

SERIALIZED SPEECH INFORMATION GUIDANCE WITH OVERLAPPED ENCODING SEPARATION FOR MULTI-SPEAKER AUTOMATIC SPEECH RECOGNITION

Hao Shi¹, Yuan Gao¹, Zhaoheng Ni², Tatsuya Kawahara¹

¹Graduate School of Informatics, Kyoto University, Kyoto, Japan

²Meta, New York, USA

ABSTRACT

Serialized output training (SOT) attracts increasing attention due to its convenience and flexibility for multi-speaker automatic speech recognition (ASR). However, it is not easy to train with attention loss only. In this paper, we propose the overlapped encoding separation (EncSep) to fully utilize the benefits of the connectionist temporal classification (CTC) and attention (CTC-Attention) hybrid loss. This additional separator is inserted after the encoder to extract the multi-speaker information with CTC losses. Furthermore, we propose the serialized speech information guidance SOT (GEncSep) to further utilize the separated encodings. The separated streams are concatenated to provide single-speaker information to guide attention during decoding. The experimental results on Libri2Mix and Libri3Mix show that the single-speaker encoding can be separated from the overlapped encoding. The CTC loss helps to improve the encoder representation under complex scenarios (three-speaker and noisy conditions), which makes the EncSep have a relative improvement of more than 8% and 6% on the noisy Libri2Mix and Libri3Mix evaluation sets, respectively. GEncSep further improved performance, which was more than 12% and 9% relative improvement for the noisy Libri2Mix and Libri3Mix evaluation sets.

Index Terms— Automatic speech recognition, multi-speaker, overlapped encoding separation, serialized output training

1. INTRODUCTION

Automatic speech recognition gets impressive performance with the development of deep learning [1–6]. The word error rate (WER) of ASR for single speaker conditions has achieved the level of human transcribers [3, 7], even when faced with many complex scenarios, such as noise [8–12]. However, compared to the additive noise and reverberation, the inference from other speakers, known as the cocktail problem, has a more severe effect on ASR [13–15]. As a result, ASR performance has dramatic degradation under multi-speaker scenarios [14].

Recently, extensive research has been conducted on multi-speaker ASR [14, 15]. It is intuitive to decompose the multi-speaker ASR task to speech separation and then recognition [16, 17]. The pipeline consists of a speech separation front-end [18, 19] followed by a recognizer [16]. However, speech separation front-ends, especially the single-channel methods, often bring speech information loss and distortion issues, which harms ASR [16, 20]. With the development of end-to-end ASR, the ASR back-end already has some abilities to handle multi-speaker conditions [21, 22]. Utterance-level permutation invariant training (uPIT) [21] is popular for multi-speaker ASR [23]. During training, all possible permutations of speakers are needed to compute loss, and the smallest loss is used for backpropagation [23]. In the uPIT-based ASR, the number of output layers constrains the maximum number of speakers [24], and training becomes computationally complex with more speakers [24].

The serialized output training (SOT) [24] is proposed to solve the above-mentioned drawbacks. SOT-based ASR is based on the attention-based encoder-decoder (AED) structure [25, 26]. It designs the training label: the overlapped speeches are serialized into a single token sequence according to the speaking start-time of each speaker [24]. The problem of the variable speaker number could also be alleviated without the performance degradation compared to uPIT-based ASR [24]. Given the success of SOT, more studies improve the model's performance by designing training targets. The timestamp of each speaker is inserted as a token to the training label [27]. Besides, the speaker diarization task also helps the SOT-based ASR [28].

The connectionist temporal classification (CTC) [29] enforces the monotonic alignment between speech and label sequences, which helps the attention by solving the misalignment issues [30]. Thus, the CTC and attention (CTC-Attention) hybrid loss [30] is widely adopted for training the ASR systems. Although it is commonly used in single-speaker conditions, it is difficult to use it in SOT-based ASR because the serialized training label is difficult to align with the overlapped speech embedding. Thus, the SOT-based ASR systems only use attention cross-entropy loss for training [24, 27, 28].

