component	Datasets used	Development strategy
S1 [30]	ASD	MCoRec, AVA-Speech, MSDWILD, M3SD	Multi-dataset ASD training
	AVTSE	LRS3, VoxCeleb2, AVSpeech, DNS-Noise	Synthetic 2–3 speaker mixtures
	AVSR	LRS2, LRS3, VoxCeleb2, AVSpeech, AVYT, MCoRec.	Large-scale AV pretraining, Whisper pseudo-labels
S2 [59]	AVSR (AV-HuBERT CTC/Attention)	LRS2, VoxCeleb2, AVYT, AVYT-mix, MCoRec	Pretraining + MCoRec fine-tuning. MCoRec also used to derive overlap templates
	AVSR (Whisper-Flamingo)	LRS2, VoxCeleb2, AVYT, AVYT-mix, MCoRec	End-to-end training with template-driven simulation and short MCoRec adaptation
S3 [1]	-	MCoRec	MCoRec dataset to tune hyperparameters
S4 [32]	AVSR	AVYT, LRS3, AMI, MCoRec	Stage-1 simulated pretraining on AVYT+LRS3. Stage-2 on AMI. Fine-tuning on MCoRec (+ simulated AMI)
S5 [27]	AVASR + auxiliary AVTSE	LRS2, AVYT, VoxCeleb2, MCoRec	Simulated mixtures from LRS2/AVYT/VoxCeleb2. Final-stage training mixed with MCoRec
S6 [39]	AVTSE	VoxCeleb2-2Mix, VoxCeleb2, MCoRec	Pretrained on simulated VoxCeleb2-2Mix and tuned with pseudo labels derived from MCoRec
S7 [37]	-	MCoRec	Uses MCoRec dataset to tune hyperparameters of UniGaze, MGaze and AVFuse

adds an acoustic speech activity fallback for passive or low-content speakers.

Visual interaction cues and multimodal fusion: S7 is the clearest example of a third strategy (conversation clustering from visual interaction cues). Instead of relying purely on temporal overlap or semantic content, it estimates gaze directions with UniGaze[51], computes mutual gaze scores between speakers, and uses these signals as evidence of conversational linkage. To improve robustness, the gaze-based similarity is fused with the baseline overlap-based score before AHC. This design is particularly interesting because it operates on interaction structure more directly than transcript-based approaches. However, this approach is sensitive to errors in gaze estimation, as head-pose variation, occlusion, and weak mutual gaze can weaken or distort interaction signals even among speakers in the same conversation. S7 therefore illustrates both the promise and the limitations of going beyond speech-only interaction cues.

4.3 Cross-Cutting Design Patterns

First, nearly all systems remain modular. Even when they use stronger backbones, explicit extraction, or LLM-based reasoning, transcription and clustering are still solved as separate stages connected by intermediate outputs such as timestamps or transcripts. End-to-end modeling of "who speaks what, when, and to whom" remains largely absent.

Second, systems differ not only by modality, but by *where* the modality is used. Visual information supports segmentation (most systems except S4), extraction (S1, S5, S6), recognition (all systems), or clustering (S7). Likewise, language models may appear as post-processors for transcript correction (S1, S2), or as LLM-based semantic clustering components that infer speaker groupings from transcripts (S1, S2, S4).

4.4 Training Data and Supervision

Beyond architectural design, the systems differ in their use of training data and development strategies. As shown in Table 3, most approaches combine large-scale external audio-visual corpora (e.g., LRS2, LRS3, VoxCeleb2, AVSpeech, AVYT) with the MCoRec dataset, but adopt distinct strategies for leveraging them. S1, S2, S4, and S5 follow a multi-stage training paradigm, where AVSR models are first pretrained on large-scale or simulated mixtures and then fine-tuned on MCoRec to adapt to real overlapping conversational conditions. S1 further extends this strategy to additional components, jointly training ASD and AVTSE, while S5 incorporates AVTSE as an auxiliary task during training. In contrast, S6 relies on a pseudo-supervised approach, using enhanced close-talk signals derived from MCoRec as supervision for training AVTSE. S3 and S7 take a different direction, avoiding model retraining and instead tuning hyperparameters directly on MCoRec.

