Computer Speech and Language 95 (2026) 101821

Contents lists available at ScienceDirect

Computer Speech & Language

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

Towards decoupling frontend enhancement and backend recognition in monaural robust ASR

Yufeng Yang [a],*, Ashutosh Pandey [a], DeLiang Wang [a,b]

[a] Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, 43210, OH, United States
[b] Center for Cognitive and Brain Science, The Ohio State University, 1835 Neil Ave, Columbus, 43210, OH, United States

ARTICLE INFO

Keywords:
CHiME-2
CHiME-4
Robust ASR
Speech distortion
Speech enhancement

ABSTRACT

It has been shown that the intelligibility of noisy speech can be improved by speech enhancement (SE) algorithms. However, monaural SE has not been established as an effective frontend for automatic speech recognition (ASR) in noisy conditions compared to an ASR model trained on noisy speech directly. The divide between SE and ASR impedes the progress of robust ASR systems, especially as SE has made major advances in recent years. This paper focuses on eliminating this divide with an ARN (attentive recurrent network) time-domain, a TF-CrossNet time–frequency domain, and an MP-SENet magnitude-phase based enhancement model. The proposed systems decouple frontend enhancement and backend ASR, with the latter trained only on clean speech. Results on the WSJ, CHiME-2, LibriSpeech, and CHiME-4 corpora demonstrate that ARN, TF-CrossNet, and MP-SENet enhanced speech all translate to improved ASR results in noisy and reverberant environments, and generalize well to real acoustic scenarios. The proposed system outperforms the baselines trained on corrupted speech directly. Furthermore, it cuts the previous best word error rate (WER) on CHiME-2 by 28.4% relatively with a 5.6% WER, and achieves 3.3/4.4% WER on single-channel CHiME-4 simulated/real test data without training on CHiME-4. We also observe consistent improvements using noise-robust Whisper as the backend ASR model.

1. Introduction

In real environments, acoustic interference is ubiquitous in speech communication, and negatively impacts the performance of speech-based applications such as smart home devices Heymann et al. (2018) and conference transcription systems Fu et al. (2021). To attenuate acoustic interference, speech enhancement (SE) algorithms estimate clean speech from noisy or reverberant speech. These algorithms have achieved remarkable success in improving the quality and intelligibility of speech with background interference, particularly with the rise of deep learning in the field Wang and Chen (2018). However, a major disappointment is that monaurally enhanced speech does not translate to improved automatic speech recognition (ASR) that is trained on clean speech, and this has been attributed to the distortion introduced by monaural SE Wang et al. (2020a). The divide between SE and ASR has persisted despite considerable research over decades to bridge enhancement and ASR (Cooke et al., 2001; Raj et al., 2004; Narayanan and Wang, 2014; Wang et al., 2020a; Zhang et al., 2021). This study represents a new effort to bridge the fields of SE and ASR.

Traditional SE approaches are based on spectral subtraction, Wiener filtering, and statistical model based approaches Loizou (2013). SE has also been approached from the perspective of traditional neural networks (Tamura, 1989; Sorensen, 1991;

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

https://doi.org/10.1016/j.csl.2025.101821
Received 1 November 2024; Received in revised form 9 February 2025; Accepted 4 May 2025
Available online 20 May 2025

Moon and Hwang, 1993), as well as modern deep neural networks (Xu et al., 2014; Wang and Wang, 2013). In modern formulations, SE is mainly conducted in the time–frequency (T–F) domain, such as short-time Fourier transform (STFT). Common practice is to enhance spectral magnitudes and reconstruct the enhanced waveform using the enhanced magnitude and noisy phase. However, this approach ignores phase information, which is important for enhancement performance Paliwal et al. (2011). A recent trend is to enhance both spectral magnitude and phase in approaches such as complex ratio masking (Williamson et al., 2016; Hu et al., 2020), which estimates the complex ideal ratio mask by predicting the real and imaginary spectrograms. Alternatively, complex spectral mapping predicts directly the real and imaginary spectrograms of clean speech from the complex STFT of noisy speech (Fu et al., 2017a; Tan and Wang, 2019; Wang et al., 2020b). Different from spectral methods, time-domain enhancement predicts clean speech samples from noisy speech samples, hence enhancing speech magnitude and phase simultaneously (Fu et al., 2017b; Pascual et al., 2017; Luo and Mesgarani, 2019). In this study, we select a time domain, a T–F domain, and a magnitude-phase based SE model as frontends. For time-domain SE, we employ the attentive recurrent network (ARN), which incorporates an RNN (recurrent neural network) with a self-attention mechanism Pandey and Wang (2022). For T–F domain SE, a recently proposed TF-CrossNet is employed, and it includes a cross-temporal module in addition to narrow-band and cross-band modules A. Kalkhorani and Wang (2024). For magnitude-phase based model, we adopt MP-SENet (Lu et al., 2023a,b), which enhances the magnitude and phase of a noisy speech signal for SE. All models achieve excellent enhancement performance.

In terms of robust ASR, prevailing approaches perform acoustic modeling directly on noisy speech for noise-dependent or noise-independent models, which is proven to be effective on series of CHiME challenges (Barker et al., 2013; Vincent et al., 2013; Barker et al., 2015; Vincent et al., 2017). The drawback of such approaches is that noise-dependent models do not generalize well to untrained noises, and noise-independent models need an enormous amount of transcribed noisy speech for training, which is not only costly but also infeasible in many real-world applications. When tested on clean speech, an ASR model trained on noisy speech also results in a performance gap compared with a corresponding model trained on clean speech, due to the mismatch between training and test conditions.

To bridge the divide between SE and ASR, attempts have been made to perform acoustic modeling on enhanced speech (Seltzer et al., 2013; Meng et al., 2018; Wang et al., 2020a) or enhanced features Wang and Wang (2019) in a distortion-independent way. However, the backends of these systems are dependent on the frontends. When a frontend is replaced, say by a better enhancement model, the backend would need to be retrained or ASR performance degrades. Conversely, SE can be designed to serve ASR. In Plantinga et al. (2021), a perceptual loss based model was proposed to guide the training of a SE frontend using senone labels from an acoustic model (AM). In Ho et al. (2023), a noise- and artifact-aware loss function is designed to train the frontend. The study in Narayanan et al. (2023) focuses on improving T–F mask quality. In Iwamoto et al. (2022), Zorilă and Doddipatla (2022), mixing enhanced speech with noisy speech mitigates processing artifacts and improves ASR, but the mixed speech is still noisy. In Kinoshita et al. (2020), Ochiai et al. (2024), Sato et al. (2022), Zorilă and Doddipatla (2022), Panchapagesan et al. (2022), Sato et al. (2021), SE is shown as an effective frontend for ASR, but the backend is trained on multi-conditioned speech, mismatched with the enhanced speech which is aimed to be clean. To combine SE and ASR, a SE model and an ASR model can be jointly trained (Xu et al., 2019; Menne et al., 2019; Zhu et al., 2023a). Similar ideas are also utilized in end-to-end (E2E) systems (Shi et al., 2022; Chang et al., 2022). Such systems often have a huge model size and are difficult to train. Furthermore, the frontend and backend inside a joint or E2E system are dependent on each other, making it problematic to benefit from an improved frontend or backend individually, hence limiting the flexibility in real applications. Results in Masuyama et al. (2023) demonstrate that, although fine-tuning an E2E system lowers word error rate (WER), the individual performance of the frontend degrades significantly when it is decoupled after joint fine-tuning.

In this paper, we investigate the most straightforward approach to robust ASR where an ASR model is trained on clean speech only and its input is frontend enhanced speech directly. Thus the proposed robust ASR system decouples the frontend and backend. In other words, the frontend performs SE without ASR considerations, and the backend ASR model is designed to recognize clean speech without considering the potential distortion of an enhancement frontend. By clean speech, we do not mean recording conditions in a sound booth or studio; as shown in the collection of the LibriSpeech corpus Panayotov et al. (2015), clean utterances can be recorded in a variety of quiet places in real environments. By combining two decoupled modules directly to recognize noisy speech, we demonstrate that the proposed system outperforms alternative robust ASR systems, including noise-independent, reverberation-independent, and interference-independent models, on noisy, reverberant, and reverberant-noisy speech. When tested on the medium vocabulary track (track 2) of the CHiME-2 corpus, our system achieves a 5.6% average WER. To our knowledge, this result represents the best on this dataset to date and outperforms the previous best by 28.4%. Results on the Wall Street Journal (WSJ) and LibriSpeech corpora show that the proposed system outperforms all baselines in noisy and reverberant conditions. Our system also generalizes well to CHiME-4 using an off-the-shelf frontend trained on LibriSpeech and a backend trained on WSJ clean speech, and achieves 3.3/4.4% WER on single-channel CHiME-4 simulated/real test data without training on CHiME-4. We observe consistent improvements when replacing the backend with the noise-robust ASR model of Whisper Radford et al. (2023) on CHiME-4. Our investigation reveals that using short-time objective intelligibility (STOI) Taal et al. (2011) as the model selection criterion is superior for SE models in terms of ASR. Our study makes three main contributions to monaural robust ASR. First, we advocate the approach of decoupling a frontend devoted to enhancing noisy speech and a backend devoted to recognizing clean speech, which contrasts with the prevailing approach that trains ASR on noisy speech. Second, we demonstrate that the decoupled approach outperforms the prevailing approach on multiple datasets. Third, we advance the state-of-the-art ASR results on CHiME-2 and CHiME-4 datasets. Unlike the previous best systems on single-channel CHiME-4, we obtain the top performance with a pre-trained speech enhancement frontend.