In this paper, we first propose the overlapped encoding

All the experiments were conducted at Kyoto University

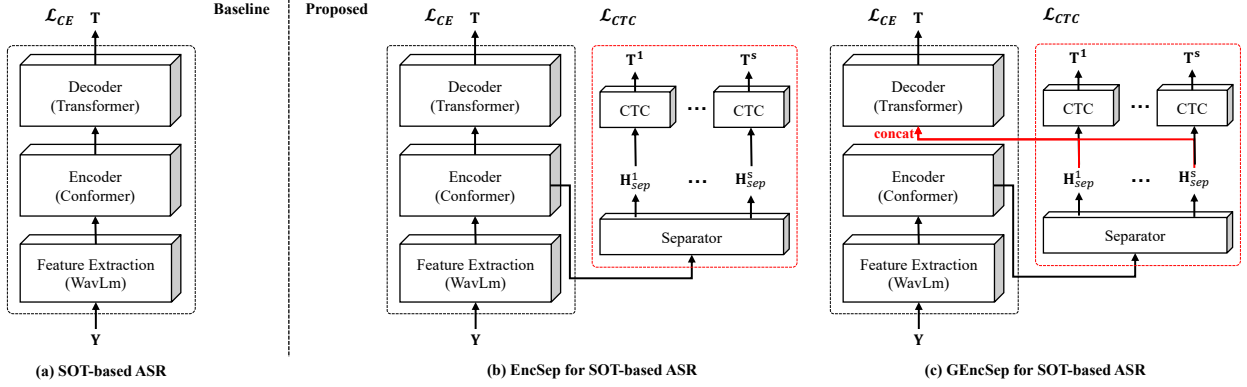


Fig. 1. Flowchart of training strategies in different multi-speaker ASR systems.

separation (EncSep) to improve the encoding representation in the SOT-based ASR by utilizing the CTC-Attention loss. An additional separator is used to get the single speaker's encoding from the encoder's overlapped speech embedding. The separated encodings are arranged according to speaking time, and the CTC loss is computed. We also propose the single-speaker information guidance SOT (GenSep) to further utilize the separated encodings, which are concatenated for decoding. The attention mechanisms are used to focus on the different single-speaker information from the concatenated embedding.

2. RELATED WORKS

2.1. Utterance-level Permutation Invariant Training (uPIT) for Multi-speaker ASR

uPIT-based multi-speaker ASR contains a shared mixed encoder, several speaker-differentiating (SD) encoders, a shared recognition encoder, a shared attention module, and a shared decoder. The mixed encoder transforms the overlapped speech feature \mathbf{Y} to embedding \mathbf{H} . Then, the extracted mixed embeddings are fed into several SD encoders. Each SD encoder only extracts one speaker information \mathbf{H}_{SD}^s , s represents the s -th speaker. The shared recognition encoder transforms the representation of the s -th speaker from SD encoders to high-level representations for recognition \mathbf{H}_{Reg}^s .

Finding a definite relationship between the separated features and multiple speakers is difficult for multi-speaker processing. uPIT was proposed to solve this permutation issue by computing the smallest loss with all different estimation-label permutations. uPIT is adopted to compute the CTC losses between the encoding of the recognition encoder \mathbf{H}_{Reg}^s and its corresponding label \mathbf{T}^s . An additional linear layer is inserted after the recognition encoder to compute the CTC loss and get the transcription \mathbf{C}^s . The number of permutations between the recognition encodings and labels is $S!$, where S represents the speaker number. The loss function of CTC with uPIT is

as follows:

$$\mathcal{L}_{CTC-uPIT} = \arg \min_{\pi \in \mathcal{P}} \sum_{s=1}^S \text{Loss}_{CTC}(\mathbf{C}^s, \mathbf{T}^{\pi(s)}), \quad (1)$$

