Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

SC-SOT: Conditioning the Decoder on Diarized Speaker Information for End-to-End Overlapped Speech Recognition

Yuta Hirano[1], *Sakriani Sakti*[1]

[1]Nara Institute of Science and Technology, Japan

yuta.hirano.ia4@naist.ac.jp, ssakti@is.naist.jp

Abstract

We propose **S**peaker-**C**onditioned **S**erialized **O**utput **T**raining (SC-SOT), an enhanced SOT-based training for E2E multi-talker ASR. We first probe how SOT handles overlapped speech, and we found the decoder performs implicit speaker separation. We hypothesize this implicit separation is often insufficient due to ambiguous acoustic cues in overlapping regions. To address this, SC-SOT explicitly conditions the decoder on speaker information, providing detailed information about "who spoke when". Specifically, we enhance the decoder by incorporating: (1) speaker embeddings, which allow the model to focus on the acoustic characteristics of the target speaker, and (2) speaker activity information, which guides the model to suppress non-target speakers. The speaker embeddings are derived from a jointly trained E2E speaker diarization model, mitigating the need for speaker enrollment. Experimental results demonstrate the effectiveness of our conditioning approach on overlapped speech.

Index Terms: overlapped speech recognition, serialized output training, decoder conditioning, speaker diarization

1. Introduction

Since the development of the first speech recognition system over half a century ago, speech recognition technology has steadily advanced. In particular, recent advances in automatic speech recognition (ASR) have been remarkable, largely due to the rise of deep learning [1, 2]. In common single-speaker tasks, especially, models surpassing human parity are being reported one after another [3, 4, 5].

Given such impressive achievements, ASR may seem like a solved problem. However, there are still many challenges to be addressed. One of the most significant challenges is overlapped speech recognition, which is referred as multi-talker ASR. Multi-talker ASR deals with the scenario where multiple speakers are talking simultaneously, resulting in overlapped speech segments. This presents a significant departure from the well-studied single-talker ASR scenario, where the input audio segment is assumed to contain only one speaker at a time. The core challenge in multi-talker ASR lies in disentangling the mixed acoustic signals to accurately transcribe the speech of each individual speaker. The ability to accurately process and transcribe multi-talker speech is crucial for a wide range of applications, including automated meeting transcription and analysis of multi-party conversations.

Driven by the rapid development of deep learning, recent research has increasingly focused on end-to-end (E2E) models for multi-talker ASR. These E2E models offer the potential for joint optimization of all components through back-propagation, leading to improved performance and simplified system design. A number of promising E2E approaches have emerged [6, 7, 8, 9, 10]. The initial approach was proposed by Yu *et al.* [6]. In this work, the permutation invariant training (PIT) was applied to a neural network with two output channels, enabling the transcription of overlapped speech. PIT was originally proposed for speech separation [11] and was later applied to E2E multi-talker ASR [6]. However, a major drawback of PIT is that the number of speakers that can be handled is limited to a predetermined number. In fact, most PIT-based methods could only handle two speakers [11, 6, 12, 13]. To address this limitation, serialized output training (SOT) [7] was proposed. In SOT, unlike PIT, transcriptions for all speakers is predicted as a single token sequence. As a result, SOT does not require a predetermined number of speakers and can theoretically handle arbitrary number of speakers.

Although SOT successfully addresses the permutation problem and allows for a variable number of speakers, handling heavily overlapped speech is still a significant challenge for SOT. This may be because SOT-based models basically learn how to transcribe overlapped speech only from labels serialized in speakers' speaking start times. To overcome this limitation, we propose to enhance SOT by explicitly providing the model with information about "who spoke when". This is achieved through our proposed approach called **S**peaker-**C**onditioned **S**erialized **O**utput **T**raining (SC-SOT), which conditions the decoder on speaker information. Specifically, SC-SOT incorporates two key components: (1) speaker embedding, derived from a jointly trained E2E neural speaker diarization with encoder-decoder based attractors (EEND-EDA) model [14], which allows the model to focus on the acoustic characteristics of the current target speaker, and (2) speaker activity information, also obtained from the EEND-EDA, which helps suppress non-target speakers or silence. By incorporating this explicit speaker information, SC-SOT aims to enhance the implicit speaker separation performed by the decoder. Moreover, our proposed approach mitigates the need for speaker enrollment by leveraging the jointly trained EEND-EDA. The main contributions of this paper can be summarized as follows.

