Speech Communication 182 (2026) 103427

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Elevating robust multi-talker ASR by decoupling speaker separation and speech recognition

Yufeng Yang [a],*, Hassan Taherian [a], Vahid Ahmadi Kalkhorani [a], DeLiang Wang [b]

[a] Department of Computer Science and Engineering, The Ohio State University, OH, United States
[b] School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, China

ARTICLE INFO

Keywords:
LibriCSS
Libri2Mix
Robust multi-talker ASR
SMS-WSJ
Speaker separation
Speech distortion

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 the backend on noisy speech to mitigate processing artifacts. In this work, we propose to decouple the training of the speaker separation frontend and the ASR backend, and evaluate the proposed system with the backend trained on clean speech only. Our decoupled system achieves 5.1% word error rates (WER) on the Libri2Mix dev/test sets, significantly outperforming other multi-talker ASR baselines. Its effectiveness is also demonstrated with the state-of-the-art 7.60%/5.74% WERs on 1-ch and 6-ch SMS-WSJ. Furthermore, on recorded LibriCSS, we achieve the speaker-attributed WER of 2.92%. These state-of-the-art results suggest that decoupling speaker separation and recognition is an effective approach to elevate robust multi-talker ASR performance. Finally, we provide insights into the acoustic conditions where the decoupled approach is expected to outperform the mainstream approach of training on noisy speech.

1. 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 (Rabiner, 1989), to hybrid systems of deep neural network and HMM (DNN-HMM) (Hinton et al., 2012) to end-to-end (E2E) systems (Prabhavalkar et al., 2023). ASR has been seamlessly integrated into our daily lives as an indispensable 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 mismatch between speech signals in such environments and the transcribed training data in clean conditions pose a persistent obstacle, demanding the development of robust ASR systems capable of overcoming this mismatch (Vincent et al., 2013, 2017).

Meanwhile, the introduction of deep learning has also dramatically improved speech separation performance (Wang and Chen, 2018), including the intelligibility and quality of degraded speech signals. To address the robust ASR problem, a straightforward approach is to employ a speech separation model as the frontend for an ASR backend model. Speech separation models operate in the time (Luo and Mesgarani, 2019; Pandey and Wang, 2022) or time–frequency (T–F) domains (Wang et al., 2023; Quan and Li, 2024; Kalkhorani and Wang, 2024) 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 (Wang et al., 2019).

To address this challenge, prevailing approaches train an ASR model directly on noisy speech (Vincent et al., 2013, 2017; Yang et al., 2022) or enhanced speech (Wang et al., 2019), or train a joint system of speech separation frontend and ASR backend (Chang et al., 2022). In multi-talker or conversational cases, training ASR on overlapped speech directly is problematic due to the large variety of mixed talkers. Although there are multi-talker ASR methods, including serialized output training (Kanda et al., 2020) and other techniques (Guo et al., 2021; Zhang and Qian, 2023; Zhang et al., 2023; H. Shi et al., 2024; M. Shi et al., 2024; Y. Shi et al., 2024), 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 attractive to address overlapped speech for robust ASR (Qian et al., 2018), and permutation invariant training (PIT) is employed in many related studies (Kolbæk et al., 2017; Chang et al.,

* Corresponding author.
E-mail address: yang.5662@osu.edu (Y. Yang).

https://doi.org/10.1016/j.specom.2026.103427
Received 9 November 2025; Received in revised form 26 May 2026; Accepted 29 May 2026
Available online 6 June 2026

2020; Lu et al., 2021). The training target of a speaker separation frontend is usually clean speech. Thus, the choice of training data for the ASR backend becomes critical to achieving high performance. The mainstream strategy trains ASR on single-talker speech with various noise augmentations as done, for example, in the baseline system of the SMS-WSJ corpus (Drude et al., 2019). However, when tested on clean speech, a 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 (Panayotov et al., 2015). As speech separation performance continues to improve, is training the ASR backend on noisy speech still the most effective approach, given the mismatch between the separated speech that is aimed to be clean and the noisy training speech?

In this work, we aim to address such mismatch and elevate robust ASR performance. We investigate robust ASR in multi-talker scenarios and propose to decouple the two stages of speaker separation and speech recognition. Our approach separates the training of a 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 is expected to be alleviated. In prior studies, such systems are considered suboptimal since frontend processing artifacts degrade the ASR trained on clean speech. However, as found in Yang et al. (2023, 2026), with modern monaural speech enhancement, ASR trained on clean speech can outperform that trained on noisy speech. Frontend and backend joint training has been reported to improve overall ASR performance (Fan et al., 2020; Shi et al., 2022; Chang et al., 2022; Masuyama et al., 2023a; Wang et al., 2024). We do not investigate a jointly trained 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 (Masuyama et al., 2023b). In comparison, the proposed system is more flexible and does not require task-specific re-training or fine-tuning when part of the system is updated.

The proposed decoupled approach is evaluated on the Libri2Mix (Cosentino et al., 2020), SMS-WSJ (Drude et al., 2019), an extension of SMS-WSJ with a larger size called SMS-WSJ-Large, and LibriCSS (Chen et al., 2020) corpora. On Libri2Mix, we achieve the state-of-the-art word error rate (WER) of 5.1% on the dev/test sets, significantly outperforming other multi-talker ASR baselines. By training the backend on clean speech only, we achieve 7.60% and 5.74% WER on 1-ch and 6-ch SMS-WSJ test sets, outperforming the previous best that trains the backend on reverberant-noisy speech with three times our training data (Wang et al., 2023; Quan and Li, 2024). Moreover, on recorded LibriCSS, we obtain the speaker-attributed WER of 2.92%, outperforming the previous best (Taherian and Wang, 2024) by 9.3% with an ASR backend trained on clean speech. We also evaluate the generalization ability of the decoupled system with different training data on SMS-WSJ-Large in 1-ch and 6-ch conditions. These state-of-the-art 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 only. The idea of this decoupled pipeline is straightforward and has been considered for single-talker speech in numerous previous studies. The novelty of this study lies in the consideration of multi-talker speech; to our knowledge, this is the first study that addresses multi-talker speaker separation in a decoupled system. Second, this work demonstrates that, with a powerful speaker separation frontend, training ASR on clean speech can actually elevate recognition performance, as shown in our advancement of state-of-the-art results on the Libri2Mix, SMS-WSJ, and LibriCSS datasets. Hence, we believe that the decoupled approach offers a strong alternative to the mainstream approach for robust multi-talker ASR.

This paper expands a preliminary study (Yang et al., 2025) in several major ways. First, we perform evaluations on more corpora, including Libri2Mix, SMS-WSJ, and SMS-WSJ-Large. Second, in addition to the analysis in multi-channel conditions, we evaluate on single-channel Libri2Mix, SMS-WSJ, and SMS-WSJ-Large. Third, we add another ASR backend on LibriCSS that is trained with self-supervised learning (SSL) features (Chen et al., 2022). Fourth, we introduce a new dataset SMS-WSJ-Large, which is easy to generate and reproduce. The evaluation on this dataset indicates that the effective decoupling is not limited to multi-channel conditions for robust multi-talker ASR, and provides insight into the conditions where the proposed approach outperforms the mainstream approach.

The remainder of the paper is organized as follows. Section 2 presents the system model and describes the DNN architectures of the frontend and backend. Section 3 describes the experimental setup and implementation details. Section 4 presents the evaluation results and comparisons. Conclusions and discussions are made in Section 5.

2. Methods

2.1. Problem formulation

2.1.1. 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. $\mathbf{Y}(t,f)$, $\mathbf{N}(t,f)$, $\mathbf{X}(c,t,f)$, $\mathbf{S}(c,t,f)$, and $\mathbf{H}(c,t,f) \in \mathbb{C}^P$ respectively denote the STFT of the received mixture, reverberant noise, reverberant speech, direct-path signal and reflections of speaker c at time t and frequency f. 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 $\mathbf{Y}$.