where \mathcal{P} represents the set of all permutations of $1, \dots, S$. $\pi(s)$ represents the s -th element in a permutation π , and \mathbf{T} is the set of transcription labels for S speakers. The recognition encodings \mathbf{H}_{Reg} are finally fed into the shared attention and the shared decoder to get the attention-based outputs \mathbf{H}_{Att}^s . The loss of the decoder is cross-entropy (CE) with the same permutation as that of the CTC-uPIT:

$$\mathcal{L}_{CE} = \sum_{s=1}^S \text{Loss}_{CE}(\mathbf{H}_{Att}^s, \mathbf{T}^{\hat{\pi}(s)}) \quad (2)$$

where $\hat{\pi}(s)$ represents the permutation obtained from the CTC-uPIT. The final training loss for uPIT-based ASR is CTC-Attention Hybrid loss:

$$\mathcal{L}_{uPIT} = \lambda \mathcal{L}_{CTC-uPIT} + (1 - \lambda) \mathcal{L}_{CE}. \quad (3)$$

λ is the hyperparameter to control the two losses.

2.2. Serialized Output Training (SOT) for Multi-speaker ASR

SOT is based on the attention-based encoder-decoder (AED) structure to solve the multi-speaker ASR, which is shown in Fig. 1-(a). It contains an encoder, an attention module, and a decoder. All three components cooperate during the decoding to output transcriptions of different speakers according to their start-time of speaking. The encoder transforms the overlapped speech feature \mathbf{Y} to embedding \mathbf{H} . Different from the uPIT-based ASR, the overlapped speech embedding is directly fed into the attention and decoder to get the transcriptions without any implicit or explicit separation. Training targets help the model achieve this ability. It innovatively arranges the transcriptions of different speakers according to their start-time of speaking to form a new transcription. A special symbol $\langle sc \rangle$ is inserted between

different speakers to represent the speaker change. For example, for a two-speaker case, the label will be given as $\mathbf{T} = \{t_1^1, \dots, t_1^{N^1}, \langle sc \rangle, t_2^1, \dots, t_2^{N^2}\}$, where t_1 and t_2 represent the transcriptions of the 1-th and 2-nd speaker, respectively. The N^1 and N^2 represent the lengths of the transcriptions. Based on this training target, the attention mechanism is able to focus on the required information in encoding the overlapped speech and decoding the transcriptions of multiple speakers \mathbf{C} according to their speaking time. The loss function of SOT-based ASR is simple as follows:

$$\mathcal{L}_{\text{SOT}} = \text{Loss}_{\text{CE}}(\mathbf{C}, \mathbf{T}), \quad (4)$$

Because the embedding obtained by the encoder does not perform any separation, it is hard to align the speech embedding with the label sequence by CTC. As a result, only CE loss is used when training the SOT-based ASR system.

2.3. Self-Supervised Learning-based Feature Extraction

Feature extraction with self-supervised learning (SSL) shows powerful performance improvement for ASR [9]. The outputs of the SSL module replace the traditional features like Mel-Frequency Cepstral Coefficients (MFCC) and Filter Bank (F-Bank). The SSL-based feature extraction module [31–33] typically consists of convolution layers and transformer, e.g., WavLM [34], HuBERT [35]. With massive pretraining data, the SSL models already contained the capability to convert the speech waveform into the acoustic hidden units. Moreover, some SSL models [34, 35] demonstrate robustness against noise or other inference speakers by incorporating noisy simulation. The other parts of the ASR system are the same as the other E2E ASR systems.

3. PROPOSED METHOD

To improve the encoder's representation, we first propose the overlapped encoding separation (EncSep) to utilize the CTC-Attention hybrid loss in the SOT-based ASR. Then, we propose the single-speaker information guidance SOT (GEncSep) to utilize the separated embeddings.