- We show that the decoder plays a crucial role in handling overlapped speech by visualizing attention weights.
- We propose a conditioning method for the decoder using speaker embeddings and speaker activity information without the need for speaker enrollment.
- We propose a method to integrate speaker counting by the diarization module into SOT-based E2E multi-talker ASR.
- Experimental results demonstrate our proposed approach improves SOT-based E2E multi-talker ASR in terms of both speech recognition accuracy and speaker counting accuracy.

10.21437/Interspeech.2025-1243

2. Related Works

2.1. Handling overlapped speech using speaker embeddings

Target speech extraction is a task that aims to isolate the speech of a specific speaker from a mixture containing multiple speakers, relying on cues about the target speaker's identity. Speaker embedding is a common type of information used in target speech extraction [15, 16, 17]. Target-speaker voice activity detection (TS-VAD) [18, 19] is another successful approach that leverages speaker embeddings to handle overlapped speech. TS-VAD is one of a mainstream technique in speaker diarization, solving the diarization problem by performing voice activity detection for each speaker, conditioned on their respective speaker embeddings. These studies suggest that using speaker embeddings as conditioning information is effective for handling overlapped speech.

Beyond speech separation and speaker diarization tasks, speaker embeddings have also been integrated into E2E multi-talker ASR. Huang *et al.* [20] proposed an E2E multi-talker ASR model based on self-supervised learning (SSL) models, using speaker embeddings. They demonstrated the effectiveness of speaker embeddings for multi-talker ASR. However, the number of speakers their model can handle is fixed due to the use of connectionist temporal classification (CTC) decoder [21], which typically requires a fixed output dimension corresponding to the number of speakers. Furthermore, this approach assumes the availability of speaker enrollment data for each speaker during both training and inference. The requirement for speaker enrollment poses a significant limitation in many real-world applications.

2.2. Handling overlapped speech using speaker activity information

Relatively few studies have explored the use of speaker activity information for handling overlapped speech, especially when compared to the extensive work on speaker embedding-based methods. Polok *et al.* [22, 23] proposed a conditioning method to adapt Whisper [3] for target-speaker ASR, utilizing speaker activity information instead of speaker embeddings to extract the target speaker's representation from intermediate encoder layers. This approach achieves target-speaker ASR without requiring the learning of a complete speaker embedding space. However, relying solely on speaker activity information is often insufficient for accurately handling overlapped speech, particularly when the overlapping speakers exhibit similar activity patterns within input speech segments. Our proposed approach addresses this limitation by leveraging not only speaker activity information but also the corresponding speaker embeddings. This serves as a more robust and discriminative representation of each speaker, even in challenging overlapping conditions.

3. Proposed Approach

3.1. Encoder and decoder behavior in handling overlapped speech

Although SOT-based models can transcribe overlapped speech without separating it, how SOT-based models achieve this has not been revealed. To understand SOT-based models' behavior, we explored attention weights in the model. Figure 1 shows the examples of visualized attention weights of both the encoder and the decoder in an SOT-based model. Looking at the source-target attention weight, it turned out that the relationship between the target token sequence and the encoder representation is not fully monotonic and overlapped segments were attended twice. Moreover, the encoder maintains temporal relationship between the frames and does not perform any separation. These findings indicate that the encoder keeps the temporal information of input overlapped speech and the decoder performs speaker separation implicitly. This finding greatly influenced the design of the proposed approach.

