FIRING-NET: A FILTERED FEATURE RECYCLING NETWORK FOR SPEECH ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Current deep neural networks for speech enhancement (SE) aim to minimize the distance between the output signal and the clean target by filtering out noise features from input features. However, when noise and speech components are highly similar, SE models struggle to learn effective discrimination patterns. To address this challenge, we propose a Filter-Recycle-Interguide framework termed **FI**lter-**R**ecycle-**IN**ter**G**uide **NET**work (FIRING-Net) for SE, which filters the input features to extract target features and recycles the filtered-out features as non-target features. These two feature sets then guide each other to refine the features, leading to the aggregation of speech information within the target features and noise information within the non-target features. The proposed FIRING-Net mainly consists of a Local Module (LM) and a Global Module (GM). The LM uses outputs of the speech extraction network as target features and the residual between input and output as non-target features. The GM leverages the energy distribution of the self-attention map to extract target and non-target features guided by the highest and lowest energy regions. Both LM and GM include interaction modules to leverage the two feature sets in an inter-guided manner for collecting speech from non-target features and filtering out noise from target features. Experiments confirm the effectiveness of the Filter-Recycle-Interguide framework, with FIRING-Net achieving a strong balance between SE performance and computational efficiency, surpassing comparable models across various SNR levels and noise environments.

1 INTRODUCTION

Speech enhancement (SE) aims to separate speech from background interference signals (Xu et al., 2013). It is a fundamental speech processing problem and has been frequently used as the preprocessor of several acoustic tasks, such as speech recognition (Peng et al., 2022), speaker verification (Rao et al., 2019), speaker diarization (Sell et al., 2018), etc. Traditional SE approaches, such as Wiener filtering (Chen et al., 2006), spectral subtraction (Vaseghi, 2008), and principle component analysis (Srinivasarao & Ghanekar, 2020), assume that the noises belong to stationary signals, which are significantly different from the speech signal. However, this assumption usually cannot be satisfied in practice and thus these approaches often fail in real-world applications.

Recently, deep networks have shown their promising performance on SE, even under highly non-stationary noise environments. Deep learning-based SE methods train SE networks on extensive noisy-clean pairs to learn to distinguish between speech and noise, aiming to effectively remove noise components (Sell et al., 2018; Xu et al., 2013). For example, we can utilize the convolutional operations to extract local speech features (Park & Lee, 2017; Pandey & Wang, 2019), the recurrent neural networks or transformers to capture global speech features (Sun et al., 2017; Kim et al., 2020), and integrating both techniques can enhance internal feature differences within speech (Abdulatif et al., 2024; Lu et al., 2023; Xu et al., 2023). Despite significant improvements, issues of speech distortion and residual noise persist (Jo & Yoo, 2010; Xia et al., 2020; Wakabayashi et al., 2018). That is, they remain ineffective against noise containing highly similar acoustic events, such as speech from non-target speakers or echoes reflected from walls (Hu et al., 2023; Zheng et al., 2021). The primary reason is the often indistinct differences between speech and noise components. As shown in Figure 1(a), SE networks struggle to learn discriminative patterns between highly similar speech and noise components, leading to the overestimation or underestimation of noise information.

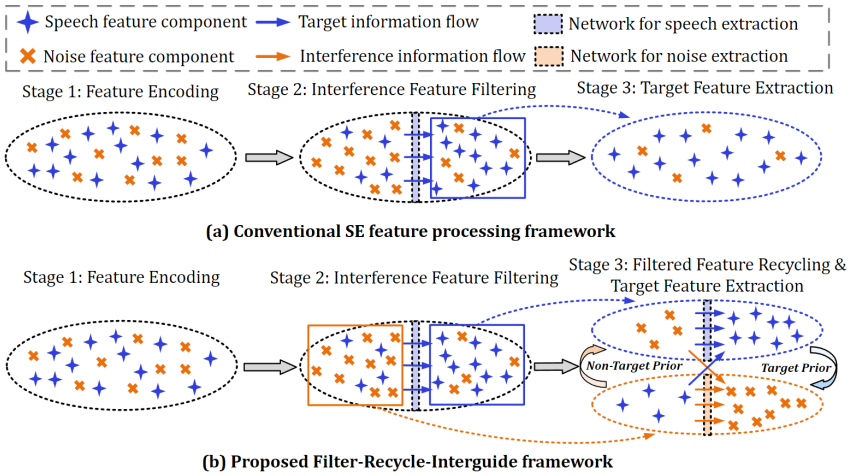


Figure 1: Comparison of (a) the conventional SE feature processing framework and (b) the proposed Filter-Recycle-Interguide framework. (a) In the conventional SE processing framework, the network module aims to filter input features with the goal of removing noise components. However, when speech and noise components are highly similar, the module may either overestimate or underestimate the noise components in the input features. (b) In the proposed Filter-Recycle-Interguide framework, the filtered features are re-collected and interactively processed with the output features. Since the output features primarily contain speech information and the filtered features mainly include noise information, this framework enables mutual guidance between these two feature types.

Some recent solutions use noise as part of the supervision to enable SE networks to establish a dual mapping from mixed signals to both speech and noise signals. A representative work is the dual-branch network structure, where the two branches aim to model the speech and noise signal, respectively, and an interacting operation is designed to explore the correlation between speech and noise features in a noisy mixture (Zheng et al., 2021; Zhao et al., 2022; Li et al., 2024). This method interactively leverages the features of speech and noise signals and using both as supervision. The goal is to ensure the relative independence of speech and noise signals in the feature space, and thereby enabling the network to extract discriminative patterns for each signal. However, due to the dynamic and highly random nature of noise, completely predicting a noise signal is infeasible (Ortega-García & González-Rodríguez, 1996). Consequently, its performance still heavily relies on the accuracy of the speech signal prediction branch, thereby diminishing the primary purpose of predicting noise signals.

To remedy this drawback, we design a Filter-Recycle-Interguide (FRI) strategy to extract more discriminative patterns between speech and noise within the mixture input. This is achieved by: (1) avoiding overemphasis on exploring feature differences within speech and (2) achieving more robust noise feature extraction for interaction with speech features without directly predicting the noise signal. As shown in Figure 1(b), our method consists of three main steps: (1) Processing input features through network modules to extract target features, similar to conventional SE models; (2) Re-collecting and redefining filtered-out features as non-target features; (3) Guiding and interacting target and non-target features, where target features, mainly composed of speech components, integrate speech from non-target features and remove noise, while non-target features, predominantly containing noise, assist in refining the target features.

By incorporating the designed FRI strategy into deep neural networks, we obtain the proposed **FI**lter-**R**ecycle-**IN**ter**G**uide **NET**work (FIRING-Net), which contains two crucial components: Local Module (LM) and Global Module (GM). The LM extracts target and non-target features by utilizing output of the convolutional module and computing the residual between the input and output. The GM derives target and non-target features based on the energy distribution from the self-attention map, employing high and low energy regions for guidance. Both LM and GM incorporate interaction modules to integrate the features in an inter-guided manner, effectively extracting speech from non-target features and filtering out noise from target features.

Our model is trained on DNS Challenge 2021 (Reddy et al., 2021) and is evaluated on two public datasets (WSJ0-SI 84 + NOISX-92 (Paul & Baker, 1992; Varga & Steeneken, 1993) and AVSpeech + AudioSet (Ephrat et al., 2018; Gemmeke et al., 2017)). The extensive results demonstrate the efficacy of the proposed method. Specifically, under the babble noise environment, it achieves a 3.03% relative improvement in terms of PESQ and a 6% dB gain in terms of SI-SNRi compared to the most competitive MP-SENet (Lu et al., 2023).

To summarize, the main contributions of this paper are:

- We design a Filter-Recycle-Interguide SE strategy, which enables mutual refinement between retained and filtered-out features, i.e., effectively eliminating noise from the retained features while recovering speech information from the filtered-out features.
- We propose FIRING-Net for SE, comprising two key components: LM and GM. The LM extracts target and non-target features by computing the residual between the input and output. The GM identifies target and non-target features according to the energy distribution within the self-attention map.
- We conduct extensive experiments on multiple popular datasets across various SNR levels and diverse noise conditions.

2 Related Works

The integration of deep learning has significantly enhanced SE performance. Recently, convolutional neural networks (CNNs) have been employed in SE tasks with notable success, as they effectively capture implicit information within the speech signal and manage small shifts in the time-frequency (TF) domain of speech features, thus adapting to varying speaker identities and acoustic environments (Park & Lee, 2017; Fu et al., 2017). However, CNNs are somewhat limited in modeling broader dependencies in low-level features (O'shea & Nash, 2015). To address this issue, many approaches have turned to transformers (Vaswani et al., 2017) to replace CNNs, enabling the capture of global information in waveforms or spectrograms (Kim et al., 2020; Dang et al., 2022a). Furthermore, several studies have sought to integrate CNNs and transformers, thereby harnessing both local and global information (Abdulatif et al., 2024; Lu et al., 2023). However, deep learning-based SE methods primarily focus on extracting speech features based on speech characteristics alone, leading to speech distortion or residual noise in the enhanced output when speech and noise components in the deep latent space are highly similar.

Rather than solely focusing on estimating target speech, some SE methods improve the performance by building noise models to account for noise prior. For instance, certain approaches (Odelowo & Anderson, 2017; Liu et al., 2021) implement spectral subtraction using deep neural networks (DNNs) through a two-stage process: noise signal estimation followed by speech signal recovery. Although outperform traditional signal processing techniques in modeling structured noise (Ortega-García & González-Rodríguez, 1996), DNNs' generalization capability remains limited. To enhance noise modeling accuracy, advanced methods (Zheng et al., 2021; Zhao et al., 2022; Li et al., 2024) introduce a two-branch framework to predict both speech and noise simultaneously, incorporating interaction modules at various layers based on the correlations between predicted speech and the residual signal. However, this approach still struggles with the unpredictability of noise, as its success largely depends on the accuracy of the speech prediction branch.

