ADAPTING DIARIZATION-CONDITIONED WHISPER FOR END-TO-END MULTI-TALKER SPEECH RECOGNITION

Martin Kocour[⋆,†] *Martin Karafiat*[⋆] *Alexander Polok*[⋆]
Dominik Klement[⋆] *Lukáš Burget*[⋆] Jan Černocký[⋆]
[⋆] Speech@FIT, Brno University of Technology, Czechia
[†] Filevine, USA

ABSTRACT

We propose a speaker-attributed (SA) Whisper-based model for multi-talker speech recognition that combines target-speaker modeling with serialized output training (SOT). Our approach leverages a Diarization-Conditioned Whisper (DiCoW) encoder to extract target-speaker embeddings, which are concatenated into a single representation and passed to a shared decoder. This enables the model to transcribe overlapping speech as a serialized output stream with speaker tags and timestamps. In contrast to target-speaker ASR systems such as DiCoW, which decode each speaker separately, our approach performs joint decoding, allowing the decoder to condition on the context of all speakers simultaneously. Experiments show that the model outperforms existing SOT-based approaches and surpasses DiCoW on multi-talker mixtures (e.g., LibriMix).

Index Terms— conversational speech recognition, multi-talker ASR, speaker-attributed ASR, serialized output training.

1. INTRODUCTION

Automatic speech recognition (ASR) has seen remarkable progress over the past decade, driven by large-scale datasets, powerful neural architectures, and self-supervised learning techniques [1], [2]. Most ASR systems, however, have traditionally assumed single-speaker, clean speech conditions typical of voice search, dictation, and other laboratory-controlled benchmarks, where recent models have achieved near-human performance, even on long-form audio.

In contrast, real-world conversations are inherently multi-speaker, often with overlapping speech, dynamic turn-taking, and background noise. These factors complicate transcription, particularly in multi-party dialogues, where speaker turns frequently interleave and overlap. Consequently, the research focus has expanded toward multi-talker ASR, often incorporating speaker diarization [3], [4], [5], [6], [7] to segment and attribute speech to individual speakers. The series of CHiME challenges [8], which involved distant-microphone recordings of conversations, underscored the limitations of conventional ASR in such settings, highlighting the need for systems that can robustly handle overlap and noise.

Recent advances in multi-talker ASR (MT-ASR) have focused on directly transcribing overlapping speech while handling speaker assignment. Permutation-invariant training [9] addresses the issue of speaker label ambiguity but suffers from factorial complexity, motivating successors such as Heuristic Error Assignment Training [10], [11], which approximates optimal assignments. Serialized output training (SOT) [12] simplifies decoding with speaker change tokens but lacks explicit speaker modeling. To better leverage speaker cues, Diarization-Conditioned Whisper (DiCoW) [13] adapts Whisper with lightweight diarization-conditioned modules, SLIDAR [14] integrates diarization and ASR within a unified model, and Microsoft's Whisper extension [15] injects speaker change tokens to improve overlap handling.

In this work, we build on prior advances by proposing a modified Whisper-based architecture that unifies target-speaker ASR (TS-ASR) and serialized output training (SOT). Central to our approach is the integration of a Diarization-Conditioned Whisper (DiCoW) encoder, pretrained for TS-ASR, which uses diarization information to focus on individual speakers. For each speaker in the recording, the encoder produces a dedicated representation - referred to as a *speaker-channel* - that captures speaker-specific acoustic features. These embeddings are then jointly fed into a single shared decoder, enabling simultaneous decoding across overlapping speakers. The decoder generates serialized transcriptions that include both speaker tags and timestamps. Unlike conventional TS-ASR systems that decode each speaker independently, our model performs joint decoding conditioned on the global conversational context. This holistic design enhances robustness in highly overlapped speech scenarios, as demonstrated through evaluations on both synthetic mixtures (e.g., Libri2Mix) [16] and real-world multi-speaker recordings such as AMI [17] and NOTSOFAR [18]. This work aims to advance multi-talker ASR by integrating diarization-based conditioning with serialized output training in a unified architecture.