2.1.2. Automatic speech recognition

An ASR system estimates a word sequence $\mathbf{W}^*$ given a sequence of acoustic features $\mathbf{A}$ of speech signal $\mathbf{a}$, which can be formulated as

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

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

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

where $p_{\mathcal{AM}}$ and $P_{\mathcal{LM}}$ are AM likelihood and LM prior probability, respectively. The AM predicts the likelihood of acoustic features of a phoneme or another linguistic unit, and the 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 $\mathbf{A}$.

2.2. Speaker separation frontend

2.2.1. SpatialNet

SpatialNet (Quan and Li, 2024) is employed as a 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

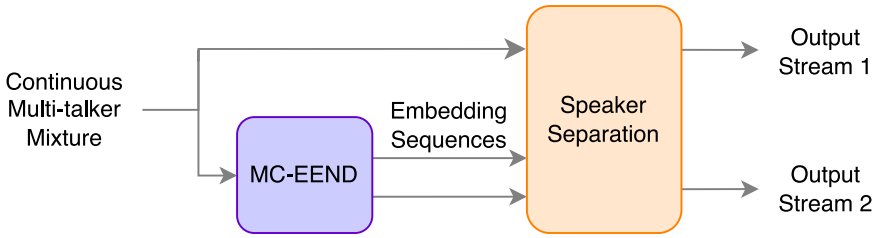


Fig. 1. Diagram of the SSND frontend.

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 among all frequencies and adjacent frequencies, respectively.

SpatialNet performs multi-channel complex spectral mapping (Wang and Wang, 2020) 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 (Williamson et al., 2016). The separated waveforms are generated by performing an inverse STFT on the estimated RI parts.

2.2.2. TF-CrossNet

We use TF-CrossNet as another T–F domain multi-channel speaker separation frontend (Kalkhorani and Wang, 2024). Different from SpatialNet, 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.

2.2.3. Speaker separation via neural diarization

We employ speaker separation via neural diarization (SSND) (Taherian and Wang, 2024) for continuous multi-talker speech recognition (Chen et al., 2020). SSND performs speaker separation by integrating speaker diarization. The diarization is performed with multi-channel end-to-end neural diarization through an encoder–decoder-based attractor module (MC-EEND), which is trained with location-based training (LBT) (Taherian et al., 2022) to resolve the permutation ambiguity issue in talker-independent speaker separation. After diarization, a sequence of speaker embeddings computed from non-overlapped speech frames is used 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.

For SSND, the processing pipeline is shown in Fig. 1. The MC-EEND module takes a continuous multi-talker signal as input and generates two streams of speaker-attributed embeddings. The embeddings are concatenated with the multi-channel mixture signal to a speaker separation model, and the separation model outputs two streams, each corresponding to separated, non-overlapped speech signals.

2.3. Speech recognition backend

2.3.1. Factorized time-delayed neural network

We utilize a factorized time-delayed neural network (TDNN-F) based on the WSJ Kaldi recipe (Povey et al., 2011) 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 Drude et al. (2019).

2.3.2. Wide-residual conformer

We utilize an E2E ASR model in Yang et al. (2026), 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 (Watanabe et al., 2018), and adapts the standard CTC and 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 Heymann et al. (2016)). 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 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.

2.4. Proposed decoupled approach

We introduce the decoupled approach by presenting a comparison of different ASR backend training data in Fig. 2. Training the ASR backend on single-talker speech with reverberation and noise augmentations aims to mitigate acoustic distortions and processing artifacts, which is the mainstream approach and is denoted as distortion-tolerant. However, it yields suboptimal results when tested on clean speech, as discussed earlier. On the other hand, the ASR backend trained on enhanced or separated speech is frontend-dependent, making the system inflexible. Retraining is needed when a better frontend or backend needs to be incorporated to achieve improved performance. To address these limitations and leverage the capabilities of powerful separation frontends, we propose to train the ASR backend on clean speech only. The proposed decoupled system significantly improves the flexibility compared to the frontend-dependent approach, and elevates the performance on clean speech relative to the distortion-tolerant approach.

3. Experimental setup

3.1. Datasets

3.1.1. Libri2Mix

We use the ESPnet version[2] of Libri2Mix (Cosentino et al., 2020), which is designed specifically for multi-talker ASR. The data generation process is based on the official implementation[3] for speaker separation and an offset is added to one speaker in each speech mixture. The dataset utilizes clean speech sourced from LibriSpeech (Panayotov et al., 2015) and noise from WHAM! (Wichern et al., 2019) datasets. The training set has 50 800 and 13 900 utterances in train-360 and

[1] https://github.com/yfyangseu/espnet.
[2] https://github.com/espnet/espnet/tree/master/egs2/librimix/sot_asr1.
[3] https://github.com/JorisCos/LibriMix.

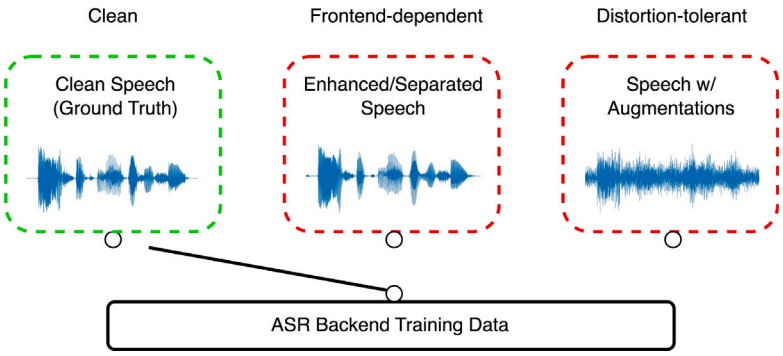


Fig. 2. Comparison of different ASR backend training approaches.

train-100 sets, respectively. The development set and test set have 3000 utterances each. The sampling rate is 16 kHz. In this paper, we focus on the challenging robust multi-talker ASR in noisy conditions for Libri2Mix.

3.1.2. SMS-WSJ and SMS-WSJ-Large

We employ SMS-WSJ (Drude et al., 2019) to evaluate ASR with multi-channel speaker separation in reverberant conditions. The dataset consists of 33 561, 982, and 1332 train, validation, and test mixtures, respectively. All utterances are drawn from the WSJ0 and WSJ1 datasets (Paul and Baker, 1992). 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 sampled from [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.

To evaluate the generalization ability and the effect of training data size on a robust multi-talker ASR system, we create a new dataset called SMS-WSJ-Large, by modifying the data generation process of SMS-WSJ.[4] We multiply the original training data size by a factor of 4, resulting in 134 244 training utterances. The numbers of utterances in the development and test sets are kept the same as SMS-WSJ. The training and development data T60 and SNR ranges are modified to [0.2, 1.1] s and [0, 40] dB, respectively. For the test set, we consider three T60 ranges: [0.2, 0.5], [0.5, 0.8], and [0.8, 1.1] s, and four SNR ranges: [0, 10], [10, 20], [20, 30], and [30, 40] dB. By combining each T60 and SNR range, 12 test sets are generated for evaluation in SMS-WSJ-Large. The random seed is the same as in Drude et al. (2019) for easy replication.

3.1.3. LibriCSS

LibriCSS is a dataset designed for evaluating continuous multi-talker speech recognition (Chen et al., 2020). The dataset has 10 one-hour sessions, each of which has 6 ten-minute mini-sessions with different overlap levels. These levels are 0S (no speaker overlap and short inter-utterance silence ranges in [0.1, 0.5] s), 0L (no speaker overlap and long 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 test-clean set (Panayotov et al., 2015) and the sampling rate is 16 kHz. The recordings are made of loudspeakers in a real 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 center microphone.

The meeting-style training data for MC-EEND and SSND is generated following a LibriCSS recipe,[5] where the LibriSpeech training set is utilized. We follow the data generation setup of Taherian and Wang (2024) for speaker diarization and speaker separation training. The center microphone is treated as the reference microphone.