3.1. Overlapped Encoding Separation (EncSep)

In the SOT-based ASR, the encoder transforms the overlapped speech feature \mathbf{Y} to embedding \mathbf{H} . The embedding \mathbf{H} here is overlapped by speakers. Thus, some module for speech separation is introduced to convert the overlapped encoding \mathbf{H} into separated single-speaker encodings \mathbf{H}_{sep} :

$$\mathbf{H}_{\text{sep}} = \text{Separator}(\mathbf{H}). \quad (5)$$

The Long Short-Term Memory (LSTM) [36] is adopted as the separator in this work:

$$\begin{aligned} \hat{\mathbf{H}} &= \text{LayerNorm}(\text{LSTM}(\mathbf{H})), \\ \mathbf{H}_{\text{sep}}^s &= \text{ReLU}(\text{Linear}^s(\hat{\mathbf{H}})), \end{aligned} \quad (6)$$

Several linear layers are used to generate the single-speaker information $\mathbf{H}_{\text{sep}}^s$. Linear layers correspond to a speaker according to the start time of speaking in a serialized manner. Then, the CTC loss is computed as follows:

$$\begin{aligned} \mathcal{L}_{\text{CTC-EncSep}} &= \sum_{s=1}^S \text{Loss}_{\text{CTC}}(\mathbf{C}^s, \mathbf{T}^s) \\ &= \sum_{s=1}^S \text{Loss}_{\text{CTC}}(\text{Linear}^s(\mathbf{H}_{\text{sep}}^s), \mathbf{T}^s) \end{aligned} \quad (7)$$

where \mathbf{T}^s represents the s -th transcription arranged in the serialized manner.

The attention-based CE loss, same as Eqn. (4), is also used for training. The final loss function for training EncSep is as follows:

$$\mathcal{L}_{\text{EncSep}} = \gamma \mathcal{L}_{\text{CTC-EncSep}} + (1 - \gamma) \mathcal{L}_{\text{SOT}}, \quad (8)$$

where the γ is the hyperparameter to balance the two losses. The flowchart of the proposed EncSep training strategy is shown in Fig. 1–(b). The separator is only used to introduce CTC information during training. When decoding, EncSep maintains the same structure as the SOT-based method (Fig. 1–(a)). Thus, EncSep does not increase any computational cost during decoding from the SOT-based method.

3.2. Single-speaker Information Guidance SOT (GEncSep)

EncSep only uses the separated embeddings \mathbf{H}_{sep} to introduce the CTC loss into the encoder. To further utilize the separated embeddings \mathbf{H}_{sep} , we also propose the single-speaker information guidance SOT (GEncSep). Compared with the EncSep, the GEncSep utilizes the separated embeddings \mathbf{H}_{sep} from the separator, which is shown in Fig. 1–(c).

GEncSep contains an encoder, which transforms the overlapped speech feature \mathbf{Y} to embedding \mathbf{H} . The separator separates the overlapped embedding \mathbf{H} into single-speaker embeddings $\mathbf{H}_{\text{sep}}^1, \dots, \mathbf{H}_{\text{sep}}^S$. Then, the separated embeddings are concatenated over time dimension as:

$$\mathbf{H}_{\text{con}} = \text{Concat}(\mathbf{H}_{\text{sep}}^1, \dots, \mathbf{H}_{\text{sep}}^S) \quad (9)$$

The attention mechanism is used to compute the attention weights with single-speaker information:

$$\mathbf{a}_{\text{con}}^n = \text{Attention}(\mathbf{H}_{\text{con}}, \mathbf{d}^{n-1}). \quad (10)$$

$\mathbf{a}_{\text{con}}^n$ represents the context vector obtained with the attention mechanism from the concatenated embedding \mathbf{H}_{con} . \mathbf{d}^{n-1} is the hidden state of the decoder. The decoder decodes from the attention feature and also the previously predicted tokens:

$$\mathbf{c}^n = \text{Decoder}(\mathbf{H}_{\text{con}}, \mathbf{a}_{\text{con}}^n, \mathbf{c}^{1:n-1}). \quad (11)$$

The corresponding output \mathbf{C} is generated iteratively by the decoder. The training loss function of GEncSep is the same as Eqn. (8).