This paper expands a preliminary study Yang et al. (2023a) in a number of significant ways. First, on the task of recognizing noisy speech, we investigate the proposed system on the WSJ, LibriSpeech, and CHiME-4 corpora in addition to CHiME-2, introducing new test conditions and speech materials. Second, we evaluate the proposed system in reverberant and reverberant-noisy conditions. Third, we add another noise-independent baseline trained on more data than the noise-independent model in Yang et al. (2023a). Fourth, we extend the AM from a hybrid system of deep neural network and hidden Markov model (DNN-HMM) in Yang et al. (2023a) to an E2E architecture. Fifth, we compare the decoupled ASR system with two more baseline ASR systems on the LibriSpeech corpus trained on dynamically generated noisy speech and multi-conditioned speech, matching the data seen by the SE frontend during training, further demonstrating the effectiveness of the proposed decoupled system. Sixth, we add experiments on CHiME-2 with a T–F domain TF-CrossNet and a magnitude-phase based MP-SENet SE frontend, and demonstrate that the benefit of decoupling the frontend and backend is not limited to time-domain SE. Last, the decoupled system achieves a 3.3/4.4% WER on the single-channel CHiME-4 simulated/real test sets, demonstrating strong generalization ability to cross-corpus real acoustic scenarios. Also on CHiME-4 we observe consistent improvements when replacing the backend with the noise-robust ASR model of Whisper Radford et al. (2023).

The remainder of the paper is organized as follows. Section 2 describes the frontend and backend models for SE and ASR. Section 3 describes the experimental setup and implementation details. Evaluation results and comparisons are presented in Section 4. Section 5 concludes the paper.

2. System description

2.1. Problem formulation

2.1.1. Monaural speech enhancement

The monaural SE problem can be formulated as follows:

$$\mathbf{y} = \mathbf{h} * \mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{y}$, $\mathbf{h}$, $\mathbf{s}$, and $\mathbf{n}$ denote noisy speech, room impulse response (RIR), clean speech, and additive noise, respectively; $*$ denotes convolution. SE computes an estimate of $\mathbf{s}$, $\hat{\mathbf{s}}$, from $\mathbf{y}$. When the speech signal is anechoic, $\mathbf{h}$ can be omitted. When the signal is reverberant, Eq. (1) can be expressed in another way as

$$\begin{aligned} \mathbf{y} &= (\mathbf{h}_d + \mathbf{h}_r) * \mathbf{s} + \mathbf{n} \\ &= \mathbf{h}_d * \mathbf{s} + \mathbf{h}_r * \mathbf{s} + \mathbf{n} \\ &= \mathbf{s}_d + \mathbf{s}_r + \mathbf{n}, \end{aligned} \tag{2}$$

where $\mathbf{s}_d$ and $\mathbf{s}_r$ denote the direct-path and reverberated speech, respectively. The RIRs of direct-path speech and reverberated speech are denoted by $\mathbf{h}_d$ and $\mathbf{h}_r$, respectively. SE aims to estimate $\mathbf{s}_d$ in this study.

2.1.2. Automatic speech recognition

An ASR system computes the optimal word sequence $\mathbf{W}^*$ given a sequence of acoustic features $\mathbf{X}$ of speech signal $\mathbf{x}$, which is formulated as a maximum *a posteriori* probability problem

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

where $\mathcal{AM}$ and $\mathcal{LM}$ denote an AM and language model (LM), respectively. Using Bayes' theorem, Eq. (3) can be written as

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} p_{\mathcal{AM}}(\mathbf{X}|\mathbf{W}) P_{\mathcal{LM}}(\mathbf{W}), \tag{4}$$

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, a word sequence is predicted directly given $\mathbf{X}$.

2.2. Frontend networks

2.2.1. Attentive recurrent network

We employ ARN as the time-domain frontend of ASR, which comprises RNN, self-attention, feedforward network, and layer normalization modules. Details of ARN building blocks can be found in Pandey and Wang (2022). In this work, the non-causal version of ARN is used, namely the RNN in ARN is bi-directional long short-term memory (BLSTM) and self-attention is unmasked.

Fig. 1 shows the diagram of ARN. After an input signal with M samples is chunked into overlapping frames, all frames are projected into a latent representation of size N by a linear layer. Then the latent representation is processed by four consecutive ARN blocks, and another linear layer projects the output of the last ARN block back to size L. The enhanced speech is finally computed using the overlap-and-add (OLA) method.

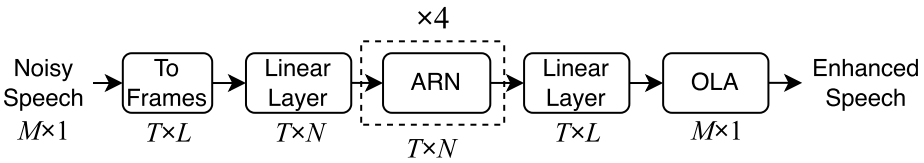

Fig. 1. Diagram of ARN for speech enhancement. T is the total number of frames and L is the frame length.

2.2.2. TF-CrossNet

We use TF-CrossNet as the T–F domain frontend for monaural SE A. Kalkhorani and Wang (2024). TF-CrossNet is motivated by SpatialNet Quan and Li (2024), but introduces a cross-temporal module after the cross-band module and positional encoding. This modification enhances temporal processing for monaural separation. TF-CrossNet performs complex spectral mapping by predicting the real and imaginary (RI) parts of the STFT of clean speech from the stacked RI parts of the STFT of noisy speech (Williamson et al., 2016; Tan and Wang, 2019). The enhanced waveform is generated by performing an inverse STFT on the enhanced RI parts.

2.3. Backend networks

2.3.1. Wide residual conformer acoustic model

We utilize a Conformer-based AM Yang et al. (2022) as the backend of the proposed system, denoted as WRConformer AM. It is built upon a wide residual BLSTM network (WRBN) (see Wang et al. 2020a) and shown to outperform it on the CHiME-4 single-channel track Yang et al. (2022). The system architecture of WRConformer AM is shown in Fig. 2, where FFN denotes a feedforward network. WRConformer AM takes as input 80-dimensional mean-normalized log-Mel filterbank features extracted from the frontend output, coupled with its delta and delta-delta features. First, the input is processed by a wide residual convolutional layer denoted as WRCNN, which passes the input signal through a convolution layer and uses three residual blocks to extract representations at different frequency resolutions Zagoruyko and Komodakis (2016). Afterwards, an utterance-wise batch normalization and a linear layer with ELU (exponential linear unit) non-linearity are utilized to project the signal onto 320 dimensions. Then a linear layer projects the signal onto the dimension of multi-head self-attention. After processing by two blocks of the Conformer encoder with absolute positional encoding, the signal is projected onto 1024 dimensions, followed by ReLU (rectified linear unit) activation and dropout. Next, a linear layer projects the signal to the final output of each frame as the posterior probability of 1965 context-dependent senone states. Then the output is sent to a decoder for text transcripts.

The Conformer encoder is a neural network architecture designed to improve ASR by capturing both local and global dependencies in audio sequences. It builds upon a Transformer encoder Vaswani et al. (2017), with an additional convolutional network, and macaron-like feed-forward layers. By modeling local features through the convolutional components and global interactions via Transformer blocks, it shows better performance compared with a standard Transformer encoder in ASR Gulati et al. (2020).

2.3.2. End-to-end wide residual conformer network

We extend WRConformer AM into a connectionist temporal classification (CTC) and attention Conformer-encoder Transformer-decoder E2E ASR model. Leveraging the ASR recipe implemented in ESPnet Watanabe et al. (2018), we adapt the standard CTC/attention Conformer-encoder Transformer-decoder ASR recipe to E2E WRConformer. In this adaptation, we replace the 2-D convolution in the subsampling module with modified 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 processing of the subsampling module, matching the default ESPnet subsampling module. In E2E WRConformer, the number of Conformer blocks is set to 10 and the other configurations are the same as the default setting. The implementation is available online.[1]

3. Experimental setup

3.1. Frontend enhancement

3.1.1. Reverberation generation

RIRs are generated in the same way as Pandey et al. (2022). Room length and width range in [5, 10] m, and height in [3, 4] m. A sound source and a microphone are sampled at least 0.5 m away from the walls. The distance between the two ranges from 0.75 m to 2 m. We use Pyroomacoustics Scheibler et al. (2018) for RIR generation, which is a hybrid method where the image method with an order of 6 is used to model early reflections and late reverberation is modeled by ray-tracing. RIRs for training, validation, and test data are randomly sampled from 319666, 40091, and 40243 generated RIRs, respectively.

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

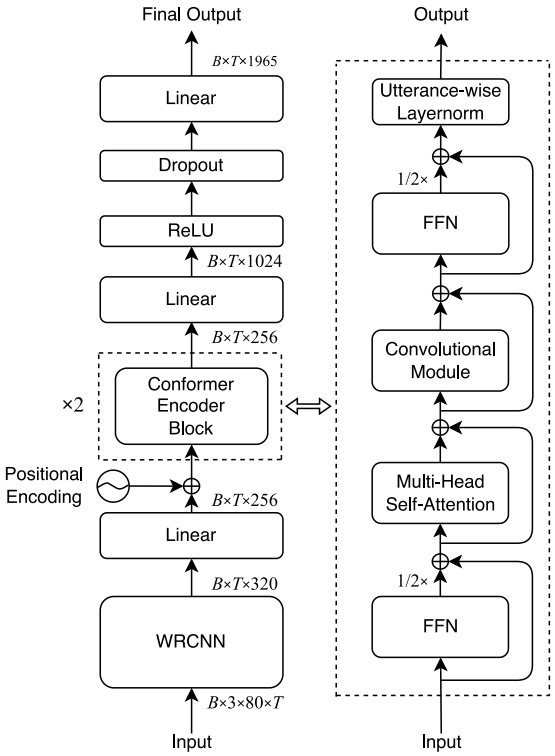

Fig. 2. System architecture of a WRConformer AM. B denotes the batch size, and T denotes the number of time frames of the longest utterance in a batch.