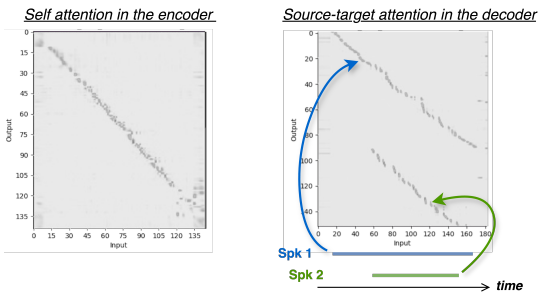


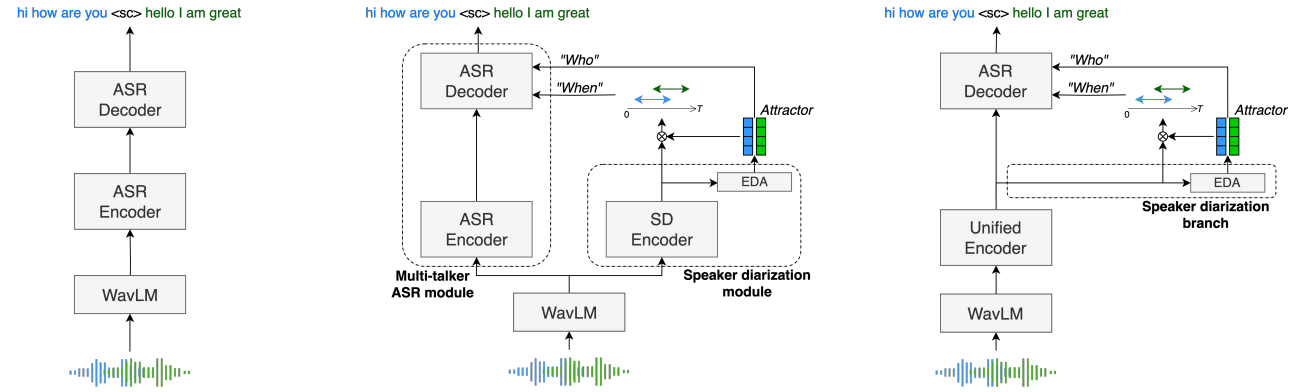
Figure 1: *The visualized attention weights of a SOT-based model in two-speaker overlap case. (Left) Self attention in the encoder and (Right) source-target attention in the decoder. In the visualized source-target attention, the attended frames in each diagonal line align approximately with the speaker's active time.*

3.2. Conditioning the decoder on speaker information

Serialized output training (SOT) [7] is a simple training strategy for E2E multi-talker ASR. Figure 2 (a) shows the overview of SOT-based multi-talker ASR model. SOT does not require a specialized model architecture to handle overlapped speech; the same model used for single-talker scenarios can be employed. However, the applicable models are generally limited to attention-based encoder-decoder architectures. This restriction is due to the fact that the relationship between input speech and the target token sequence is not strictly monotonic in SOT. SOT models multi-talker ASR as follows. First, SOT introduces a special token $\langle sc \rangle$ that represents speaker change. Second, the target sequence is created by serializing multiple references from different speakers into a single sequence and inserting $\langle sc \rangle$ between them. For example, for a two-speaker case, the target sequence is $\{y_1^1, \cdots, y_{N_1}^1, \langle sc \rangle, y_1^2, \cdots, y_{N_2}^2, \langle eos \rangle\}$, where N_1 and N_2 are the lengths of the references from the first and second speakers, respectively. References from multiple speakers are usually serialized in first-in first-out (FIFO) order based on their speaking start time.

We propose to enhance SOT by conditioning the decoder on speaker embeddings and speaker activity information. To obtain those two information from input overlapped speech, SC-SOT has a speaker diarization module (branch) based on E2E neural diarization with encoder-decoder based attractors (EEND-EDA) [14]. This speaker diarization module can predict speaker activities and local speaker embeddings called attractor from overlapped speech as well as EEND-EDA. This speaker diarization module is jointly trained with multi-talker ASR module, and the FIFO speaker order is used for the label permutation of the speaker diarization labels instead of the permutation given by PIT [11]. SC-SOT-based model is optimized using the following loss function:

$$\mathcal{L} = \mathcal{L}_{asr} + \alpha \mathcal{L}_{diar}, \tag{1}$$



(a) SOT-based E2E Multi-talker ASR

(b) Proposed E2E Multi-talker ASR (**SC-SOT**)

(c) Proposed E2E Multi-talker ASR with Multi-task learning (**SC-SOT + MTL**)

Figure 2: *Overview of the conventional SOT-based end-to-end multi-talker ASR and the proposed approach.*

where $\mathcal{L}_{asr}$ and $\mathcal{L}_{diar}$ are the loss function of SOT and EEND-EDA, respectively. α is a hyperparameter to adjust the loss for speaker diarization task. Figure 2 (b) and (c) provide the overview of the proposed SC-SOT with and without multi-task learning.

