ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888074

Elevating Robust ASR By Decoupling Multi-Channel Speaker Separation and Speech Recognition

Yufeng Yang[1], Hassan Taherian[1], Vahid Ahmadi Kalkhorani[1], DeLiang Wang[1,2]
[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{yang.5662, taherian.1, ahmadikalkhorani.1, wang.77}@osu.edu

***Abstract*—Despite the tremendous success of automatic speech recognition (ASR) with the introduction of deep learning, its performance is still unsatisfactory in many real-world multi-talker scenarios. Speaker separation excels in separating individual talkers but, as a frontend, it introduces processing artifacts that degrade the ASR backend trained on clean speech. As a result, mainstream robust ASR systems train on noisy speech to avoid processing artifacts. In this work, we propose to decouple the training of the multi-channel speaker separation frontend and the ASR backend, with the latter trained only on clean speech. On SMS-WSJ, the proposed approach achieves a word error rate (WER) of 5.74%, outperforming the previous best by 14.3%. Furthermore, on recorded LibriCSS, we achieve the speaker-attributed WER of 3.86%, outperforming the previous best system trained on the same data by 24.8%. These state-of-the-art results suggest that decoupling speech separation and recognition is a potentially effective approach to robust ASR.**

***Index Terms*—LibriCSS, multi-channel, robust ASR, SMS-WSJ, speaker separation**

I. Introduction

In the era of deep learning, automatic speech recognition (ASR) has made tremendous strides, progressing from conventional Gaussian mixture model plus hidden Markov model (GMM-HMM) approaches [1], to hybrid systems of deep neural network and HMM (DNN-HMM) [2] to end-to-end (E2E) systems [3]. ASR has been seamlessly integrated into our daily lives as a fundamental component in personal assistants and home devices, significantly boosting human-computer interaction. Despite its success, ASR is prone to acoustic interference, such as reverberation, background noise, and overlapping speakers. The unavoidable mismatch between speech signals in such environments and the transcribed training data in clean conditions poses a persistent obstacle, demanding the development of robust ASR systems capable of overcoming this mismatch [4], [5].

Meanwhile, the introduction of deep learning has also dramatically improved speech separation performance [6], including the intelligibility and quality of degraded speech signals. To address the robust ASR problem, a straightforward approach is to leverage a speech separation model as the frontend for an ASR backend model. Speech separation models operate in the time [7], [8] or time-frequency (T-F) domains [9]–[11] for speech enhancement (speech-noise separation) or multi-talker speaker separation. Despite these effective frontends, the processing artifacts introduced in separation can be detrimental to the ASR backend trained on clean speech [12].

To address this problem, prevailing approaches train the ASR model directly on noisy speech [4], [5], or enhanced speech [12], or train a joint system of speech separation frontend and ASR backend [13]. In multi-talker cases, training ASR on overlapped speech directly is problematic due to the large variety of mixed talkers. Although there are multi-talker ASR approaches, such as serialized output training [14], their performance degrades on single-talker ASR. Also, the amount of annotated data for conversational ASR is much smaller than that for single-talker ASR. Hence, speaker separation becomes necessary to address overlapped speech for robust ASR [15]. The selection of training data for the ASR backend has now become critical to achieving high performance. The mainstream strategy trains ASR on single-talker speech with various noise augmentations, as in the baseline system of the SMS-WSJ corpus [16]. However, when tested on clean speech, an inevitable performance gap arises between the ASR model trained on noisy speech and that trained on clean speech. By clean speech, we include single-talker utterances recorded in a variety of quiet places in the real environment, as done in the collection of the LibriSpeech corpus [17]. As speech separation performance continues to improve, is training the ASR backend on noisy speech still the most effective approach? This approach creates a mismatch between the separated speech that is aimed to be clean and the noisy training data.