5 Results

Table 4 reveals several consistent trends across systems. First, JointError is largely driven by WER once clustering F1 exceeds approximately 0.95. For example, S1 and S4 achieve similar clustering performance on the evaluation set (0.9572 vs. 0.9522), yet their JointError closely follows their WER (0.3018 vs. 0.3034), indicating that transcription dominates in this range. While clustering becomes less influential for top systems at this level, it remains a differentiating factor for lower-performing systems where F1 drops below 0.90 (e.g., S7.1: 0.7878 vs. S7.2: 0.8948, with corresponding JointError 0.3581 vs. 0.2885). Second, systems leveraging LLM-based semantic clustering achieve consistently high F1 scores (e.g., S1, S2, S4 all above 0.95), substantially outperforming the baseline clustering (0.8194). Third, stronger AVSR modeling, whether through highly optimized pipelines (S1) or long-context architectures (S4),

Table 4: Development and evaluation results. Ranks are assigned within each split based on JointError.

System	DEV				EVAL			
	Rank	WER ↓	Conv F1 ↑	JointError ↓	Rank	WER ↓	Conv F1 ↑	JointError ↓
S1	1	**0.3140** ± 0.0264	**1.0000** ± 0.0000	**0.1570** ± 0.0132	1	**0.3018** ± 0.0208	**0.9572** ± 0.0448	**0.1597** ± 0.0173
S2.1	3	0.4288 ± 0.0198	0.9778 ± 0.0333	0.2264 ± 0.0258	4	0.4525 ± 0.0231	0.9545 ± 0.0452	0.2360 ± 0.0187
S2.2	3	0.4288 ± 0.0197	0.9778 ± 0.0333	0.2264 ± 0.0246	3	0.4451 ± 0.0219	0.9545 ± 0.0469	0.2323 ± 0.0181
S2.3	7	0.4877 ± 0.0226	0.9778 ± 0.0333	0.2559 ± 0.0258	5	0.4952 ± 0.0270	0.9442 ± 0.0471	0.2689 ± 0.0255
S3	8	0.4931 ± 0.0165	0.9028 ± 0.0691	0.2899 ± 0.0379	9	0.5306 ± 0.0332	0.8137 ± 0.0606	0.3639 ± 0.0365
S4	2	0.3369 ± 0.0271	0.9667 ± 0.0433	0.1804 ± 0.0255	2	0.3034 ± 0.0219	0.9522 ± 0.0478	0.1626 ± 0.0200
S5	10	0.4955 ± 0.0183	0.8153 ± 0.0924	0.3459 ± 0.0518	10	0.5133 ± 0.0227	0.8194 ± 0.0606	0.3697 ± 0.0353
S6.1	9	0.4803 ± 0.0274	0.8153 ± 0.0916	0.3383 ± 0.0537	8	0.5010 ± 0.0227	0.8194 ± 0.0628	0.3635 ± 0.0382
S6.2	12	0.6059 ± 0.0256	0.8153 ± 0.0943	0.4011 ± 0.0562	12	0.5152 ± 0.0240	0.7845 ± 0.0725	0.3955 ± 0.0415
S7.1	5	0.4699 ± 0.0198	0.9846 ± 0.0231	0.2431 ± 0.0162	7	0.4867 ± 0.0267	0.7878 ± 0.0732	0.3581 ± 0.0414
S7.2	6	0.4699 ± 0.0193	0.9624 ± 0.0453	0.2551 ± 0.0286	6	0.4867 ± 0.0272	0.8948 ± 0.0547	0.2885 ± 0.0297
Baseline	11	0.4991 ± 0.0157	0.8153 ± 0.0921	0.3477 ± 0.0530	11	0.5199 ± 0.0223	0.8194 ± 0.0612	0.3730 ± 0.0350

consistently outperforms approaches relying on explicit target-speaker extraction (S6), with large WER gaps on evaluation (e.g., S1: 0.3018 vs. S6.1: 0.5010), indicating that improving robustness within the recognition model is more effective than front-end separation. Fourth, LLM-based post-correction provides limited gains under high-WER conditions, as seen in the small differences between S2 variants (e.g., S2.1: 0.4525 vs. S2.2: 0.4451 WER). Fifth, the multimodal cue-based approache (S7) performs well on the development set (F1 up to 0.9846) but degrades on evaluation (as low as 0.7878), suggesting weaker generalization compared to semantic clustering approaches. Finally, reducing reliance on strict short-segment ASD, either through stronger detection (S1) or longer-context inference (S4), appears beneficial, although ASD remains a core component in all systems. Overall, these results indicate that improvements in AVSR modeling are the main driver of performance once clustering performance is high, while clustering and post-processing play a secondary but still necessary role.