3.2.1. *Conditioning the decoder on speaker embeddings*

The speaker embeddings predicted by the speaker diarization block are used to allow the model to focus on the acoustic characteristics of the target speaker. To condition the decoder on the speaker embeddings, the position-wise feed-forward network (FFN) in the l-th transformer decoder layer is modified as follows:

$$\begin{aligned} \mathrm{FFN}_l(\mathbf{z}_{n-1,s} + \mathbf{w}^{spk} \cdot \mathbf{a}_s) \quad (l = 1) \\ \mathrm{FFN}_l(\mathbf{z}_{n-1,s}) \quad (l > 1), \end{aligned} \tag{2}$$

where $\mathbf{z}_{n-1,s}$ is the decoder representation calculated from $n-1$th token for the s-th speaker, $\mathbf{a}_s$ is the attractor of the s-th speaker, and $\mathbf{w}^{spk}$ is a learnable weight matrix. This formulation shares an idea with Transcribe-to-Diarize [24]. However, the objectives differ, and several aspects of our approach are also distinct: the method for acquiring speaker embeddings, multi-task learning with E2E speaker diarization, and functions described in remaining sections.

3.2.2. *Conditioning the decoder on speaker activity information*

The speaker activity information predicted by the speaker diarization block is used to guide the model to suppress non-target segments. We selectively apply attention penalty to source-target attention in the decoder. This method is originally proposed by Polok *et al.* [25] to convert Whisper into a target speaker ASR model. We adopt this method for SOT-based E2E multi-talker ASR. In each attention head of the source-target attention, we penalize the source-target attention scores in non-target speaker segments by modifying them as follows:

$$Penalized_Attention = \mathbf{Q}\mathbf{K}^T - \mathbf{c}, \tag{3}$$

where the penalty, $\mathbf{c}$=$\{c_1, c_2, \ldots, c_T\}$, is calculated as:

$$c_t = c \odot \mathbb{1}(p_{t,s}^{diar} < \theta), \tag{4}$$

where $\mathbf{Q}\mathbf{K}^T$ is the original source-target attention score before normalization, c is a hyperparameter controlling the strength of the penalty, $p_{t,s}^{diar}$ represents the posterior probability for speaker s at time frame t predicted by the speaker diarization module, and θ is a threshold determining whether a speaker is considered active. $\odot$ and $\mathbb{1}$ denote the Hadamard product and the indicator function, respectively.

3.2.3. *Speaker counting by EEND-EDA*

In addition to the two conditioning methods described above, we propose incorporating speaker counting by EDA module into decoding. In conventional SOT-based models, decoding terminates when the end-of-sequence ($\langle eos \rangle$) token is output. While SOT-based models have been shown to have high speaker counting accuracy, it is still not perfect. Furthermore, because the speaker counting task is not explicitly trained, there may be room for improvement in speaker counting through supervised learning. EEND-EDA optimizes speaker counting through discriminative training. Therefore, enforcing the number of speakers predicted by the speaker diarization module to the decoding may improve speaker counting of SOT-based models. To enforce the number of speakers predicted by the speaker diarization module during decoding, we made the following modifications. First, we replaced the $\langle eos \rangle$ token at the end of the label with the $\langle sc \rangle$ token. This prevents the decoding process from stopping automatically. Second, decoding is terminated when the number of generated $\langle sc \rangle$ tokens equals the number of speakers predicted by the speaker diarization module.

4. Experimental Setup

4.1. Dataset

The training set consists of LibriSpeech-360h [27], Libri2Mix, and Libri3Mix [28]. Libri2Mix and Libri3Mix were modified by introducing utterance start time offsets. For Libri2Mix, the simulating scripts and metadata provided by ESPnet [26] were used. The simulating scripts and metadata for generating the training set from Libri3Mix were created by following the same simulation protocol as Libri2Mix. We used the test-set of Libri2Mix and Libri3Mix for evaluation.