3.1.2. WSJ

We utilize the WSJ0 SI-84 corpus Paul and Baker (1992) to train ARN for SE and WRConformer AM for ASR. This dataset comprises English utterances from 42 male and 41 female speakers, providing 7138, 1206, and 330 utterances for training, validation, and test, respectively. For ARN training on WSJ, we design three ARN systems aiming at denoising, dereverberation, or both, denoted as DN-ARN, DR-ARN, and NR-ARN, respectively.

Training data for DN-ARN is generated by randomly mixing 7138 anechoic-clean WSJ training utterances with noise segments randomly picked from the 10k non-speech sounds of a sound effect library (http://www.sound-ideas.com) with signal-to-noise ratio (SNR) uniformly selected in [−7, 0] dB and [0, 10] dB ranges, with 50% probability for each range. Validation data is generated by mixing 1206 anechoic-clean WSJ validation utterances with the factory noise from the NOISEX-92 dataset Varga and Steeneken (1993) at −6 dB SNR. Once trained, DN-ARN is tested on noisy speech generated by mixing 330 anechoic-clean WSJ test utterances with ADTBabble and ADTCafeteria noises (http://www.auditec.com/) at {−6, −3, 0, 3, 6, 9} dB SNR levels.

Training data for DR-ARN is generated by randomly convolving 7138 anechoic-clean WSJ training utterances with training RIRs in the T60 (room reverberation time) range of [0.2, 1.0] s. DR-ARN is validated on reverberant-clean utterances generated by convolving 1206 anechoic-clean WSJ validation utterances with validation RIRs in the T60 range of [0.8, 1.0] s. Once trained, DR-ARN is evaluated on reverberant speech generated by convolving 330 anechoic-clean WSJ test utterances with test RIRs in the T60 range of [0.2, 0.4] s, [0.4, 0.6] s, [0.6, 0.8] s, and [0.8, 1.0] s separately. The training target of DR-ARN is direct-path speech, which is generated by convolving the anechoic-clean speech with the direct-path RIR. When generating a test utterance, an anechoic-clean speech signal with L_s samples is convolved with a direct-path RIR whose peak value is at the pth sample to generate reverberant-clean speech. Then the first $p-1$ samples of the reverberant-clean speech signal are removed and the next L_s samples are retained for testing. This technique addresses the alignment issue for ASR decoding, such that the direct-path speech signals of different RIRs have the same WER when tested on the backend trained on anechoic-clean speech.

Training data for NR-ARN is generated by randomly convolving 7138 anechoic-clean WSJ training utterances with training RIRs in the same way as for DR-ARN and then mixing with 10k noises in the same way as for DN-ARN. Validation data is generated in the same way as that for DR-ARN but mixed with factory noise at −6 dB SNR. Test data for NR-ARN is generated in the same way as for DR-ARN but mixed with noise in the same way as for DN-ARN such that test speech is both reverberant and noisy. For better comparison, noise segments added to the reverberant-clean speech for each level of T60 are the same.

DN-ARN, DR-ARN, and NR-ARN share the same model architecture. The sampling rate for all utterances is 16 kHz. All training samples are generated randomly and dynamically for all ARNs. We apply root mean square normalization to noisy mixtures, and clean speech is scaled to produce a specified SNR level. During training, the number of samples for each utterance is set to 64000. Input and output frame size is set to 16 ms with a 2 ms frame shift. Dimension N for BLSTM is set to 1024 such that forward and

backward hidden states are sized at 512. The dropout rate is set to 0.05 for feedforward blocks. The ARNs are trained using the PCM (phase-constrained magnitude) loss Pandey and Wang (2021), which is computed in terms of both estimated speech and estimated noise in the T–F domain defined as

$$L_{PCM}(\mathbf{s},\hat{\mathbf{s}}) = \frac{1}{2}L_{SM}(\mathbf{s},\hat{\mathbf{s}}) + \frac{1}{2}L_{SM}(\mathbf{n},\hat{\mathbf{n}}), \tag{5}$$

where $\hat{\mathbf{n}} = \mathbf{y} - \hat{\mathbf{s}}$ is the estimated noise. The STFT magnitude (SM) loss L_{SM} is defined as

$$\begin{aligned} L_{SM}(\mathbf{s},\hat{\mathbf{s}}) = &\frac{1}{TF}\Sigma_{t=0}^{T-1}\Sigma_{f=0}^{F-1}|(|S_r(t,f)| + |S_i(t,f)|) \\ &- (|\hat{S}_r(t,f)| + |\hat{S}_i(t,f)|)|, \end{aligned} \tag{6}$$

where $\mathbf{S}$ and $\hat{\mathbf{S}}$ denote the STFT of $\mathbf{s}$ and $\hat{\mathbf{s}}$, respectively. Subscript r and i respectively denote the real and imaginary part. The number of frequency bins is denoted by F.

One epoch of training consists of 157036 utterances, and the ARNs are trained for 100 epochs with batch size 16. The Adam optimizer Kingma and Ba (2015) is utilized. The learning rate of the first 33 epochs is fixed to $2e^{-4}$ and then exponentially decays every epoch till the final learning rate of $2e^{-5}$. All ARNs are trained on two NVIDIA V100 GPUs. Because STOI is shown to relate to WER Moore et al. (2017), we also use validation STOI as a model selection criterion in addition to validation PCM loss. Validation STOI is computed by averaging the STOI scores on all utterances in the validation set for each epoch. Once the training is done, the checkpoint corresponding to the best validation score is selected.

3.1.3. CHiME-2

Our experiments are conducted on the medium vocabulary track (track 2) of the CHiME-2 corpus Vincent et al. (2013). It is a commonly used dataset for robust ASR evaluation and is generated by convolving WSJ clean speech with binaural RIRs (BRIRs) and mixing with non-stationary family home noise Barker et al. (2013). Although the speech materials are from the WSJ corpus, due to the BRIRs used in the data generation process Christensen et al. (2010), speech signal alignments are altered, which makes it infeasible to use WSJ anechoic-clean speech as the training target for SE. Therefore, we treat reverberant-clean speech as the target signal for training DN-ARN. Two channels are averaged to produce single-channel speech. We employ ARN for time-domain and TF-CrossNet for T–F domain SE on denoising. To further analyze the effect of SE network architecture on the effectiveness of decoupling SE frontend and ASR backend, we add magnitude-phase based MP-SENet (Lu et al., 2023b,a), which is the top performing model on VoiceBank+DEMAND Valentini-Botinhao et al. (2016) and the Deep Noise Suppression (DNS) challenge datasets Reddy et al. (2020).

Training data for DN-ARN, TF-CrossNet, and MP-SENet is generated by randomly mixing 7138 reverberant-clean training utterances from CHiME-2 with 10k noises in the same way as for DN-ARN for WSJ. Validation data is generated by mixing 409 reverberant-clean validation utterances from CHiME-2 with the factory noise at −6 dB SNR. Once trained, the frontends are tested on reverberant-noisy test data, which has six SNR levels with each containing 330 utterances. The training setup of DN-ARN on CHiME-2 is the same as DN-ARN on WSJ.

For TF-CrossNet, the network configurations are kept the same as in A. Kalkhorani and Wang (2024). The learning rate schedule and batch sizes are the same as those of DN-ARN trained on WSJ. One TF-CrossNet is trained with the PCM loss, and the other is trained with scale-invariant signal-to-distortion ratio (SI-SDR) Le Roux et al. (2019) training and validation loss.

For MP-SENet, we use the same network configuration as in Lu et al. (2023a). One model is trained with the same setting as in Lu et al. (2023a) on VoiceBank+DEMAND with a composite loss, which is a combination of magnitude and phase spectrum loss, complex spectrum loss, STFT consistency loss, and MetricGAN-based loss. We only change the batch size to 24 and the network is trained on 4 NVIDIA H100 GPUs. We also train another MP-SENet with the same training configuration as DN-ARN on CHiME-2 with PCM loss, with a batch size of 12 on 4 H100 GPUs.

3.1.4. LibriSpeech

We evaluate the denoising performance of ARN on the LibriSpeech corpus Panayotov et al. (2015), which contains around 1000 h of read English speech, sourced from audiobooks available in the public domain, through the LibriVox project.[2] LibriSpeech has different speech materials from WSJ or CHiME-2. The training data is generated by mixing all 960 hr training utterances from train-clean-100, train-clean-360, and train-other-500, with 10k noises at SNR uniformly selected within the [−5, 0] dB and [0, 10] dB ranges, with 50% probability for each range. Validation data is generated by mixing factory noise with all dev-clean and dev-other utterances at −5 dB SNR. The test-clean and test-other utterances are mixed with ADTBabble and ADTCafeteria noise at {−5, −2, 0, 2, 5, 10} dB SNR levels for evaluation.

DN-ARN on LibriSpeech is trained on two NVIDIA A100 GPUs, with a batch size of 32. The learning rate schedule, optimizer, and loss function are the same as those of DN-ARN on WSJ.

The SE performance is evaluated using standard STOI and perceptual evaluation of speech quality (PESQ) Rix et al. (2001) metrics. STOI ranges typically between [0, 1] and indicates speech intelligibility, usually in percentage. PESQ ranges between [−0.5, 4.5] and a higher score denotes higher speech quality.

[2] https://librivox.org

3.2. Backend recognition

3.2.1. WSJ

The backend of the proposed DN-ARN, DR-ARN, and NR-ARN systems is the WRConformer AM trained on anechoic-clean WSJ speech, with 7138, 1206, and 330 utterances for training, validation, and test, respectively.