Tables 5 and 6 provide complementary views of system behavior under increasing conversational complexity. WER generally increases as the number of speakers grows across all systems, reflecting the increasing impact of overlap and interference. In contrast, varying the number of conversations has a limited and less consistent effect on WER, which remains relatively stable across 2–3 conversations (e.g., S1: 0.294 → 0.304). Comparing these two trends suggests that transcription difficulty is more sensitive to speaker interference than to the number of conversational groups. Clustering performance shows the opposite pattern. F1 remains near-saturated for top systems in simpler settings (fewer speakers or a single conversation), but degrades as conversational structure becomes more complex, particularly in multi-conversation scenarios. Systems with weaker AVSR quality or non-semantic clustering approaches (e.g., S3, S5, S6, S7) show clear drops in F1 as the number of conversations increases (e.g., S3: 1.00 → 0.77 → 0.83), while LLM-based approaches (S1, S2, S4) maintain consistently high clustering performance. Interestingly, S2 consistently achieves high clustering performance despite high WER. Overall, increasing the number of speakers primarily affects WER, while increasing the number of conversations mainly impacts clustering performance. Strong systems (S1, S4) remain stable across both conditions, whereas systems with weaker AVSR or non-semantic clustering degrade more noticeably, particularly in multi-conversation settings.

Table 7 analyzes system performance under different levels of speaking activity. The speaking activity ratio is defined as the proportion of time a speaker is active relative to the session duration. Speakers are categorized as low (< 0.45), mid (0.45–0.60), or high (≥ 0.60) activity, with 24.1%, 26.9%, and 49.1% of speakers in each group, respectively. A consistent trend across systems is that WER decreases as speaking activity increases (e.g., S1: 0.371 → 0.271; S2.2: 0.554 → 0.386), indicating that low-activity speakers with more fragmented and sparse speech are more challenging for transcription. Clustering performance does not show a consistent trend with speaking activity. While some systems improve with more speech (e.g., S3, S7), others degrade (e.g., S5, S6), indicating that higher activity introduces not only more evidence but also more overlap and cross-speaker interference. In contrast, top systems (S1, S4) maintain both low WER and high F1 across all activity levels, suggesting greater robustness to fragmented conversational structure. Overall, these results show that beyond speaker density and conversational structure, temporal continuity of speech is another key factor influencing both recognition and clustering performance in the MCoRec setting.

To better understand system failures beyond simple overlap effects, we analyze the top AVSR systems (S1 and S4) at the session. In session_12, session_14, session_147, and session_151, all speakers exhibit extreme overlap (90–95% of speaking time overlapping with at least three others), and all systems perform poorly (WER $\approx$ 0.45–0.6), indicating a global failure case under uniformly high interference. However, in session_123, session_125, and session_13, speakers also experience similarly high overlap (mean ~0.97–1.0), yet performance varies widely across speakers (e.g., WER ~0.095 to ~0.457). This shows that overlap alone does not explain performance differences. Instead, high-error speakers exhibit stronger

Table 5: Performance on the evaluation set by number of speakers.