3 Method

Figure 2 is an overall illustration of our method. The noisy signal $\mathbf{y} \in \mathbb{R}^{1\times L}$ undergoes a short-time Fourier transform (STFT) to produce a complex spectrogram $\mathbf{Y} \in \mathbb{R}^{2\times T\times F}$. We extract the magnitude spectrogram $\mathbf{Y}_M \in \mathbb{R}^{1\times T\times F}$ and the wrapped phase spectrogram $\mathbf{Y}_P \in \mathbb{R}^{1\times T\times F}$, where T and F denote time and frequency. Before processing with FIRING-Net, a power-law compression (Wisdom et al., 2019) is applied to $\mathbf{Y}_M$ using a compression factor of 0.3 (Braun & Tashev, 2021), to enhance alignment with human auditory perception (Lee et al., 2018). The resulting compressed magnitude spectrogram $\mathbf{Y}_M^c$ is then concatenated with $\mathbf{Y}_P$ and used as the input for FIRING-Net.

Encoder. Given the input feature $\mathbf{Y}_{in} \in \mathbb{R}^{2\times T\times F}$, the encoder includes two convolution blocks, together with three densely connected Local Modules for encoder (LM-Es) positioned between the two

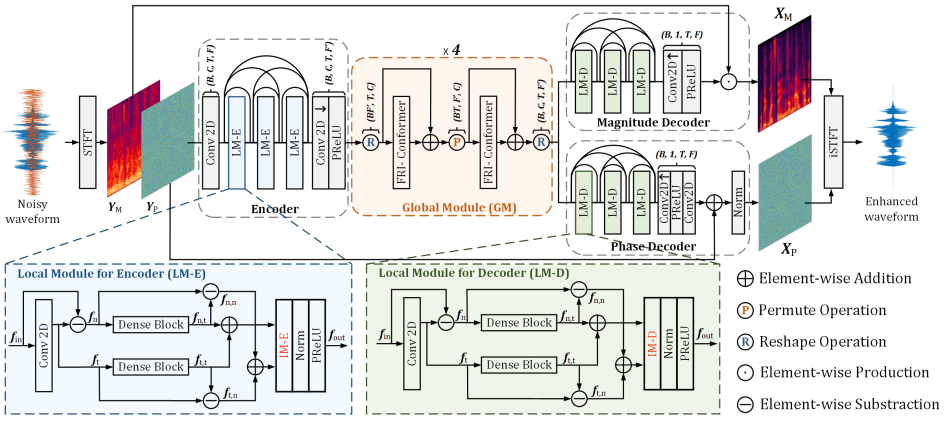


Figure 2: Overview of our FIRING-Net, which is an encoder-decoder architecture that leverages the Filter-Recycle-Interguide strategy for SE by dividing input features into target (speech) and non-target (noise) components. The encoder, through LMs, filters the input to isolate target features while recycling the filtered-out elements as non-target information. These two sets of features guide each other via interaction modules to refine speech extraction and noise suppression. GMs, utilizing multi-head self-attention, further optimize this process by capturing global time-frequency dependencies. Finally, the decoder employs separate magnitude and phase decoders to reconstruct the enhanced speech, applying masks for amplitude correction and residuals for phase refinement.

convolutional blocks. The first convolution block maps $\mathbf{Y}_{in}$ to an intermediate space with C channels. Three LM-Es aim to extract more completed speech information and accurately suppress noise by using the Filter-Recycle-Interguide framework (see Sec 3.1), while dense connections combine features from all layers to capture different levels of detail. The final convolution block decreases the frequency dimension to F' to reduce complexity. Subsequently, the encoder output is then processed by N Global Modules (GMs), with N set to 4, as described in Sec 3.2. The architecture, inspired by TF-Conformer (Abdulatif et al., 2024; Lu et al., 2023), alternately captures time and frequency dependencies within the Filter-Recycle-Interguide framework.

Magnitude and Phase Decoders. The decoder processes the output from the N GMs in a decoupled manner, involving two branches: the magnitude decoder and the phase decoder. The former is designed to predict a mask, which is then element-wise multiplied by the input magnitude spectrogram to enhance the signal. In contrast, the phase decoder predicts a residual that refines the noisy phase to improve phase accuracy (Afouras et al., 2018). Both decoders contain three densely connected LMs for decoder (LM-Ds). In both branches, a deconvolutional block is employed to upsample the frequency dimension back to F, while reducing the channel number to 1. For the magnitude decoder, a PReLU activation function is used to predict the final mask, allowing it to learn different slopes for each frequency band (Abdulatif et al., 2024). In the phase decoder, the predicted phase residual is added to the noisy phase and then L2-normalized to produce a clean phase prediction (Afouras et al., 2018).

3.1 Local Module

The core technique of LM is mainly based on back-projection that was initially introduced in the context of image super-resolution (Haris et al., 2018; 2019; Liu et al., 2019), where it iteratively employs feedback residuals to enhance high-resolution images. In our work, we adapt and extend the back-projection technique to address SE problems, leading to the formulation of our building blocks. Specifically, we treat the residual obtained through back projection as non-target features, representing the information filtered out by the network. We extract speech components from this residual to address speech distortion. Conversely, the features retained by the network represent target features, where we focus on eliminating noise components to mitigate noise residual issues.

We replace the conventional convolutional modules in the Encoder and Magnitude and Phase Decoders with Local Module for Encoder (LM-E) and Local Module for Decoder (LM-D). As shown

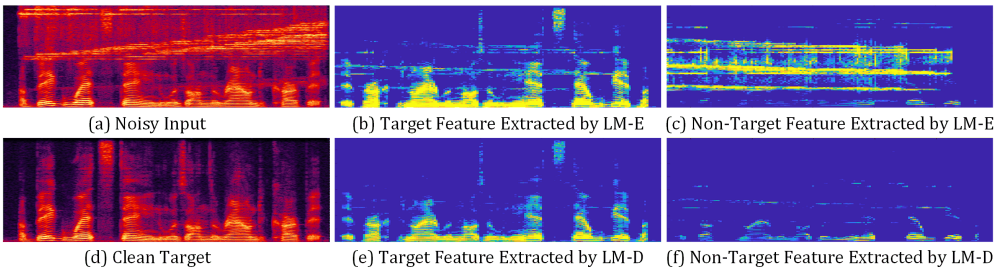


Figure 3: Visualizations of target and non-target features extracted by LM-E and LM-D, along with their corresponding spectrograms of noisy mixtures and clean target speech.

in Figure 2, LM-E and LM-D have identical structures, with the only difference being that LM-E contains an Interaction Module specifically designed for the Encoder part (IM-E), while LM-D includes an Interaction Module tailored for the Decoder part (IM-D). The LM refines a target feature $\mathbf{f}_t$ processed from the input feature $\mathbf{f}_{in}$ by applying a convolutional layer and filter out the non-target feature $\mathbf{f}_n$,

$$\mathbf{f}_t = \mathcal{H}_{conv}(\mathbf{f}_{in}), \qquad \mathbf{f}_n = \mathbf{f}_{in} - \mathbf{f}_t, \tag{1}$$

where $\mathcal{H}_{conv}(\cdot)$ denotes the convolution operation.

A simple convolutional layer, due to its limited capacity, only provides a coarse separation of target and non-target features within the $\mathbf{f}_{in}$. Consequently, we further refine both the target and non-target features separately. For the non-target features, we use a dilated DenseNet (Pandey & Wang, 2020) to further extract the ignored target features $\mathbf{f}_{n,t}$ and calculate the residual between $\mathbf{f}_{n,t}$ and $\mathbf{f}_n$ to obtain more purified non-target feature $\mathbf{f}_{n,n}$. Similarly, we apply the same process to the target features, resulting in more purified target feature $\mathbf{f}_{t,t}$ as well as contributing to the refinement of non-target feature $\mathbf{f}_{t,n}$,

$$\mathbf{f}_{n,t} = \mathcal{H}_{dense}(\mathbf{f}_n), \qquad \mathbf{f}_{n,n} = \mathbf{f}_n - \mathbf{f}_{n,t}, \tag{2}$$

$$\mathbf{f}_{t,t} = \mathcal{H}_{dense}(\mathbf{f}_t), \qquad \mathbf{f}_{t,n} = \mathbf{f}_t - \mathbf{f}_{t,t}, \tag{3}$$

where $\mathcal{H}_{dense}(\cdot)$ denotes the dilated DenseNet that contains three convolution blocks with dense connections, the dilation factors of each block are set to $\{1, 2, 4\}$. We aggregate all target and non-target features through an addition operation to obtain the refined target features $\hat{\mathbf{f}}_t = \mathbf{f}_{t,t} + \mathbf{f}_{n,t}$ and non-target feature $\hat{\mathbf{f}}_n = \mathbf{f}_{t,n} + \mathbf{f}_{n,n}$, respectively.

The purpose of the Interaction Module is to enable mutual guidance between target and non-target features, allowing for a more refined extraction of speech information from non-target features to supplement the target features, while filtering out noise information from the target features. However, since the information contained in non-target features differs between the Encoder and Decoder, we designed IM-E and IM-D to specifically handle the target and non-target features in LM-E and LM-D, respectively.

LM-E: In the Encoder, input features typically contain more complete noise components. As shown in Figure 3(c), the non-target features extracted by LM-E are predominantly noise. The role of IM-E is to extract speech components from these non-target features and filter out noise from the target features. As depicted in Figure 4(a), IM-E concatenates $\hat{\mathbf{f}}_t$ and $\hat{\mathbf{f}}_n$ and feeds them into a convolutional module to generate a mask $\mathbf{M}_n$. This mask identifies the preserved areas of non-target features to extract speech information, which is then added to $\hat{\mathbf{f}}_t$ to produce a more speech information completed target feature $\mathbf{f}'_t$. Additionally, $|\mathit{Sigmoid}(\hat{\mathbf{f}}_n) - \mathbf{M}_n|$ isolates purer noise information from the non-target features (The reason for using "$|\mathit{Sigmoid}(\hat{\mathbf{f}}_n) - \mathbf{M}_n|$" instead of "$1 - \mathbf{M}_n$" is explained in the Appendix B.3), which is combined with $\mathbf{f}'_t$ to guide noise extraction from $\mathbf{f}'_t$, before feeding into a convolutional module to generate a mask $\mathbf{M}_t$ that is subtracted by "$\mathit{Sigmoid}(\hat{\mathbf{f}}_t)$", and the absolute value is taken afterward, i.e., $|\mathit{Sigmoid}(\hat{\mathbf{f}}_t) - \mathbf{M}_t|$, to filter out remaining noise.