2. METHODS

2.1. Speech recognition with Whisper

We build on the Whisper model [19], which is a multilingual encoder-decoder ASR system trained on a large collection of weakly labeled data. Whisper follows the attention-based encoder-decoder (AED) architecture [20], [21], with both the encoder and decoder composed of Transformer blocks [22].

The audio encoder takes as input log mel-filterbank features $\mathbf{X} \in \mathbb{R}^{d_f \times 2T}$ and maps them into hidden embeddings:

$$\mathbf{H} = \text{Encoder}(\mathbf{X}), \quad \mathbf{H} \in \mathbb{R}^{d_m \times T} \tag{1}$$

where d_f and $2T$ are the feature dimension and frame count of the input, and d_m and T are the corresponding dimensions of the encoder output, since the input is subsampled in initial convolutional layers by factor of 2.

The decoder generates the output tokens autoregressively, conditioning on the previously predicted tokens $Y_{1:n-1}$, a task-specific prefix $\mathbf{t}$, and the encoder output $\mathbf{H}$:

$$\mathbf{o}_n = \text{Decoder}(\mathbf{t}, Y_{1:n-1}, \mathbf{H}), \quad \mathbf{o}_n \in \mathbb{R}^{|\mathcal{V} \cup \mathcal{W}|} \tag{2}$$

where $\mathbf{o}_n$ is the output distribution over the vocabulary $\mathcal{V}$ and set of timestamps $\mathcal{W}$ at step n.

Although Whisper performs well on a wide range of single-talker domains, it is not explicitly trained for multi-talker ASR and lacks the ability to perform speaker attribution. Consequently, its performance deteriorates in overlapping speech scenarios [13]. To

Submitted to *ICASSP2026, May 04-08, 2026, Barcelona, Spain*

address this, we build on DiCoW and fine-tuned Whisper with components designed for speaker-aware modeling and multi-talker output generation.

2.2. Target-speaker conditioning in DiCoW

In DiCoW, the encoder is adapted for target-speaker ASR using diarization masks. These masks encode frame-level speaker activity using four classes (STNO): $\mathcal{S}$ (silence), $\mathcal{T}$ (only target speaker active), $\mathcal{N}$ (only non-target speaker active), and $\mathcal{O}$ (target speaker overlaps with other speaker).

To effectively condition the model on these masks, DiCoW introduces the frame-level diarization-dependent transformation (FDDT) layer [13]. Let $\mathbf{h}^l \in \mathbb{R}^{d_m \times T}$ denote the input to Encoder's l-th Transformer layer. FDDT applies four class-specific affine transformations, weighted by the corresponding STNO probabilities:

$$\begin{aligned}\hat{\mathbf{h}}_t^l = &\left(\mathbf{W}_{\mathcal{S}}^l \mathbf{h}_t^l + \mathbf{b}_{\mathcal{S}}^l\right) p_{\mathcal{S}}^t + \left(\mathbf{W}_{\mathcal{T}}^l \mathbf{h}_t^l + \mathbf{b}_{\mathcal{T}}^l\right) p_{\mathcal{T}}^t \\ &+ \left(\mathbf{W}_{\mathcal{N}}^l \mathbf{h}_t^l + \mathbf{b}_{\mathcal{N}}^l\right) p_{\mathcal{N}}^t + \left(\mathbf{W}_{\mathcal{O}}^l \mathbf{h}_t^l + \mathbf{b}_{\mathcal{O}}^l\right) p_{\mathcal{O}}^t, \end{aligned} \tag{3}$$

where $p_{\mathcal{S}}^t, p_{\mathcal{T}}^t, p_{\mathcal{N}}^t$ and $p_{\mathcal{O}}^t$ correspond to STNO probabilities derived from standalone diarization at time t. This convex combination allows the model to adjust its internal representations depending on the speaker context. See [13] for additional details. In this work, we used only the oracle diarization derived from human annotations, i.e. the transformation simplifies to selecting the corresponding class-specific transformation.