System	2 spk		3 spk		4 spk		5 spk		6 spk		7 spk		8 spk	
	WER ↓	F1 ↑	WER ↓	F1 ↑	WER ↓	F1 ↑	WER ↓	F1 ↑	WER ↓	F1 ↑	WER ↓	F1 ↑	WER ↓	F1 ↑
S1	0.219	**1.00**	0.295	**1.00**	0.249	**1.00**	**0.354**	**1.00**	0.282	**0.97**	**0.509**	**1.00**	**0.436**	**1.00**
S2.1	0.441	**1.00**	0.509	**1.00**	0.398	**1.00**	0.502	**1.00**	0.424	0.96	0.618	**1.00**	0.582	**1.00**
S2.2	0.434	**1.00**	0.491	**1.00**	0.390	**1.00**	0.493	**1.00**	0.417	0.96	0.613	**1.00**	0.574	**1.00**
S2.3	0.414	0.92	0.519	**1.00**	0.427	**1.00**	0.533	**1.00**	0.472	0.96	0.705	0.80	0.713	0.85
S3	0.548	0.92	0.559	**1.00**	0.555	0.87	0.576	0.79	0.480	0.79	0.646	**1.00**	0.627	0.70
S4	**0.214**	**1.00**	**0.290**	**1.00**	**0.248**	**1.00**	**0.354**	**1.00**	**0.281**	0.95	0.514	**1.00**	0.468	**1.00**
S5	0.543	0.92	0.573	**1.00**	0.465	0.92	0.575	0.82	0.476	0.78	0.657	0.80	0.621	0.56
S6.1	0.488	0.92	0.524	**1.00**	0.448	0.92	0.547	0.82	0.472	0.78	0.685	0.80	0.637	0.56
S6.2	0.498	0.92	0.552	**1.00**	0.467	0.75	0.564	0.88	0.482	0.79	0.681	0.40	0.668	0.36
S7.1	0.453	0.83	0.497	0.50	0.428	0.90	0.555	0.68	0.453	0.75	0.635	**1.00**	0.643	0.90
S7.2	0.453	0.92	0.497	0.50	0.428	0.95	0.555	0.87	0.453	0.92	0.635	0.62	0.643	**1.00**
Baseline	0.544	0.92	0.557	**1.00**	0.477	0.92	0.576	0.82	0.483	0.78	0.636	0.80	0.641	0.56

Table 6: Benchmarking evaluation set by number of conversations.

System	1 conv		2 conv		3 conv	
	WER ↓	F1 ↑	WER ↓	F1 ↑	WER ↓	F1 ↑
S1	0.225	**1.00**	0.294	**1.00**	**0.304**	**0.98**
S2.1	0.456	**1.00**	0.457	0.99	0.433	**0.98**
S2.2	0.453	**1.00**	0.447	0.99	0.427	**0.98**
S2.3	0.425	**1.00**	0.473	0.97	0.505	0.94
S3	0.572	**1.00**	0.562	0.77	0.490	0.83
S4	**0.215**	**1.00**	**0.290**	**1.00**	0.312	0.96
S5	0.569	**1.00**	0.522	0.82	0.485	0.79
S6.1	0.502	**1.00**	0.501	0.82	0.486	0.79
S6.2	0.514	**1.00**	0.513	0.76	0.502	0.77
S7.1	0.466	**1.00**	0.490	0.71	0.470	0.80
S7.2	0.466	**1.00**	0.490	0.86	0.470	0.93
Baseline	0.565	**1.00**	0.529	0.82	0.493	0.79

Table 7: Benchmarking the evaluation set by speaking activity, where Low denotes activity < 0.45, Mid denotes activity 0.45–0.60, and High denotes activity ≥ 0.60.

System	Low		Mid		High	
	WER ↓	F1 ↑	WER ↓	F1 ↑	WER ↓	F1 ↑
S1	0.371	0.98	0.296	**0.99**	**0.271**	**0.98**
S2.1	0.563	0.97	0.462	**0.99**	0.393	**0.98**
S2.2	0.554	0.97	0.456	**0.99**	0.386	**0.98**
S2.3	0.596	0.97	0.510	0.95	0.438	0.96
S3	0.680	0.73	0.544	0.79	0.450	0.85
S4	**0.364**	**0.99**	**0.292**	**0.99**	0.280	0.97
S5	0.637	0.80	0.542	0.78	0.437	0.75
S6.1	0.593	0.80	0.503	0.75	0.445	0.64
S6.2	0.601	0.80	0.516	0.78	0.443	0.75
S7.1	0.592	0.74	0.502	0.74	0.427	0.80
S7.2	0.592	0.90	0.502	0.91	0.427	0.91
Baseline	0.634	0.80	0.543	0.78	0.451	0.75

cross-speaker confusion, with 31–53% of substitution words matching other speakers, compared to ∼14% for easier speakers. They also tend to have more fragmented speech, with shorter segments and fewer reference words. Overall, systems fail either under uniformly high interference across all speakers, or when the target speech is fragmented and easily confused with competing speakers.

