A Multi-view Fusion Approach for Enhancing Speech Signals via Short-time Fractional Fourier Transform

Zikun Jin[1,2], **Yuhua Qian**[1,2*], **Xinyan Liang**[1,2] and **Haijun Geng**[3]
[1]Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China
[2]Key Laboratory of Evolutionary Science Intelligence of Shanxi ProvinceShanxi University, Taiyuan 030006, Shanxi,China
[3]School of Automation and Software Engineering, Shanxi University, Taiyuan 030006, Shanxi,China
{jinzikun, liangxinyan48, ghj123025449}@163.com, jinchengqyh@126.com,

Abstract

Deep learning-based speech enhancement (SE) methods focus on reconstructing speech from the time or frequency domain. However, these domains cannot provide enough information to capture the dynamics of non-stationary signals accurately. To enrich information, this work proposes a multi-view fusion SE method (MFSE). Specifically, MFSE extends the representation space of speech to the dynamic domain (also called fractional domain) between the time and frequency domains by using the short-time fractional Fourier transform (STFrFT). Subsequently, we construct inputs as modes of the primary short-time Fourier transform (STFT) spectrum and the auxiliary STFrFT spectrum views and adaptively identify the optimal fractional STFrFT spectrum from the infinitely continuous fractional domain by leveraging the average spectral centroids. The framework extracts potential features through multiple designed convolutional modules and captures the correlation between different speech frequencies through multi-granularity attention. Experimental results show that the proposed method significantly improves performance in several metrics compared to existing single-channel SE methods based on time and frequency domains. Furthermore, the results of its generalizability evaluation show that the multi-view method outperforms the single-view method under a wide range of SNR conditions.

1 Introduction

Speech enhancement (SE), commonly applied in various scenarios, aims to improve the intelligibility and quality of speech signals degraded by noise or other forms of distortion. It is a crucial task in modern communication and accessibility technologies, helping people with hearing impairments by filtering background noise in hearing aids, improving speech clarity for accurate recognition in voice-controlled systems, and improving call quality by reducing ambient noise in telecommunications [Li *et al.*, 2022].

*Corresponding author: jinchengqyh@126.com

The speech representation space plays a key role in SE base on deep learning [Yu *et al.*, 2020; Tian *et al.*, 2024]. The existing methods mainly depend on time-domain and frequency-domain. (1) *Time domain methods* take raw speech signals as input and apply deep neural network architectures for processing. Some approaches [Défossez *et al.*, 2020; Strauss and Edler, 2021; Kong *et al.*, 2022] leverage UNet architectures to extract multiscale features, while RNNs or LSTMs capture long-term dependencies in the time-domain signals. These methods preserve the time-domain characteristics of speech, which is beneficial for enhancing the speech SNR [Luo and Mesgarani, 2019]. (2) *Frequency-domain methods*, such as the Short-Time Fourier Transform (STFT), provide flexible signal representations for speech and noise separation. The models learn how to extract speech components from the noise spectrum while suppressing the noise. The enhanced spectrum is converted back to a time domain signal by the Inverse STFT (ISTFT) [Hu *et al.*, 2020; Fu *et al.*, 2021; Shin *et al.*, 2023]. These methods can suppress noise in a finer space than the time domain, which is advantageous for improving speech intelligibility.

Recent works argue that relying solely on time or frequency domain information only captures partial features of the audio, thus limiting the comprehensiveness of the data analysis [Tang *et al.*, 2021; Lin *et al.*, 2024]. To take full advantage of the information in both domains and to integrate their representation spaces, recent approaches have focused on joint optimization of the time and frequency domains as multi-view inputs. These methods [Nareddula *et al.*, 2021; Dang *et al.*, 2023] utilize different networks to map time and frequency domain views to the same underlying representation thus fusing them, extending the representation space of speech and noise to improve the performance of speech enhancement. Some recent work such as [Liang *et al.*, 2024; Guo *et al.*, 2024a] have carried out some architectural innovations for multi-view convergence. There are also works such as [Liang *et al.*, 2022; Liang *et al.*, 2025] that utilize associative relationships and trusted fusion to enhance the interpretability of multi-view fusion.

However, these approaches still focus on analyzing signals from only two perspectives, time and frequency, and they have a fixed resolution for time and frequency, making it difficult to achieve a fine-grained representation of speech and

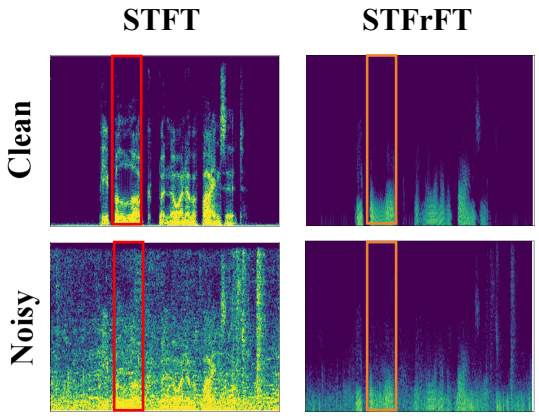

Figure 1: Visualize STFT, and STFrFT of the sample speech, respectively.