4. EXPERIMENTS

4.1. Datasets and Experimental Settings

We used the LibriMix dataset [37] to evaluate the model performance. It used the train-clean-100, train-clean-360, dev-clean, and test-clean subsets from the LibriSpeech dataset [38] as the clean speech. The noise samples were taken from WHAM! dataset [39]. We used the official scripts¹ to synthesize Libri2Mix and Libri3Mix. We used the offset file to make different speaking start times for multiple speakers. The two-speaker offset files followed the official ESPnet setting², and the three-speaker offset files were created by ourself.

Conformer [3] was used as the ASR back-end. Except for the proposed modules, the other modules followed the official ESPnet settings³. The baselines and EncSep used the original Transformer layer. For GEncSep, the structure of the decoder layer is shown in Fig. 1–(c). A character-based vocabulary was used with the size 32, including a $\langle sc \rangle$ for speaker change symbol. We used the WavLM-Large [34] for feature extraction, which has been shown effective for noise-robust ASR [40]. During training, all parameters of the WavLM-Large were frozen. In the proposed method, the input and output dimensions of the separator were 256 and 512; the number of LSTM layers were 2. The dimensions of linear layers were 512 and 256 for input and output, respectively. We also tried bidirectional LSTM layers. For bidirectional LSTM, the dimensions of linear layers were 1024 and 256 for input and output, respectively.

For baselines, the “SOT” followed the ESPnet official setting; only the attention CE loss was used with Eqn. (4). “SOT-H” was selected as another baseline. Different from the “SOT”, it used the CTC-Attention hybrid loss instead of only Attention loss with the serialized label to train the model. Its neural network structure followed the ESPnet setting.

The training epochs were 60, and the final evaluation model was averaged over ten checkpoints, according to the loss in the validation set. It should be emphasized that pre-training is unnecessary for the two-mixed condition, but a few epochs of original SOT training are needed for the three-mixed condition. The all parameters of GEncSep was pretrained with EncSep.

4.2. Experimental Results

Table 1 shows the experimental results on noisy Libri2Mix and Libri3Mix sets. Compared with “SOT”, “SOT-H” had a significant performance degradation on both the development and evaluation sets of Libri2Mix with the CTC-Attention

Table 1. Experimental Results on Noisy Libri2Mix and Libri3Mix. “Bi” represents the Bi-directional.

Exp.	Systems	Bi	Libri2Mix		Libri3Mix	
			Dev	Eval	Dev	Eval
1	SOT	-	19.4	17.1	30.5	28.2
2	SOT-H	-	24.6	22.0	-	-
3	EncSep	✗	18.4	15.9	28.5	26.5
4	EncSep	✓	18.0	15.7	28.6	26.4
5	GEncSep	✗	18.3	15.3	28.8	25.7
6	GEncSep	✓	17.2	15.0	28.0	25.9

Table 2. Experimental Results on Clean Libri2Mix and Libri3Mix. “Bi” represents the Bi-directional.

Exp.	Systems	Bi	Libri2Mix		Libri3Mix	
			Dev	Eval	Dev	Eval
7	SOT	-	6.8	7.0	15.0	14.7
8	SOT-H	-	10.4	10.3	-	-
9	EncSep	✗	7.0	7.3	13.9	13.4
10	EncSep	✓	7.0	7.2	13.9	13.5
11	GEncSep	✗	6.7	6.8	13.0	14.3
12	GEncSep	✓	6.4	6.6	13.3	13.1