All WRConformer AMs share the same model architecture. The configuration of WRCNN is kept the same as in Yang et al. (2022). The attention dimension is set to 256. The kernel size of the 1-D depthwise convolution is set to 16. We use the learning rate schedule from Vaswani et al. (2017), with 2k warm-up steps and a learning rate factor k of 100. The Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e^{-9}$ is utilized for model training. The batch size is set to 3 and short utterances are padded with zeros to match the length of the longest utterance in each batch. The dropout rate is set to 0.15 for network and attention weights. All WRConformer AMs are trained for 25 epochs and model selection is based on the cross-entropy loss on the validation set. Training labels are 1965 senones in our experiments, and they are generated as in Wang and Wang (2016). For log-Mel feature extraction, we follow the approach in Wang et al. (2020a), but skip pre-emphasizing, dithering, and direct-current offset removal steps. A Hamming window is applied to the input waveform for STFT. Then a small value of e^{-40} is added to prevent the underflow of logarithmic operation.

The decoder is the same as in Wang and Wang (2016). AM outputs are first subtracted by the log priors and then fed to the decoder, which is based on the CMU pronunciation dictionary and the official 5k close-vocabulary tri-gram language model. The decoding beamwidth is set to 13, and the lattice beamwidth is 8. The number of active tokens ranges from 200 to 700. Language model weights ranging from 4 to 25 are utilized.

We train two baseline WRConformer AMs for each system. One is trained on 157036 utterances (denoted as 160k) following the approach in Wang et al. (2020a), and another AM is trained on 320k utterances (denoted as 320k) in the same way for a potentially better baseline. Each baseline is trained using the same inputs as those to the corresponding frontend. That is, the baselines for DN-ARN, DR-ARN, and NR-ARN are trained on anechoic-noisy, reverberant-clean, and reverberant-noisy speech, respectively. The value of the learning rate factor k for this model training is divided by 10 after the 6-th epoch and by another 10 after the 12-th epoch. Decreasing the value of k plays the role of fine-tuning and yields better training. Model trained on 320k reverberant-noisy speech for the DR-ARN system overfits during training, so k is initially set to 50 and divided by 10 after the 12-th epoch. We believe that the smaller variation of reverberation than additive noise accounts for the overfitting in the DR-ARN system.

To match the learning rate with the backend trained on 160k utterances, we modify the learning rate for the backend trained on 320k utterances. One learning rate value is used twice before changing to the next value in the learning rate schedule of the backend trained on 160k utterances. For example, if $[lr1, lr2, lr3]$ are three consecutive learning rates for the backend trained on 160k utterances, the backend trained on 320k utterances will use the learning rates of $[lr1, lr1, lr2, lr2, lr3, lr3]$ for six consecutive training steps. This modification makes the learning rates of the two backends epoch-wise consistent.

3.2.2. CHiME-2

The backend of the proposed DN-ARN system is the WRConformer AM trained and validated on CHiME-2 7138 and 409 reverberant-clean utterances, respectively. It has the same training setup with the backend on WSJ. We train two baseline WRConformer AMs on 160k and 320k utterances with the same settings as WSJ on reverberant-noisy speech. Following the official CHiME-2 recipe, We also train and validate a noise-dependent WRConformer AM (denoted as noise-dependent backend) on 7138 and 409 CHiME-2 reverberant-noisy utterances, respectively. It is trained with the same settings as other backends except for a learning rate factor k of $1e^4$ and 5k warm-up steps.

3.2.3. LibriSpeech

The backend utilized for LibriSpeech is the E2E WRConformer that has 10 Conformer encoder blocks, 6 Transformer decoder blocks, 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. STFT frame and shift sizes are 512 and 160, respectively. The CTC weight is set to 0.3, and the label smoothing weight is 0.1. The E2E WRConformer is trained for 50 epochs on the LibriSpeech 960 hr dataset on four NVIDIA A100 GPUs.

We create a training set by mixing the LibriSpeech 960 hr training set and 10k noises in the same way as for DN-ARN on LibriSpeech. To ensure robust learning, each utterance in the LibriSpeech training set is mixed with 4 randomly selected noise segments, resulting in a total of 3840 hr noisy speech (denoted as 4k hr). The validation set is the same as for DN-ARN on LibriSpeech.

As the baselines for WSJ and CHiME-2 do not see the same amount of noise as the frontend training, we add two dynamic baselines on LibriSpeech to ensure that the E2E WRConformer trained on noisy speech observes the same amount of noise as that seen by the frontend during training. The DN-ARN takes 4 s long utterances during training, resulting in 312 hr noisy speech per epoch. In contrast, the E2E WRConformer sees 960 hr of noisy speech per epoch. Thus, for the first dynamic baseline, we generate noisy speech dynamically to train E2E WRConformer for 35 epochs, exactly matching the amount of noise seen by DN-ARN. For the second dynamic baseline, we train it in the same way as the first dynamic baseline but add clean speech to the training set to create a multi-conditioned training set. The setup for dynamic baselines is the same as for E2E WRConformer on clean speech.

Table 1
ASR (%WER) results of the proposed DN-ARN system and comparison systems on WSJ. 'Enh.' denotes 'Enhanced'. All ASR models are trained on anechoic speech.

Test noise	Row	SE network	ASR network	ASR train data	SNR						Avg
					−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	
ADT	Unproc.	–	WRConformer	Clean	97.3	89.5	70.8	46.1	27.5	15.6	57.8
	1 Wang et al. (2020a)	–	WRBN	Noisy-160k	54.0	31.5	17.9	10.8	7.2	4.9	21.1
	2	–	WRConformer	Noisy-160k	37.2	19.8	10.7	6.1	4.2	3.5	13.6
	3	–	WRConformer	Noisy-320k	32.0	16.8	9.0	5.4	4.1	3.0	11.7
	4 Wang et al. (2020a)	GRN	WRBN	Enh.-160k	39.4	20.9	11.0	6.2	4.4	3.2	14.2
	5	ARN (PCM)	WRConformer	Clean	**18.4**	10.0	**5.4**	**3.8**	3.4	**2.9**	**7.3**
	6	ARN (STOI)	WRConformer	Clean	18.8	**9.8**	**5.4**	4.0	3.3	3.0	7.4
	7	ARN (LibriSpeech, STOI)	WRConformer	Clean	21.9	10.9	5.9	3.9	**3.2**	**2.9**	8.1

3.3. Cross-corpus generalization to CHiME-4

To validate the cross-corpus generalization of the proposed system, we take the CHiME-4 Vincent et al. (2017) corpus for evaluation, which comprises simulated and real (recorded) noisy speech using WSJ text material. There are 1640 real and 1640 simulated noisy utterances from 4 different speakers in the development set. The test set consists of 1320 real and 1320 simulated noisy utterances from 4 other speakers. Real data is recorded speech in noisy environments (bus, cafe, pedestrian area, and street junction) uttered by real talkers, and simulated data is generated by mixing recorded speech via a close-talking microphone and separately recorded noises. The training set is not used in this study. We select the fifth channel for single-channel track evaluation. In our decoupled system, we take the ARN trained on LibriSpeech with STOI validation as the frontend, and WRConformer AM trained on WSJ0 SI-84 as the backend. We follow the same procedure as in Yang et al. (2022) for LM rescoring and unsupervised speaker adaptation for a fair comparison between the proposed decoupled system and the ASR model trained on noisy speech. We also compare enhanced and unenhanced noisy speech using the noise-robust backend of Whisper.

4. Results and discussions

This section presents and analyzes the evaluation and comparison results of the decoupled systems and the corresponding baselines on three corpora. As the focus of this study is on ASR performance, we only provide a summary of the SE performance for each system. Since background noise and room reverberation are the two main distortions to speech signals that are of different kinds, we analyze the performance of each distortion separately and then both in this section. The columns of each table follow the order of test condition, frontend network, backend network, backend training data, and results.

4.1. Results on WSJ

4.1.1. Denoising only

On average, DN-ARN with PCM and STOI validation improves STOI by 19.2% and 19.3%, and PESQ by 1.4, respectively, across different SNRs. DN-ARN with STOI validation slightly outperforms DN-ARN with PCM validation.

ASR evaluation results are shown in Table 1. Results on babble and cafeteria noise are averaged and denoted as ADT. When tested on the WSJ anechoic-clean utterances, the WERs for the WRConformer AM trained on clean, noisy-160k, and noisy-320k are 2.0%, 2.4%, and 2.0%, respectively. We first evaluate and compare AMs trained on noisy speech. The results in row 2 and row 3 outperform those of row 1 by 35.5% and 44.4%, respectively. The results demonstrate the effectiveness of the noise-independent WRConformer AM baselines.

We next evaluate the decoupled system and compare it with the WRBN-based system that is trained on GRN (gated residual network) Tan et al. (2018) enhanced speech Wang et al. (2020a). The latter system also applies the same GRN model for SE, which creates matched training and test conditions. As a result, the WER score in row 4 is the best among the different kinds of ASR training data generation investigated in Wang et al. (2020a), including noise-independent training. The decoupled system in row 5 achieves 7.3% WER on average, outperforming that in row 4 by 48.4% relatively. The large WER improvement can be attributed to the decoupled system with a more powerful SE frontend, in addition to the use of WRConformer in row 5. The decoupled system with DN-ARN with STOI validation performs closely in row 6. Both decoupled systems substantially outperform all baselines. These results demonstrate that SE translates to improved ASR results for anechoic-noisy speech. We also present cross-corpus evaluation results with ARN trained on LibriSpeech (Section 3.1.4) in row 7. The ASR backend is the same as in the unprocessed (unproc.) row, as well as rows 5 and 6. Row 7 has lower WERs than the noise-robust backend in row 3, and achieves best performance in 6 dB and 9 dB conditions.

Table 2
ASR (%WER) results of the proposed DR-ARN system and comparison systems on WSJ. All ASR models are WRConformer AM trained on clean speech.