3.2. Frontend configurations

3.2.1. Libri2Mix

We employ TF-CrossNet (Kalkhorani and Wang, 2024) for 1-ch speaker separation on Libri2Mix. The network is configured with 12 TF-CrossNet blocks where the number of hidden channels is set to 192 and the cross-band hidden dimension is 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 16 and 8 ms, respectively. Additionally, random chunk positional encoding is employed for training. A combination of RI-Mag loss (Wang et al., 2020) and the scale-invariant signal-to-distortion ratio (SI-SDR) loss (Le Roux et al., 2019) is utilized for model training. These loss functions are defined as

$$\mathcal{L} = \mathcal{L}_{\text{RI-Mag}} + \mathcal{L}_{\text{SI-SDR}}, \tag{4a}$$

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

$$\mathcal{L}_{\text{SI-SDR}} = -\sum_{c=1}^{C} 10\log_{10}\frac{\|\mathbf{s}_c\|_2^2}{\|\hat{\mathbf{s}}_c - \alpha_c\mathbf{s}_c\|_2^2}, \tag{4c}$$

$$\alpha_c = \frac{\mathbf{s}_c^T\hat{\mathbf{s}}_c}{\mathbf{s}_c^T\mathbf{s}_c}, \tag{4d}$$

where $\mathbf{s}$ and $\hat{\mathbf{s}}$ denote the ground-truth and estimated speech, and $\mathbf{S}$ and $\hat{\mathbf{S}}$ denote their STFTs, respectively. Subscript r and i denote the real and imaginary parts of STFT, respectively. $|\cdot|$ denotes magnitude, $\|\cdot\|_1$ denotes the L_1 norm, and $\|\cdot\|_2$ denotes the L_2 norm. Term α_c is the scaling factor. The loss function is computed via utterance-level PIT (Kolbæk et al., 2017). The network training is performed on 4 NVIDIA H100 GPUs with 94 GB memory for 200 epochs, employing automatic mixed precision for accelerated training for the first 150 epochs and full precision training for the last 50 epochs. We utilize the Adam optimizer (Kingma and Ba, 2015) with a maximum learning rate of 0.001. We use a cosine warm-up scheduler that increases the learning rate from $1e^{-4}$ to $1e^{-3}$ in the first 10 epochs. Afterwards, the ReduceLROnPlateau scheduler in PyTorch is utilized, with patience of 3 epochs and a factor of 0.8. The batch size is set to the maximum number that can fit into the GPU memory. The final model with the best validation SI-SDR value is selected for evaluation.

3.2.2. SMS-WSJ and SMS-WSJ-Large

For SMS-WSJ, we also employ TF-CrossNet and keep the same training configuration as in Kalkhorani and Wang (2024) and Section 3.2.1. The differences from Section 3.2.1 are that the STFT frame size and shift are 32 and 16 ms, respectively. Additionally, long short-term memory (LSTM) and random chunk positional encoding are employed for 1-ch

[4] https://github.com/fgnt/sms_wsj.
[5] https://github.com/jsalt2020-asrdiar/jsalt2020_simulate.

and 6-ch training, respectively. The training is performed on 4 NVIDIA A100 GPUs with 80 GB memory.

For training on 6-ch SMS-WSJ-Large, we follow the same procedure as that for 6-ch SMS-WSJ and train a TF-CrossNet model from scratch. For 1-ch SMS-WSJ-Large, we take model weights from the pre-trained model on SMS-WSJ 1-ch, and then fine-tune it with SI-SDR loss $\mathcal{L}_{\text{SI-SDR}}$ on the 1-ch SMS-WSJ-Large data. In this fine-tuning, the learning rate of the Adam optimizer is lowered to $1e^{-5}$, and in the learning rate schedule, the initial and peak learning rate is lowered to $1e^{-5}$ and $1e^{-4}$, respectively. To make fair comparisons, we also fine-tune the same pre-trained model on 1-ch SMS-WSJ with the same loss function and learning rate schedule, and take this fine-tuned model on the same dataset as a pre-trained model in Section 4.3.

3.2.3. LibriCSS

We employ the same setup as Taherian and Wang (2024) 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. The STFT window size and window shift are set to 25 ms and 10 ms, respectively. A 23-dimensional log-Mel filterbank is used for feature extraction, and extracted features are concatenated with those of the seven preceding and seven succeeding frames. For spatial features, inter-channel phase differences (IPDs) are extracted and passed through seven convolutional blocks, having channel configurations of (8,8,16,16,32,32,32). The initial block has a kernel size of (15, 1), and all later blocks have a kernel size of (3, 5). All blocks have a stride of (1, 2). The log-Mel features are concatenated with the spatial features for training. We use the Adam optimizer for training, with a learning rate of 0.001, coupled with the Noam scheduler (Vaswani et al., 2017) with 125k warm-up steps. The number of speakers is set to 8 for both training and testing. For speech activity decisions, a threshold of $\tau = 0.5$ is used. Finally, during post-processing, we apply a 31-frame median filter to avoid generating exceedingly short segments.

For speaker separation, we use SpatialNet-large (Quan and Li, 2024) for fair comparisons, consisting of 12 blocks, $D = 192$ channels, narrowband hidden dimensionality of 384, and cross-band hidden dimensionality 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. RI-Mag loss $\mathcal{L}_{\text{RI-Mag}}$ is utilized to train SpatialNet. The Adam optimizer is used with an initial learning rate set to 0.001, which is halved if no improvement in validation loss is observed over three epochs.

3.3. Backend configurations

3.3.1. Libri2Mix

We utilize WRConformer as the ASR backend for Libri2Mix. It 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. All WRConformers are trained on WavLM extracted features (Chen et al., 2022) for Libri2Mix.

We train several WRConformers for comparison. For models trained on clean speech, one is trained on the train-clean-100 and train-clean-360 sets of LibriSpeech, and validated on the LibriSpeech dev set, denoted as "LibriSpeech Clean". Another is trained on single-speaker clean speech of Libri2Mix (s1 and s2 in train-100 and train-360 sets), and validated on the Libri2Mix clean dev set, denoted as "Libri2Mix Clean".

For distortion-tolerant models trained on noisy speech, one is trained on Libri2Mix single-speaker noisy speech (mix_single in train-100 and train-360) denoted as "Libri2Mix Noise-dependent". For another model, we dynamically and randomly mix the noise (noise in train-100 and train-360) and single-speaker clean speech (s1 and s2 in train-100 and train-360) as training data, and the trained model is denoted as "Libri2Mix Noise-independent". The training SNR is sampled from a normal distribution of $\mathcal{N}(-2, 3.6)$ dB according to Cosentino et al. (2020). We also mix clean speech from LibriSpeech train-100 and train-360 with noise in Libri2Mix train-100 and train-360 dynamically and randomly with the same SNR as in Libri2Mix Noise-independent, denoted as "LibriSpeech Noise-independent". All distortion-tolerant WRConformers are validated on the Libri2Mix noisy single-speaker dev set to ensure their noise robustness. We also employ a generic robust ASR model of Whisper Large-v3 (Radford et al., 2023) for comparison. During the ASR decoding stage, no LM is used.

To assess performance upper bounds, we train and compare to a distortion-dependent WRConformer on the TF-CrossNet separated Libri2Mix training data, denoted as "Libri2Mix Distortion-dependent", and a jointly fine-tuned system, denoted as "Jointly Fine-tuned". The Libri2Mix Distortion-dependent system is trained with the same setup as Libri2Mix Clean. The jointly fine-tuned system combines pre-trained TF-CrossNet and Libri2Mix Clean WRConformer, and is fine-tuned with an ASR loss only, following Masuyama et al. (2023b). We freeze WavLM inside the joint system and update the parameters of TF-CrossNet, WRConformer encoder, and WRConformer decoder. The joint fine-tuning is done on 4 NVIDIA A100 GPUs with a learning rate of $1e^{-5}$ for 20 epochs, and the model that has the best validation loss is selected for testing. The batch size can only be set to 1 due to the large size of the joint system. Note that both systems have an interdependency between the frontend and backend, requiring re-training when the frontend or backend is updated.

3.3.2. SMS-WSJ and SMS-WSJ-Large

The ASR backend is based on the TDNN-F AM (Drude et al., 2019). We train five AMs on different types of data. The task-standard AM is trained on reverberant-noisy speech from the first, third, and fifth microphones. We train four additional AMs on WSJ 8 kHz (denoted as WSJ) clean speech, direct-path speech, 1-ch TF-CrossNet separated training set, and 6-ch TF-CrossNet separated training set, all from the first microphone only. The alignment is based on the WSJ 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 used only as a reference of performance upper bound because this will create a dependency between the frontend and backend. For evaluation on SMS-WSJ-Large, we employ the task-standard AM and the AM trained on direct-path speech, representing the mainstream approach and the proposed decoupled approach, respectively.