hybrid loss, which confirms that the encoder’s implicit separation performed poorly (comparison between Exp. 1 and Exp. 2). With the separator, the proposed “EncSep” significantly improved from “SOT” (p-value < 0.01) for both the development and evaluation sets of Libri2Mix and Libri3Mix (comparison between Exp. 1 and Exp. 3). It was shown that the CTC losses with separator benefited the encoder representation since “EncSep” has the same structure as “SOT” during decoding. The bi-directional separator did not bring further improvement (comparison between Exp. 3 and Exp. 4). For “GEncSep”, the experimental results suggested that providing the separated embedding helped the decoding, significantly improving the performance on both the development and evaluation sets of Libri2Mix and Libri3Mix compared with “SOT” (p-value < 0.01, comparison between Exp. 1 and Exp. 5). Furthermore, the bi-directional separator showed significant improvement (p-value < 0.01, comparison between Exp. 5 and Exp. 6) on development sets of Libri2Mix and Libri3Mix, which were used to select the evaluation model.

Table 2 shows the experimental performance on clean Libri2Mix and Libri3Mix sets. Similar to the noisy conditions, “SOT-H” had a significant performance degradation

¹<https://github.com/JorisCos/LibriMix>

²https://github.com/espnet/espnet/tree/master/egs2/librimix/sot_asr1

³https://github.com/espnet/espnet/blob/master/egs2/librimix/sot_asr1/conf/tuning/train_sot_asr_conformer_wavlm.yaml

Table 3. Comparison of ASR systems on Libri2Mix and Libri3Mix. “SSL” represents the self-supervised learning.

Systems	w/ Front-end	w/ SSL	Libri2Mix		Libri3Mix	
			Dev	Eval	Dev	Eval
Noisy						
PIT-Conformer ⁴	✗	✗	23.7	23.5	-	-
Conditional-Conformer [41]	✗	✗	24.5	24.9	-	-
WavLM Base [42]	✗	✓	-	27.5	-	-
TS-HuBERT [42]	✗	✓	-	24.8	-	-
SOT-Conformer ⁵	✗	✓	19.4	17.1	30.5	28.2
TSE-Whisper [43]	✓	✗	-	12.0	-	-
GEncSep (this study)	✗	✓	17.2	15.0	28.0	25.9
Clean						
W2V-baseline [44]	✗	✓	11.6	12.3	-	-
W2V-Sidecar-ft [44]	✗	✓	7.7	8.1	-	-
WavLM Base+ [45]	✗	✓	-	8.3	-	-
WavLM Base+ [45]	✓	✓	-	7.6	-	-
TSE-CLN [45]	✓	✓	7.1	7.6	-	-
C-HuBERT LARGE [35]	✗	✓	6.6	7.8	-	-
SOT-Conformer	✗	✓	6.8	7.0	15.0	14.7
GEncSep (this study)	✗	✓	6.4	6.6	13.3	13.1

from “SOT” (comparison between Exp. 7 and Exp. 8). The “EncSep” did not show any improvement on Libri2Mix (comparison between Exp. 7 and Exp. 9). Compared with the noisy two-speaker condition, the “SOT” already had strong abilities for encoding the clean two-speaker features. CTC helped the encoder improve its encoding capabilities in more complex scenarios. The proposed “EncSep” still significantly improved the three-speaker conditions (p-value < 0.01, comparison between Exp. 7 and Exp. 9). The experimental results for “GEncSep” suggested that providing the separated embedding also helped the decoding under clean conditions, especially for three-speaker conditions (all the Libri2Mix and Libri3Mix evaluation sets are significantly improved from “SOT”, p-value < 0.01, comparison between Exp. 7 and Exp. 12).

We compared several noise robust end-to-end ASR systems in the literature (after 2020) in Table 3. It should be emphasized that although the proposed method showed compelling performance in the end-to-end style for multi-speaker ASR, there is still a gap compared to the systems with pipeline systems (speech separation front-end with recognizer, TSE-Whisper) under noisy two-speaker conditions. Another possible reason for such a gap is that whisper’s training data is massive, while other ASR back-ends only use the noisy Libri2Mix training set. For clean sets, the advantages of the front-end did not perform as well as the “SOT-Conformer”.