LM-D: In the Decoder, input features typically contain minimal or no noise components. As illustrated in Figure 3(f), non-target information often contains more speech components that are difficult to use for target speech reconstruction. Therefore, IM-D ensures mutual guidance between

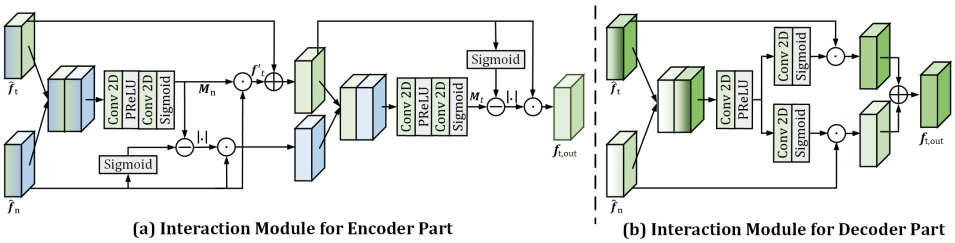


Figure 4: Network architecture of the Interaction Module (IM): (a) The IM in the encoder extracts speech from non-target features guided by target features and removes noise from target features guided by non-target features. (b) The IM in the decoder, where noise is largely suppressed, refines discarded speech information from non-target features using target feature guidance for a more complete speech reconstruction.

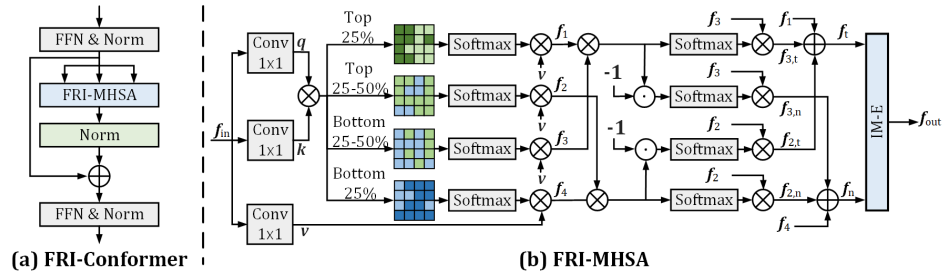


Figure 5: Network architecture of the Filter-Recycle-Interguide Conformer (FRI-Conformer) in the global module, featuring the crucial Filter-Recycle-Interguide Multi-Head Self-Attention (FRI-MHSA). FRI-MHSA classifies input features into four parts based on attention map energy distribution, using the highest and lowest energy regions to guide speech extraction from mid-low energy features and noise extraction from mid-high energy features.

target and non-target features, allowing speech information from non-target features to contribute effectively to the reconstruction of the target speech, thus reducing speech distortion. As shown in Figure 4(b), IM-D concatenates $\hat{\mathbf{f}}_t$ and $\hat{\mathbf{f}}_n$ and processes them through a convolutional module to generate an intermediate feature. This intermediate feature is then processed by two separate convolutional modules to produce masks for $\hat{\mathbf{f}}_t$ and $\hat{\mathbf{f}}_n$. The masked $\hat{\mathbf{f}}_t$ and $\hat{\mathbf{f}}_n$ are then combined to produce the output target feature with enhanced speech information.

3.2 Global Module

The GM is designed to overcome the limitations of the LM, which can only extract target and non-target features within a limited receptive field. By integrating the long-range contextual modeling capabilities of Transformers (Vaswani et al., 2017), the Global Module facilitates the extraction of these features from a global perspective of the speech signal. As depicted in Figure 2, the GM adopts a dual-path attention-based structure (Abdulatif et al., 2024; Wang et al., 2021) and employs two FRI-Conformer blocks in sequence. The first stage captures time dependencies with an input shape of $BF' \times T \times C$, where B represents the batch size, while the second stage captures frequency dependencies with an input shape of $BT \times F' \times C$. As depicted in Figure 5(a), each FRI-Conformer employs two half-step feedforward networks (FFNs) with a Filter-Recycle-Interguide Multi-head Self-attention (FRI-MHSA) module in between.

The design of FRI-MHSA is based on the energy distribution in the attention feature map, where high-energy regions guide the extraction of speech features in low-energy regions, and low-energy regions guide the extraction of noise features in high-energy regions. Given a query $\mathbf{q}$, key $\mathbf{k}$, and value $\mathbf{v}$, the output of dot-product attention is generally formulated as:

$$\text{Att}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}(\mathbf{q}\mathbf{k}^\top)\mathbf{v}. \tag{4}$$

In our work, FRI-MHSA segments the feature map obtained from $\mathbf{qk}^\top$ based on energy levels, extracting the top 25% highest energy regions, the upper-middle 25%, the lower-middle 25%, and the bottom 25% lowest energy regions. We draw inspiration from the top-k operation (Chen et al., 2023b; Xiao et al., 2024) to implement this extraction and assign an infinitesimal value to the unextracted portions of each feature map. Subsequently, we apply softmax to each of the four feature maps, generating four masks that extract four types of features from $\mathbf{v}$. The mask generated from the top 25% energy feature map extracts target features $\mathbf{f}_1$ from $\mathbf{v}$, primarily containing speech information, which guides the extraction of speech information from the lower-middle 25% energy feature map $\mathbf{f}_3$. Conversely, the mask corresponding to the bottom 25% energy feature map extracts non-target features $\mathbf{f}_4$ from $\mathbf{v}$, primarily containing noise information, which in turn guides the extraction of noise information from the upper-middle 25% energy feature map $\mathbf{f}_2$. This guided extraction method is based on calculating the cross-similarity between two types of features through matrix multiplication, allowing the extraction of the parts from the guided features that are most similar to the guiding features. Therefore, the above process can be expressed as:

$$\mathbf{f}_{3,t} = \text{softmax}(\mathbf{f}_1\mathbf{f}_3^\top)\mathbf{f}_3, \qquad \mathbf{f}_{3,n} = \text{softmax}(-1 * (\mathbf{f}_1\mathbf{f}_3^\top))\mathbf{f}_3, \tag{5}$$

$$\mathbf{f}_{2,n} = \text{softmax}(\mathbf{f}_4\mathbf{f}_2^\top)\mathbf{f}_2, \qquad \mathbf{f}_{2,t} = \text{softmax}(-1 * (\mathbf{f}_4\mathbf{f}_2^\top))\mathbf{f}_2, \tag{6}$$

where $\mathbf{f}_{2,t}$ and $\mathbf{f}_{2,n}$ represent the target and non-target features extracted from $\mathbf{f}_2$, and $\mathbf{f}_{3,t}$ and $\mathbf{f}_{3,n}$ represent the target and non-target features extracted from $\mathbf{f}_3$. Finally, the target feature is obtained by $\mathbf{f}_t = \mathbf{f}_1 + \mathbf{f}_{2,t} + \mathbf{f}_{3,t}$, while the non-target feature is obtained by $\mathbf{f}_n = \mathbf{f}_4 + \mathbf{f}_{2,n} + \mathbf{f}_{n,t}$. These features, $\mathbf{f}_t$ and $\mathbf{f}_n$ are fed into IM-E for further processing.

4 Experimental Setup

4.1 Dataset

We trained FIRING-Net using the Interspeech 2021 DNS Challenge dataset (Reddy et al., 2021), sampled at 16 kHz. A total of 60,000 reverberant speech clips, approximately 500 hours in duration, were generated, with 55,000 clips designated for training and 5,000 for validation. The noise clips were mainly sourced from Audioset (Gemmeke et al., 2017), DEMAND (Thiemann et al., 2013a), and Freesound (Fonseca et al., 2017). During training, the audio was randomly segmented into 4-second clips and processed with randomly selected room impulse responses (RIRs) from OpenSLR26 and OpenSLR28 (Ko et al., 2017) (T_{60} in the range from 0.3s to 1.3s). The noisy speech was created by mixing reverberant speech with noise, with the SNR range set between -5 dB and 5 dB. We selected two datasets as the test sets for performance evaluation under various unknown noise conditions:

- **WSJ0-SI 84 + NOISEX-92:** We selected 651 utterances from 8 speakers in the WSJ0-SI 84 dataset (Paul & Baker, 1992). Noise samples were taken from the NOISEX-92 dataset (Varga & Steeneken, 1993), and test mixtures were created by combining these noise samples with the speech at SNR levels of -5 dB, 0 dB, and 5 dB in reverberant conditions, with T_{60} values randomly selected between 0.3s and 1.3s.
- **AVSpeech + AudioSet:** The AVSpeech dataset consists of public instructional YouTube videos, from which 3-10s clips were automatically extracted, ensuring that the only audible sound in each clip is from a single speaker (Ephrat et al., 2018). For our experiments, we downloaded 1,199 clips from the test set, utilizing only the audio portions. Four representative noise types: babble, engine, baby cry and laughter, and traffic, from the Audioset dataset (Gemmeke et al., 2017) were introduced, and noisy speech is generated through a weighted linear combination of clean utterances from AVSpeech and noise segments from AudioSet,
$$x_i = s_j^{AVSpeech} + 0.3 \cdot v_k^{AudioSet}, \tag{7}$$
where $s_j^{AVSpeech}$ and $v_k^{AudioSet}$ represent 4-second randomly sampled speech and noise segments, respectively.

4.2 Model Settings and Evaluation Metrics

We trained the proposed model for 120 epochs with a batch size of 2, utilizing the Adam optimizer with an initial learning rate of 0.001. If the best model was not identified for 15 consecutive epochs,

Table 1: Comparison with selected baseline models on WSJ0-SI 84+NOISEX-92 with different SNR levels. Bold and underline indicate the best and second-best results.