6 Conclusion

This paper analyzed the CHiME-9 MCoRec submissions from a system-level perspective, focusing on how different design choices address the core challenge of recognizing and organizing speech in multi-party, overlapping conversations. While the task is often framed as a severe overlap problem, our results show that overlap alone does not explain performance differences. Instead, the main difficulty lies in correctly associating speech with the target speaker and maintaining consistency across fragmented conversational structure. Across systems, we observe three distinct strategies: (i) strengthening audio-visual representations for target-speaker recognition, (ii) explicitly handling interference through extraction or auxiliary objectives, and (iii) leveraging higher-level reasoning, especially with large language models, to resolve conversation structure. These strategies address complementary failure modes, and no single approach is sufficient. Systems integrating multiple such components achieve up to 57% relative reduction in the joint error rate. This suggests that the cocktail-party problem is not dominated by a single factor such as overlap, but by multiple interacting sources of ambiguity that require coordinated modeling across components.

Acknowledgments

The authors gratefully acknowledge support from the EU's Horizon research & innovation programme (101135798 – Meetween; 101213369 – DVPS). We thank Interactive-AI LLC for providing the database for research purposes.

References

[1] Sean Ackermann, Michael Becker, Erol Celik, Magdalena Eggers, Johannes Anton Kaiser, Hendrik Leddin, Adrian Nowak, Alexander Reimann, Dagmar Schönenberg, Johanna Schulze, Alison Vanzetta, Alexandra Wasilkow, and Stefan Goetze. 2026. The AUVIS System for the CHiME-9 Multi-Modal Context-aware Recognition (MCoRec) Challenge. In *9th CHiME Speech Separation and Recognition Challenge (CHiME 2026)*. 31–34. doi:10.21437/CHiME.2026-7

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition. arXiv:1809.00496 [cs.CV]

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1723, 21 pages.

[4] Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation. In *Interspeech 2023*. 4064–4068. doi:10.21437/Interspeech.2023-2279

[5] Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation. In *Interspeech 2023*. 4064–4068. doi:10.21437/Interspeech.2023-2279

[6] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland. 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4353–4356. doi:10.1109/ICASSP.2008.4518619

[7] Udo Bub, Martin Hunke, and Alex Waibel. 1995. Knowing who to listen to in speech recognition: Visually guided beamforming. In *1995 ICASSP*, Vol. 1. IEEE.

[8] Jean Carletta. 2007. Unleashing the Killer Corpus: Experiences in Creating the Multi-Everything AMI Meeting Corpus. *Language Resources and Evaluation* 41, 2 (2007), 181–190. http://www.jstor.org/stable/30200580

[9] Kent K. Chang, Mackenzie Hanh Cramer, Anna Ho, Ti Ti Nguyen, Yilin Yuan, and David Bamman. 2026. Multimodal Conversation Structure Understanding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vera Demberg, Kentaro Inui, and Lluís Marquez (Eds.). Association for Computational Linguistics, Rabat, Morocco, 7437–7458. doi:10.18653/v1/2026.eacl-long.349

[10] Guan-Lin Chao, William Chan, and Ian Lane. 2016. Speaker-Targeted Audio-Visual Models for Speech Recognition in Cocktail-Party Environments. In *Proc. Interspeech 2016*. 2120–2124.