3.3.3. LibriCSS

We utilize WRConformer as the backend for LibriCSS. We train two WRConformers, one is based on log-Mel features (denoted as Our E2E) and the other is based on WavLM extracted features (denoted as Our E2E-SSL). WRConformers are trained on LibriSpeech for 50 epochs on 4 A100 GPUs with the same configuration as the ESPnet LibriSpeech recipe in Watanabe et al. (2018). We compute the concatenated minimum-permutation WER (cpWER) (Watanabe et al., 2020), which is calculated 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. Tested on the LibriSpeech test-clean/test-other sets, our E2E and our E2E-SSL respectively achieve 1.9%/4.1% and 1.8%/3.5% WERs.

Table 1
ASR (%WER) results of the proposed and comparison systems on Libri2Mix.

Systems	w/ Frontend	w/ SSL	WER		
			Dev	Test	Ground truth
SOT-Conformer (Watanabe et al., 2018)	✗	✗	24.7	23.3	–
SOT-WavLM-Conformer (Watanabe et al., 2018)	✗	✓	19.4	17.1	–
Conditional-Conformer (Guo et al., 2021)	✗	✗	24.5	24.9	–
TS-HuBERT (Zhang and Qian, 2023)	✗	✓	–	24.8	–
TSE-Whisper (Zhang et al., 2023)	✓	✗	–	12.0	–
GEncSep (H. Shi et al., 2024)	✗	✓	17.2	15.0	–
WavLM LLM (M. Shi et al., 2024)	✗	✓	11.4	10.2	–
WavLM LLM w/ Fine-tune (M. Shi et al., 2024)	✗	✓	10.3	9.0	–
PIT-SOT (Y. Shi et al., 2024)	✗	✗	–	7.0	–
MUSE-TSASR (Seong et al., 2025)	✗	✓	11.0	10.5	–
UME-ASR (Shakeel et al., 2025)	✗	✗	–	6.4	–
Libri2Mix Clean	✓	✓	5.1	5.1	2.3
Libri2Mix Noise-dependent	✓	✓	5.0	4.9	2.9
Libri2Mix Noise-independent	✓	✓	5.2	5.0	2.8
LibriSpeech Clean	✓	✓	5.2	5.2	4.1
LibriSpeech Noise-independent	✓	✓	5.2	5.2	3.9
Whisper Large-v3 (Radford et al., 2023)	✓	✗	5.6	5.0	2.5
Libri2Mix Distortion-dependent	✓	✓	4.5	4.7	2.8
Jointly Fine-tuned	✓	✓	4.5	4.6	–

4. Results and discussions

4.1. Results on Libri2Mix

In terms of speaker separation metrics, TF-CrossNet achieves 14.70 in SDR and 14.19 in scale-invariant SNR (SI-SNR) on the Libri2Mix development set, and 14.57 in SDR and 14.04 in SI-SNR on the test set. The WER results of the proposed approach and the comparison with other systems are presented in Table 1. We compare the WERs on the development and test set of Libri2Mix with other baselines. For the proposed approach, we also evaluate and compare the WERs of different ASR backends on the Libri2Mix single-speaker test clean speech, denoted as "Ground Truth" in Table 1.

For comparison systems, the official ESPnet baseline using serialized output training (SOT) (Kanda et al., 2020) and Conformer (Gulati et al., 2020), and another baseline utilizing WavLM to extract features, achieve 23.3% and 17.1% WER on the test set, respectively. In Guo et al. (2021), a non-autoregressive model based on a conditional chain model with Conformer and CTC loss achieves 24.9% WER. Target-speaker HuBERT (TS-HuBERT) based method (Zhang and Qian, 2023) achieves 24.8% WER on the test set, and target speaker extraction (TSE) based method (Zhang et al., 2023) achieves 12.0% WER on the test set. In H. Shi et al. (2024), overlapped encoding separation is utilized, and the serialized speech information guidance SOT (GEncSep) based method achieves 15.0% WER on the test set. Large language models (LLMs) are employed in M. Shi et al. (2024) to help multi-talker ASR, and with fine-tuning on Libri2Mix, the best system achieves 9.0% WER on the test set. The system in Y. Shi et al. (2024) utilizes a multi-talker ASR model trained on the speech material of the full LibriSpeech dataset and achieves 7.0% WER. Be leveraging multiple speaker embedding models for target speaker ASR, the system in Seong et al. (2025) achieves 10.5% WER on the test set. To the best of our knowledge, the 6.4% WER in Shakeel et al. (2025) represents the current best result on the Libri2Mix dataset, and this system utilizes an architecture that jointly learns representations for speaker diarization, speech separation, and multi-talker ASR.

Now, we analyze and compare the results of systems that follow the proposed decoupled approach with different ASR backends. We first use TF-CrossNet to separate the noisy Libri2Mix two-talker mixtures into two streams of separated speech. Then, with the resulting permutation, WERs are computed given the transcripts of Ground Truth. We report the WERs on the development and test sets with several ASR backends described in Section 3.3.1. All ASR backends have similar performances, with the best of 4.9% from the Libri2Mix Noise-dependent model. When the ASR backend is trained on clean speech only, Libri2Mix Clean and LibriSpeech Clean yield very close 5.1% and 5.2% WERs, respectively. Even for the Whisper Large-v3, the 5.0% WER is the same as the Libri2Mix Noise-independent model. Our systems significantly outperform the current state-of-the-art result of 6.4% WER by over 20% relatively, demonstrating the effectiveness of the decoupled approach.

Since several systems perform similarly on Libri2Mix, which ASR backend should we select? To answer this question, we evaluate each ASR backend on the Ground Truth, and find that Libri2Mix Clean yields the best WER. It outperforms the Libri2Mix Noise-dependent and Libri2Mix Noise-independent models, indicating that training the backend on single-speaker noisy speech degrades the performance on clean speech, as mentioned in Sections 1 and 2.4. Comparing with the ASR backend that yields the best performance on the Libri2Mix test set, i.e., Libri2Mix Noise-dependent, Libri2Mix Clean underperforms by 3.9% relatively on the Libri2Mix test set but outperforms on Ground Truth by 20.7%. The second best system on Ground Truth comes from Whisper Large-v3, which underperforms by 2.0% on the Libri2Mix test set compared to Libri2Mix Noise-dependent, and outperforms on Ground Truth by 13.8%. Taking the difference of the training data sizes of Libri2Mix Clean and Whisper into consideration, we conclude that the answer to the earlier question is Libri2Mix Clean. These results suggest that, when a powerful speaker separation frontend is available, training the ASR backend on clean speech is an effective option for robust multi-talker ASR, which stands in sharp contrast to the mainstream approach that trains ASR backend on noisy speech.

Next, we report the results of Libri2Mix Distortion-dependent and Jointly Fine-tuned models. The distortion-dependent model achieves 4.7% WER on the test set, outperforming 5.1% WER from the decoupled Libri2Mix Clean by 7.8% relatively. However, Libri2Mix Clean outperforms Libri2Mix Distortion-dependent on Ground Truth by 17.8% relatively, showing that optimizing the backend on separated speech degrades ASR performance on clean speech. The jointly fine-tuned model achieves 4.6% WER on the test set, outperforming all other systems in Table 1, showing the advantage of joint training. As pointed out in Section 1, these top-performing systems sacrifice the modularity of the frontend and backend for better performance. In comparison, the proposed decoupled approach is more balanced in terms of flexibility and performance.