Row	SE network	ASR train data	T60				Avg
			0.2–0.4 s	0.4–0.6 s	0.6–0.8 s	0.8–1.0 s	
Unproc.	–	Anechoic	15.2	33.6	50.4	57.9	39.3
1	–	Reverberant-160k	3.5	4.3	4.5	5.3	4.4
2	–	Reverberant-320k	3.3	4.4	4.5	5.1	4.3
3	ARN (PCM)	Anechoic	2.5	**3.0**	**3.2**	**3.3**	**3.0**
4	ARN (STOI)	Anechoic	**2.4**	**3.0**	**3.2**	3.5	**3.0**

Table 3
ASR (%WER) results of the proposed NR-ARN system and comparison systems on WSJ. Clean data denotes anechoic clean and noisy data denotes reverberant noisy. All ASR models are WRConformer AM. ASR training data is anechoic in the clean case and reverberant in the noisy case.

Test noise	Test T60	SE network	ASR train data	SNR						Avg
				−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	
ADT	0.2–0.4 s	–	Clean	96.0	93.0	84.5	67.9	49.7	35.3	71.1
		–	Noisy-160k	52.4	31.4	17.7	10.2	6.5	5.0	20.5
		–	Noisy-320k	48.6	28.2	15.6	9.3	6.3	4.9	18.8
		ARN (PCM)	Clean	**38.0**	**20.8**	12.0	7.7	5.4	**4.1**	14.7
		ARN (STOI)	Clean	**38.0**	20.9	**11.4**	**7.2**	**5.1**	**4.1**	**14.5**
	0.4–0.6 s	–	Clean	96.7	95.0	89.5	79.1	66.7	53.6	80.1
		–	Noisy-160k	55.3	34.7	19.8	12.1	8.0	6.0	22.7
		–	Noisy-320k	51.6	31.4	18.6	11.3	7.4	5.6	21.0
		ARN (PCM)	Clean	46.5	27.7	15.1	9.4	6.0	**4.7**	18.2
		ARN (STOI)	Clean	**46.0**	**27.4**	**15.0**	**9.4**	**6.0**	**4.7**	**18.1**
	0.6–0.8 s	–	Clean	96.9	95.8	91.7	85.2	75.6	66.6	85.3
		–	Noisy-160k	58.9	37.9	21.9	14.0	9.5	7.0	24.8
		–	Noisy-320k	55.5	34.4	20.5	12.8	8.3	6.0	22.9
		ARN (PCM)	Clean	52.7	32.3	18.4	11.3	7.4	5.6	21.3
		ARN (STOI)	Clean	**52.6**	**31.1**	**17.9**	**10.9**	**7.2**	**5.5**	**20.9**
	0.8–1.0 s	Mixture	Clean	97.1	96.3	93.9	88.6	81.8	74.2	88.6
		–	Noisy-160k	61.0	39.6	24.6	15.4	10.2	7.4	26.4
		–	Noisy-320k	58.5	37.1	21.9	13.7	9.3	7.2	24.6
		ARN (PCM)	Clean	57.6	34.8	20.4	12.2	**8.3**	**6.2**	23.2
		ARN (STOI)	Clean	**57.1**	**34.5**	**20.0**	**12.1**	**8.3**	**6.2**	**23.0**

4.1.2. Dereverberation only

DR-ARN with PCM validation and STOI validation are evaluated on reverberant-clean WSJ speech in four T60 ranges. The two DR-ARNs improve STOI by 17.5% on average. The two DR-ARNs have the same PESQ scores in the four T60 ranges, and on average improve PESQ by 1.4.

Table 2 provides the ASR results of the proposed system and baselines. Tested on the anechoic-clean WSJ utterances, the WRConformer AMs trained on anechoic, reverberant-160k, and reverberant-320k have 2.0%, 3.0%, and 3.1% WER, respectively. The results in row 2 outperform those in row 1 by 2.3% relatively, showing a slight effect of training data size. In rows 3 and 4, DR-ARN with PCM and STOI validation produces 3.0% WER. The results in row 3 outperform those of the row 2 by 30.6%. This shows that DR-ARN is capable of effective speech dereverberation, and improved dereverberation translates to better ASR with the corresponding AM trained on anechoic speech.

4.1.3. Denoising and dereverberation

NR-ARN is evaluated on babble and cafeteria noise at six SNR levels in four T60 ranges. NR-ARN with PCM validation performs similarly to that with STOI validation. In the T60 range of [0.8, 1.0] s, NR-ARN with STOI validation yields 28.5% and 27.6% improvement in STOI and 1.1 improvement in PESQ, for babble and cafeteria noise, respectively. The results illustrate the ability of NR-ARN in joint denoising and dereverberation.

Table 3 presents the ASR results of the proposed decoupled system of NR-ARN and comparisons with baselines. Results on babble and cafeteria noise are averaged and denoted as ADT. Boldface entries indicate the best results for each T60 range. When tested on the WSJ anechoic-clean utterances, the WER for the WRConformer trained on clean, noisy-160k, and noisy-320k are 2.0%, 3.6%, and 3.1%, respectively. AM with NR-ARN with STOI validation outperforms that with PCM validation and the noise- and reverberation-independent baselines in all test conditions. For four T60 ranges in increasing order, AM with NR-ARN with STOI validation outperforms the baseline trained on noisy-320k by 23.2%, 13.9%, 9.0%, and 6.4%, respectively. These results demonstrate that NR-ARN enhanced speech results in improved ASR performance for reverberant-noisy speech.

Table 4
ASR (%WER) results of the proposed DN-ARN system and comparison systems on CHiME-2. All ASR models are trained on reverberant speech. 'Enh.' and 'Feat.' denote 'Enhanced' and 'Feature', respectively.

Row	SE network	ASR network	ASR train data	SNR						Avg
				−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	
Unproc.	–	WRConformer	Clean	73.8	64.9	57.6	45.1	36.1	28.5	51.0
1 Wang et al. (2020a)	–	WRBN	Noisy-160k	17.5	13.1	10.7	8.8	7.7	6.6	10.7
2	–	WRConformer	Noisy-160k	19.8	13.3	10.9	9.3	7.3	6.6	11.2
3	–	WRConformer	Noisy-320k	16.3	10.2	9.3	7.1	6.5	5.6	9.2
4 Wang et al. (2020a)	–	WRBN	CHiME-2 Noisy	14.8	10.0	9.0	6.8	6.3	5.5	8.7
5	–	WRConformer	CHiME-2 Noisy	14.4	10.3	7.9	6.7	6.0	5.5	8.5
6 Wang et al. (2020a)	GRN	WRBN	Enh.-160k	15.5	11.0	9.7	7.1	6.5	5.5	9.2
7 Wang and Wang (2019)	GRN	WRBN	Enh. Feat.-160k	13.1	9.4	7.9	6.2	5.5	4.5	7.8
8 Plantinga et al. (2021)	Wide ResNet	MLP	Multi-conditioned	15.2	10.9	8.3	6.7	5.8	5.2	8.7
9	ARN (PCM)	WRConformer	Clean	13.3	9.7	7.8	6.4	5.3	4.6	7.8
10	ARN (STOI)	WRConformer	Clean	9.9	7.0	6.5	5.5	4.5	4.2	6.3
11	TF-CrossNet (PCM)	WRConformer	Clean	8.5	**6.5**	5.5	**4.8**	4.5	**4.1**	**5.6**
12	TF-CrossNet (STOI)	WRConformer	Clean	**7.8**	6.7	**5.2**	5.2	**4.4**	4.2	**5.6**
13	TF-CrossNet (SI-SDR)	WRConformer	Clean	9.1	7.5	6.2	5.4	5.2	4.6	6.4
14	MP-SENet (Composite Loss)	WRConformer	Clean	9.3	7.1	6.0	4.8	4.8	4.4	6.1
15	MP-SENet (PCM)	WRConformer	Clean	11.8	8.3	6.8	6.2	4.9	4.8	7.1
16	MP-SENet (STOI)	WRConformer	Clean	12.0	8.0	6.4	6.0	5.1	4.3	7.0

Table 5
ASR (%WER) results of different E2E WRConformer models tested on LibriSpeech test sets. 'Dyn.' and 'cond.' Denote 'Dynamic' and 'conditioned', respectively.

Train data	Test data			
	dev-clean	dev-other	test-clean	test-other
Clean	1.7	3.9	1.9	4.1
Noisy-4k-hr	4.3	8.4	4.6	8.5
Dyn. Noisy	2.3	5.5	2.6	5.7
Dyn. Multi-cond.	6.0	10.7	6.1	10.7

4.2. Results on CHiME-2

On average, DN-ARN with PCM and STOI validation improves STOI by 12.6% and 12.1%, and PESQ by 1.0, respectively. For TF-CrossNet with PCM, STOI, and SI-SDR validation, the STOI improvements are 9.4%, 9.0%, and 9.1% and PESQ improvements are 0.8, 0.7, and 0.8, respectively. MP-SENet with composite loss, PCM and STOI validation improves STOI by 10.4%, 12.8%, and 12.6% and PESQ by 1.1%, respectively.

ASR results are presented in Table 4. When tested on CHiME-2 reverberant-clean utterances, the WERs for the WRConformer AM trained on clean, CHiME-2 noisy, noisy-160k, and noisy-320k are 2.9%, 4.4%, 3.5%, and 3.3%, respectively.

Trained on noisy speech directly, the baseline in row 3 improves WER by 14.7% compared with those of row 1. The noise-dependent WRConformer AM in row 5 outperforms the noise-dependent WRBN in row 4 by 2.6%. WRBN-based models perform consistently worse than WRConformer AMs, demonstrating the effectiveness of the WRConformer baselines.