noise. Coupled with the fact that the STFT originally contained time variations, the effect of optimizing the two jointly is not obvious. The key to improving SE is to find a more flexible auxiliary view for STFT that extends the representation space to better decouple speech and noise. The short-time fractional order Fourier transform (STFrFT), on the other hand, allows the adjustment of time and frequency resolution by introducing a fractional order parameter, resulting in better time resolution in the high-frequency bands and better frequency resolution in the low-frequency bands, which can be adapted to the characteristics of different signals. This flexible resolution adjustment capability allows the model to capture richer time-frequency information, thus improving the effectiveness of SE. As shown in Figure 1, STFrFT spectrum is capable of obtaining a more focused spectral representation, which serves as an auxiliary view to direct the STFT spectrum to focus on important information at the same time frame (red and orange box).

In this paper, we present a novel multi-view single-channel SE method that employs the short-time fractional Fourier transform (STFrFT), which provides a richer representation space than methods that use only the time and frequency domains.

The main contributions of this work include:

- We extend speech to the dynamic domain by STFrFT. Determining the best auxiliary STFrFT view from the dynamic domain using spectral centroids makes STFT more focused on useful frequency information.
- We propose a lightweight framework containing a feature extraction module and an attention module for better separation of speech and noise.
- In VoiceBank+Demand dataset, our proposed algorithm has significant advantages over other advanced algorithms.

2 Related Work

The current inputs for deep learning-based single-channel SE are mainly classified into three cases: (i) Time domain based inputs. (ii) Frequency domain based inputs. (iii) Multi-view inputs optimized jointly in both the time and frequency domains. Inputs based on STFT variations are commonly referred to as the time-frequency (T-F) domain in the field of SE, which is abbreviated as the frequency domain in this paper to better distinguish it from multi-view inputs.

Time domain based SE models operate directly on time domain waveforms without domain transformation. SEGAN [Pascual *et al.*, 2017] is a model that transforms noisy speech into clear speech, with its generator handling the transformation and its discriminator distinguishing between synthetic and real clear speech. Building on this idea, DSEGAN [Phan *et al.*, 2020] introduces multiple generator chains to perform multi-stage augmented mapping, which enhances the coherence of the representation space compared to SEGAN. In contrast, SEflow [Strauss and Edler, 2021] proposes a normalized flow framework that directly models the enhancement process by estimating the density of clean speech utterances. Additionally, DEMUCS [Défossez *et al.*, 2020] and CleanUNet [Kong *et al.*, 2022] both introduce causal SE models that operate directly on raw waveforms, further advancing the capabilities of speech enhancement technologies.

STFT domain based SE methods process speech signals in the frequency domain, where the input signals are typically amplitude spectrum, complex spectrum, and phase spectrum. MMSEGAN [Soni *et al.*, 2018] integrates GAN within an augmented framework, using STFT domain masks to implicitly learn and predict clean STFT domain representations. In a similar vein, PSMGAN [Routray and Mao, 2022] tackles the challenge of phase information by designing a phase-sensitive speech enhancement technique, employing conditionally generated adversarial networks. Meanwhile, PR-WaveGlow [Maiti and Mandel, 2020] and Glance [Li *et al.*, 2022] take a different approach, using the UNet architecture to enhance the representation space of the information flow through specialized modules developed in each of their frameworks. On the other hand, S4ND [P-J *et al.*, 2023] and Dual-S4D [Sun *et al.*, 2024] employ Structured State Spaces for Sequences, a linear time-invariant system, to provide more granular continuous characterizations of long sequences. In contrast, these works [Lu *et al.*, 2021; Tai *et al.*, 2023a; Tai *et al.*, 2023b; Guo *et al.*, 2024b] propose a different approach by modeling natural images and raw audio waveforms through paired diffusion and inversion processes, offering a novel method for both image and audio enhancement.

Multi-view based SE models typically combine time and frequency domain features to expand the representation space. In an effort to unify the feature spaces in both the time and frequency domains, [Wang *et al.*, 2024] introduces a feature-fused two-branch SE method, which leverages the encoder from [Luo and Mesgarani, 2019]. Building on this idea, [Tang *et al.*, 2021] proposes a two-channel attention spanning framework called TFTNet, designed to fully exploit the correlation between the time and frequency axes. TFTNet takes time-frequency spectrograms as inputs and produces time-domain waveforms as outputs, enhancing the overall performance of speech enhancement tasks by more effectively capturing cross-domain dependencies.