In this work, we aim to eliminate such mismatch and elevate the robust ASR performance. We address multi-channel robust ASR in multi-talker scenarios and propose to decouple the stages of speaker separation and speech recognition. Our approach separates the training of a multi-channel speaker separation frontend and an ASR backend trained on clean speech only, i.e. each stage is trained separately without considering the other. In this way, the mismatch between the frontend output and the backend training data could be alleviated. As shown in [18], [19], with monaural speech enhancement, ASR trained on clean speech outperforms that trained on noisy speech. We do not investigate a joint model as it has been shown that, after joint fine-tuning, the performance of each part degrades on its original task for a robust ASR system designed with a speaker separation frontend and ASR backend [20].

The proposed decoupled system is evaluated on the SMS-WSJ and LibriCSS corpora. On SMS-WSJ, we achieve a word error rate (WER) of 5.74% by training the backend on clean speech only, outperforming by 14.3% the previous best [10] that trains the backend on reverberant-noisy speech with three times our training data. Moreover, on LibriCSS, we achieve the speaker-attributed WER of 3.86%, outperforming the previous best [21] by 24.8% with an ASR backend trained on the same clean speech data. These results with different frontend and backend combinations show that the proposed decoupled approach can substantially elevate the performance of robust ASR, and the ASR backend does not have to be trained on noisy speech as done in the mainstream approach.

The main contributions are summarized as follows. First, we develop a decoupled robust ASR system where the speaker separation frontend and ASR backend are independently trained, with the backend trained on clean speech. Second, this work demonstrates that, with a strong speaker separation frontend, training ASR on clean speech can elevate recognition performance; we have advanced the state-of-the-art results on the SMS-WSJ and LibriCSS datasets. The decoupled approach offers a strong alternative to the mainstream approach for robust conversational ASR.

The remainder of the paper is organized as follows. Section II formulates the system model and describes the DNN architectures of the frontend and backend. Section III describes the experimental setup and implementation details. Section IV presents the results and comparisons. Conclusions and discussions are made in Section V.

II. Methods

A. Problem Formulation

1) Speaker Separation: The physical model of a P-microphone mixture can be formulated in the short-time Fourier transform (STFT) domain as

$$\begin{aligned}\mathbf{Y}(t,f) &= \sum_{c=1}^{C}\mathbf{X}(c,t,f)+\mathbf{N}(t,f)\\ &= \sum_{c=1}^{C}(\mathbf{S}(c,t,f)+\mathbf{H}(c,t,f))+\mathbf{N}(t,f),\end{aligned} \tag{1}$$

where C is the number of speakers, which is set to 2 in this study, assuming at most 2 speakers talk simultaneously. $\mathbf{Y}(t,f)$, $\mathbf{X}(c,t,f)$, $\mathbf{N}(t,f)$, $\mathbf{S}(c,t,f)$, and $\mathbf{H}(c,t,f) \in \mathbb{C}^P$ respectively denote the STFT of the received mixture, reverberant image, reverberant noise, direct-path signal, and early reflections plus late reverberation at time t and frequency f of speaker c. We drop the index of t and f in later notations. Speaker separation aims to estimate $S_q(c)$ for each source at a reference microphone q given input $\mathbf{Y}$.

2) Automatic Speech Recognition: An ASR system estimates a word sequence $\mathbf{W}^*$ given a sequence of acoustic features $\mathbf{X}$ of speech signal $\mathbf{x}$, which can be formulated as

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P_{\mathcal{AM},\mathcal{LM}}(\mathbf{W}|\mathbf{X}), \tag{2}$$

where $\mathcal{AM}$ and $\mathcal{LM}$ denote an acoustic model (AM) and language model (LM), respectively. Using Bayes' theorem, Eq. 2 can be written as

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} p_{\mathcal{AM}}(\mathbf{X}|\mathbf{W})P_{\mathcal{LM}}(\mathbf{W}), \tag{3}$$

where $p_{\mathcal{AM}}$ and $P_{\mathcal{LM}}$ are AM likelihood and LM prior probability, respectively. An AM predicts the likelihood of acoustic features of a phoneme or another linguistic unit, and an LM provides a probability distribution over words or sequences of words in a speech corpus. In an E2E ASR system, the word sequence is predicted directly given $\mathbf{X}$.