We next evaluate the proposed decoupled system and compare it with the other systems that incorporate SE. Row 6 of Table 4 is WRBN trained on GRN enhanced speech signals Wang et al. (2020a), and row 7 is WRBN trained on GRN enhanced features (specifically, magnitude spectra) Wang and Wang (2019). A perceptual loss based model Plantinga et al. (2021) is included in row 8, and it employs a Wide ResNet trained with a perceptual loss as the frontend and its backend use an off-the-shelf Kaldi CHiME-2 recipe trained on multi-conditioned speech (WSJ0 SI-84 clean + CHiME-2 reverberant-clean + CHiME-2 reverberant-noisy). The proposed system with DN-ARN with STOI validation achieves 6.3% WER, which outperforms the previous best Wang and Wang (2019) by 19.3% relatively. In row 12, the decoupled system with TF-CrossNet with STOI validation achieves a 5.6% WER, outperforming the previous best by 28.3% relatively. Also, STOI validation produces better results than PCM or SI-SDR validation. These results clearly demonstrate that either the time or T–F domain frontend can deal with reverberant-noisy speech for decoupled robust ASR, expanding our previous findings on time-domain frontend only Yang et al. (2023a). In row 14, MP-SENet trained with the composite loss outperforms DN-ARN with a 6.1% WER. In row 15 and 16, MP-SENet trained with the PCM loss outperforms the ASR baseline trained on noisy speech in row 3 and 5, as well as the previous best in row 7. The slight improvement of row 16 over 15 indicates the effectiveness of model selection based on STOI validation. That row 14 outperforms rows 15 and 16 suggests the composite loss originally designed for MP-SENet Lu et al. (2023a), although complicated, appears most effective for this architecture.

4.3. Results on LibriSpeech

As in the evaluation on WSJ, the decoupled system with DN-ARN on LibriSpeech is evaluated with babble and cafeteria noises. Results are averaged across two noise types and denoted as ADT in Table 6. On the test-clean set, DN-ARN with PCM and STOI

Table 6
ASR (%WER) results of the proposed DN-ARN and comparison systems on LibriSpeech. All ASR models use E2E WRConformer.

Test noise	Test data	Row	SE network	ASR train data	SNR						Avg
					−5 dB	−2 dB	0 dB	2 dB	5 dB	10 dB	
ADT	test-clean	Unproc.	–	Clean	100.4	88.7	71.5	48.5	20.2	5.1	55.7
		1	–	Noisy-4k-hr	41.1	20.3	13.1	9.4	6.8	6.3	16.1
		2	–	Dynamic Noisy	39.0	17.7	10.8	7.1	4.6	3.2	13.7
		3	–	Dynamic Multi-conditioned	55.3	29.7	19.2	13.4	9.1	6.5	22.2
		4	ARN (PCM)	Clean	28.7	**13.0**	**7.9**	5.3	**3.5**	**2.4**	**10.1**
		5	ARN (STOI)	Clean	**28.6**	**13.0**	**7.9**	**5.2**	**3.5**	**2.4**	**10.1**
	test-other	Unproc.	–	Clean	102.8	97.6	87.3	71.3	44.4	15.7	69.8
		6	–	Noisy-4k-hr	63.9	41.1	29.8	22.0	15.3	12.0	30.7
		7	–	Dynamic Noisy	62.2	38.4	27.2	18.8	12.4	8.00	27.9
		8	–	Dynamic Multi-conditioned	77.3	52.7	39.6	29.9	20.9	14.1	39.3
		9	ARN (PCM)	Clean	52.4	31.0	**21.3**	**14.9**	**9.7**	**6.2**	22.6
		10	ARN (STOI)	Clean	**52.1**	**30.7**	21.4	15.0	9.8	**6.2**	**22.5**

validation improves STOI by 21.5%, and PESQ by 1.5. On the test-other set, DN-ARN with PCM and STOI validation shows improvements in STOI by 20.5% and 20.6%, and in PESQ by 1.2, respectively.

For ASR performance, the WERs of each system tested on clean speech are presented in Table 5. The E2E WRConformer trained on clean speech outperforms all baseline models. The E2E WRConformer trained on dynamic noisy speech outperforms the other two baselines. The WERs of the ASR model trained on dynamic multi-conditioned speech are the highest among all the baselines, due to its training to match the data observed for SE frontend training which is quite different from clean test data.

Tested under ADT noises, the proposed decoupled systems outperform all baselines. The decoupled systems with DN-ARN with PCM and STOI validation have very close performances. The system with STOI validation in row 5 and row 10 of Table 6 outperforms the best baseline trained on dynamic noisy speech in row 2 and row 7 by 26.3% and 19.4% relatively on test-clean and test-other datasets, respectively. This demonstrates that, using the same amount of noisy speech, training an SE frontend is more effective than training an ASR model for robust ASR.

4.4. Results on CHiME-4

Table 7 presents CHiME-4 evaluation results, as well as comparisons with many baseline systems. The comparison systems include six trained on CHiME-4 noisy speech directly (Chen et al., 2018; Guo et al., 2021; Heymann et al., 2016; Du et al., 2016; Wang et al., 2020b; Yang et al., 2022), and five that leverage additional speech materials for pretraining and employ CHiME-4 training data for fine-tuning (Zhu et al., 2023a; Yang et al., 2023b; Hu et al., 2024; Wang et al., 2022; Chang et al., 2022). Without training on CHiME-4, the proposed decoupled system achieves 3.3/4.4% WER on simulated/real test data. Our system on the simulated data dramatically outperforms the previous best Chang et al. (2022) by 45.8%, utilizing around 100 times less cross-corpus speech data. With the same amount of cross-corpus speech material, the decoupled system outperforms the previous best Hu et al. (2024) on the real data by 16.2%. There is a WER gap between our result and the current state-of-the-art Chang et al. (2022) on real data, and this is because the training data for the ASR model in Chang et al. (2022) contains matched-condition real data, while our ASR model is trained on WSJ only, making the unseen real condition more difficult. It is worth noting that all of the previous state-of-the-art methods on the single-channel CHiME-4 evaluation utilize CHiME-4 training data, making ours the first state-of-the-art method that benefits from pretrained speech enhancement on a different corpus and ASR. Another interesting observation regards the relative WER scores between the simulated and real data: the trend of our system reverses that of all the other baselines. It seems more reasonable to expect better results on simulated data, as exhibited in our system, and this trend is more consistent with two-channel and six-channel CHiME-4 evaluation results Wang et al. (2020b). This demonstrates that the decoupled system generalizes well to cross-corpus real acoustic scenarios.

To gain more insight into how unprocessed noisy speech and enhanced speech perform on a robust ASR model, we employ Whisper Radford et al. (2023), an open-source ASR model trained on 680k hours of multi-conditioned speech data. The WER results are shown in Table 8. We experiment with Whisper models of varying sizes, and compare the WERs on unprocessed speech and enhanced speech by ARN trained on LibriSpeech (see Section 3.1.4). The results show that ARN enhanced speech consistently improves ASR performance across all Whisper models, except for a slight decrease in the real development set for Whisper-Large. Notably, Whisper-Small with enhanced speech outperforms Whisper-Medium on unprocessed speech with only around 40% of parameters.

It is noteworthy that the WER scores of even Whisper-Large on enhanced speech are worse than those of our decoupled system and several other baselines in Table 7. This mainly reflects the fact that the baseline models in Table 7 all make use of the CHiME-4 training data. Although the decoupled system does not use the training data, its ASR model is trained on WSJ, which shares the same text corpus with CHiME-4. As described in Section 3.3, the ASR model in the decoupled system includes the steps of language model rescoring and iterative speaker adaptation. Without these processing steps, the decoupled system will underperform Whisper-Large.

Table 8 also reports the combined parameter count of the frontend and backend for each model size and speech type. The decoupled nature of SE frontend and ASR backend allows for independent modification of each component, eliminating the need for costly re-training. This flexibility is a key advantage of the proposed decoupled system.

Table 7
ASR (%WER) results of the proposed and comparison systems on CHiME-4 (Single-channel).

System	Cross-corpus Speech Hour (hr)/ Train on CHiME-4?	Dev. set		Test set	
		Simu.	Real	Simu.	Real
Kaldi Baseline Chen et al. (2018)	0/✓	6.8	5.6	12.2	11.4
ESPnet Conformer Guo et al. (2021)		9.1	7.9	14.2	13.4
WRBN Heymann et al. (2016)		6.7	5.2	11.1	9.3
Du et al. (2016)		6.6	4.6	11.8	9.2
Wang et al. (2020b)		5.0	3.5	9.4	6.8
Yang et al. (2022)		5.0	3.4	8.6	6.3
Zhu et al. (2023b)	1k/✓	–	3.1	–	5.8
Yang et al. (2023b)	1k/✓	–	3.1	–	5.7
Hu et al. (2024)	1k/✓	–	2.6	–	5.3
Wang et al. (2022)	60k/✓	–	2.7	–	5.5
Chang et al. (2022)	94k/✓	3.2	**2.0**	6.1	**3.9**
Ours	1k/✗	**3.1**	3.0	**3.3**	4.4

Table 8
ASR (%WER) comparisons on CHiME-4 with different whisper models and their numbers of parameters.

Whisper model	Speech type	Dev. Set		Test set		#Params (M)
		Simu.	Real	Simu.	Real	
Tiny	Unproc.	22.8	16.7	29.8	28.6	39
	Enhanced	10.5	9.9	11.4	13.9	94
Base	Unproc.	15.6	10.5	20.9	18.0	74
	Enhanced	8.3	7.8	9.3	11.2	129
Small	Unproc.	10.7	7.0	14.5	11.0	244
	Enhanced	6.6	6.1	7.1	8.3	299
Medium	Unproc.	8.6	5.7	12.1	8.6	769
	Enhanced	5.7	5.5	6.4	7.3	824
Large	Unproc.	7.0	4.5	10.2	7.0	1550
	Enhanced	4.8	4.7	5.5	6.4	1605

4.5. *Why does the decoupled approach work?*

The experimental results on denoising, dereverberation, and both demonstrate the effectiveness of the decoupled approach for robust ASR. In this section, we analyze factors contributing to the success of the proposed decoupled approach over previous methods that achieve suboptimal performance with ASR trained on clean speech. We investigate the impacts of training data configuration, loss functions, and model architectures, and suggest that the strength of SE frontend is the main reason.