2.3. Multi-talker ASR with speaker-attributed Whisper

To adapt Whisper for speaker-attributed (SA) ASR, we fine-tune the model using Serialized Output Training (SOT) [12], [23], which enables a single-decoder AED model to produce interleaved transcriptions of multiple speakers. Unlike in previous SOT-based Whisper work [15], we explicitly model speaker identities and timing by extending Whisper's standard tokens $\mathcal{V}$ with speaker-aware timestamp tokens $\mathcal{W} \times \mathcal{U}$, e.g. $\langle$|s1_2.2|$\rangle$ denote segment related to speaker 1 starting/ending at 2.2 seconds (relative to 30 s segment). We refer to those special tokens as *speaker-timestamps*. This resembles the approach of SLIDAR [14], with the key difference that timestamps and speaker labels are jointly encoded as unified tokens.

The order of speaker-attributed segments follows their onset time (FIFO). Speaker labels are assigned consistently throughout the recording based on external diarization, i.e. transcriptions for speaker u correspond always to the same speaker in the entire recording. Within each speaker's turn, timestamps are constrained to progress monotonically, while transitions to another speaker allow rolling back in time, effectively modeling overlapping speech.

2.4. Speaker-attributed Diarization-Conditioned Whisper

This section introduces our proposed Speaker-Attributed Diarization-Conditioned Whisper (SA-DiCoW) model, as shown in Figure 1. Given an input mixture $\mathbf{X}$ and a diarization-derived STNO mask $\mathbf{M}_u$ for each speaker u, we obtain speaker-specific encoder representations by running the DiCoW encoder separately for each speaker:

$$\hat{\mathbf{H}}_u = \text{DiCoW Encoder}(\mathbf{X}, \mathbf{M}_u), \quad \hat{\mathbf{H}}_u \in \mathbb{R}^{d_m \times T}. \tag{4}$$

These outputs, which we refer to as *speaker-channel* embeddings, provide time-aligned, speaker-conditioned representations of the mixture. To incorporate speaker identity into these embeddings,

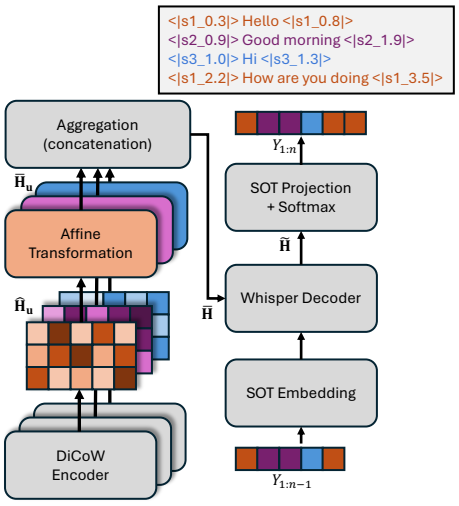

Fig. 1: Overall architecture of proposed SA-DiCoW.

we apply a learned affine transformation to each speaker-channel embedding:

$$\bar{\mathbf{H}}_u = \mathbf{W}_u \hat{\mathbf{H}}_u + \mathbf{b}_u, \quad \mathbf{W}_u \in \mathbb{R}^{d_m \times d_m}, \mathbf{b}_u \in \mathbb{R}^{d_m}. \tag{5}$$

This transformation effectively injects global speaker information into the embeddings, which should help the model assign correct speaker labels throughout the recording.