B. Speaker Separation Frontend

1) SpatialNet: SpatialNet [10] is employed as one T-F domain multi-channel speaker separation frontend. It has interleaved narrow-band and cross-band blocks to exploit narrow-band and across-frequency spatial information, respectively. The narrow-band blocks process frequencies independently, and use a self-attention mechanism and temporal convolutional layers to perform spatial-feature-based speaker clustering and temporal smoothing and filtering, respectively. The cross-band blocks process frames independently, and use full-band linear layer and frequency convolutional layers to learn the correlation between all frequencies and adjacent frequencies, respectively.

SpatialNet performs complex spectral mapping by predicting the real and imaginary (RI) parts of the STFT of each talker from the stacked RI parts of the STFT of overlapped speech [22]. The separated waveforms are generated by performing an inverse STFT on the estimated RI parts.

2) TF-CrossNet: We use TF-CrossNet as another T-F domain multi-channel speaker separation frontend [11]. Motivated by SpatialNet [10], TF-CrossNet introduces positional encoding and a cross-temporal module after the cross-band module. This modification enhances temporal processing. TF-CrossNet also performs complex spectral mapping for speaker separation.

3) Speaker Separation via Neural Diarization: We leverage speaker separation via neural diarization (SSND) [21] for multi-talker speech recognition [23]. SSND performs speaker separation with the integration of speaker diarization. The diarization is performed with a multi-channel end-to-end neural diarization with an encoder-decoder-based attractor module (MC-EEND) trained with location-based training (LBT) [24] to resolve the permutation ambiguity. After diarization, a sequence of speaker embeddings computed from the non-overlapped speech is leveraged to facilitate the assignment of speakers to the output streams of the speaker separation model. In this way, speaker assignment is accomplished during the diarization process, instead of the speaker separation process.

C. Speech Recognition Backend

1) Factorized Time-delayed Neural Network: We utilize a factorized time-delayed neural network (TDNN-F) based on the WSJ Kaldi recipe [25] as one ASR backend. The AM consists of 8 TDNN-F layers, and the final word sequence is obtained by decoding the state posteriors with a default WSJ tri-gram Kaldi language model without additional N-best restoring. More details can be found in [16].

2) Wide-residual Conformer: We utilize an E2E ASR model in [19], which is a connectionist temporal classification (CTC) and attention Conformer-encoder Transformer-decoder E2E ASR model, denoted as wide-residual Conformer (WRConformer) [1]. It leverages the ASR recipe in ESPnet [26], and adapts the standard CTC/attention Conformer-encoder Transformer-decoder ASR recipe to WRConformer. In this adaptation, the 2-D convolution in the subsampling module is replaced by a modified wide-residual convolutional neural network (WRCNN), which comprises two ResBlocks (see [27]). The first ResBlock projects an input log-Mel feature to 512 dimensions, while the second Resblock maintains the same input and output dimensions. Each ResBlock subsamples time frames by a factor of 2, resulting in the total number of frames reduced by a factor of 4 after the processing of the subsampling module, matching the default ESPnet subsampling module. In WRConformer, the number of Conformer encoders is set to 10 and the other configurations are the same as the default ESPnet setting.

III. Experimental Setup

A. Datasets

1) SMS-WSJ: We employ SMS-WSJ [16] to evaluate ASR with multi-channel speaker separation in reverberant conditions. The dataset consists of 33561, 982, and 1332 train, validation, and test mixtures, respectively. All utterances are drawn from the WSJ0 and WSJ1 datasets [28]. The sampling rate is 8 kHz and the longer utterance determines the length of the mixture. The sensor array is a circle with a radius of 10 cm. T60s are samples in [0.2, 0.5] s, and the distance between the array and the speaker ranges in [1, 2] m. Additive white sensor noise is added with signal-to-noise ratio (SNR) ranges in [20, 30] dB. The first microphone is selected as the reference microphone.