Table 1: *WER, SCA and DER (dev / test) of our proposed SC-SOT with different conditioning configurations compared to the conventional SOT on Libri2Mix and Libri3Mix. MTL denotes multitask learning with the ASR objective and the speaker diarization objective. "*" indicates the use of oracle speaker diarization information.*

		Information to condition		**Libri2Mix**			**Libri3Mix**		
	Model	**Speaker Emb.**	**Speaker Activity Info.**	**WER(%)**	**SCA(%)**	**DER(%)**	**WER(%)**	**SCA(%)**	**DER(%)**
1	SOT [26]	-	-	19.7 / 17.1	-	-	-	-	-
2	SOT (ours)	-	-	19.2 / 16.8	98.8 / 99.0	-	36.0 / 34.5	89.2 / 88.1	-
3	SOT + MTL	-	-	19.8 / 17.7	99.6 / 99.5	3.6 / 3.7	34.5 / 32.9	93.4 / 93.4	8.2 / 8.1
4	SC-SOT	✓	×	18.0 / 15.6	99.6 / 99.4	4.6 / 4.3	34.3 / 33.1	88.4 / 86.2	21.9 / 22.1
5	SC-SOT + MTL	✓	×	**17.5 / 15.2**	**99.9 / 99.9**	8.2 / 8.9	**33.2 / 30.9**	95.4 / 95.2	22.6 / 23.4
6	SC-SOT + MTL	×	✓	17.8 / 15.7	99.8 / **99.9**	3.1 / 3.1	34.5 / 32.5	93.7 / 93.9	7.2 / 7.4
7	SC-SOT + MTL	✓	✓	18.7 / 16.2	99.2 / 99.5	7.0 / 7.3	36.2 / 34.6	**96.0 / 95.3**	20.6 / 21.2
8	SC-SOT + MTL	✓	✓*	17.4 / 15.1	-	-	31.1 / 29.2	-	-

4.2. Model configuration

The baseline SOT-based model employs a Conformer encoder [29] and a Transformer decoder [30]. The Conformer encoder has 12 layers, while the Transformer decoder is structured with 6 layers. Both the encoder and decoder utilize 4 attention heads and a hidden dimension of 256 and a 2048-dimensional FFN. The kernel size of the convolution layer in the Conformer encoder was set to 31. Input features were derived from the WavLM large [31] representation, using a weighted sum with a learnable weight. The multi-talker ASR module of all the proposed model employs the same architecture as the baseline model. When an independent speaker diarization module is used, its encoder is built with 4 layers of Conformer, incorporating 4 attention heads, a hidden dimension of 256, and a 1024-dimensional FFN. The EDA module consists of one unidirectional LSTM layer with 256 hidden units. During training, the ground truth speaker diarization label were used for conditioning. The hyperparameter α in equation 1 was set to 0.1. The hyperparameter c and θ in equation 3 and 4 were set to 50 and 0.5, respectively. All models were trained using the Adam optimizer [32]. The learning rate was set to 0.001 and warmed up for 10000 iterations. The batch bin size was set to 2.4e7 and trained all models for 50 epochs. Character was used as the output unit. Word error rate (WER) and speaker counting accuracy (SCA) were used for evaluation matrices. We report diarization error rate (DER) with collar tolerance of 0.0 sec and median filtering of 11 frames as a supplemental result.

5. Experimental Results

Table 1 presents the performance of our proposed SC-SOT-based model on the Libri2Mix and Libri3Mix datasets, alongside baseline comparisons and variations in conditioning configurations. Our proposed SC-SOT models, particularly when combined with multi-task learning (MTL), demonstrate promising performance gains over the conventional SOT-based model. On Libri2Mix, the best performing SC-SOT+MTL model (using speaker embeddings only) achieves a WER of 17.5% / 15.2%, a substantial improvement over the baseline SOT's 19.2% / 16.8%. Similarly, on Libri3Mix, the same SC-SOT+MTL configuration yields a WER of 33.2% / 30.9%, outperforming the baseline SOT. Introducing speaker embeddings alone (rows 5 and 6) consistently lowers WER compared to unconditioned SOT models, highlighting the utility of speaker identity information for improved implicit speaker separation by the decoder. It is important to note that these WER improvements were achieved without any additional inputs.

Incorporating only speaker activity information (row 6) also contributes to WER reduction. However, the combination of conditioning on both speaker embeddings and speaker activity information (row 7) showed WER increase from SC-SOT with a single condition. Particularly, SC-SOT earned worse WER than SOT on Libri3Mix. This WER increase seems to be caused by many errors in speaker activity information. We consistently observed that the accuracy of the diarization module is distorted when speaker embeddings are used for conditioning during training (row 4, 5, and 7). This may indicate the optimal representation of speaker embeddings is different between ASR module and diarization module, and the gradient from SOT objective guided the representation of speaker embeddings to be more optimal for ASR task than for diarization task.