Next, we stack (concatenate) all speaker-channel embeddings into a unified encoder representation:

$$\bar{\mathbf{H}} = \mathop{\Big\|}_{u=1}^{|\mathcal{U}|} \bar{\mathbf{H}}_u, \quad \bar{\mathbf{H}} \in \mathbb{R}^{d_m \times T \cdot |\mathcal{U}|} \tag{6}$$

We explored also other aggregation strategies, including weighted sum, average, masked average (based on the diarization mask). Empirically, time-wise concatenation performs the best, since it preserves original representations across speakers.

During decoding, each standard input token (i.e. $y_n \in \mathcal{V}$) is embedded as usual. If the token represents a speaker-timestamp (i.e. $y_n \in \mathcal{W} \times \mathcal{U}$), it is first embedded as the standard Whisper timestamp, then passed through a speaker-specific affine transformation to encode the speaker identity implicitly in the decoder (cf. (5)). The Whisper decoder, largely unchanged, then processes the modified token embeddings along with encoder embeddings $\bar{\mathbf{H}}$ to produce decoder hidden states $\tilde{\mathbf{H}} \in \mathbb{R}^{d_m \times N}$, which are projected into three output distributions:

$$\mathbf{o}_n^{(t)} = \mathbf{W}^{(t)} \tilde{\mathbf{H}}_n + \mathbf{b}^{(t)}, \qquad t \in \{\text{lex}, \text{time}, \text{spk}\}, \tag{7}$$

where $\mathbf{o}_n^{\text{lex}}$ are the logits over the standard tokens, $\mathbf{o}_n^{\text{spk}}$ are the logits over speaker identities, $\mathbf{o}_n^{\text{time}}$ are the logits over timestamps, and $\mathbf{o}_{n,uw}^{\text{spk-time}} = \mathbf{o}_{n,u}^{\text{spk}} + \mathbf{o}_{n,w}^{\text{time}}$ are the combined logits used for speaker-timestamp tokens for speaker u and time-stamps w at time step n. This formulation allows the model to generate speaker-attributed transcripts with accurate timestamps while maintaining compatibility with Whisper's original decoding mechanisms, aside from minimal extensions necessary to encode speaker identities.

Overall, this architecture effectively leverages diarization to structure the encoder input and guide the decoding process, enabling robust speaker-attributed transcription with only minimal modifications to the original Whisper architecture. The newly introduced model parameters are initialized to identity mappings, ensuring that the model behaves like Whisper at the beginning of training.

3. EXPERIMENTAL SETUP

3.1. Datasets

All experiments are conducted using single-channel audio and oracle diarization derived from the available annotations. We train our proposed model using only English multi-talker datasets: NOTSOFAR [24], AMI [17], and LibriMix [16]. Data preparation is handled using Lhotse recipes [25], with minor modifications to ensure that all segments comply with the 30-second input constraint of Whisper.

NOTSOFAR is a recently released Microsoft multi-speaker meeting data set. Each meeting averages 6 minutes in duration, involving 4-8 participants and totaling 35 unique speakers. The recordings reflect a wide range of real-world acoustic conditions and conversational styles. The audio was captured using a proprietary device that integrates speech enhancement techniques such as beamforming, dereverberation, and denoising. Although the dataset comes also with simulated training data, we only use real recordings.

AMI is a well-established benchmark for meeting transcription. We use audio from the single distant microphone (SDM) and also individual headset mix (IHM). Only the first channel of the microphone array is used to maintain consistency with our single-channel setup.

LibriMix is a synthetic dataset derived from LibriSpeech [26], containing artificially mixed speech from two (Libri2Mix) or three (Libri3Mix) speakers in a left-aligned manner. Thus, the shorter source speech entirely overlaps with the longer one from the start.

3.2. Evaluation and metrics

We report performance using cpWER (Concatenated minimum-Permutation Word Error Rate), as implemented in the meeteval toolkit [27]. The cpWER metric extends standard WER by accounting for both word recognition and diarization errors, making it particularly suitable for speaker-attributed ASR evaluation. This metric has been widely adopted in the multi-talker ASR community, including in recent CHiME evaluations [8]. The evaluation is conducted using long-form decoding with beam size of 10 via Hugging Face's Transformers library [28]. Although the model is trained on 30-second segments, at inference we process continuous long-form audio by decoding it in consecutive 30-second chunks. For a fair comparison with the DiCoW model, we do not apply CTC rescoring, as our proposed SA-DiCoW model does not use it either.