[1] https://github.com/yfyangseu/espnet

2) LibriCSS: LibriCSS is a dataset designed for evaluating multi-talker speech recognition [23]. The dataset has 10 one-hour sessions each has 6 ten-minute mini-sessions with different overlap levels. These levels are 0S (no overlap and inter-utterance silence ranges in [0.1, 0.5] s), 0L (no overlap and inter-utterance silence ranges in [2.9, 3.0] s), and speaker overlap ratio at 10%, 20%, 30%, and 40%. The utterances are drawn from the LibriSpeech [17] test-clean set and the sampling rate is 16 kHz. The recordings are made in a meeting room with a seven-channel circular microphone array, which has six microphones evenly placed on a circle with a radius of 4.25 cm and an additional central microphone.

The meeting-style training data for MC-EEND and SSND is generated following a LibriCSS recipe[2], where the LibriSpeech training set is utilized. We follow the same data generation setup as [21] for speaker diarization and speaker separation training. The central microphone is treated as the reference microphone.

B. Frontend Configurations

1) SMS-WSJ: For SMS-WSJ, we employ TF-CrossNet [11] for multi-channel speaker separation. The model is configured with 12 TF-CrossNet blocks where the number of channels is set to 192 and the cross-band hidden dimension of 16. In addition, the narrow-band hidden dimension is set to 384, and the number of attention heads is 4. The STFT frame size and shift are 32 and 16 ms, respectively. Additionally, a random chunk positional encoding is employed. TF-CrossNet is trained and validated on all six-channel training and validation mixtures. The RI-Mag loss [29] is utilized for training, which is defined as

$$\mathcal{L}_{\text{RI-Mag}}(\mathbf{S}, \hat{\mathbf{S}}) = \left\| \mathbf{S}_r - \hat{\mathbf{S}}_r \right\|_1 + \left\| \mathbf{S}_i - \hat{\mathbf{S}}_i \right\|_1 + \left\| |\mathbf{S} - \hat{\mathbf{S}}| \right\|_1, \tag{4}$$

where $\mathbf{S}$ and $\hat{\mathbf{S}}$ denote the STFT of the ground truth and estimated speech. Subscript r and i denote the real and imaginary parts of STFT, respectively. $|\cdot|$ denotes magnitude and $||\cdot||_1$ denotes the L_1 norm. The loss function is computed via permutation-invariant training [30]. Training is performed on 4 NVIDIA A40 GPUs for 125 epochs, employing automatic mixed precision for accelerated training. The gradient clipping is set to 2.

2) LibriCSS: We employ the same setup as [21] for MC-EEND and SSND training. The MC-EEND encoder for diarization utilizes eight Transformer blocks, each with 16 attention heads and a hidden dimension of 256. For speaker separation, we use SpatialNet-large [10], consisting of 12 blocks, D = 192 channels, narrowband hidden dimensions of 384, and cross-band hidden dimensions of 16. STFT window size and shift are 32 and 16 ms, respectively. SpatialNet incorporates speaker embedding sequences with multi-channel speech mixtures, processed through separate encoders and stacked for subsequent SpatialNet blocks. Same as the SMS-WSJ setup, we use automatic mixed precision during training, and RI-Mag loss is utilized to train SpatialNet.

C. Backend Configurations

1) SMS-WSJ: The ASR backend is based on the TDNN-F AM [16]. We train four AMs on different types of training data. The task-standard AM is trained on reverberant-noisy speech from the first, third, and fifth microphones. We train three additional AMs on WSJ 8 kHz (denoted as WSJ) clean speech, direct-path speech, and TF-CrossNet separated training set, all from the first microphone only. The alignment is based on the clean speech for the AM trained on WSJ, and on the first channel of direct-path speech for all other AMs. The training and decoding setup for all AMs follow the task-standard settings. The TF-CrossNet-separated training set is only used as a reference of the performance because this will create a dependency between the frontend and backend.

2) LibriCSS: WRConformer has 10 Conformer encoders, 6 Transformer decoders, and an attention dimension of 512 with 8 attention heads. The feedforward layer operates with a dimension of 2048. A dropout rate of 0.1 is applied. The CTC weight is set to 0.3, and the label smoothing weight is 0.1. The STFT frame size and shift are 512 and 160, respectively. WRConformer is trained on LibriSpeech for 50 epochs on 4 NVIDIA A100 GPUs. Tested on LibriSpeech test-clean and test-other sets, WRConformer achieves 1.9% and 4.1% WER. We denote WRConformer as *Our E2E* in Section IV-B and compute the concatenated minimum-permutation WER (cpWER) [31]. cpWER is computed by concatenating all utterances of each speaker for both reference and hypothesis, then computing the WER between the reference and all possible speaker permutations of the hypothesis, and finally picking the lowest WER value.