Method	Param.	-5 dB			0 dB			5 dB		
		STOI	PESQ	SI-SNRi	STOI	PESQ	SI-SNRi	STOI	PESQ	SI-SNRi
Unprocessed	-	0.6172	1.48	-	0.7813	1.76	-	0.8669	1.92	-
PHASEN	6.41M	0.7886	2.25	12.84	0.8981	2.85	11.82	0.9238	3.09	9.45
SN-Net	8.14M	0.8157	2.31	13.18	0.9031	2.93	12.10	0.9305	3.14	10.05
Inter-SubNet	2.29M	0.8038	2.29	13.41	0.8968	2.89	11.95	0.9287	3.13	10.78
CMGAN	1.83M	0.8205	2.38	14.29	0.9085	2.97	12.53	0.9324	3.17	11.65
MP-SENet	2.05M	0.8342	2.43	15.68	0.9112	3.03	13.69	0.9384	3.21	11.74
FIRING-Net	**1.74M**	**0.8414**	**2.51**	**16.11**	**0.9195**	**3.09**	**14.20**	**0.9430**	**3.28**	**12.06**

Table 2: Comparison with selected baseline models on AVSpeech + AudioSet with different types of noise. Bold and underline indicate the best and second-best results.

Method	Babble			Engine			Baby Cry and Laughter			Traffic		
	STOI	PESQ	SI-SNRi	STOI	PESQ	SI-SNRi	STOI	PESQ	SI-SNRi	STOI	PESQ	SI-SNRi
Unprocess	0.7971	1.56	-	0.7992	1.62	-	0.7876	1.75	-	0.8032	1.88	-
PHASEN	0.8956	2.85	11.04	0.9142	2.94	11.46	0.8974	2.71	11.37	0.9076	2.90	11.69
SN-Net	0.9214	2.94	12.11	0.9362	3.02	11.98	0.9198	2.94	11.86	0.9298	3.00	12.06
Inter-SubNet	0.9226	2.92	11.96	0.9301	2.99	11.83	0.9213	2.99	12.01	0.9185	2.96	12.10
CMGAN	0.9354	3.03	12.36	0.9472	3.08	12.57	0.9387	3.08	12.34	0.9323	3.07	12.39
MP-SENet	0.9437	3.07	13.24	0.9521	3.13	13.32	0.9453	3.10	12.96	0.9453	3.14	13.50
FIRING-Net	**0.9558**	**3.17**	**13.85**	**0.9643**	**3.21**	**14.12**	**0.9582**	**3.18**	**14.32**	**0.9541**	**3.23**	**14.16**

the learning rate was halved. Early stopping was employed, terminating the training if the best model was not found after 30 consecutive epochs. The STFT was applied using a Hanning window with a 32 ms window length and a 16 ms frame shift to convert the signal into the frequency domain. The number of channels $\mathbf{C}$ for all convolutional layers was set to 32. In the FRI-MHSA module, we used 8 attention heads. To mitigate overfitting, the dropout probability for all layers was set to 0.1. Four loss functions are utilized for model training: (1) Magnitude Loss: The mean squared error (MSE) loss is computed between the target and enhanced magnitude spectra. (2) Phase Loss: Phase loss is determined using the anti-wrapping function (Ai & Ling, 2023a). (3) Complex Loss: The MSE loss is calculated between the real and imaginary parts of the target and enhanced complex spectra. (4) Time Loss: The L1 loss is applied between the target and enhanced waveforms. We use the following three commonly seen SE metrics for evaluation purposes. **PESQ:** Perceptual Evaluation of Speech Quality (Rix et al., 2001). **STOI:** Short-Time Objective Intelligibility (Taal et al., 2011). **SI-SNRi:** Scale-Invariant Signal-to-Noise Ratio improvement (Le Roux et al., 2019). For all the metrics, the higher the score, the better the performance.

More details of experiential settings can be found in Appendix A.

5 RESULTS

5.1 MODEL COMPARISON

We conducted extensive experiments to quantitatively compare the SE performance of our proposed FIRING-Net with some existing SE models. We selected some typical models that have shown good performance, including PHASEN (Yin et al., 2020), SN-Net (Zheng et al., 2021), Inter-SubNet (Chen et al., 2023a), CMGAN (Abdulatif et al., 2024), and MP-SENet (Lu et al., 2023). Note that these baseline models and the proposed FIRING-Net are trained on the same training set and evaluated on the same test set.

WSJ0-SI 84+NOISEX-92: This dataset provides a controlled environment combining speech with various noise types, making it ideal for evaluating the performance of SE models under standardized conditions. As detailed in Table 1, FIRING-Net outperforms baseline models, particularly in challenging scenarios such as -5 dB SNR. FIRING-Net achieved a PESQ score of 2.51, which is notably higher than that of the best-performing baseline models, including MP-SENet (2.43) and CMGAN (2.38). This indicates that FIRING-Net is more effective in maintaining speech clarity under severe noise conditions. Furthermore, FIRING-Net's STOI score of 0.8414 demonstrates its superior ability to preserve speech intelligibility compared to other models in noisy environments.

Table 3: Ablative analysis of LM-E and LM-D by measuring STOI, PESQ, SI-SNRi, and number of trainable parameters.

Encoder	Decoder	STOI	PESQ	SI-SNRi	Params.
LM-E	LM-D	0.9135	3.03	13.96	1.81M
DCCM	DCCM	0.8657	2.84	12.64	1.89M
LM-E	DCCM	**0.8854**	2.93	**12.89**	**1.87M**
DCCM	LM-D	0.8782	**2.94**	12.71	**1.87M**
FGCM	FGCM	0.8738	2.88	12.84	**1.78M**
LM-E	FGCM	0.8927	**2.95**	**13.27**	1.80M
FGCM	LM-D	**0.9003**	2.93	13.14	1.79M

Table 4: Ablation study on GM by measuring STOI, PESQ, SI-SNRi, number of trainable parameters, and real-time factor.

Method	STOI	PESQ	SI-SNRi	Params.	CPU RTF
GM (*FRI-MHSA with 4 parts*)	0.9135	3.03	13.96	1.81M	0.60s
-FRI-MHSA (1 Part)	0.8594	2.80	12.57	**1.74M**	**0.39s**
-FRI-MHSA (2 Parts)	0.8862	2.89	12.89	1.80M	0.41s
-FRI-MHSA (3 Parts)	0.8987	2.97	13.14	1.81M	0.52s
-FRI-MHSA (5 Parts)	0.9152	3.04	13.85	1.83M	0.72s
-FRI-MHSA (6 Parts)	**0.9201**	**3.06**	**13.99**	1.84M	0.87s
w/o Frequency Processing	**0.8254**	2.75	11.97	1.38M	0.37s
w/o Time Processing	0.8303	**2.77**	**12.14**	1.29M	0.39s
w/o GM	0.7631	2.64	10.52	**1.03M**	**0.24s**

These findings highlight the robustness and effectiveness of FIRING-Net across a range of noise levels, making it suitable for applications requiring high-quality SE in challenging settings.

AVSpeech+AudioSet: This dataset presents a more complex and varied set of challenges for speech enhancement models compared to the controlled WSJ0-SI 84 + NOISEX-92 environment. It includes a wide range of real-world noise types and speaker variations. As shown in Table 2, FIRING-Net consistently outperforms baseline models across all tested noise conditions, including babble, engine, baby cry and laughter, and traffic noise. The strong generalization capability of FIRING-Net on this dataset is particularly impressive, as it is essential for effectively managing the diverse noise conditions encountered in real-world scenarios, especially speech-like noises such as babble, baby cry, and baby laughter. For instance, FIRING-Net achieves a PESQ score of 3.16 and an STOI score of 0.9556 with babble noise, surpassing other models. This demonstrates FIRING-Net's effectiveness in enhancing speech in the presence of complex, non-stationary noises. Additionally, we further highlight the advantages and flexibility of FIRING-Net by evaluating its performance in environments with multiple types of background noise.

5.2 Ablation Study

We present ablative experiments to analyze the contribution of each component of our model. We ablated the design choices and measured the average increase on the WSJ0-SI 84 + NOISEX-92 dataset. The following can be summarized from the ablation results:

LM: To assess the effectiveness of the proposed LM, we introduced two alternative modules: the densely connected convolutional module (DCCM), consisting of 6 cascaded convolutional layers Pandey & Wang (2020), and the fully gated convolutional module (FGCM), with 3 ELU-activated gated convolutional blocks Tan & Wang (2019). Both modules were configured to have a similar number of trainable parameters for a fair comparison. The most significant performance drop occurred when LM-E and LM-D were removed. FGCM, with its gating mechanism, effectively emphasized target features and outperformed DCCM. Overall, substituting LM-E or LM-D with DCCM or FGCM in the encoder or decoder led to a notable decline in performance across key metrics like STOI, PESQ, and SI-SNRi.

GM: To validate the effectiveness of the GM, we conducted two sets of comparative experiments, as presented in Table 4. The first set examines the FRI-MHSA module, where the $\mathbf{qk}^\top$ feature

Table 5: Ablation study on FRI Strategy in terms of STOI, PESQ, and SI-SNRi.

Method	STOI	PESQ	SI-SNRi
GM	**0.9135**	**3.03**	**13.96**
-*Target Feature*	0.8677	2.86	12.88
-*Target and Original Features*	0.8961	2.95	13.16
-*Non-Target and Original Features*	0.8306	2.69	9.52
LM-E	**0.9135**	**3.03**	**13.96**
-*Target Feature*	0.8738	2.84	12.83
-*Target and Original Features*	0.9003	2.92	13.05
-*Non-Target and Original Features*	0.7972	2.61	9.17
LM-D	**0.9135**	**3.03**	**13.96**
-*Target Feature*	0.8832	2.90	13.08
-*Target and Original Features*	0.9033	2.97	**13.29**
-*Non-Target and Original Features*	0.8992	2.93	13.38

map is initially divided into four segments based on energy levels. We adjusted the number of segments and observed a significant performance drop when reduced to one segment (equivalent to standard MHSA). As the number of segments increases from 1 to 4, performance improves, demonstrating the advantage of subdividing the feature map and using high- and low-energy regions to guide intermediate feature extraction. However, performance plateaus beyond four segments, suggesting that further subdivision yields no additional benefit. Additionally, increasing the number of segments affects real-time processing efficiency. The second set of experiments assesses the GM's effectiveness in handling features across the time and frequency dimensions. Removing the processing of either dimension results in a significant performance decline.