[11] Hang Chen, Jun Du, Yusheng Dai, Chin-Hui Lee, Sabato Marco Siniscalchi, Shinji Watanabe, Odette Scharenborg, Jingdong Chen, Baocai Yin, and Jia Pan. 2022. Audio-Visual Speech Recognition in MISP2021 Challenge: Dataset Release and Deep Analysis. In *Interspeech 2022*. 1766–1770. doi:10.21437/Interspeech.2022-10483

[12] Hang Chen, Jun Du, Yu Hu, Li-Rong Dai, Bao-Cai Yin, and Chin-Hui Lee. 2021. Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement. *Neural Networks* 143 (2021), 171–182. doi:10.1016/j.neunet.2021.06.003

[13] Luyao Cheng, Hui Wang, Chong Deng, Siqi Zheng, Yafeng Chen, Rongjie Huang, Qinglin Zhang, Qian Chen, Xihao Li, and Wen Wang. 2025. Integrating Audio, Visual, and Semantic Information for Enhanced Multimodal Speaker Diarization on Multi-party Conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 19914–19928. doi:10.18653/v1/2025.acl-long.977

[14] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip Reading Sentences in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3444–3453. doi:10.1109/CVPR.2017.367

[15] Alfred Dielmann, Giulia Garau, and Hervé Bourlard. 2010. Floor holder detection and end of speaker turn prediction in meetings. In *Interspeech 2010*. 2306–2309. doi:10.21437/Interspeech.2010-632

[16] Paul Duchnowski, Martin Hunke, Dietrich Busching, Uwe Meier, and Alex Waibel. 1995. Toward movement-invariant automatic lip-reading and speech recognition. In *ICASSP*, Vol. 1. IEEE, 109–112.

[17] Paul Duchnowski, Uwe Meier, and Alex Waibel. 1994. See me, hear me: integrating automatic speech recognition and lip-reading. In *3rd ICSLP*. doi:10.21437/ICSLP.1994-139

[18] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. In *ACM SIGGRAPH*.

[19] J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. 347–354. doi:10.1109/ASRU.1997.659110

[20] Ruohan Gao and Kristen Grauman. 2021. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 15490–15500.

[21] Israel D. Gebru, Silèye Ba, Xiaofei Li, and Radu Horaud. 2018. Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 5 (2018), 1086–1099. doi:10.1109/TPAMI.2017.2648793

[22] Ralph Gross, Michael Bett, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. 2000. Towards a multimodal meeting record. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, Vol. 3. IEEE, 1593–1596.

[23] Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3682–3692.

[24] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*. 5036–5040. doi:10.21437/Interspeech.2020-3015

[25] Daya Guo, Dejian Yang, Haowei Zhang, and ... et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (Sept. 2025), 633–638. doi:10.1038/s41586-025-09422-z

[26] Alexandros Haliassos, Andreas Zinonos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2024. BRAVEn: Improving Self-supervised pre-training for Visual and Auditory Speech Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 11431–11435. doi:10.1109/ICASSP48485.2024.10448473

[27] Roland Hartanto, Daichi Nitsu, Nhu Minh Phuong Dinh, and Koichi Shinoda. 2026. Science Tokyo CHiME-9 MCoRec System Description. In *9th CHiME Speech Separation and Recognition Challenge (CHiME 2026)*. 21–25. doi:10.21437/CHiME.2026-5

[28] Mao-Kui He, Jun Du, and Chin-Hui Lee. 2022. End-to-End Audio-Visual Neural Speaker Diarization. In *Interspeech 2022*. 1461–1465. doi:10.21437/Interspeech.2022-10106

[29] Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. 2025. An LLM Benchmark for Addressee Recognition in Multi-modal Multi-party Dialogue. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*. 330–334.

[30] Ya Jiang, Ruoyu Wang, Jingxuan Zhang, Jun Du, Yi Han, Zihao Quan, Hang Chen, Yeran Yang, Kongzhi Zheng, Zhuo Chen, Yanhui Tu, Shutong Niu, Changfeng Xi, Mengzhi Wang, Zhongbin Wu, Jieru Chen, Henghui Zhi, Weiyi Shi, Shuhang Wu, Genshun Wan, Jia Pan, and Jianqing Gao. 2026. The USTC-NERCSLIP Systems for the CHiME-9 MCoRec Challenge. In *9th CHiME Speech Separation and Recognition Challenge (CHiME 2026)*. 1–5. doi:10.21437/CHiME.2026-1

[31] Natasa Jovanovic and Rieks op den Akker. 2004. Towards Automatic Addressee Identification in Multi-party Dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. Association for Computational Linguistics, Cambridge, Massachusetts, USA, 89–92. https://aclanthology.org/W04-2317/