4.2. Results on SMS-WSJ

4.2.1. Task-standard evaluations

In Tables 2 and 3, we compare the task-standard evaluation of TF-CrossNet with other baseline systems on the SMS-WSJ corpus in 1-ch

Table 2
Speaker separation results on SMS-WSJ (1-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
DPRNN-TasNet (Luo et al., 2020b)	6.5	–	2.28	0.734	38.10
$SISO_1$ (Wang et al., 2021a)	5.7	–	2.4	0.748	28.70
DNN_1+ (FCP+DNN_2) × 2 (Wang et al., 2021a)	12.7	14.1	3.25	0.899	12.80
DNN_1+ (msFCP+DNN_2) × 2 (Wang et al., 2021b)	13.4	–	3.41	–	10.90
TF-GridNet (1-stage) (Wang et al., 2023)	16.2	17.2	3.45	0.924	9.49
TF-GridNet (2-stage) (Wang et al., 2023)	18.4	19.6	3.70	0.952	7.91
TF-CrossNet	19.2	20.2	3.74	0.953	8.13

Table 3
Speaker separation 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 (Luo et al., 2020a)	8.6	–	2.37	0.771	29.80
MC-ConvTasNet (Zhang et al., 2020)	10.8	–	2.78	0.844	23.10
$MISO_1$ (Wang et al., 2021a)	10.2	–	3.05	0.859	14.0
LBT (Taherian et al., 2022)	13.2	14.8	3.33	0.910	9.60
$MISO_1$-BF-$MISO_3$ (Wang et al., 2021a)	15.6	–	3.76	0.942	8.30
TF-GridNet (1-stage) (Wang et al., 2023)	19.9	21.2	3.89	0.966	6.92
TF-GridNet (2-stage) (Wang et al., 2023)	22.8	24.9	4.08	0.980	6.76
SpatialNet (Quan and Li, 2024)	25.1	27.1	4.08	0.980	6.70
TF-CrossNet	25.8	27.6	4.20	0.987	6.30

and 6-ch conditions, in standard speech separation metrics of SDR, SI-SDR, perceptual evaluation of speech quality (PESQ) (Rix et al., 2001), and extended short-time objective intelligibility (eSTOI) (Taal et al., 2011), as well as WER. TF-CrossNet outperforms all baseline systems. Related separation results can also be found in Kalkhorani and Wang (2024).

4.2.2. Comparison of different AMs

In Table 4, we compare the ASR performance among AMs trained on different data. The reverberant-noisy speech is denoted as *reverb-noisy* and *mixture* denotes unprocessed two-talker mixtures. In the penultimate row, given the 1-ch TF-CrossNet output, the reverb-noisy AM produces a 8.13% WER, and if the training data is switched to direct-path speech, the WER gets lowered to 7.60%, outperforming the previous best of 7.91% (Wang et al., 2023) with two-stage training. For 6-ch TF-CrossNet, the switch from reverb-noisy to direct-path AM lowers the WER from 6.30% to 5.74%, outperforming the previous best (Quan and Li, 2024) 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.

It is worth noting that the AM trained on TF-CrossNet separated training set achieves 6.26% WER for 1-ch and 5.49% for 6-ch. Although such matched training produces lower WERs, these backends depend on the data from pre-trained TF-CrossNets. When a better frontend is available, retraining these backends is required to achieve optimal performance, making these models less practical. The results in Table 4 show that training the ASR backend on clean speech to elevate recognition performance should be preferred to an ASR backend trained on noisy speech.

4.3. Results on SMS-WSJ-Large

We compare the systems trained or fine-tuned on SMS-WSJ-Large and the pre-trained models on SMS-WSJ to investigate the effects of frontend training data on the overall performance of decoupled systems, and to find out which ASR backend is most promising in certain conditions. For the ASR backends, we select the task-standard reverberant-noisy speech trained AM, and the direct-path speech trained AM as described in Section 3.3.2.

4.3.1. Evaluation in single-channel setup

The evaluation results on 1-ch SMS-WSJ-Large are presented in Table 5. We compare the separated speech by the TF-CrossNets fine-tuned on SMS-WSJ and on SMS-WSJ-Large, in three T60 and four SNR ranges. For each test condition and system, two WERs are computed based on the reverb-noisy speech trained and the direct-path speech trained AM. By comparing the WERs, we observe that, by increasing the size of the training dataset, WERs are lowered on both AMs. This meets the expectation that more training data translates to improved frontend performance, and reduces the processing artifacts in separated speech that degrade ASR backend performance.

To better understand relative WERs in Table 5, we visualize the percentage improvement comparing the direct-path speech trained AM over the reverberant-noisy speech trained AM in Fig. 3. Each block denotes a test condition, and the number in each block denotes the relative WER reduction by switching the ASR backend from reverberant-noisy speech trained AM to direct-path speech trained AM. A color bar is used to denote the value of relative WER reduction. Positive numbers in green denote that the system produces lower WER with direct-path speech trained AM, and negative numbers in red denote that the system produces lower WER with reverberant-noisy speech trained AM. For example, the 5.5% in green for 30–40 dB SNR and 0.2–0.5 s T60 denotes that the WER produced by the direct-path speech trained AM outperforms the reverberant-noisy speech trained AM by 5.5%. For TF-CrossNet pre-trained on 1-ch SMS-WSJ, only two blocks are green in Fig. 3(a), with maximum improvement of 5.5%. For TF-CrossNet trained on 1-ch SMS-WSJ-Large, with more frontend training data, the blocks for all test conditions when the T60 range is 0.2–0.5 s are green, and the maximum improvement increases to 6.9%. This finding indicates that with more frontend training data, decoupling the frontend and backend will more likely benefit robust multi-talker ASR. However, for 1-ch systems, when the test condition is highly adverse, say with low SNR or high T60, the mainstream approach to train ASR backend on noisy speech is still advantageous, and the direct-path speech trained AM elevates the overall performance in moderately and mildly adverse test conditions.

4.3.2. Evaluation in six-channel setup

Table 6 presents the WER comparisons this array setup. The improvements achieved by training TF-CrossNet on 6-ch SMS-WSJ-Large instead of SMS-WSJ are consistently and significantly larger compared with those in Table 5. On the task-standard test condition with T60 of 0.2–0.5 s and SNR of 20–30 dB, the WER with direct-path speech trained AM is lowered to 5.63% compared with the 5.74% result in Section 4.2.2.

Fig. 4 visualizes the percentage improvement of the direct-path speech trained AM relative to the reverberant-noisy speech trained AM. With 6-ch inputs instead of 1-ch, the separation frontend leverages

Table 4
ASR (%WER) results of different AMs on different test data on SMS-WSJ.

AM Train data	Test data					
	Mixture	Reverb-noisy	WSJ	Direct-path	TF-CrossNet (1-ch)	TF-CrossNet (6-ch)
Reverb-noisy	79.11	8.52	6.45	6.16	8.13	6.30
WSJ	90.97	50.70	5.15	5.56	7.86	5.94
Direct-path	90.05	48.18	5.26	5.23	7.60	5.74
TF-CrossNet (1-ch)[a]	90.37	46.61	5.46	5.30	6.26	5.48
TF-CrossNet (6-ch)[a]	90.65	49.03	5.19	5.26	6.74	5.49

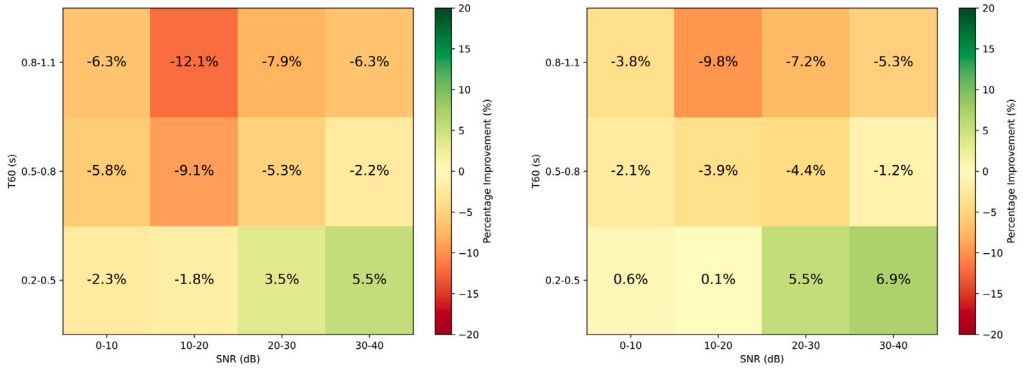
[a] Denotes performance upper bound for each condition.