FRI Framework: The core innovation of FIRING-Net lies in its FRI framework, where target and non-target features guide each other to extract highly similar information. This approach enhances the aggregation of speech information within target features and noise information within non-target features. Table 5 demonstrates the effectiveness of this FRI strategy for SE. In the evaluation of LM, we observe that LM-E using non-target and original features performs worse than LM-E using target and original features. Additionally, LM-E -*Target Feature*, which processes only target features, highlights the importance of target features. LM-E also surpasses LM-E -*Target + Original Features*, confirming the value of non-target features and the effectiveness of the FRI framework in leveraging both feature types. For LM-D, results differ slightly. The performance of LM-D -*Target + Original Features* and LM-D -*Non-Target + Original Features* is similar and both exceed LM-D -*Target Feature*. This suggests that non-target information in the decoder contains useful speech components not involved in reconstruction, improving performance when both feature types are used. The evaluation of GM mirrors LM-E, reinforcing the effectiveness of the FRI framework. Notably, GM -*Target Feature* significantly outperforms GM -*FRI-MHSA (1 Part)* as shown in Table 4, indicating the superior efficacy of target feature extraction.

6 Conclusion

In this work, we propose an FRI framework that separates input features into target and non-target sets. These two sets guide each other to refine information, leading to the clustering of speech information in the target features and noise in the non-target features. We introduce FIRING-Net, a speech enhancement network comprising two main components: LM and GM. Both modules integrate interaction mechanisms to enable mutual guidance between the features, where speech is extracted from the non-target features and noise is filtered out from the target features. We conducted extensive experiments to validate the effectiveness of our method. From the results, we mainly conclude that: 1) Non-target features filtered by the SE network still contain speech information, and recycling them with the FRI strategy significantly boosts performance; 2) Mutual guidance between target and non-target features is crucial for filtering noise from the target and recovering speech from the non-target features. Future research focuses on designing lightweight model and enhancing real-time performance to enable deployment across various devices.

REFERENCES

Sherif Abdulatif, Ruizhe Cao, and Bin Yang. Cmgan: Conformer-based metric-gan for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The Conversation: Deep Audio-Visual Speech Enhancement. In *Proc. Interspeech 2018*, pp. 3244–3248, 2018. doi: 10.21437/Interspeech.2018-1400.

Yang Ai and Zhen-Hua Ling. Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Yang Ai and Zhen-Hua Ling. Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Sebastian Braun and Ivan Tashev. A consolidated view of loss functions for supervised deep learning-based speech enhancement. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 72–76. IEEE, 2021.

Jingdong Chen, Jacob Benesty, Yiteng Huang, and Simon Doclo. New insights into the noise reduction wiener filter. *IEEE Transactions on audio, speech, and language processing*, 14(4):1218–1234, 2006.

Jun Chen, Wei Rao, Zilin Wang, Jiuxin Lin, Zhiyong Wu, Yannan Wang, Shidong Shang, and Helen Meng. Inter-subnet: Speech enhancement with subband interaction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5896–5905, 2023b.

Feng Dang, Hangting Chen, and Pengyuan Zhang. DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6857–6861. IEEE, 2022a.

Feng Dang, Hangting Chen, and Pengyuan Zhang. Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6857–6861. IEEE, 2022b.

Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real Time Speech Enhancement in the Waveform Domain. In *Proc. Interspeech 2020*, pp. 3291–3295, 2020. doi: 10.21437/Interspeech.2020-2409.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 37, 2018.

Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai. Raw waveform-based speech enhancement by fully convolutional networks. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 006–012. IEEE, 2017.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1664–1673, 2018.

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3897–3906, 2019.

Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.

Yuchen Hu, Chen Chen, Heqing Zou, Xionghu Zhong, and Eng Siong Chng. Unifying speech enhancement and separation with gradient modulation for end-to-end noise-robust speech separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Seokhwan Jo and Chang D Yoo. Psychoacoustically constrained and distortion minimized speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 18(8):2099–2110, 2010.

Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6649–6653. IEEE, 2020.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5220–5224. IEEE, 2017.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630. IEEE, 2019.

Jinkyu Lee, Jan Skoglund, Turaj Shabestary, and Hong-Goo Kang. Phase-sensitive joint learning algorithms for deep learning-based speech enhancement. *IEEE Signal Processing Letters*, 25(8): 1276–1280, 2018.

Yihao Li, Meng Sun, Xiongwei Zhang, et al. Scale-aware dual-branch complex convolutional recurrent network for monaural speech enhancement. *Computer Speech & Language*, 86:101618, 2024.

Hsin-Yi Lin, Huan-Hsin Tseng, Xugang Lu, and Yu Tsao. Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport. *Advances in Neural Information Processing Systems*, 34:19935–19946, 2021.

Wenzhe Liu, Andong Li, Yuxuan Ke, Chengshi Zheng, and Xiaodong Li. Know Your Enemy, Know Yourself: A Unified Two-Stage Framework for Speech Enhancement. In *Interspeech*, pp. 186–190, 2021.

Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, and Wan-Chi Siu. Hierarchical back projection network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra. In *Proc. INTERSPEECH 2023*, pp. 3834–3838, 2023. doi: 10.21437/Interspeech.2023-1441.

Jianfen Ma, Yi Hu, and Philipos C Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405, 2009.

Babafemi O Odelowo and David V Anderson. A noise prediction and time-domain subtraction approach to deep neural network based speech enhancement. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 372–377. IEEE, 2017.

Javier Ortega-García and Joaquín González-Rodríguez. Overview of speech enhancement techniques for automatic speaker recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pp. 929–932. IEEE, 1996.

Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Ashutosh Pandey and DeLiang Wang. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6875–6879. IEEE, 2019.

Ashutosh Pandey and DeLiang Wang. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6629–6633. IEEE, 2020.

Se Rim Park and Jin Won Lee. A fully convolutional neural network for speech enhancement. *Proc. Interspeech 2017*, pp. 1993–1997, 2017.

Douglas B Paul and Janet Baker. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning*, pp. 17627–17643. PMLR, 2022.

Wei Rao, Chenglin Xu, Eng Siong Chng, and Haizhou Li. Target speaker extraction for multi-talker speaker verification. *Proc. Interspeech 2019*, pp. 1273–1277, 2019.

Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. Icassp 2021 deep noise suppression challenge. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6623–6627. IEEE, 2021.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, et al. Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural DIHARD challenge. In *Interspeech*, pp. 2808–2812, 2018.

V Srinivasarao and Umesh Ghanekar. Speech enhancement-an enhanced principal component analysis (epca) filter approach. *Computers & Electrical Engineering*, 85:106657, 2020.

Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Multiple-target deep learning for LSTM-RNN based speech enhancement. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pp. 136–140. IEEE, 2017.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7):2125–2136, 2011.

Ke Tan and DeLiang Wang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:380–390, 2019.

Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Wenxuan Xie, and Wenjun Zeng. Joint time-frequency and time domain learning for speech enhancement. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3816–3822, 2021.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19. AIP Publishing, 2013a.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, pp. 035081. Acoustical Society of America, 2013b.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pp. 146–152, 2016.

Andrew Varga and Herman JM Steeneken. 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 communication*, 12(3):247–251, 1993.

Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA*, pp. 1–4. IEEE, 2013.

Yukoh Wakabayashi, Takahiro Fukumori, Masato Nakayama, Takanobu Nishiura, and Yoichi Yamashita. Single-channel speech enhancement with phase reconstruction based on phase distortion averaging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9): 1559–1569, 2018.

Kai Wang, Bengbeng He, and Wei-Ping Zhu. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7098–7102. IEEE, 2021.

Scott Wisdom, John R Hershey, Kevin Wilson, Jeremy Thorpe, Michael Chinen, Brian Patton, and Rif A Saurous. Differentiable consistency constraints for improved deep speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 900–904. IEEE, 2019.

Yangyang Xia, Sebastian Braun, Chandan KA Reddy, Harishchandra Dubey, Ross Cutler, and Ivan Tashev. Weighted speech distortion losses for neural-network-based real-time speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 871–875. IEEE, 2020.

Yi Xiao, Qiangqiang Yuan, Kui Jiang, Jiang He, Chia-Wen Lin, and Liangpei Zhang. TTST: A top-k token selective transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, 2024.

Xinmeng Xu, Weiping Tu, and Yuhong Yang. CASE-Net: Integrating local and non-local attention operations for speech enhancement. *Speech Communication*, 148:31–39, 2023.

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1):65–68, 2013.

Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9458–9465, 2020.

Dacheng Yin, Zhiyuan Zhao, Chuanxin Tang, Zhiwei Xiong, and Chong Luo. TridentSE: Guiding Speech Enhancement with 32 Global Tokens. In *Proc. INTERSPEECH 2023*, pp. 3839–3843, 2023. doi: 10.21437/Interspeech.2023-565.

Guochen Yu, Andong Li, Chengshi Zheng, Yinuo Guo, Yutian Wang, and Hui Wang. Dual-branch attention-in-attention transformer for single-channel speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7847–7851. IEEE, 2022.

Haoran Zhao, Nan Li, Runqiang Han, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu. A deep hierarchical fusion network for fullband acoustic echo cancellation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9112–9116. IEEE, 2022.

Chengyu Zheng, Xiulian Peng, Yuan Zhang, Sriram Srinivasan, and Yan Lu. Interactive speech and noise modeling for speech enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14549–14557, 2021.