4.5.1. *Effects of training data configuration*

To analyze the impact of training data configuration, we compare different SNR settings for DN-ARN training on the CHiME-2 corpus, and the corresponding WER results are presented in Table 9. We evaluate five different configurations, and denote them as Normal, Low, High, Quarter High, and Half-and-Half. Training SNR in dB is sampled from a uniform distribution $\mathcal{U}(-7, 10)$ for Normal, $\mathcal{U}(-7, 0)$ for Low, $\mathcal{U}(0, 10)$ for High, respectively. For Quarter High, we sample SNR from uniform distributions of $\mathcal{U}(-7, 10)$ and $\mathcal{U}(0, 10)$, with 75% and 25% probability respectively for each distribution to be selected. For Half-and-Half, the SNR is sampled from $\mathcal{U}(-7, 10)$ and $\mathcal{U}(0, 10)$, with 50% and 50% probability for each distribution to be selected. For each SNR setting, a DN-ARN is trained and two checkpoints are selected based on PCM loss and STOI validation. Half-and-Half and Low with STOI validation produce the best performance with the ASR backend trained on clean speech, and for the model selected by PCM validation, Half-and-Half is slightly better than Low. This justifies our adoption of the Half-and-Half configuration for subsequent SE frontend training. Interestingly, as shown in Table 9, the Normal, Low, and Quarter High demonstrate comparable performance with Half-and-Half, and the WER for High with STOI validation is higher than those of the other settings. This suggests that incorporating training data with lower SNR ranges benefits ASR systems trained on clean speech.

4.5.2. *Effects of loss functions*

To investigate the impact of loss functions on SE frontend training, We employ AECNN which is a time-domain model Pandey and Wang (2019). We compare four loss functions: PCM loss, time-domain mean squared error (MSE) loss denoted as time-domain,

Table 9
ASR (%WER) results of DN-ARN trained with different training data configurations on CHiME-2. The backend WRConformer is trained on clean speech.

Data configuration	Model selection criterion	SNR						Avg
		−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	
Normal	PCM	12.5	9.2	7.6	6.3	5.3	4.8	7.6
	STOI	10.2	7.3	**6.2**	5.7	4.7	**4.2**	6.4
Low	PCM	13.8	8.9	8.3	6.5	5.7	5.2	8.1
	STOI	**9.7**	**7.0**	6.6	**5.3**	4.8	**4.2**	**6.3**
High	PCM	12.6	8.3	7.4	5.7	4.9	4.6	7.2
	STOI	11.7	8.1	7.2	5.5	4.8	4.6	7.0
Quarter High	PCM	13.3	9.5	8.4	6.2	5.6	5.1	8.0
	STOI	10.2	6.9	6.5	5.5	**4.5**	4.7	6.4
Half and Half	PCM	13.3	9.7	7.8	6.4	5.3	4.6	7.8
	STOI	9.9	**7.0**	6.5	5.5	**4.5**	**4.2**	**6.3**

Table 10
ASR (%WER) results of DN-AECNN trained on different loss functions on WSJ. All ASR models are trained on anechoic clean speech.

Test noise	Loss function	SNR						Avg
		−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	
ADT	Time-domain	60.9	33.7	18.1	10.3	6.9	5.0	22.5
	STFT MAE	52.8	28.9	15.1	8.7	5.8	4.5	19.3
	STFT MSE	55.8	30.2	15.7	8.9	6.2	**4.4**	20.2
	PCM	**50.9**	**27.1**	**14.3**	**8.5**	**5.5**	**4.4**	**18.5**

STFT mean absolute error (MAE) loss denoted as STFT-MAE, and STFT MSE loss denoted as STFT-MSE. The time-domain loss is defined as

$$L_{\text{time-domain}}(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{M} \sum_{k=0}^{M-1} (\mathbf{s}[k] - \hat{\mathbf{s}}[k])^2, \tag{7}$$

where M is the number of speech signal samples. The STFT-MAE loss is the same as L_{SM} in Eq. (6) and the STFT-MSE loss is defined as

$$L_{STFT-MSE}(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} \left(\sqrt{S_r(t,f)^2 + S_i(t,f)^2 + \alpha} - \sqrt{\hat{S}_r(t,f)^2 + \hat{S}_i(t,f)^2 + \alpha} \right)^2, \tag{8}$$

where α is a small positive constant added to stabilize the training ($\alpha = 1e^{-7}$). All AECNN models are trained on WSJ with the same configuration as DN-ARN and are denoted by DN-AECNN. Evaluation is performed in the same way as in Section 4.1.1. As shown in Table 10, the model trained with the PCM loss achieves the best results, which we attribute to the joint modeling of both speech and noise characteristics of the PCM loss. Thus, PCM has been adopted as the loss function to train all SE frontends. The evaluations on WSJ, CHiME-2, CHiME-4, and LibriSpeech corpora consistently demonstrate that the PCM loss is effective in improving robust ASR performance with a clean-trained ASR backend.

4.5.3. Effects of model architectures

The impact of model architectures on the decoupled approach can be assessed from the results in Table 4, which evaluates three different SE frontends on the CHiME-2 corpus: time-domain ARN, time–frequency domain TF-CrossNet, and magnitude-phase based MP-SENet. All three decoupled systems consistently outperform the strong baseline ASR model trained on noisy speech (see rows 3 and 5), despite their distinct architectures. This empirical evidence suggests that the effectiveness of the decoupled framework is architecture-agnostic, if the SE frontend provides sufficient enhancement quality.

The above analyses also indicate that the main reason why previous decoupled methods fail to outperform ASR trained on noisy speech: the lack of sufficiently powerful SE methods or speech enhancement quality. That the decoupled approach works in this study may be attributed to our employment of very recent, high-performance SE frontends of ARN, TF-CrossNet, and MP-SENet. These models with a suitable loss function, trained under sufficiently adverse conditions, and coupled with a downstream ASR model trained on clean speech only, can now show better results than an ASR model trained on noisy speech signals. With the rapid advance of monaural SE algorithms, we expect that the decoupled approach will be advantageous in more and more adverse acoustic conditions.

5. Concluding remarks

This study aims to eliminate the divide between speech enhancement frontend and recognition backend in monaural robust ASR. The time-domain ARN, T–F domain TF-CrossNet, and magnitude-phase based MP-SENet are employed as SE frontends to WRConformer based ASR models trained on clean speech only. The proposed system decouples SE and ASR. Results on the WSJ, CHiME-2, LibriSpeech, and CHiME-4 corpora on denoising, dereverberation, or both, show that SE gains translate to ASR gains. The proposed system outperforms the interference-independent baselines for all test conditions. Our CHiME-2 results have updated the previous best WER by 28.3% relatively in the standard evaluation, and we achieve 3.3/4.4% WER on simulated/real data without training on CHiME-4. These CHiME-4 results cut the previous best WER on simulated data by a large margin, and represent the best WER on real data using the same amount of out-of-corpus speech materials. Further evaluation of enhanced speech on Whisper demonstrates the potential of ARN to serve as a generic SE frontend in a decoupled system.

The largest STOI and PESQ improvements as well as the ASR improvements come from DN-ARN on WSJ, showing that ARN excels in mapping anechoic-noisy speech into anechoic-clean speech, the task that ARN is originally designed for Pandey and Wang (2022). For reverberant and reverberant-noisy speech, the proposed system using DR-ARN, NR-ARN, and TF-CrossNet performs very well.

The model selection criterion of maximum STOI tends to produce better ASR performance than PCM validation, especially for reverberant-noisy speech. For DN-ARN and DR-ARN on WSJ, the system with PCM validation outperforms the system with STOI validation. This gets reversed for DN-ARN on CHiME-2, TF-CrossNet on CHiME-2, and NR-ARN on WSJ, where input speech to ASR is reverberant-noisy. But the WER improvements in DN-ARN on CHiME-2 and NR-ARN on WSJ are larger than the slight improvements in DN-ARN and DR-ARN on WSJ, and TF-CrossNet on CHiME-2. Consistent with Moore et al. (2017), our observations suggest that STOI validation is a strong criterion for SE model selection for downstream ASR tasks. If an ASR model trained on clean speech is available, another possibility is to perform SE model selection directly using ASR accuracy, which is expected to better match the ASR task. Further research is needed to compare ASR and STOI criteria.

In future work, we plan to improve the performance of the frontend for joint denoising and dereverberation, and develop a generic frontend SE model. In addition, we plan to extend our approach to multi-talker speaker separation to test its ability to eliminate interfering speakers, and multi-channel robust ASR tasks.

CRediT authorship contribution statement

Yufeng Yang: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Ashutosh Pandey:** Formal analysis, Conceptualization. **DeLiang Wang:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yufeng Yang, Ashutosh Pandey, DeLiang Wang reports financial support was provided by National Institutes of Health. Yufeng Yang, Ashutosh Pandey, DeLiang Wang reports equipment, drugs, or supplies was provided by Ohio Supercomputer Center. Yufeng Yang, Ashutosh Pandey, DeLiang Wang reports financial support was provided by National Science Foundation. 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 research was supported by an NIH, United States grant (R01DC012048), the Ohio Supercomputer Center, and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

Data availability

Data will be made available on request.

References

A. Kalkhorani, V., Wang, D., 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.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines. In: Proc. IEEE ASRU. pp. 504–511.

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The PASCAL CHiME speech separation and recognition challenge. Comput. Speech Lang. 27, 621–633.

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.