Table 5
ASR (%WER) comparison results on SMS-WSJ-Large (1-ch) test set. Each entry denotes the WER generated from reverb-noisy speech trained AM (left) and direct-path speech trained AM (right) separated by a slash.

System	T60	SNR			
		0–10 dB	10–20 dB	20–30 dB	30–40 dB
SMS-WSJ Pre-trained	0.2–0.5 s	42.11/43.10	12.49/12.72	8.06/7.80	7.25/6.85
	0.5–0.8 s	51.80/54.98	16.15/17.77	9.89/10.44	8.94/9.14
	0.8–1.1 s	64.97/69.32	24.40/27.76	14.65/15.90	13.02/13.90
SMS-WSJ-Large	0.2–0.5 s	34.43/34.24	11.86/11.85	8.05/7.61	7.21/6.71
	0.5–0.8 s	40.54/41.40	14.12/14.70	9.17/9.59	8.38/8.48
	0.8–1.1 s	47.88/49.78	17.45/19.35	11.60/12.50	10.38/10.96

Table 6
ASR (%WER) comparison results on SMS-WSJ-Large (6-ch) test set. Each entry denotes the WER generated from reverb-noisy speech trained AM (left) and direct-path speech trained AM (right) separated by a slash.

System	T60	SNR			
		0–10 dB	10–20 dB	20–30 dB	30–40 dB
SMS-WSJ Pre-trained	0.2–0.5 s	21.92/24.51	7.97/8.05	6.38/5.74	6.19/5.41
	0.5–0.8 s	29.25/34.53	8.77/9.46	6.59/6.18	6.28/5.75
	0.8–1.1 s	40.47/48.31	10.06/11.90	7.13/6.97	6.92/6.79
SMS-WSJ-Large	0.2–0.5 s	13.98/14.32	7.51/6.76	6.45/5.63	6.28/5.42
	0.5–0.8 s	16.12/16.83	7.72/7.24	6.56/5.84	6.30/5.55
	0.8–1.1 s	18.86/20.58	8.28/8.14	6.83/6.22	6.46/5.85



(a) SMS-WSJ pre-trained (b) SMS-WSJ-Large trained

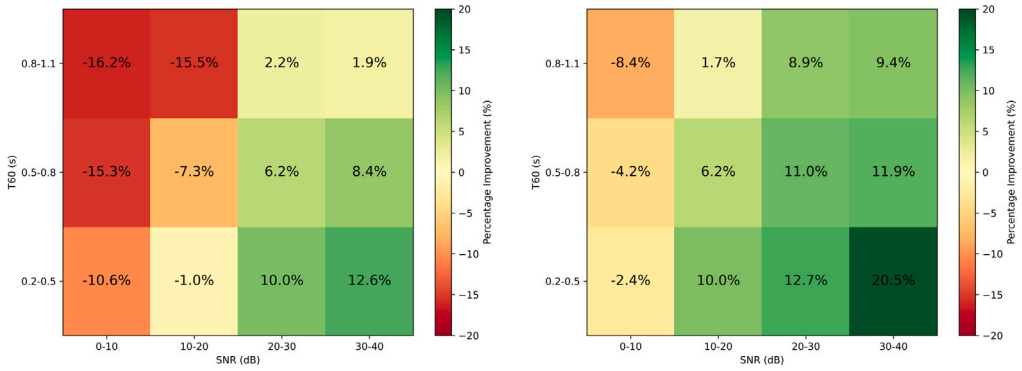
Fig. 3. Visualization of the percentage improvement of the direct-path over reverb-noisy speech trained AM on SMS-WSJ-Large test set (1-ch). Negative numbers indicate ASR degradation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

spatial information and has better separation quality compared with the 1-ch model. This leads to more green blocks in Fig. 4 compared with Fig. 3. When the frontend is trained on SMS-WSJ, the direct-path trained AM produces lower WER relative to the reverberant-noisy speech trained AM for all test conditions when SNR is 20 dB or higher as shown in Fig. 4(a). When the frontend training data is changed to SMS-WSJ-Large, all test conditions with SNR at 10 dB or higher show better performance with the direct-path speech trained AM as shown in Fig. 4(b). The largest improvement also increases from 12.6% to 20.5% by changing the frontend training data from SMS-WSJ to SMS-WSJ-Large.

4.3.3. When to decouple separation and recognition?

The results in Sections 4.3.1 and 4.3.2 provide insights into the acoustic conditions where the proposed decoupled approach performs better than the mainstream approach:

- In the single-channel setup shown in Fig. 3, with the pre-trained frontend on SMS-WSJ, the decoupled approach outperforms the mainstream approach for T60 in 0.2–0.5 s and SNR in 20–40 dB. With the frontend trained on SMS-WSJ-Large, the decoupled approach performs better for T60 in 0.2–0.5 s and SNR in 0–40 dB. The mainstream approach achieves lower WER at higher T60 and lower SNR values.



(a) SMS-WSJ pre-trained (b) SMS-WSJ-Large trained

Fig. 4. Visualization of the percentage improvement of the direct-path over reverb-noisy speech trained AM on SMS-WSJ-Large test set (6-ch). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
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
MIMO-BF-MIMO (Taherian et al., 2024)	Oracle	E2E	3.45	3.87	3.15	4.46	5.35	5.76	4.44
SSND (Taherian and Wang, 2024)	Oracle	E2E	4.04	3.97	3.37	3.54	4.51	4.66	4.04
SSND (Taherian and Wang, 2024)	Oracle	E2E-SSL	2.28	2.38	2.36	2.27	2.68	2.47	2.42
SSND	Oracle	Our E2E	3.62	3.49	3.40	3.56	4.19	4.11	3.77
SSND	Oracle	Our E2E-SSL	2.27	2.48	2.22	2.17	2.53	2.48	2.36
SSND	Oracle (w/ relaxation)	Our E2E	2.47	2.38	2.31	2.48	3.12	3.15	2.69
SSND	Oracle (w/ relaxation)	Our E2E-SSL	2.02	1.92	1.95	2.00	2.31	2.47	2.13
Unprocessed (Taherian and Wang, 2024)	X-vector + SC (Raj et al., 2021)	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
Mask-based MVDR (Chen et al., 2020)	X-vector + SC	E2E	8.78	13.07	10.51	15.37	17.54	17.63	14.13
MIMO-BF-MIMO (Taherian et al., 2024)	X-vector + SC	E2E	7.05	8.05	7.31	8.48	10.81	10.03	8.76
SSND (Taherian and Wang, 2024)	MC-EEND	E2E	5.56	3.52	3.98	4.76	5.58	6.55	5.13
GSS (Cord-Landwehr et al., 2025)	Spatio-spectral Diarization	E2E	5.87	5.90	5.90	7.46	8.16	8.79	6.53
SSND	MC-EEND	Our E2E	4.27	2.41	3.25	3.38	4.21	4.95	3.86
SSND	MC-EEND	Our E2E-SSL	3.43	2.77	2.15	2.29	3.06	3.69	2.92

- In the six-channel setup shown in Fig. 4, with T60 in 0.2–1.1 s and SNR in 20–40 dB, the decoupled approach with the pre-trained frontend outperforms the mainstream approach. When the frontend is trained on SMS-WSJ-Large, the decoupled approach performs better for T60 in 0.2–1.1 s and SNR in 10–40 dB.
- With the same test SNR, T60, and training data, there are more conditions in the six-channel setup than in the single-channel setup under which the decoupled approach performs better. Furthermore, under the same conditions, the six-channel setup obtains higher amounts of improvement than the single-channel setup. We attribute this to better speaker separation of the multi-channel system relative to the single-channel system (Kalkhorani and Wang, 2024).

These findings suggest that decoupling the frontend and backend is an effective approach when the separation frontend produces a quality output. Such an output can be obtained at relatively low T60s and high SNRs, or with a powerful separation frontend.