IV. Results and Discussions

A. Results on SMS-WSJ

TABLE I: Results on SMS-WSJ (6-channel).

Model	SI-SDR	SDR	PESQ	eSTOI	WER
Unprocessed	-5.5	-0.4	1.50	0.441	79.11
Oracle direct-path	∞	∞	4.50	1.000	6.16
FasNet+TAC [32]	8.6	-	2.37	0.771	29.80
MC-ConvTasNet [33]	10.8	-	2.78	0.844	23.10
MISO_1 [34]	10.2	-	3.05	0.859	14.0
LBT [24]	13.2	14.8	3.33	0.910	9.60
MISO_1-BF-MISO_3 [34]	15.6	-	3.76	0.942	8.30
TF-GridNet [9]	22.8	24.9	4.08	0.980	6.76
SpatialNet [10]	25.1	27.1	4.08	0.980	6.70
TF-CrossNet	**25.8**	**27.6**	**4.20**	**0.987**	**6.30**

In Table I, we compare the task-standard evaluation of the proposed system with other baseline systems on the SMS-WSJ corpus in terms of signal-to-distortion ratio (SDR), scale-invariant SDR (SI-SDR) [35], perceptual evaluation of speech quality (PESQ) [36], extended short-time objective intelligibility (eSTOI) [37], and WER. TF-CrossNet outperforms all baseline systems.

TABLE II: ASR (%WER) results of different AMs on different test data on SMS-WSJ. † denotes performance upperbound.

Test Data	AM Train Data			
	Reverb-noisy	WSJ	Direct-path	TF-CrossNet†
Mixture	79.11	90.97	90.05	90.65
Reverb-noisy	8.52	50.70	48.18	49.03
WSJ	6.45	5.15	5.26	5.19
Direct-path	6.16	5.56	5.23	5.26
TF-CrossNet	6.30	5.94	**5.74**	5.49

In Table II, we compare the ASR performance among AMs trained on different data. The reverberant-noisy speech is denoted as *reverb-noisy* and *mixture* denotes the unprocessed two-talker mixture. In the last row, given the TF-CrossNet output, the task-standard AM produces 6.30% WER, while if the training data is switched to direct-path speech, the WER gets lowered to 5.74%, outperforming the

[2]https://github.com/jsalt2020-asrdiar/jsalt2020_simulate

TABLE III: cpWER (in %) results for different separation and diarization methods on LibriCSS.

Separation Method	Diarization Method	ASR	Overlap Ratio						Avg.
			0S	0L	10%	20%	30%	40%	
Unprocessed	Oracle	Our E2E	3.80	3.61	9.60	16.67	24.97	34.29	17.08
SSND [21]	Oracle	E2E	4.04	3.97	3.37	3.54	4.51	4.66	4.04
SSND	Oracle	Our E2E	3.62	3.49	3.40	3.56	4.19	4.11	3.77
SSND	Oracle (w/ relaxation)	Our E2E	**2.47**	**2.38**	**2.31**	**2.48**	**3.12**	**3.15**	**2.69**
Unprocessed [21]	X-vector + SC [38]	E2E	13.95	12.20	20.12	29.64	35.06	41.81	27.01
Unprocessed	X-vector + SC	Our E2E	11.59	11.26	19.01	26.68	32.24	39.48	24.83
SSND [21]	MC-EEND	E2E	5.56	3.52	3.98	4.76	5.58	6.55	5.13
SSND	MC-EEND	Our E2E	**4.27**	**2.41**	**3.25**	**3.38**	**4.21**	**4.95**	**3.86**

previous best [10] by 14.3% relatively. This switch elevates the ASR performance with only a third of training utterances. It demonstrates that with a strong separation frontend, the backend does not have to be trained on noisy speech. The mismatch between frontend output and backend noisy training data degrades the recognition performance of the mainstream approach. Note that the AM trained on TF-CrossNet separated training set achieves 5.49% WER. However, this backend depends on the pre-trained TF-CrossNet, so it is treated as our performance upperbound with TF-CrossNet. When a better frontend is available, retraining this backend is required to achieve optimal performance, making it less preferred. The results suggest training the ASR backend on clean speech with a strong multi-talker speaker separation frontend to elevate recognition performance over an ASR backend trained on noisy speech.