[32] Dominik Klement, Alexander Polok, Nguyen Hai Phong, Prachi Singh, and Lukáš Burget. 2026. BUT System Description for CHiME-9 MCoRec Challenge. In *9th CHiME Speech Separation and Recognition Challenge (CHiME 2026)*. 6–10. doi:10.21437/CHiME.2026-2

[33] Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who Is Speaking to Whom? Learning to Identify Utterance Addressee in Multi-Party Conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1909–1919. doi:10.18653/v1/D19-1199

[34] Suyeon Lee, Chaeyoung Jung, Youngjoon Jang, Jaehun Kim, and Joon Son Chung. 2024. Seeing through the conversation: Audio-visual speech separation based on diffusion model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12632–12636.

[35] Jiahong Li, Chenda Li, Yifei Wu, and Yanmin Qian. 2024. Unified Cross-Modal Attention: Robust Audio-Visual Speech Recognition and Beyond. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 32 (March 2024), 1941–1953. doi:10.1109/TASLP.2024.3375641

[36] Zhengyang Li, Thomas Graave, Björn Möller, Zehang Wu, Matthias Franz, and Tim Fingscheidt. 2026. Noise-Robust AV-ASR Using Visual Features Both in the Whisper Encoder and Decoder. arXiv:2601.18396 [eess.AS] https://arxiv.org/abs/2601.18396

[37] Zhengyang Li, Aziz Hakiri, Zehang Wu, Thomas Graave, Ernst Seidel, Yihui Fu, Björn Möller, Patrick Blumenberg, and Tim Fingscheidt. 2026. Conversation Clustering by Mutual Gaze Estimation and AV-ASR by Dual Model Output Fusion. In *9th CHiME Speech Separation and Recognition Challenge (CHiME 2026)*. 16–20. doi:10.21437/CHiME.2026-4

[38] Junhua Liao and et al. 2023. A Light Weight Model for Active Speaker Detection. In *CVPR*. 22932–22941. doi:10.1109/CVPR52729.2023.02196

[39] Tongtao Ling, Pengjie Shen, and Zhong-Qiu Wang. 2026. The SUSTech AILab System Description for CHiME-9 MCoRec Challenge. In *9th CHiME Speech Separation and Recognition Challenge (CHiME 2026)*. 11–15. doi:10.21437/CHiME.2026-3

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11976–11986.

[41] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49357.2023.10096889

[42] Uwe Meier, Rainer Stiefelhagen, Jie Yang, and Alex Waibel. 2000. Towards unrestricted lip reading. *International Journal of Pattern Recognition and Artificial Intelligence* 14, 05 (2000), 571–585.

[43] Djamel Mostefa, Nicolas Moreau, Khalid Choukri, Gerasimos Potamianos, Stephen M Chu, Ambrish Tyagi, Josep R Casas, Jordi Turmo, Luca Cristoforetti, Francesco Tobia, et al. 2007. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language resources and evaluation* 41, 3 (2007), 389–407.

[44] Thai-Binh Nguyen, Ngoc Quan Pham, and Alexander Waibel. 2025. Cocktail-Party Audio-Visual Speech Recognition. In *Interspeech*.

[45] Thai-Binh Nguyen, Ngoc-Quan Pham, and Alexander Waibel. 2025. Cocktail-Party Audio-Visual Speech Recognition. In *Interspeech 2025*. 1828–1832. doi:10.21437/Interspeech.2025-676

[46] Thai-Binh Nguyen, Katerina Zmolikova, Pingchuan Ma, Ngoc Quan Pham, Christian Fuegen, and Alexander Waibel. 2026. A Cocktail-Party Benchmark: Multi-Modal dataset and Comparative Evaluation Results. arXiv:2510.23276 [cs.CL] https://arxiv.org/abs/2510.23276

[47] Thai-Binh Nguyen, Katerina Zmolikova, Pingchuan Ma, Ngoc Quan Pham, Christian Fuegen, and Alexander Waibel. 2026. A Cocktail-Party Benchmark: Multi-Modal Dataset and Comparative Evaluation Results. arXiv:2510.23276 [cs.CL]

[48] Zexu Pan, Gordon Wichern, François G Germain, Aswin Subramanian, and Jonathan Le Roux. 2024. Late audio-visual fusion for in-the-wild speaker diarization. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 174–178.