Compared with “SOT-Conformer”, the proposed GEncSep had significant performance improvement (p-value < 0.05). Compared with other methods, the SOT-based multi-speaker ASR system has advantages for more speakers.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we focused on improving the serialized output training (SOT) for ASR. We first proposed the overlapped encoding separation (EncSep) to fully utilize the benefits of the CTC and attention hybrid loss. The experimental results on Libri2Mix and Libri3Mix datasets show that the single-speaker encoding can be separated from the overlapped encoding. The CTC losses with separator benefit the encoder representation. Then, we proposed the serialized speech information guidance SOT (GEncSep) to further utilize the separated information. GEncSep further improved performance with the separated embeddings for decoding. Compared with the clean Libri2Mix and Libri3Mix, the proposed method has more significant advantages over the more complex, noisy Libri2Mix and Libri3Mix. As a result, the proposed GEncSep had more than 12% and 9% relative improvement for the noisy Libri2Mix and Libri3Mix evaluation sets compared to the original SOT. In the future, we will make the system use different information sources by fusing the overlapped and separated embeddings.

6. REFERENCES

- [1] L. Deng and X. Li, “Machine learning paradigms for speech recognition: An overview,” *IEEE TASLP*, vol. 21, no. 5, pp. 1060–1089, 2013.
- [2] H. Shi, M. Mimura, and T. Kawahara, “Waveform-domain speech enhancement using spectrogram encoding for robust speech recognition,” *IEEE/ACM TASLP*, vol. 32, pp. 3049–3060, 2024.
- [3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *Proc. Interspeech*, 2020, pp. 5036–5040.
- [4] H. Sun, S. Zhao, X. Kong, X. Wang, H. Wang, J. Zhou, and Y. Qin, “Iterative prototype refinement for ambiguous speech emotion recognition,” in *Interspeech 2024*, 2024, pp. 3200–3204.
- [5] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, “Fine-grained disentangled representation learning for multimodal emotion recognition,” in *Proc. ICASSP. IEEE*, 2024, pp. 11051–11055.
- [6] H. Sun, S. Zhao, S. Li, X. Kong, X. Wang, A. Kong, J. Zhou, Y. Chen, W. Zeng, and Y. Qin, “Enhancing emotion recognition in incomplete data: A novel cross-modal alignment, reconstruction, and refinement framework,” *arXiv preprint arXiv:2407.09029*, 2024.
- [7] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 745–777, 2014.
- [8] H. Shi, M. Mimura, L. Wang, J. Dang, and T. Kawahara, “Time-domain speech enhancement assisted by multi-resolution frequency encoder and decoder,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [9] X. Chang, T. Maekaku, Y. Fujita, and S. Watanabe, “End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation,” in *Proc. Interspeech*, 2022, pp. 3819–3823.
- [10] H. Shi, L. Wang, S. Li, C. Fan, J. Dang, and T. Kawahara, “Spectrograms fusion-based end-to-end robust automatic speech recognition,” in *Proc. APSIPA ASC*, 2021, pp. 438–442.
- [11] J. Li, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, pp. –, 2022.
- [12] H. Shi and T. Kawahara, “Exploration of adapter for noise robust automatic speech recognition,” *arXiv preprint arXiv:2402.18275*, 2024.
- [13] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, “Using Semi-supervised Learning for Monaural Time-domain Speech Separation with a Self-supervised Learning-based SI-SNR Estimator,” in *Proc. INTERSPEECH*, 2023, pp. 3759–3763.
- [14] N. Kanda, S. Horiguchi, R. Takashima, Y. Fujita, K. Nagamatsu, and S. Watanabe, “Auxiliary Interference Speaker Loss for Target-Speaker Speech Recognition,” in *Proc. Interspeech*, 2019, pp. 236–240.
- [15] T. v. Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, “Multi-Talker ASR for an Unknown Number of Sources: Joint Training of Source Counting, Separation and ASR,” in *Proc. Interspeech*, 2020, pp. 3097–3101.
- [16] A. Narayanan and D. Wang, “Investigation of speech separation as a front-end for noise robust speech recognition,” *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 826–835, 2014.
- [17] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, “A separation priority pipeline for single-channel speech separation in noisy environments,” in *Proc. ICASSP*, 2024, pp. 12511–12515.
- [18] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [19] Y. Wang, J. Zhang, S. Chen, W. Zhang, Z. Ye, X. Zhou, and L. Dai, “A study of multichannel spatiotemporal features and knowledge distillation on robust target speaker extraction,” in *Proc. ICASSP*, 2024, pp. 431–435.
- [20] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, “How does end-to-end speech recognition training impact speech enhancement artifacts?,” *arXiv preprint arXiv:2311.11599*, 2023.
- [21] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM TASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.
- [22] A. Tripathi, H. Lu, and H. Sak, “End-to-end multi-talker overlapping speech recognition,” in *Proc. ICASSP*, 2020, pp. 6129–6133.
- [23] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *Proc. ICASSP*, 2017, pp. 241–245.