A Supplementary Experimental Setups

A.1 Loss Function

We employ multi-level loss functions to train the proposed FIRING-Net. Following the approach in (Braun & Tashev, 2021; Lu et al., 2023; Abdulatif et al., 2024), we utilize time-domain loss ($\mathcal{L}_T$), magnitude loss ($\mathcal{L}_M$), and complex loss ($\mathcal{L}_C$), which are defined as:

$$\mathcal{L}_T = \mathbb{E}_{\mathbf{s},\hat{\mathbf{s}}}[||\mathbf{s}-\hat{\mathbf{s}}||_1], \qquad \mathcal{L}_M = \mathbb{E}_{\mathbf{S}_m,\hat{\mathbf{S}}_m}[||\mathbf{S}_m-\hat{\mathbf{S}}_m||_2^2], \tag{8}$$

$$\mathcal{L}_C = \mathbb{E}_{\mathbf{S}_r,\hat{\mathbf{S}}_r}[||\mathbf{S}_r-\hat{\mathbf{S}}_r||_2^2] + \mathbb{E}_{\mathbf{S}_i,\hat{\mathbf{S}}_i}[||\mathbf{S}_i-\hat{\mathbf{S}}_i||_2^2], \tag{9}$$

where $(\mathbf{S}_r, \mathbf{S}_i)$ and $(\hat{\mathbf{S}}_r, \hat{\mathbf{S}}_i)$ represent the real and imaginary parts of the clean and enhanced complex spectrogram. Previous works optimize the phase spectrogram within the complex spectrogram, since the absolute distance between two phases may not be their actual distance. Following the (Lu et al., 2023; Ai & Ling, 2023b), we adopt anti-wrapping phase loss to optimize the phase spectrogram. The anti-wrapping phase loss includes three sub-losses, i.e., instantaneous phase loss $\mathcal{L}_{IP}$, group delay loss $\mathcal{L}_{GD}$, and instantaneous angular frequency loss $\mathcal{L}_{IAF}$, which are defined as:

$$\mathcal{L}_{IP} = \mathbb{E}_{\mathbf{S}_P,\hat{\mathbf{S}}_P}[||f_{AW}(\mathbf{S}_P-\hat{\mathbf{S}}_P)||_1], \tag{10}$$

$$\mathcal{L}_{GD} = \mathbb{E}_{\Delta_{DF}(\mathbf{S}_P,\hat{\mathbf{S}}_P)}[||f_{AW}(\Delta_{DF}(\mathbf{S}_P-\hat{\mathbf{S}}_P))||_1], \tag{11}$$

$$\mathcal{L}_{IAF} = \mathbb{E}_{\Delta_{DT}(\mathbf{S}_P,\hat{\mathbf{S}}_P)}[||f_{AW}(\Delta_{DT}(\mathbf{S}_P-\hat{\mathbf{S}}_P))||_1], \tag{12}$$

where $f_{AW}(t) = |t - 2\pi \cdot \text{round}(\frac{1}{2\pi})|, t \in \mathbb{R}$ is an anti-wrapping function, which is used to avoid the error expansion issue caused by phase wrapping. Δ_{DF} and Δ_{DT} represent the differential operators along the frequency axis and time axis, respectively. The anti-wrapping phase loss $\mathcal{L}_P$ is defined as:

$$\mathcal{L}_P = \mathcal{L}_{IP} + \mathcal{L}_{GD} + \mathcal{L}_{IAF}. \tag{13}$$

Finally, the total loss $\mathcal{L}_{Total}$ is the linear combination of $\mathcal{L}_T$, $\mathcal{L}_M$, $\mathcal{L}_C$, and $\mathcal{L}_P$,

$$\mathcal{L}_{Total} = \alpha_T\mathcal{L}_T + \alpha_M\mathcal{L}_M + \alpha_C\mathcal{L}_C + \alpha_P\mathcal{L}_P, \tag{14}$$

where α_T, α_M, α_C, and α_P are hyperparameter and we follow (Lu et al., 2023) to set them to 0.2, 0.9, 0.1, and 0.3.

Table 6: Comparison with other methods on VoiceBank + DEMAND dataset. "-" denotes the result that is not provided in the original paper. Bold and underline indicate the best and second-best results.

Models	Param. (M)	FLOPs (G)	PESQ	CSIG	CBAK	COVL	STOI
Noisy	-	-	1.97	3.35	2.44	2.63	0.91
DEMCUS (Défossez et al., 2020)	33.53	77.8	3.07	4.31	3.40	3.63	<u>0.95</u>
TFT-Net (Tang et al., 2021)	5.81	295.0	2.75	3.93	3.44	3.34	-
SN-Net (Zheng et al., 2021)	-	-	3.12	4.39	3.60	3.77	-
DB-AIAT (Yu et al., 2022)	2.81	68.0	3.31	4.61	3.75	3.96	-
DPT-FSNet (Dang et al., 2022b)	**0.88**	**55.7**	3.33	4.58	3.72	4.00	**0.96**
CMGAN (Abdulatif et al., 2024)	1.83	81.3	3.41	4.63	3.94	4.12	**0.96**
TridentSE (Yin et al., 2023)	3.03	59.8	3.47	4.70	3.81	4.10	**0.96**
MP-SENet (Lu et al., 2023)	2.05	84.7	<u>3.50</u>	<u>4.73</u>	<u>3.95</u>	<u>4.22</u>	**0.96**
FIRING-Net(Proposed)	<u>1.74</u>	<u>64.2</u>	**3.57**	**4.79**	**3.98**	**4.33**	**0.96**

A.2 Evaluation Metrics

PESQ: Perceptual Evaluation of Speech Quality (PESQ) is a standard method for objectively assessing how speech quality is perceived by listeners (Rix et al., 2001). It provides an estimate of the subjective mean opinion score (MOS) for normal-hearing individuals, specifically evaluating audio quality in noisy or distorted telephone networks (Ma et al., 2009). The PESQ scale typically ranges from 1.0 to 4.5, making it a widely-used metric for measuring the performance of speech enhancement algorithms and the clarity of processed speech.

STOI: The Short-Term Objective Intelligibility (STOI) metric, which ranges from 0 to 1, provides an objective measure of speech intelligibility. It is particularly effective in evaluating speech in environments with temporally modulated noise or time-frequency processed signals, and is designed for normal-hearing listeners (Taal et al., 2011).

SI-SNRi: The Scale-Invariant Signal-to-Noise Ratio Improvement (SI-SNRi) is a metric used to assess the quality of enhanced speech. It is derived from the scale-invariant signal-to-noise ratio (SNR) (Le Roux et al., 2019), with higher values indicating better performance. SI-SNRi is defined as follows:

$$\text{SI-SNRi}(\mathbf{x}, \mathbf{s}, \hat{\mathbf{s}}) = \text{SI-SNR}(\mathbf{s}, \hat{\mathbf{s}}) - \text{SI-SNR}(\mathbf{s}, \mathbf{x}). \tag{15}$$

B Additional Experiments

B.1 Evaluation on VoiceBank + DEMAND Dataset

To ensure a fair comparison between the proposed model and other SOTA models, we trained FIRING-Net on the training set of the public VoiceBank+DEMAND dataset (Valentini-Botinhao et al., 2016) and evaluated its performance on the test set. The VoiceBank + DEMAND test set includes two unseen speakers from the VoiceBank dataset (Veaux et al., 2013), with 20 distinct noise conditions. These conditions consist of five noise types from the DEMAND dataset (Thiemann et al., 2013b), each tested at four SNR levels (17.5, 12.5, 7.5, and 2.5 dB), resulting in a total of 824 test samples. Each test speaker has approximately 20 different sentences per condition. To evaluate and compare the enhanced speech quality across methods, we use mean opinion score (MOS) predictors: signal distortion (CSIG), background intrusiveness (CBAK), and overall quality (COVL), with scores ranging from 1 to 5 (Hu & Loizou, 2007). Additionally, PESQ and STOI are also employed. The averaged SE results on the VoiceBank + DEMAND dataset are presented in Table 6, in which we observe that the proposed FIRING-Net achieves superior performance than several SOTA SE models across all performance measures.

B.2 Evaluation on More Complex Noise Conditions

We further show the advantages and flexibility of the proposed FIRING-Net by exploring FIRING-Net performance in environments where multiple types of noise are contained in the target domain (Lin et al., 2021), Figure 6 presents the PESQ and STOI results on multiple target noise types, where B, E, and T indicate "Babble", "Engine", and "Traffic", respectively. We observe that the