To isolate the potential of combining both conditioning inputs, we conducted an evaluation using oracle speaker diarization information alongside the predicted speaker embeddings (row 8). This configuration reduced WER, particularly on Libri3Mix. This result strongly indicates that the concept of combining speaker embeddings and activity information is highly beneficial, and that the performance limitations observed in row 7 are primarily due to the accuracy of the speaker diarization module, rather than an inherent incompatibility between the two conditioning types. The gap between the performance with predicted (row 7) and oracle (row 8) speaker activity information highlights the substantial gains achievable with improved speaker diarization accuracy.

6. Conclusion

This work investigated and enhanced the implicit speaker separation capabilities of SOT-based models. Our initial analysis, through visualization of source-target attention weights in an SOT-based model, revealed that the decoder performs a degree of speaker separation. Building upon this key insight, we proposed an approach: conditioning the decoder on speaker embeddings and speaker activity information to explicitly augment this implicit separation process. We achieved relative reductions in WER of up to 9.5% on Libri2Mix and 10.4% on Libri3Mix. Furthermore, experiments utilizing oracle speaker diarization information as speaker activity input yielded even greater accuracy. This highlights the significant potential of our approach and points towards future research directions focused on refining the speaker diarization component.

7. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681.

8. References

[1] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[2] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[4] F. M. Ramirez, L. Chkhetiani, A. Ehrenberg, R. McHardy, R. Botros, Y. Khare, A. Vanzo, T. Peyash, G. Oexle, M. Liang *et al.*, "Anatomy of industrial scale multilingual asr," *arXiv preprint arXiv:2404.09841*, 2024.

[5] K. C. Puvvada, P. Żelasko, H. Huang, O. Hrinchuk, N. R. Koluguri, K. Dhawan, S. Majumdar, E. Rastorgueva, Z. Chen, V. Lavrukhin *et al.*, "Less is more: Accurate speech recognition & translation without web-scale data," *arXiv preprint arXiv:2406.19674*, 2024.

[6] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint arXiv:1704.01985*, 2017.

[7] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," *arXiv preprint arXiv:2003.12687*, 2020.

[8] J. Shi, X. Chang, P. Guo, S. Watanabe, Y. Fujita, J. Xu, B. Xu, and L. Xie, "Sequence to multi-sequence learning via conditional chain mapping for mixture signals," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3735–3747, 2020.

[9] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming multi-talker asr with token-level serialized output training," *arXiv preprint arXiv:2202.00842*, 2022.

[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*, 2023.

[11] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[12] H. Seki, T. Hori, S. Watanabe, J. L. Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," *arXiv preprint arXiv:1805.05826*, 2018.

[13] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 237–244.

[14] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," *arXiv preprint arXiv:2005.09921*, 2020.

[15] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[16] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[17] J. Janskỳ, J. Málek, J. Čmejla, T. Kounovskỳ, Z. Koldovskỳ, and J. Žd'ánskỳ, "Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 676–680.

[18] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny *et al.*, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," *arXiv preprint arXiv:2005.07272*, 2020.

[19] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] Z. Huang, D. Raj, P. García, and S. Khudanpur, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2023, pp. 1–5.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[22] A. Polok, D. Klement, M. Wiesner, S. Khudanpur, J. Černockỳ, and L. Burget, "Target speaker asr with whisper," *arXiv preprint arXiv:2409.09543*, 2024.

[23] A. Polok, D. Klement, M. Kocour, J. Han, F. Landini, B. Yusuf, M. Wiesner, S. Khudanpur, J. Černockỳ, and L. Burget, "Dicow: Diarization-conditioned whisper for target speaker automatic speech recognition," *arXiv preprint arXiv:2501.00114*, 2024.

[24] N. Kanda, X. Xiao, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8082–8086.

[25] A. Polok, D. Klement, J. Han, Š. Sedláček, B. Yusuf, M. Maciejewski, M. S. Wiesner, and L. Burget, "But/jhu system description for chime-8 notsofar-1 challenge," in *Proc. CHiME 2024*, pp. 18–22.

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.

[28] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[30] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[31] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[32] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.