3.3. Model settings and training

Our proposed SA-DiCoW can model up to 8 speakers, i.e. the vocabulary contains $|\mathcal{V}| = 50\,364$ standard Whisper tokens and $|\mathcal{U}| \times |\mathcal{W}| = 8 \times 1501$ speaker-timestamp tokens. The proposed model is initialized from a publicly available Diarization-Conditioned Whisper (DiCoW) checkpoint [29], which is based on Whisper-large-v3-turbo [19]. Overall, the model comprises approximately 918M trainable parameters. Our codebase, together with the configurations of each experiment, is available on our GitHub[1].

Training is conducted in two stages. In the first stage, all original Whisper encoder and decoder parameters are frozen, and only the newly introduced components (cf. Section 2.4) are trained for 1 000 steps using the AdamW optimizer with a learning rate of 2e-4 and a linear warm-up over the first 500 steps. In the second stage, the full model is fine-tuned by unfreezing the Whisper parameters and applying a reduced learning rate of 2e-6 to the pre-trained weights. This two-phase training strategy helps retain Whisper's original linguistic capabilities while adapting it to multi-talker scenarios. All experiments were conducted using four AMD MI250x GPUs. Each device processed a batch size of 1, and gradient accumulation was used to achieve an effective batch size of 192 per model update. The model typically converged after approximately 5 000 training steps.

To prevent the model from memorizing specific mappings between speaker tags and speaker identities, we propose a speaker-order augmentation strategy. During training, we randomly permute the speaker labels assigned by diarization, encouraging the model to disassociate token identities from fixed speaker roles. This forces the model to learn speaker-specific representations solely based on which encoder embeddings the model currently attend to (cf. Section 2.4).

[1] https://github.com/BUTSpeechFIT/SOT-DiCoW.git

4. RESULTS

4.1. Impact of encoder embedding aggregation

We analyze the impact of encoder embedding aggregation strategies on cpWER. These strategies correspond to different instantiations of the aggregation function in (6), listed in Table 1. In Libri2Mix, the differences are minor, with cpWER ranging from 4.6 % to 4.8 %. Concatenation performs best at 3.9 %, due to its ability to preserve temporal and speaker-specific patterns even in clean, fully overlapped conditions.

Table 1: cpWER comparison of different aggregation methods of Encoder's embeddings on LibriMix Clean with 2 speakers and NOTSOFAR with 4-8 speakers.

Aggregation	Libri2Mix	NOTSOFAR
weighted sum	4.8	59.1
average	4.6	50.2
masked average	4.6	47.4
concatenation	3.9	21.0

In NOTSOFAR, which contains realistic conversations of 4 to 8 speakers per meeting, the aggregation strategy plays a critical role. Concatenation achieves a cpWER of 21.0 %, significantly outperforming the the other approaches. This represents a relative reduction of 64 % in cpWER w.r.t. the baseline weighted sum. The superior performance of concatenation is due to its ability to preserve the temporal and speaker-specific structure of the embeddings. In contrast, averaging-based methods attenuate the acoustic information across multiple speakers, especially in meetings with higher speaker counts, leading to a loss of speaker identity cues and degraded recognition accuracy.

4.2. Ablation study on improving speaker label assignment

A major source of errors in multi-talker ASR arises from incorrect speaker label assignment. Table 2 breaks down these errors for the original DiCoW and our SOT-based approach, showing that most cpWER arises from omission (missed speech segments) and leakage (incorrect speaker attributions) related errors (cf. [10] for details). While DiCoW avoids label confusion by decoding each speaker independently relying on diarization. Our proposed SA-DiCoW, on the contrary, decodes all speakers in a single sequence, theoretically allowing better overlap handling.