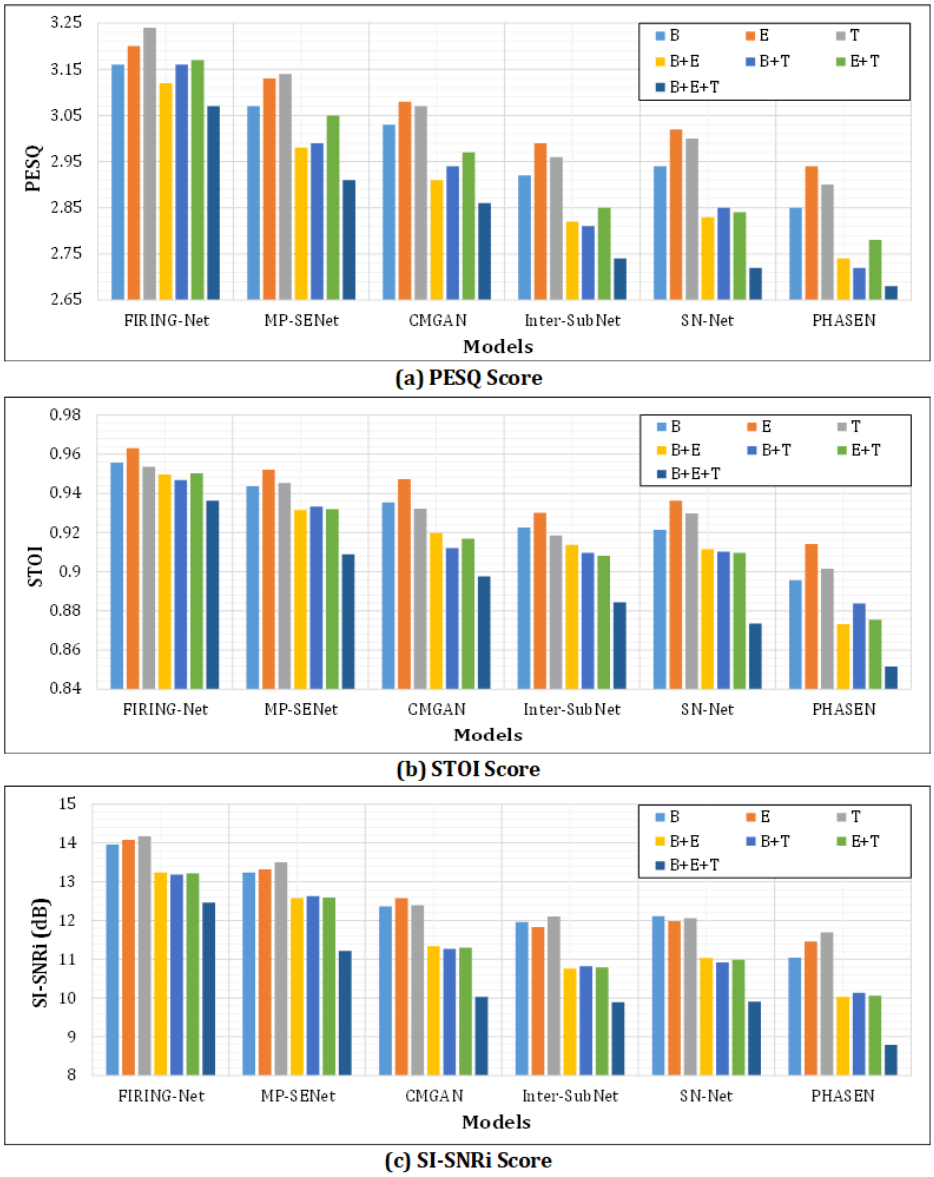


Figure 6: Comparison in PESQ, STOI, and SI-SNRi average for cases with multiple noise types, in which B, E, and T denote babble, engine, and traffic noisy types, respectively.

results using 1, 2, and 3 noise types in the background environment are comparable. Figure 7 shows the enhanced spectrograms of selected baseline models and the proposed FIRING-Net. One can observe that the proposed model sufficiently preserves the spectral details while suppressing the residual noise over the selected baselines.

B.3 Evaluation on IM-E and IM-D

As discussed earlier, the design concepts of IM-E and IM-D are distinct. The design of IM-E focuses on the mutual guidance between target and non-target features, aiming to extract speech information with high similarity to the target features from the non-target features and supplement it into the target features. At the same time, noise information that is highly similar to the non-target features is filtered out from the target features, as the features in the encoder often contain a substantial amount of noise. In contrast, the design of IM-D is based on the observation that the features in the

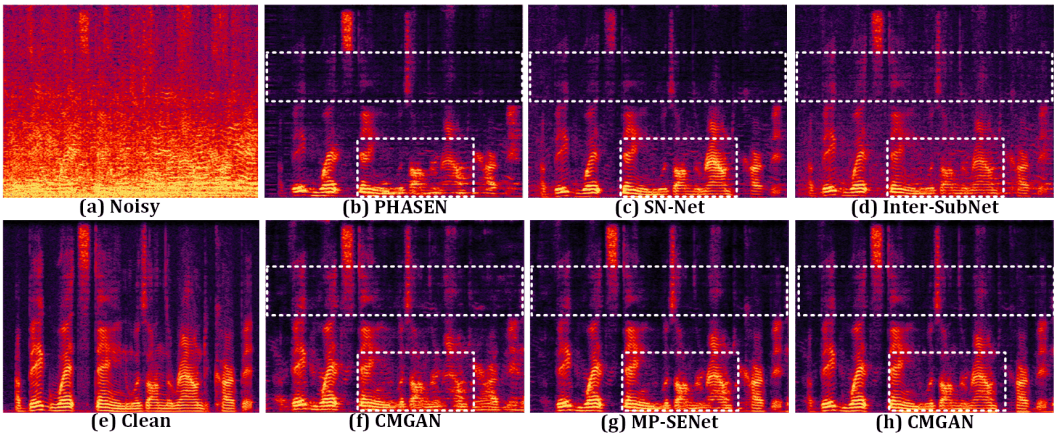


Figure 7: Visualization of spectrograms of enhanced speech generated from FIRING-Net and selected baseline models. The noisy sample contains two noise types, i.e., babble and engine, under -5dB SNR condition.

Table 7: Ablation study on IM-E and IM-D by replacing them with other modules. Bold and underline indicate the best and second-best results.

LM-E	GM	LM-D	STOI	PESQ	SI-SNRi	Param.(M)
IM-E	IM-E	IM-D	**0.9135**	**3.03**	**13.96**	<u>1.74</u>
IM-E	IM-E	IM-E	<u>0.9103</u>	<u>3.01</u>	<u>13.85</u>	1.85
IM-D	IM-D	IM-D	0.8762	2.96	12.87	**1.55**
ConvBlock	ConvBlock	ConvBlock	0.8226	2.77	11.63	1.82
ConvBlock	ConvBlock	IM-D	0.8307	2.76	11.72	1.96
IM-E	IM-E	ConvBlock	0.8812	2.98	12.94	1.75

decoder typically contain little to no noise. In this stage, the non-target features usually consist of speech information that has not been involved in the reconstruction process. Therefore, the design of IM-D primarily considers how to better fuse the speech information from both target and non-target features through mutual guidance, ensuring their effective integration in the subsequent speech reconstruction process. Therefore, we validate the effectiveness of IM-E and IM-D by replacing these modules and evaluating the performance of their internal structures when disassembled.

Replacing IM-E and IM-D: For fairness in comparison, we replace IM-E and IM-D with a configuration consisting of four convolutional blocks (each block containing a convolutional layer, batch normalization, and a PReLU activation function) to avoid performance differences arising from variations in the number of trainable parameters. The evaluation results are shown in Table 7. Firstly, we replaced all IM-D modules in LM-D with IM-E, which resulted in a slight performance decrease for FIRING-Net, along with a noticeable increase in model complexity. Next, we replaced all IMs in the network with IM-D and ConvBlock. This led to a significant drop in model performance, with ConvBlock performing worse than IM-D. To investigate the cause, we compared the relevant and irrelevant feature maps of LM-E using IM-E, IM-D, and ConvBlock, as shown in Figure 8. The figure reveals that with IM-E, speech and noise information are distinctly aggregated into the target and non-target features, respectively. In contrast, with IM-D and ConvBlock, the separation of speech and noise information within target and non-target features is not as clear, and ConvBlock exhibits more noise in both target and non-target features compared to IM-D. This suggests that the ability of LM-E to extract speech information into target features and noise information into non-target features is closely related to the structure of the interaction modules, highlighting the effectiveness of IM-E.

Internal structure disassembling for IM-E: IM-E incorporates three unique structures. First, it reverses the features emphasized in $\mathbf{M}_n$ and $\mathbf{M}_t$(for example, $\mathbf{M}_n$ emphasizes speech information in $\mathbf{f}_n$, and after reversal, we want it to emphasize noise information). Instead of subtracting these masks

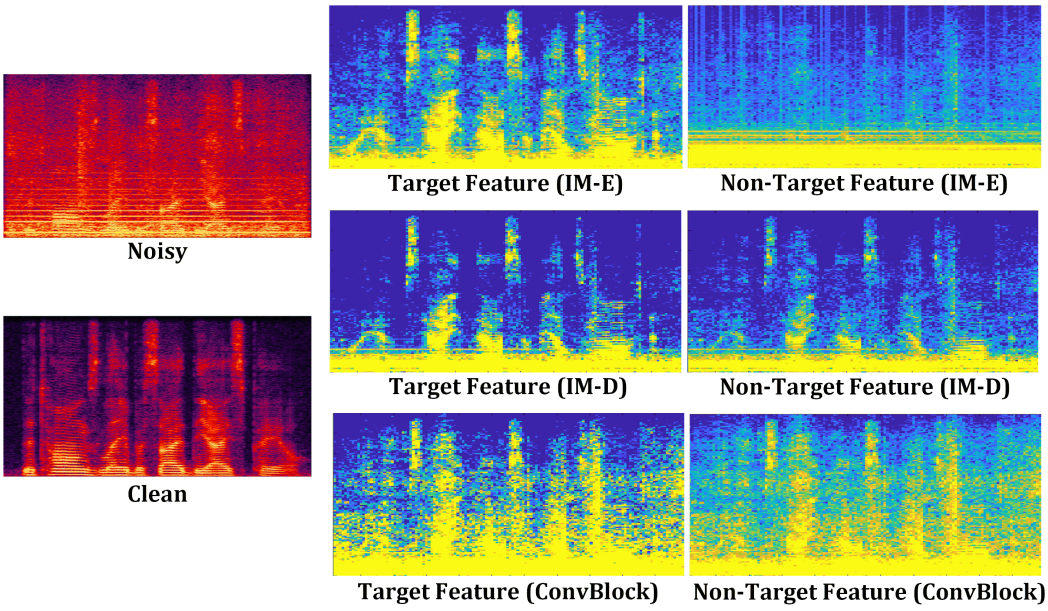


Figure 8: Visualization of feature maps captured from the first LM-E block and the LM-E when IM-E is replaced by IM-D and ConvBlock.