B. Results on LibriCSS

We report the cpWER results of the proposed system on the LibriCSS corpus in Table III. On unprocessed multi-talker speech mixture, with x-vector and SC (spectral clustering) diarization method [38], E2E ASR model achieves 27.01% cpWER and our E2E model lowers it to 24.83%. With MC-EEND and SSND, we achieve 3.86% cpWER, with only 0.09% gap to the result from oracle diarization. Compared with the 1.09% cpWER gap from 5.13% to 4.04% in [21], WRConformer shows strong robustness to diarization error as a backend for CSS task.

Further comparing the cpWER results of the proposed system with speaker diarization and the system with oracle utterance boundaries, we noticed that the proposed system outperforms the oracle system in $0L$, 10%, 20% overlap ratio conditions, which means that the oracle decision boundaries can be further relaxed, since they may introduce extra insertion and deletion errors. We apply a relaxation collar (250 ms by tradition for speaker boundary [39]) to both sides of the oracle utterance boundaries for each speaker and achieve 2.69% average cpWER, much lower than the 3.77% result from non-relaxed boundaries. This finding indicates room for further improvement despite the subtle performance gap between our system and the system with oracle diarization.

The comparison of the proposed system with other systems on speaker-attributed ASR is shown in Table IV. The E2E ASR model of baselines is based on Transformer and WRComformer is based on Conformer. According to the ESPnet LibriCSS recipe [26], the Transformer-based ASR model outperforms the Conformer-based model. Our system with WRConformer successfully outperforms the previous best with the Transformer-based ASR model by 24.8% relatively, updating the ESPnet findings. The 3.86% cpWER represents the state-of-the-art result on LibriCSS with ASR backend trained on LibriSpeech, without leveraging self-supervised learning features extracted by models such as WavLM [44]. The results demonstrate that separately improved ASR focusing on clean speech elevates the overall performance in a decoupled system.

TABLE IV: Performance comparisons of speaker-attributed ASR systems on LibriCSS.

Ref.	Separation Method	Diarization Method	ASR	cpWER (%)
[40]	CSS	DOA-based	TDNN-F [38]	12.98
[38]	CSS	X-vector + SC	E2E	12.7
[41]	-	-	SA-ASR	11.6
[42]	Speakerbeam	TS-VAD	E2E	18.8
[42]	GSS	TS-VAD	E2E	11.2
[43]	TS-SEP		E2E	6.42
[21]	SSND		E2E	5.13
Ours	SSND		Our E2E	**3.86**

V. Concluding Remarks

We have proposed a decoupled approach to elevate the performance of multi-talker ASR through a multi-channel speaker separation frontend. With powerful separation frontends available, an ASR backend trained on noisy speech may be suboptimal due to the mismatch between backend training data and separated speech. The proposed decoupled approach trains the frontend and backend separately, with the backend focusing on clean speech only. On SMS-WSJ, we achieve a word error rate of 5.74%, which outperforms the previous best trained on reverberant-noisy speech by 14.3% relatively, with a backend trained on clean speech using a third of training utterances. On LibriCSS, we elevate the ASR performance to a 3.86% cpWER, outperforming the previous best by 24.8% with the same training data. In the proposed approach, the capability of frontend separation can be readily evaluated by the backend recognition performance. Future work includes extending the decoupled approach to robust ASR systems with restricted resources and reduced model sizes, and to more challenging far-field environments.

VI. Acknowledgements

This work was supported in part by an NSF grant (ECCS-2125074), the Ohio Supercomputer Center, the NCSA Delta Supercomputer Center (OCI 2005572), and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

References

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

[2] G. Hinton, L. Deng, D. Yu *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, pp. 82–97, 2012.