Table 2: Error decomposition on NOTSOFAR. Leakage (#L) and omission (#O) related errors depicts absolute speaker-attributed errors w.r.t to number of reference words in %.

	cpWER	#S	#D	#I	#L	#O
DiCoW	18.0	9.4	3.4	5.1	3.5	2.1
SA-DiCoW	21.0	10.0	5.9	5.1	11.0	3.1

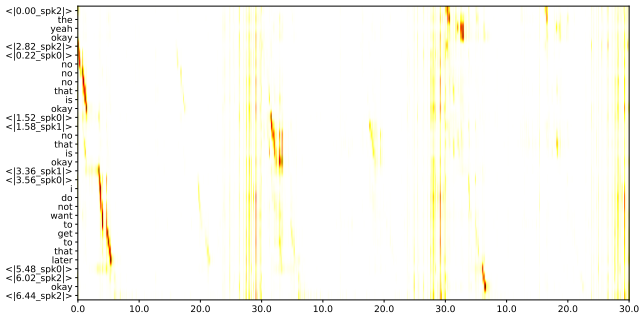

Fig. 2: Example of cross-attentions from last decoder layer: x-axis depicts time in seconds, y-axis shows tokens.

To address speaker assignment errors, we increase the cross-entropy loss weight 5-times on speaker timestamp tokens, encouraging the model to more reliably detect speaker changes and segment boundaries. This approach leads to consistent improvements in cpWER, as shown in Table 3.

Table 3: Impact of improved speaker lables on cpWER for LibriSpeech Test-Other, LibriMix Test-Clean and the NOTSOFAR.

	LS-Other	**LibriMix**		**NSF**
	1spk	2spk	3spk	4-8spk
DiCoW	4.9	4.8	32.1	18.0
SA-DiCoW	5.1	3.9	18.0	21.0
+ spk loss	5.0	3.4	17.2	20.8

For reference, DiCoW, that is, a target-speaker model that decodes each speaker independently, achieves 18.4 % cpWER in NOTSOFAR, outperforming our best SA-DiCoW model (20.8 %). However, on LibriMix with 3 speakers, DiCoW obtains 32.1 % cpWER, while our model achieves 17.2 %. This drop in DiCoW's performance on LibriMix is expected: in fully overlapped conditions, the STNO masks for different speakers become ambiguous, giving the model little guidance on which speaker to transcribe. While this limitation becomes apparent in synthetic mixtures with dense overlap, such extreme conditions are rare in natural conversations.

4.3. Analysis of cross-attention weights

To understand how the model uses speaker information during decoding, we visualize the cross-attention weights between the decoder and the speaker-channel embeddings from DiCoW encoder. Figure 2 reveals structured and interpretable attention patterns, suggesting that the model dynamically switches between relevant speaker channels when emitting tokens. The attention visualization was generated from a random utterance sampled from the AMI corpus.

Specifically, the visualization shows the cross-attention from the last decoder layer, with all attention heads averaged to produce a single heatmap. This aggregated view highlights how the decoder's focus shifts across different parts of the encoder sequence as it attends to different speakers. Because the encoder was constructed by concatenating speaker-channel embeddings—one per diarized segment—the attention weights jump between different speaker segments as the decoder aligns itself to the correct speaker when generating each token. This behavior underscores the model's ability to dynamically incorporate speaker information from multiple speaker-channel embeddings, even in highly overlapped segments.

4.4. Comparison with prior speaker-attributed ASR systems

Table 4 compares our proposed SA-DiCoW model with existing speaker-attributed ASR systems on AMI-SDM and AMI-IHM-MIX. On AMI-SDM, our method achieves (18.1 %) cpWER, outperforming previous speaker-attributed works such as Cornell et al. (21.1 %) [14] and Li et al. (21.2 %) [15]. This underscores the effectiveness of our approach in handling challenging far-field conversational data. For reference, the original target-speaker DiCoW model achieves the lowest cpWER on both datasets.