Chen, S.J., Subramanian, A., Xu, H., Watanabe, S., 2018. Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline. In: Proc. Interspeech. pp. 1571–1575.
Christensen, H., Barker, J., Ma, N., Green, P.D., 2010. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In: Proc. Interspeech. pp. 1918–1921.
Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34, 267–285.
Du, J., Tu, Y.H., Sun, L., et al., 2016. The USTC-iFlytek system for CHiME-4 challenge. In: Proc. CHiME-4. pp. 36–38.
Fu, Y., Cheng, L., Lv, S., et al., 2021. AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. In: Proc. Interspeech. pp. 3665–3669.
Fu, S.W., Hu, T.y., Tsao, Y., Lu, X., 2017a. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: Proc. IEEE Int. Workshop Mach. Learn. Signal Process. pp. 1–6.
Fu, S.W., Tsao, Y., Lu, X., Kawai, H., 2017b. Raw waveform-based speech enhancement by fully convolutional networks. In: Proc. Asia-Pacific Signal Inf. process Assoc. Annu. Summit Conf. pp. 006–012.
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., Boyer, F., Chang, X., et al., 2021. Recent developments on ESPnet toolkit boosted by conformer. In: Proc. IEEE Int. Conf. Acoust. Speech Signal process. pp. 5874–5878.
Heymann, J., Bacchiani, M., Sainath, T.N., 2018. Performance of mask based statistical beamforming in a smart home scenario. In: Proc. IEEE Int. Conf. Acoust. Speech Signal process. pp. 6722–6726.
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.
Ho, K.H., Yu, E.L., Hung, J.w., Chen, B., 2023. NAaLoss: Rethinking the objective of speech enhancement. In: Proc. IEEE Int. Workshop Mach. Learn. Signal process. pp. 1–6.
Hu, Y., Chen, C., Zhu, Q., Chng, E.S., 2024. Wav2code: Restore clean speech representations via codebook lookup for noise-robust ASR. IEEE/ ACM Trans. Audio Speech Lang. Process. 32, 1145–1156.
Hu, Y., Liu, Y., Lv, S., et al., 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In: Proc. Interspeech. pp. 2472–2476.
Iwamoto, K., Ochiai, T., Delcroix, M., et al., 2022. How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR. In: Proc. Interspeech. pp. 5418–5422.
Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proc. Int. Conf. Learn. Representations. pp. 1–15.
Kinoshita, K., Ochiai, T., Delcroix, M., Nakatani, T., 2020. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 7009–7013.
Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R., 2019. SDR–half-baked or well done? In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 626–630.
Loizou, P.C., 2013. Speech Enhancement: Theory and Practice, second ed. CRC Press, Boca Raton, FL, USA.
Lu, Y.X., Ai, Y., Ling, Z.H., 2023a. Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement. arXiv:2308.08926.
Lu, Y.X., Ai, Y., Ling, Z.-H., 2023b. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra. In: Proc. Interspeech. pp. 3834–3838.
Luo, Y., Mesgarani, N., 2019. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ ACM Trans. Audio Speech Lang. Process. 27, 1256–1266.
Masuyama, Y., Chang, X., Zhang, W., et al., 2023. 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. pp. 1–5.
Meng, Z., Li, J., Gong, Y., Juang, B.H.F., 2018. Adversarial feature-mapping for speech enhancement. In: Proc. Interspeech. pp. 3259–3263.
Menne, T., Schlüter, R., Ney, H., 2019. Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 6660–6664.
Moon, S., Hwang, J.N., 1993. Coordinated training of noise removing networks. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process, vol. 1, pp. 573–576.
Moore, A.H., Parada, P.P., Naylor, P.A., 2017. Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures. Comput. Speech Lang. 46, 574–584.
Narayanan, A., Walker, J., Panchapagesan, S., Howard, N., Koizumi, Y., 2023. Learning mask scalars for improved robust automatic speech recognition. In: Proc. IEEE Spoken Language Technology Workshop. pp. 317–323.
Narayanan, A., Wang, D., 2014. Investigation of speech separation as a front-end for noise robust speech recognition. IEEE/ ACM Trans. Audio Speech Lang. Process. 22, 826–835.
Ochiai, T., Iwamoto, K., Delcroix, M., Ikeshita, R., Sato, H., Araki, S., Katagiri, S., 2024. Rethinking processing distortions: Disentangling the impact of speech enhancement errors on speech recognition performance. arXiv:2404.14860.
Paliwal, K., Wójcicki, K., Shannon, B., 2011. The importance of phase in speech enhancement. Speech Commun. 53, 465–494.
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.
Panchapagesan, S., Narayanan, A., Shabestary, T.Z., Shao, S., Howard, N., Park, A., Walker, J., Gruenstein, A., 2022. A conformer-based waveform-domain neural acoustic echo canceller optimized for ASR accuracy. In: Proc. Interspeech. pp. 2583–2542.
Pandey, A., Wang, D., 2019. A new framework for CNN-based speech enhancement in the time domain. IEEE/ ACM Trans. Audio Speech Lang. Process. 27 (7), 1179–1188.
Pandey, A., Wang, D., 2021. Dense CNN with self-attention for time-domain speech enhancement. IEEE/ ACM Trans. Audio Speech Lang. Process. 29, 1270–1279.
Pandey, A., Wang, D., 2022. Self-attending RNN for speech enhancement to improve cross-corpus generalization. IEEE/ ACM Trans. Audio Speech Lang. Process. 30, 1374–1385.
Pandey, A., Xu, B., Kumar, A., et al., 2022. TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 6497–6501.
Pascual, S., Bonafonte, A., Serrà, J., 2017. SEGAN: Speech enhancement generative adversarial network. In: Proc. Interspeech. pp. 3642–3646.
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.
Plantinga, P., Bagchi, D., Fosler Lussier, E., 2021. Perceptual loss with recognition model for single-channel enhancement and robust ASR. arXiv:2112.06068.
Quan, C., Li, X., 2024. SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation. IEEE/ ACM Trans. Audio Speech Lang. Process. 32, 1310–1323.
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, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Commun. 43, 275–296.
Reddy, C.K., Gopal, V., Cutler, R., Beyrami, E., Cheng, R., Dubey, H., Matusevych, S., Aichner, R., Aazami, A., Braun, S., et al., 2020. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In: Proc. Interspeech. pp. 2492–2496.

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. 2, pp. 749–752.
Sato, H., Ochiai, T., Delcroix, M., Kinoshita, K., Kamo, N., Moriya, T., 2022. Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 6287–6291.
Sato, H., Ochiai, T., Delcroix, M., Kinoshita, K., Moriya, T., Kamo, N., 2021. Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition. In: Proc. Interspeech. pp. 1149–1153.
Scheibler, R., Bezzam, E., Dokmanić, I., 2018. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 351–355.
Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 7398–7402.
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.
Sorensen, H.B., 1991. A cepstral noise reduction multi-layer neural network. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 933–936.
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.
Tamura, S., 1989. An analysis of a noise reduction neural network. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 2001–2004.
Tan, K., Chen, J., Wang, D., 2018. Gated residual networks with dilated convolutions for monaural speech enhancement. IEEE/ ACM Trans. Audio Speech Lang. Process. 27, 189–198.
Tan, K., Wang, D., 2019. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ ACM Trans. Audio Speech Lang. Process. 28, 380–390.
Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J., 2016. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In: Proc. SSW. pp. 146–152.
Varga, A., Steeneken, H.J., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12, 247–251.
Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In: Proc. Adv. Neural Inf. Process Syst. pp. 5998–6008.
Vincent, E., Barker, J., Watanabe, S., et al., 2013. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 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., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ ACM Trans. Audio Speech Lang. Process. 26, 1702–1726.
Wang, H., Qian, Y., Wang, X., et al., 2022. Improving noise robustness of contrastive speech representation learning with speech reconstruction. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 6062–6066.
Wang, P., Tan, K., Wang, D.L., 2020a. 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, Y., Wang, D., 2013. Towards scaling up classification-based speech separation. IEEE/ ACM Trans. Audio Speech Lang. Process. 21, 1381–1390.
Wang, Z.Q., Wang, D., 2016. A joint training framework for robust automatic speech recognition. IEEE/ ACM Trans. Audio Speech Lang. Process. 24, 796–806.
Wang, P., Wang, D.L., 2019. Enhanced spectral features for distortion-independent acoustic modeling. In: Proc. Interspeech. pp. 476–480.
Wang, Z.-Q., Wang, P., Wang, D., 2020b. Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. IEEE/ ACM Trans. Audio Speech Lang. Process. 28, 1778–1787.
Watanabe, S., Hori, T., Karita, S., et al., 2018. ESPnet: End-to-end speech processing toolkit. In: Proc. Interspeech. pp. 2207–2211.
Williamson, D.S., Wang, Y., Wang, D., 2016. Complex ratio masking for monaural speech separation. IEEE/ ACM Trans. Audio Speech Lang. Process. 24, 483–492. http://dx.doi.org/10.1109/TASLP.2015.2512042.
Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2014. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21, 65–68.
Xu, Y., Weng, C., Hui, L., et al., 2019. Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 6745–6749.
Yang, Y., Pandey, A., Wang, D., 2023a. Time-domain speech enhancement for robust automatic speech recognition. In: Proc. Interspeech. pp. 4913–4917.
Yang, D., Wang, W., Qian, Y., 2023b. FAT-HuBERT: Front-end adaptive training of hidden-unit BERT for distortion-invariant robust speech recognition. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. pp. 1–8.
Yang, Y., Wang, P., Wang, D., 2022. A conformer based acoustic model for robust automatic speech recognition. arXiv:2203.00725.
Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. arXiv:1605.07146.
Zhang, W., Shi, J., Li, C., Watanabe, S., Qian, Y., 2021. Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions. In: Proc. IEEE Workshop on Appl. of Signal process To Audio and Acoustics. pp. 146–150.
Zhu, Q.S., Zhang, J., Zhang, Z.Q., Dai, L.R., 2023a. Joint training of speech enhancement and self-supervised model for noise-robust ASR. IEEE/ ACM Trans. Audio Speech Lang. Process. 31, 1927–1939.
Zhu, Q.S., Zhou, L., Zhang, J., et al., 2023b. Robust data2vec: Noise-robust speech representation learning for ASR by combining regression and improved contrastive learning. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 1–5.
Zorilă, C., Doddipatla, R., 2022. Speaker reinforcement using target source extraction for robust automatic speech recognition. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 6297–6301.