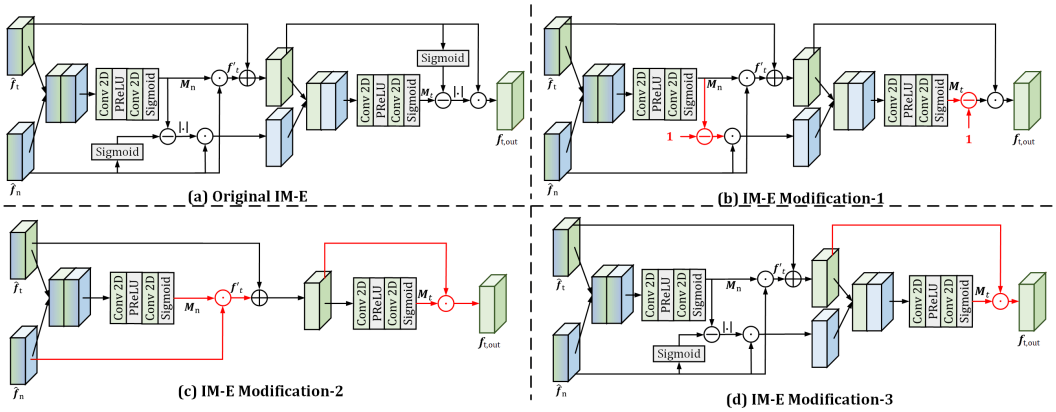
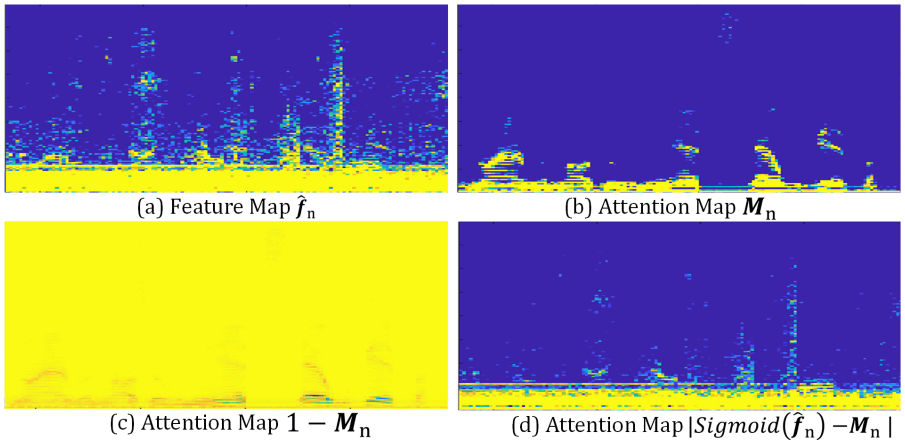


Figure 9: Network architecture of three modifications of IM-E. (a) We replace the "$|Sigmoid(\hat{\mathbf{f}}_n) - \mathbf{M}_n|$" and "$|Sigmoid(\hat{\mathbf{f}}_t) - \mathbf{M}_t|$" with "$1 - \mathbf{M}_n$" and "$1 - \mathbf{M}_t$". (b) We remove the non-target feature to guide the target feature for noise information extraction. (c) We remove the "$|Sigmoid(\hat{\mathbf{f}}_t) - \mathbf{M}_t|$". These modifications are marked in red.

directly from 1, we subtract them from $|Sigmoid(\hat{\mathbf{f}}_n)$ and $|Sigmoid(\hat{\mathbf{f}}_t)$ respectively, and then take the absolute value (This modified structure is shown in Figure 9(a)). Second, it uses $\mathbf{M}_n$, which is the complement of the one used to extract speech information from non-target features, to instead extract noise information. This noise information is then employed to guide the extraction of noise from the target features, i.e., $|Sigmoid(\hat{\mathbf{f}}_n) - \mathbf{M}_n|$ (This modified structure is shown in Figure 9(b)). Third, during noise filtering of the target feature, another mask $\mathbf{M}_t$, is reversed; initially used for filtering noise and extracting speech, it is reversed to instead extract noise, i.e., $|Sigmoid(\hat{\mathbf{f}}_t) - \mathbf{M}_t|$ (This modified structure is shown in Figure 9(c)). The results are presented in Table 8. First, we observed that the performance of IM-E Modification-1 is significantly lower than that of IM-E. Both structures attempt to reverse the features emphasized by the sigmoid activation function. To explain this

Table 8: Evaluation on the structure of IM-E. Bold and underline indicate the best and second-best results.

Module	STOI	PESQ	SI-SNRi	Param.(M)
IM-E	**0.9135**	**3.03**	**13.96**	1.74
IM-E Modification-1	0.8884	2.94	13.25	1.74
IM-E Modification-2	<u>0.9008</u>	<u>2.99</u>	13.51	**1.52**
IM-E Modification-3	0.8987	2.96	<u>13.67</u>	<u>1.64</u>



(a) Feature Map $\hat{f}_{\mathrm{n}}$ (b) Attention Map $\boldsymbol{M}_{\mathrm{n}}$

(c) Attention Map $1-\boldsymbol{M}_{\mathrm{n}}$ (d) Attention Map $|Sigmoid(\hat{f}_{\mathrm{n}})-\boldsymbol{M}_{\mathrm{n}}|$

Figure 10: Visualizations of feature and attention maps of $\hat{\mathbf{f}}_n$, $\mathbf{M}_n$, $1-\mathbf{M}_n$, and $|Sigmoid(\hat{\mathbf{f}}_n)-\mathbf{M}_n|$.

phenomenon, we present the feature spectrograms of $|Sigmoid(\hat{\mathbf{f}}_n)-\mathbf{M}_n|$ and $1-\mathbf{M}_n$ in Figure 10. From this, we can see that the $1-\mathbf{M}_n$ operation amplifies all features that are not emphasized by $\mathbf{M}_n$, and this amplification primarily boosts regions in the feature spectrum with originally weaker energy, introducing new interfering features into the feature processing. Additionally, we observe a performance drop in IM-E after modifications 2 and 3, with IM-E Modification-3 performing worse than Modification-2. This is because, in both modifications, $\mathbf{M}_t$ aggregates speech information. However, in Modification-3, $\mathbf{M}_t$ is generated from fused target and non-target features, allowing noise from the non-target features to interfere with the precision of $\mathbf{M}_t$. In the original IM-E, the operation $|Sigmoid(\hat{\mathbf{f}}_t)-\mathbf{M}_t|$ shifts the mask's function from aggregating speech to aggregating noise, thus enhancing the non-target features' contribution.

Internal structure disassembling for IM-D: The structure of IM-D can be roughly summarized as a weighted summation of target and non-target features. To validate the effectiveness of this design, we made three modifications to IM-D, as shown in Figure 10. In the first modification, we directly sum the target and non-target features without applying any weights. In the second modification, only the non-target features are weighted, while in the third modification, only the target features are weighted. To ensure a fair comparison, we increased the number of convolutional blocks in the second and third modifications so that the number of trainable parameters matches that of the original IM-D. The results, as shown in Table 9, indicate that the performance of IM-D Modification-1, where target and non-target features are summed without weighting, is noticeably worse compared to IM-D Modification-2 and IM-D Modification-3, where only one feature is weighted. Additionally,

Table 9: Evaluation on the structure of IM-D. Bold and underline indicate the best and second-best results.

Module	STOI	PESQ	SI-SNRi	Param.(M)
IM-D	**0.9135**	**3.03**	**13.96**	1.81
IM-D Modification-1	0.8762	2.95	13.23	1.23
IM-D Modification-2	<u>0.8987</u>	<u>2.98</u>	<u>13.82</u>	1.81
IM-D Modification-3	0.8934	2.97	13.76	1.81

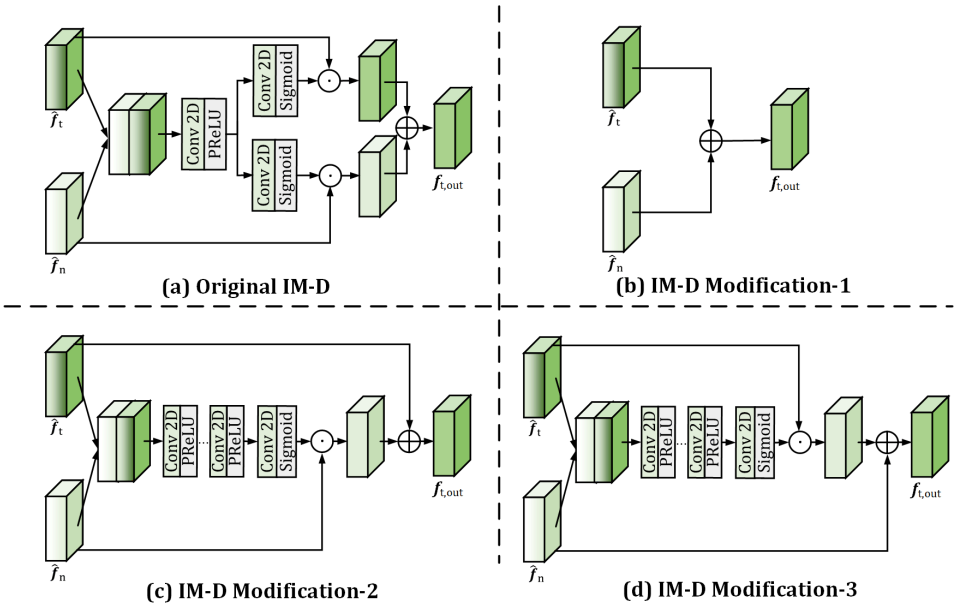


Figure 11: Network architecture of three modifications of IM-D. (a) We directly add the target and non-target features. (b) We process the non-target feature only to add with target feature. (c) We process the target feature only to add with non-target feature.

Table 10: Evaluation on Phase Decoder. Bold and underline indicate the best and second-best results.

Phase Decoder	STOI	PESQ	SI-SNRi
LM-D	**0.9135**	**3.03**	**13.96**
DCCM	0.9074	2.98	13.10
FGCM	<u>0.9103</u>	<u>2.99</u>	<u>13.54</u>
Noisy Phase	0.8621	2.89	12.62
Clean Phase	0.9296	3.09	14.83

while the performance difference between IM-D Modification-2 and IM-D Modification-3 is not significant, both modifications still perform significantly worse than the original IM-D.

B.4 Evaluation on Phase Decoder

Since phase information, unlike magnitude information, lacks a clear structural pattern, we conducted an evaluation to validate the effectiveness of the proposed FIRING-Net's Phase Decoder. For comparison, we replaced LM-D with DCCM and FGCM, and compared the performance of the phase output from the Phase Decoder with both the noisy input phase and the clean target phase. As shown in Table 10, although the performance of the speech synthesized with the Phase Decoder-generated phase is weaker than that of the clean phase, it significantly outperforms both the noisy phase and the phase produced by DCCM and FGCM-based structures. However, it is important to note that the mapping relationship between noisy and clean phase is less explicit compared to that of noisy and clean magnitude. Thus, while our proposed FRI framework theoretically does not fully apply to phase recovery, the use of IM-D in the Phase Decoder yields performance gains. Whether these gains align with the theoretical improvements described by the FRI framework remains uncertain and will be explored in future work, with a focus on developing models better suited for phase recovery.