On AMI-IHM-MIX, our model achieves 14.4 % cpWER, outperforming Wang et al. (26.6 %), and approaching the performance of SLIDAR (11.5 %) from [14]. While SLIDAR performs best on IHM, it benefits from extensive training on synthetic mixtures, including augmented AMI data. In contrast, our model is fine-tuned on a combination of real conversational data (AMI-SDM and NOTSOFAR) and artificial data (Libri2Mix), which may contribute to stronger generalization to real-world scenarios, as suggested by these results. Overall, these findings confirm that our SA-DiCoW is competitive with the current state-of-the-art and particularly effective on realistic conversational benchmarks.

Table 4: cpWER comparison with related works. Note, the cpWER from Wang et al. [30] marked with $\star$ was achieved on 10-20s long segments, i.e. the model is not penalized for speaker confusion across segments.

	AMI-SDM	**AMI-IHM-MIX**
Cornell et al. [14]	21.1	11.5
Wang et al. [30]	-	22.8*
Li et al. [15]	21.2	-
DiCoW	16.3	13.1
SA-DiCoW	18.1	14.4

5. CONCLUSION

In this paper, we present a modified Whisper-based architecture for multi-talker speech recognition that integrates target-speaker modeling with serialized output training. Using a DiCoW encoder to extract target-speaker embeddings and concatenating them into a unified representation, our SA-DiCoW enables joint decoding of overlapping speech streams. This design allows the decoder to condition on the context of all speakers simultaneously, producing serialized transcriptions enriched with speaker tags and timestamps.

Our experimental results demonstrated that the proposed model outperforms existing SOT-based approaches on synthetic mixtures (e.g., LibriMix), achieving lower cpWER by effectively modeling speaker transitions in highly overlapped conditions. However, on real-world conversational datasets like AMI and NOTSOFAR, DiCoW established target-speaker ASR system yields superior performance, indicating that separate decoding per speaker remains advantageous in highly challenging meeting scenarios. This highlights the ongoing trade-offs between joint and separate speaker modeling in multi-talker ASR, depending on the level of overlap, background noise, and conversational complexity.

Future work will focus on improving speaker label assignment, particularly in challenging meeting datasets, by exploring speaker-aware training objectives. We also plan to investigate the potential of semi-supervised learning to reduce the reliance on oracle diarization, as well as explore adaptive speaker embedding mechanisms that can dynamically adjust to different conversational contexts. Overall, our work highlights the promise of combining target-speaker modeling and serialized output training to advance the state of MT-ASR.

References

[1] W.-N. Hsu et al., "HuBERT: How much can a bad teacher benefit ASR pre-training?" In *2021 ICASSP*, IEEE, 2021, pp. 6533–6537.

[2] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[3] L. Meng et al., "Unified modeling of multi-talker overlapped speech recognition and diarization with a sidecar separator," in *Interspeech 2023*, 2023, pp. 3467–3471. DOI: 10.21437/Interspeech.2023-1422

[4] L. Meng et al., "Empowering Whisper as a joint multi-talker and target-talker speech recognition system," in *Interspeech 2024*, 2024, pp. 4653–4657. DOI: 10.21437/Interspeech.2024-971

[5] N. Kanda et al., "End-to-end speaker-attributed ASR with transformer," in *Interspeech 2021*, 2021, pp. 4413–4417. DOI: 10.21437/Interspeech.2021-101

[6] J. Han et al., "Leveraging self-supervised learning for speaker diarization," in *ICASSP 2025*, 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10889475

[7] K. Žmolíková et al., "But system for chime-6 challenge," in *Proceedings of CHiME 2020 Virtual Workshop*, Barcelona: University of Sheffield, 2020, pp. 1–3. DOI: 10.21437/CHiME.2020-13 [Online]. Available: https://www.isca-speech.org/archive/CHiME_2020/pdfs/CHiME_2020_paper_zmolikova.pdf

[8] S. Cornell et al., "The CHiME-8 DASR challenge for generalizable and array agnostic distant automatic speech recognition and diarization," in *8th International Workshop CHiME 2024*, 2024, pp. 1–6. DOI: 10.21437/CHiME.2024-1

[9] D. Yu et al., "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 ICASSP*, IEEE, 2017, pp. 241–245.