4.4. Results on LibriCSS

Before reporting speaker-attributed ASR results, we refer the reader to Table 1 of Taherian and Wang (2024) for the diarization results of the SSND frontend. The cpWER results of the proposed system on the LibriCSS corpus are given in Table 7. On unprocessed multi-talker speech mixtures, with x-vector and SC diarization method (Raj et al., 2021), the 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 our E2E model, with only 0.09% gap from the result with oracle diarization. Compared with the 1.09% cpWER gap (from 5.13% to 4.04%) in Taherian and Wang (2024), our E2E model shows strong robustness to diarization errors as a backend for the CSS task. With WavLM-based feature extraction, we achieve 2.92% cpWER with MC-EEND and SSND.

Further comparing the cpWER results of the proposed system with speaker diarization and the system with oracle utterance boundaries with our E2E model, we observe that the proposed system outperforms the oracle system in 0L, 10%, 20% overlap ratio conditions, which means that oracle decision boundaries can be further relaxed, since they may introduce extra insertion and deletion errors. We apply a relaxation collar – 250 ms as traditionally chosen for speaker boundaries (Kalda and Alumäe, 2022) – to both sides of oracle utterance boundaries for each speaker. With this relaxation collar, we achieve 2.69% average cpWER, much lower than the 3.77% result without relaxed boundaries. Boundary relaxation also reduces cpWER from 2.36% to 2.13% for our E2E-SSL model. This finding provides further room for progress towards oracle diarization.

A comprehensive comparison of the proposed system with other systems on speaker-attributed ASR is shown in Table 8. The E2E ASR model in the baselines is Transformer based and WRComformer is Conformer based. According to the ESPnet LibriCSS recipe (Watanabe et al., 2018), the Transformer-based ASR model outperforms the Conformer-based model. Our E2E system, based on WRConformer, successfully outperforms the previous best with the Transformer-based ASR model by 24.8% relatively, upgrading the ESPnet findings. The 3.86% cpWER

Table 8
Performance comparisons of speaker-attributed ASR systems on LibriCSS.

Ref.	Separation method	Diarization method	ASR	cpWER (%)
Wang and Wang (2022)	CSS	DOA-based	TDNN-F (Raj et al., 2021)	12.98
Raj et al. (2021)	CSS	X-vector + SC	E2E	12.7
Kanda et al. (2022)	–	–	SA-ASR	11.6
Delcroix et al. (2021)	Speakerbeam	TS-VAD	E2E	18.8
Delcroix et al. (2021)	GSS	TS-VAD	E2E	11.2
Boeddeker et al. (2024)	TS-SEP		E2E	6.42
Boeddeker et al. (2024)	TS-SEP		E2E-SSL	5.36
Taherian and Wang (2024)	SSND		E2E	5.13
Taherian and Wang (2024)	SSND		E2E-SSL	3.22
Cord-Landwehr et al. (2025)	GSS	Spatio-spectral diarization	E2E	6.53
Vieting et al. (2025)	CSS-AD		Whisper	5.8
Niu et al. (2025)	DCF-DS (single-channel)		E2E-SSL	4.43
Ours	SSND		Our E2E	3.86
Ours	SSND		Our E2E-SSL	2.92

obtained by our system represents the state-of-the-art result on LibriCSS with ASR backend trained on LibriSpeech, without leveraging SSL features. The results with our E2E-SSL outperform the 3.86% result with our E2E model, and advance the state-of-the-art result 3.22% in Taherian and Wang (2024) to 2.92%. The results clearly demonstrate that separately improved ASR focusing on clean speech elevates the overall performance in a decoupled system.

5. Concluding remarks

We have proposed a decoupled approach to elevate the performance of robust multi-talker ASR. With modern separation frontends available, an ASR backend trained on noisy speech may actually become 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 trained on clean speech only. The evaluation results demonstrate that our decoupled approach achieves state-of-the-art results on several corpora. On Libri2Mix, we achieve a state-of-the-art 5.1% WER by leveraging a speaker separation frontend with an ASR backend trained on Libri2Mix clean speech, outperforming the previous best by 27.1%. On 1-ch and 6-ch SMS-WSJ, we achieve WERs of 7.60% and 5.74% respectively, outperforming the previous best systems trained on reverberant-noisy speech, and our backend is trained with a third of training utterances used in the baselines. On recorded LibriCSS, we elevate the ASR performance to a 2.92% cpWER, outperforming the previous best by 9.3%. As a byproduct of the proposed approach, the capability of future frontend separation can be readily evaluated in terms of ASR by the backend acoustic model pre-trained on clean speech.

A further study on SMS-WSJ-Large suggests that the proposed decoupled approach is a strong alternative to the mainstream approach. In the single-channel setup, the decoupled approach shows better ASR performance than the mainstream approach with T60 in 0.2–0.5 s and SNR in 0–40 dB. In the six-channel setup, the decoupled approach performs better with T60 in 0.2–1.1 s and SNR in 10–40 dB. These findings show that the decoupled approach can elevate robust multi-talker ASR performance when the acoustic environment is not very reverberant and noisy, or a powerful separation algorithm is available. As speaker separation continues to advance, we expect that the decoupled approach becomes advantageous in more and more acoustic environments.

CRediT authorship contribution statement

Yufeng Yang: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hassan Taherian:** Software, Methodology, Investigation. **Vahid Ahmadi Kalkhorani:** Software, Methodology, Data curation, Conceptualization. **DeLiang Wang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: DeLiang Wang reports financial support was provided by National Science Foundation. DeLiang Wang reports equipment, drugs, or supplies was provided by Ohio Supercomputer Center. DeLiang Wang reports equipment, drugs, or supplies was provided by Pittsburgh Supercomputing Center. DeLiang Wang reports equipment, drugs, or supplies was provided by National Center for Supercomputing Applications. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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

Data availability

Data will be made available on request.

References

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

Chang, X., Maekaku, T., Fujita, Y., Watanabe, S., 2022. End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation. In: Proc. INTERSPEECH. pp. 3819–3823.

Chang, X., Zhang, W., Qian, Y., Le Roux, J., Watanabe, S., 2020. End-to-end multi-speaker speech recognition with transformer. In: Proc. IEEE ICASSP. pp. 6134–6138.

Chen, S., Wang, C., Chen, Z., et al., 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. 1505–1518.

Chen, Z., Yoshioka, T., Lu, L., et al., 2020. Continuous speech separation: Dataset and analysis. In: Proc. IEEE ICASSP. pp. 7284–7288.

Cord-Landwehr, T., Gburrek, T., Deegen, M., Haeb-Umbach, R., 2025. Spatio-spectral diarization of meetings by combining TDOA-based segmentation and speaker embedding-based clustering. In: Proc. Interspeech. pp. 5223–5227.

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., Vincent, E., 2020. LibriMix: An open-source dataset for generalizable speech separation. arXiv:2005.11262.

Delcroix, M., Zmolikova, K., Ochiai, T., Kinoshita, K., Nakatani, T., 2021. Speaker activity driven neural speech extraction. In: Proc. IEEE ICASSP. pp. 6099–6103.

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

Fan, C., Yi, J., Tao, J., Tian, Z., Liu, B., Wen, Z., 2020. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 198–209.

Gulati, A., Qin, J., Chiu, C.-C., et al., 2020. Conformer: Convolution-augmented transformer for speech recognition. In: Proc. INTERSPEECH. pp. 5036–5040.

Guo, P., Chang, X., Watanabe, S., Xie, L., 2021. Multi-speaker ASR combining non-autoregressive Conformer CTC and conditional speaker chain. In: Proc. INTERSPEECH. pp. 3720–3724.

Heymann, J., Drude, L., Haeb-Umbach, R., 2016. Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition. In: Proc. CHiME-4 Workshop. Vol. 78, p. 79.

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

Kalda, J., Alumäe, T., 2022. Collar-aware training for streaming speaker change detection in broadcast speech. In: Proc. Speaker Odyssey. pp. 141–147.

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

Kanda, N., Gaur, Y., Wang, X., Meng, Z., Yoshioka, T., 2020. Serialized output training for end-to-end overlapped speech recognition. In: Proc. INTERSPEECH. pp. 2797–2801.

Kanda, N., Xiao, X., Gaur, Y., et al., 2022. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR. In: Proc. IEEE ICASSP. pp. 8082–8086.