[49] Zexu Pan, Gordon Wichern, Yoshiki Masuyama, François G. Germain, Sameer Khurana, Chiori Hori, and Jonathan Le Roux. 2023. Scenario-Aware Audio-Visual TF-Gridnet for Target Speech Extraction. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 1–8. doi:10.1109/ASRU57964.2023.10389618

[50] Zexu Pan, Shengkui Zhao, Tingting Wang, Kun Zhou, Yukun Ma, Chong Zhang, and Bin Ma. 2025. Plug-and-Play Co-Occurring Face Attention for Robust Audio-Visual Speaker Extraction. In *Proc. Interspeech 2025*. 1933–1937.

[51] Jiawei Qin, Xucong Zhang, and Yusuke Sugano. 2026. Unigaze: Towards universal gaze estimation via large-scale pre-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5809–5820.

[52] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[53] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1182, 27 pages.

[54] Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast Conformer With Linearly Scalable Attention For Efficient Speech Recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 1–8. doi:10.1109/ASRU57964.2023.10389701

[55] Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation. In *Proc. Interspeech 2024*. 2420–2424.

[56] Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation. In *Interspeech 2024*. 2420–2424. doi:10.21437/Interspeech.2024-322

[57] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022. Robust Self-Supervised Audio-Visual Speech Recognition. In *Proc. Interspeech 2022*. 2118–2122.

[58] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022. Robust Self-Supervised Audio-Visual Speech Recognition. In *Interspeech 2022*. 2118–2122. doi:10.21437/Interspeech.2022-99

[59] Zeyan Song, Yushi Wang, and Jing Lu. 2026. The NJU-AALAB Systems for the CHiME-9 MCoRec Challenge. In *9th CHiME Speech Separation and Recognition Challenge (CHiME 2026)*. 26–30. doi:10.21437/CHiME.2026-6

[60] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. 1999. Modeling focus of attention for meeting indexing. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. 3–10.

[61] Bernhard Suhm, Brad Myers, and Alex Waibel. 1999. Model-based and empirical evaluation of multimodal interactive error correction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 584–591.

[62] Ruijie Tao, Xinyuan Qian, Yidi Jiang, Junjie Li, Jiadong Wang, and Haizhou Li. 2025. Audio-Visual Target Speaker Extraction With Selective Auditory Attention. *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), 797–811. doi:10.1109/TASLPRO.2025.3527766

[63] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 301–308. doi:10.1145/365024.365119

[64] Alex Waibe11, Hartwig Steusloff, Rainer Stiefelhagen, et al. 2005. CHIL: Computers in the human interaction loop. (2005).

[65] Zhe Wang, Shilong Wu, Hang Chen, Mao-Kui He, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Siniscalchi, Odette Scharenborg, et al. 2023. The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[66] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1240–1253. doi:10.1109/JSTSP.2017.2763455

[67] Yifei Wu, Chenda Li, Song Yang, Zhongqin Wu, and Yanmin Qian. 2021. Audio-Visual Multi-Talker Speech Recognition in a Cocktail Party. In *Interspeech 2021*. 3021–3025. doi:10.21437/Interspeech.2021-2128

[68] Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023. Efficient sequence transduction by jointly predicting tokens and durations. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1602, 23 pages.

[69] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, and et al. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[70] Jie Yang, Rainer Stiefelhagen, Uwe Meier, and Alex Waibel. 1998. Visual tracking for multimodal human computer interaction. In *SIGCHI*.

[71] Jie Yang, Xiaojin Zhu, Ralph Gross, John Kominek, Yue Pan, and Alex Waibel. 1999. Multimodal people ID for a multimedia meeting browser. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. 159–168.

[72] Kateřina Žmolíková, Simone Merello, Kaustubh Kalgaonkar, Ju Lin, Niko Moritz, Pingchuan Ma, Ming Sun, Honglie Chen, Antoine Saliou, Stavros Petridis, Christian Fuegen, and Michael Mandel. 2024. The CHiME-8 MMCSG Challenge: Multimodal conversations in smart glasses. In *8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024)*. 7–12. doi:10.21437/CHiME.2024-2