[10] D. Raj, D. Povey, and S. Khudanpur, "SURT 2.0: Advances in transducer-based multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3800–3813, Sep. 2023, ISSN: 2329-9290. DOI: 10.1109/TASLP.2023.3318398 [Online]. Available: https://doi.org/10.1109/TASLP.2023.3318398

[11] L. Lu et al., "Streaming end-to-end multi-talker speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 803–807, 2021. DOI: 10.1109/LSP.2021.3070817

[12] N. Kanda et al., "Serialized output training for end-to-end overlapped speech recognition," in *Interspeech 2020*, 2020, pp. 2797–2801. DOI: 10.21437/Interspeech.2020-999

[13] A. Polok et al., "Dicow: Diarization-conditioned whisper for target speaker automatic speech recognition," *Computer Speech & Language*, vol. 95, p. 101 841, 2026, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2025.101841

[14] S. Cornell et al., "One model to rule them all? Towards end-to-end joint speaker diarization and speech recognition," in *2024 ICASSP*, IEEE, 2024, pp. 11 856–11 860.

[15] C. Li et al., "Adapting multi-lingual ASR models for handling multiple talkers," in *Interspeech 2023*, 2023, pp. 1314–1318. DOI: 10.21437/Interspeech.2023-1276

[16] J. Cosentino et al., "LibriMix: An open-source dataset for generalizable speech separation," *arXiv: Audio and Speech Processing*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218862876

[17] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," Jul. 2005, ISBN: 978-3-540-32549-9. DOI: 10.1007/11677482_3

[18] I. Abramovski et al., *Summary of the NOTSOFAR-1 Challenge: Highlights and learnings*, 2025. arXiv: 2501.17304 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2501.17304

[19] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[20] J. K. Chorowski et al., "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, C. Cortes et al., Eds., vol. 28, Curran Associates, Inc., 2015.

[21] W. Chan et al., "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 ICASSP*, 2016, pp. 4960–4964. DOI: 10.1109/ICASSP.2016.7472621

[22] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017.

[23] N. Kanda et al., "Large-scale pre-training of end-to-end multi-talker ASR for meeting transcription with single distant microphone," in *Interspeech 2021*, 2021, pp. 3430–3434. DOI: 10.21437/Interspeech.2021-102

[24] A. Vinnikov et al., "NOTSOFAR-1 challenge: New datasets, baseline, and tasks for distant meeting transcription," in *Interspeech 2024*, 2024, pp. 5003–5007. DOI: 10.21437/Interspeech.2024-1788

[25] P. Żelasko et al., "Lhotse: A speech data representation library for the modern deep learning ecosystem," *Proceedings of NeurIPS Workshop on Data-Centric AI*, 2021.

[26] V. Panayotov et al., "Librispeech: An ASR corpus based on public domain audio books," in *2015 ICASSP*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964

[27] T. v. Neumann et al., "MeetEval: A toolkit for computation of word error rates for meeting transcription systems," in *Proceedings of the 7th International Workshop CHiME 2023*, 2023, pp. 27–32. DOI: 10.21437/CHiME.2023-6

[28] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[29] A. Polok et al., "Target speaker ASR with Whisper," in *2025 ICASSP*, 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10887683

[30] J. Wang et al., "META-CAT: Speaker-informed speech embeddings via meta information concatenation for multi-talker ASR," in *ICASSP 2025 - 2025 ICASSP*, 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10889841