Kingma, D.P., Ba, J., 2015. Adam: A method for Stochastic optimization. In: Proc. Int. Conf. Learn. Representations. pp. 1–15.

Kolbæk, M., Yu, D., Tan, Z.-H., Jensen, J., 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 25, 1901–1913.

Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R., 2019. SDR–half-baked or well done? In: Proc. IEEE ICASSP. pp. 626–630.

Lu, L., Kanda, N., Li, J., Gong, Y., 2021. Streaming end-to-end multi-talker speech recognition. IEEE Signal Process. Lett. 28, 803–807.

Luo, Y., Chen, Z., Mesgarani, N., Yoshioka, T., 2020a. End-to-end microphone permutation and number invariant multi-channel speech separation. In: Proc. IEEE ICASSP. pp. 6394–6398.

Luo, Y., Chen, Z., Yoshioka, T., 2020b. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: Proc. IEEE ICASSP. pp. 46–50.

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

Masuyama, Y., Chang, X., Cornell, S., Watanabe, S., Ono, N., 2023a. End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation. In: Proc. IEEE Spoken Language Technology Workshop. SLT, pp. 260–265.

Masuyama, Y., Chang, X., Zhang, W., et al., 2023b. Exploring the integration of speech separation and recognition with self-supervised learning representation. In: Proc. IEEE WASPAA. pp. 1–5.

Niu, S.-T., Du, J., Wang, R.-Y., Yang, G.-B., Gao, T., Pan, J., Hu, Y., 2025. DCF-DS: Deep cascade fusion of diarization and separation for speech recognition under realistic single-channel conditions. IEEE Trans. Audio Speech Lang. Process.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. LibriSpeech: an ASR corpus based on public domain audio books. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. pp. 5206–5210.

Pandey, A., Wang, D.L., 2022. Self-attending RNN for speech enhancement to improve cross-corpus generalization. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 1374–1385.

Paul, D.B., Baker, J.M., 1992. The design for the wall street journal-based CSR corpus. In: Proc. Workshop on Speech and Natural Language. pp. 357–362.

Povey, D., Ghoshal, A., Boulianne, G., et al., 2011. The Kaldi speech recognition toolkit. In: Proc. IEEE ASRU.

Prabhavalkar, R., Hori, T., Sainath, T.N., Schlüter, R., Watanabe, S., 2023. End-to-end speech recognition: A survey. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 325–351.

Qian, Y., Chang, X., Yu, D., 2018. Single-channel multi-talker speech recognition with permutation invariant training. Speech Commun. 104, 1–11.

Quan, C., Li, X., 2024. SpatialNet: Extensively learning spatial information for multi-channel joint speech separation, denoising and dereverberation. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 1310–1323.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision. In: Proc. Int. Conf. Mach. Learn.. pp. 28492–28518.

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

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. Vol. 2, pp. 749–752.

Seong, J.-S., Choi, J.-H., Jeoung, Y.-R., Kim, I., Chang, J.-H., 2025. Enhancing target-speaker automatic speech recognition using multiple speaker embedding extractors with virtual speaker embedding. In: Proc. Interspeech. pp. 4918–4922.

Shakeel, M., Sudo, Y., Peng, Y., Lin, C.-J., Watanabe, S., 2025. Unifying diarization, separation, and ASR with multi-speaker encoder. In: Proc. IEEE ASRU. pp. 1–7.

Shi, J., Chang, X., Watanabe, S., Xu, B., 2022. Train from scratch: Single-stage joint training of speech separation and recognition. Comput. Speech Lang. 76, 101387.

Shi, H., Gao, Y., Ni, Z., Kawahara, T., 2024. Serialized speech information guidance with overlapped encoding separation for multi-speaker automatic speech recognition. In: Proc. IEEE SLT. pp. 193–199.

Shi, M., Jin, Z., Xu, Y., Xu, Y., Zhang, S.-X., Wei, K., Shao, Y., Zhang, C., Yu, D., 2024. Advancing multi-talker ASR performance with large language models. In: Proc. IEEE SLT. pp. 14–21.

Shi, Y., Li, L., Wang, D., Han, J., 2024. Keyword guided target speech recognition. IEEE Signal Process. Lett. 31, 1945–1949.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE/ACM Trans. Audio Speech Lang. Process. 19, 2125–2136.

Taherian, H., Pandey, A., Wong, D., Xu, B., Wang, D., 2024. Leveraging sound localization to improve continuous speaker separation. In: Proc. IEEE ICASSP. pp. 621–625.

Taherian, H., Tan, K., Wang, D.L., 2022. Multi-channel talker-independent speaker separation through location-based training. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 2791–2800.

Taherian, H., Wang, D., 2024. Multi-channel conversational speaker separation via neural diarization. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 2467–2476.

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In: Proc. Adv. Neural Inf. Process. Syst.. pp. 5998–6008.

Vieting, P., Berger, S., von Neumann, T., Boeddeker, C., Schlüter, R., Haeb-Umbach, R., 2025. Error analysis in a modular meeting transcription system. arXiv preprint arXiv:2509.10143.

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

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

Wang, D.L., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26, 1702–1726.

Wang, Z.-Q., Cornell, S., Choi, S., et al., 2023. TF-GridNet: Integrating full- and sub-band modeling for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 3221–3236.

Wang, W., Mo, S., Dong, L., Yu, Z., Guo, J., Huang, Y., 2024. DGSRN: Noise-robust speech recognition method with Dual-Path Gated Spectral Refinement Network. In: Proc. INTERSPEECH. pp. 5018–5022.

Wang, P., Tan, K., Wang, D.L., 2019. Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 39–48.

Wang, Z.-Q., Wang, D., 2020. Multi-microphone complex spectral mapping for speech dereverberation. In: Proc. IEEE ICASSP. pp. 486–490.

Wang, Z.-Q., Wang, D., 2022. Localization based sequential grouping for continuous speech separation. In: Proc. IEEE ICASSP. pp. 281–285.

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

Wang, Z.-Q., Wang, P., Wang, D.L., 2021a. Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 2001–2014.

Wang, Z.-Q., Wichern, G., Le Roux, J., 2021b. Convolutive prediction for reverberant speech separation. In: Proc. IEEE WASPAA. pp. 56–60.

Watanabe, S., Hori, T., Karita, S., et al., 2018. ESPnet: End-to-end speech processing toolkit. In: Proc. INTERSPEECH. pp. 2207–2211.

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

Wichern, G., Antognini, J., Flynn, M., Zhu, L.R., McQuinn, E., Crow, D., Manilow, E., Roux, J.L., 2019. WHAM!: Extending speech separation to noisy environments. In: Proc. INTERSPEECH. pp. 1368–1372.

Williamson, D.S., Wang, Y., Wang, D.L., 2016. Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 24, 483–492.

Yang, Y., Pandey, A., Wang, D.L., 2023. Time-domain speech enhancement for robust automatic speech recognition. In: Proc. INTERSPEECH. pp. 4913–4917.

Yang, Y., Pandey, A., Wang, D.L., 2026. Towards decoupling frontend enhancement and backend recognition in monaural robust ASR. Comput. Speech Lang. 95, 101821.

Yang, Y., Taherian, H., Kalkhorani, V.A., Wang, D.L., 2025. Elevating robust ASR by decoupling multi-channel speaker separation and speech recognition. In: Proc. IEEE ICASSP. pp. 1–5.

Yang, Y., Wang, P., Wang, D.L., 2022. A conformer based acoustic model for robust automatic speech recognition. arXiv:2203.00725.

Zhang, W., Qian, Y., 2023. Weakly-supervised speech pre-training: A case study on target speech recognition. In: Proc. INTERSPEECH. pp. 3517–3521.

Zhang, W., Yang, L., Qian, Y., 2023. Exploring time-frequency domain target speaker extraction for causal and non-causal processing. In: Proc. IEEE ASRU. pp. 1–6.

Zhang, J., Zorilă, C., Doddipatla, R., Barker, J., 2020. On end-to-end multi-channel time domain speech separation in reverberant environments. In: Proc. IEEE ICASSP. pp. 6389–6393.