[3] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 325–351, 2023.

[4] E. Vincent, J. Barker, S. Watanabe *et al.*, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. IEEE ICASSP*, 2013, pp. 126–130.

[5] E. Vincent, S. Watanabe, A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.

[6] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, pp. 1702–1726, 2018.

[7] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, pp. 1256–1266, 2019.

[8] A. Pandey and D. L. Wang, "Self-attending RNN for speech enhancement to improve cross-corpus generalization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1374–1385, 2022.

[9] Z.-Q. Wang, S. Cornell, S. Choi *et al.*, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3221–3236, 2023.

[10] C. Quan and X. Li, "SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 1310–1323, 2024.

[11] V. A. Kalkhorani and D. L. Wang, "TF-CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single- and multi-channel speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 4999–5009, 2024.

[12] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 39–48, 2019.

[13] 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.

[14] 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.

[15] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Commun.*, vol. 104, pp. 1–11, 2018.

[16] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv:1910.13934*, 2019.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[18] Y. Yang, A. Pandey, and D. L. Wang, "Time-domain speech enhancement for robust automatic speech recognition," in *Proc. INTERSPEECH*, 2023, pp. 4913–4917.

[19] Y. Yang, A. Pandey, and D. Wang, "Towards decoupling frontend enhancement and backend recognition in monaural robust ASR," *arXiv preprint arXiv:2403.06387*, 2024.

[20] Y. Masuyama, X. Chang, W. Zhang *et al.*, "Exploring the integration of speech separation and recognition with self-supervised learning representation," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoustics*. IEEE, 2023, pp. 1–5.

[21] H. Taherian and D. Wang, "Multi-channel conversational speaker separation via neural diarization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2467–2476, 2024.

[22] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, pp. 483–492, 2016.

[23] Z. Chen, T. Yoshioka, L. Lu *et al.*, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE ICASSP*, 2020, pp. 7284–7288.

[24] H. Taherian, K. Tan, and D. L. Wang, "Multi-channel talker-independent speaker separation through location-based training," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2791–2800, 2022.

[25] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.

[26] S. Watanabe, T. Hori, S. Karita *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[27] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. CHiME-4*, vol. 78, 2016, p. 79.

[28] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop on Speech and Natural Language*, 1992, pp. 357–362.

[29] Z.-Q. Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1778–1787, 2020.

[30] 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 Trans. Audio, Speech, Language Process.*, vol. 25, pp. 1901–1913, 2017.

[31] S. Watanabe, M. Mandel, J. Barker *et al.*, "CHiME-6 challenge: tackling multispeaker speech recognition for unsegmented recordings," in *Proc. CHiME-6*, 2020, pp. 1–7.

[32] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 6394–6398.

[33] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. IEEE ICASSP*, 2020, pp. 6389–6393.

[34] Z.-Q. Wang, P. Wang, and D. L. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2001–2014, 2021.

[35] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. IEEE ICASSP*, 2019, pp. 626–630.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, 2001, pp. 749–752.

[37] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, pp. 2125–2136, 2011.

[38] D. Raj, P. Denisov, Z. Chen *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. IEEE SLT*, 2021, pp. 897–904.

[39] J. Kalda and T. Alumäe, "Collar-aware training for streaming speaker change detection in broadcast speech," in *Proc. Speaker Odyssey*, 2022, pp. 141–147.

[40] Z.-Q. Wang and D. Wang, "Localization based sequential grouping for continuous speech separation," in *Proc. IEEE ICASSP*, 2022, pp. 281–285.

[41] N. Kanda, X. Xiao, Y. Gaur *et al.*, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr," in *Proc. IEEE ICASSP*, 2022, pp. 8082–8086.

[42] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, and T. Nakatani, "Speaker activity driven neural speech extraction," in *Proc. IEEE ICASSP*, 2021, pp. 6099–6103.

[43] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. Le Roux, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 1185–1197, 2024.

[44] S. Chen, C. Wang, Z. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, pp. 1505–1518, 2022.