- [24] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized Output Training for End-to-End Overlapped Speech Recognition," in *Proc. Interspeech*, 2020, pp. 2797–2801.
- [25] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *Proc. NIPS*, 2014.
- [26] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell," *arXiv preprint arXiv:1508.01211*, 2015.
- [27] N. Makishima, K. Suzuki, S. Suzuki, A. Ando, and R. Masumura, "Joint Autoregressive Modeling of End-to-End Multi-Talker Overlapped Speech Recognition and Utterance-level Timestamp Prediction," in *Proc. INTERSPEECH*, 2023, pp. 2913–2917.
- [28] P. Shen, X. Lu, and H. Kawai, "Speaker mask transformer for multi-talker overlapped speech recognition," *arXiv preprint arXiv:2312.10959*, 2023.
- [29] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.
- [30] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [31] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, vol. 2023, pp. 765–769.
- [32] J. He, X. Shi, X. Li, and T. Toda, "Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction," *arXiv preprint arXiv:2401.13260*, 2024.
- [33] J. Tian, D. Hu, X. Shi, J. He, X. Li, Y. Gao, T. Toda, X. Xu, and X. Hu, "Semi-supervised multimodal emotion recognition with consensus decision-making and label correction," in *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*, 2023, pp. 67–73.
- [34] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "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.
- [35] M. Fazel-Zarandi and W.-N. Hsu, "Cocktail hubert: Generalized self-supervised pre-training for mixture and single-source speech," in *Proc. ICASSP*, 2023, pp. 1–5.
- [36] A. Graves, *Long Short-Term Memory*, pp. 37–45, Springer Berlin Heidelberg, 2012.
- [37] 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, Available: arXiv:2005.11262.
- [38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [39] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019.
- [40] X. Chang, T. Maekaku, Y. Fujita, and S. Watanabe, "End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation," *arXiv:2204.00540*, 2022.
- [41] P. Guo, X. Chang, S. Watanabe, and L. Xie, "Multi-Speaker ASR Combining Non-Autoregressive Conformer CTC and Conditional Speaker Chain," in *Proc. Interspeech*, 2020, pp. 3720–3724.
- [42] W. Zhang and Y. Qian, "Weakly-Supervised Speech Pre-training: A Case Study on Target Speech Recognition," in *Proc. Interspeech*, 2023, pp. 3517–3521.
- [43] W. Zhang, L. Yang, and Y. Qian, "Exploring time-frequency domain target speaker extraction for causal and non-causal processing," in *Proc. ASRU*, 2023, pp. 1–6.
- [44] L. Meng, J. Kang, M. Cui, Y. Wang, X. Wu, and H. Meng, "A sidecar separator can convert a single-talker speech recognition system to a multi-talker one," in *Proc. ICASSP*, 2023, pp. 1–5.
- [45] Z. Huang, D. Raj, P. García, and S. Khudanpur, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *Proc. ICASSP*, 2023, pp. 1–5.