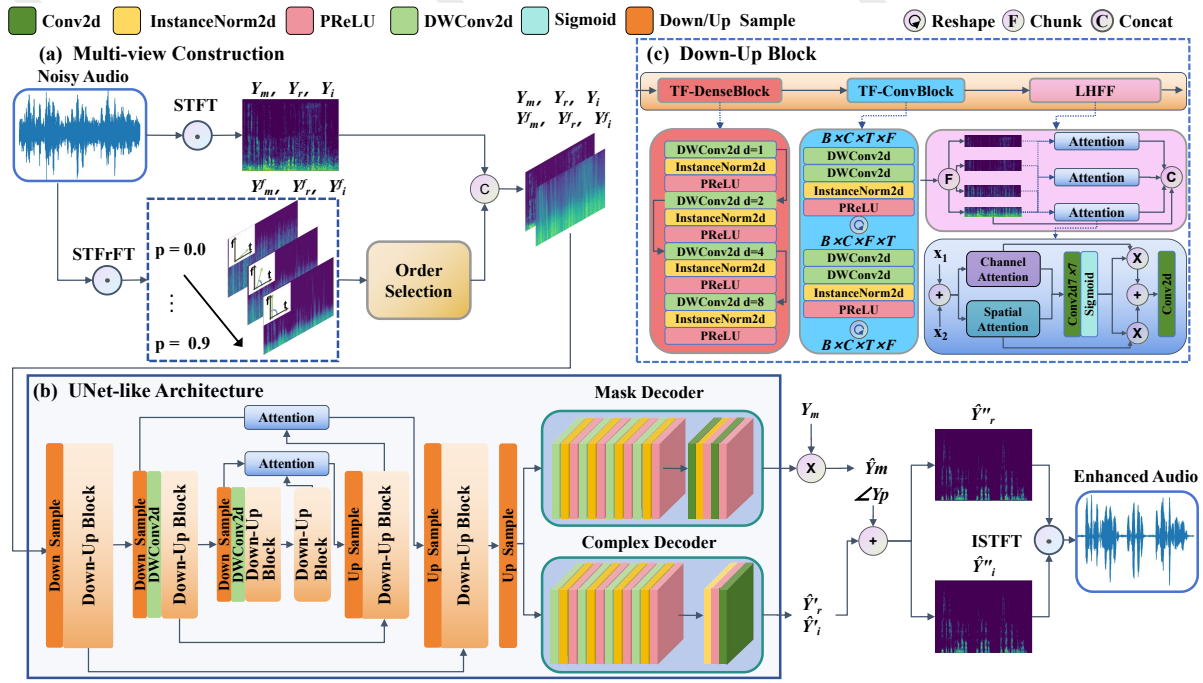

Figure 2: MFSE framework. (a) Multi-view Construction: The original speech signal is reconstructed by STFT as Y_m, Y_r, and Y_i denoted as the magnitude, real, and imaginary parts of the STFT spectrum, respectively. Similarly, the original signal is converted into multiple STFrFT spectrum by STFrFT with p from 0 to 0.9 in steps of 0.1, and then Y_m^f, Y_r^f, and Y_i^f, corresponding to the optimal order, are selected. (b) UNet-like Architecture: Contains a UNet network and MaskDecoder and ComplexDecoder. (c) Down-Up Block: consists of two convolutional modules and Low-High frequency fusion (LHFF).

3 Methodology

This section describes how to convert speech signals into STFrFT spectrum and how to pick the optimal fractional counterpart STFrFT spectrum for constructing speech-enhanced multiviews. Finally, we illustrate the design principles and loss functions of the network architecture. The MFSE is described in Figure 2.

3.1 Short-time Fractional Fourier Transform

The Fractional Fourier Transform (FrFT) generalizes the classical Fourier transform by introducing a fractional parameter that allows continuous interpolation between time and frequency domains, offering greater flexibility in adjusting the time-frequency balance and making it especially effective for analyzing signals with complex, non-stationary behaviors.

The STFrFT extends this concept by combining the advantages of the FrFT with the short-time analysis of the signal, allowing for a localized time-frequency representation. In the STFrFT, the signal is segmented into short-time intervals, with the fractional transform applied to each segment, allowing for more precise and adaptive feature extraction from non-stationary signals. This enables SE models to better capture the intricate, non-stationary characteristics of speech, enhancing performance in complex acoustic environments.

FRFT is based on the mathematical properties of the classical Fourier transform, and the fractional rotation of a signal in the time-frequency domain is achieved by introducing fractional powers of order. The definition of the FRFT for a signal

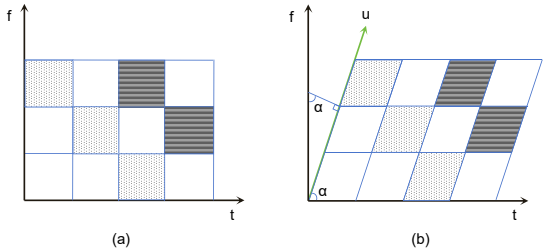

Figure 3: Visualization of the time, frequency and fractional domains: (a) STFT, (b) STFrFT.

$x(t)$ is given by

$$F_\alpha(u) = \mathcal{F}^\alpha\{x(t)\}(u) = \int_R x(t)\mathcal{K}_\alpha(u,t)dt, \quad (1)$$

where α is the angle, and when the $\alpha \neq c\pi$ ($c \in Z$), for the transformation kernel $\mathcal{K}_\alpha(u,t)$ is defined as follows

$$\mathcal{K}_\alpha(u,t) = \sqrt{(1-j\cot\alpha)/2\pi}e^{j\frac{u^2+t^2}{2}\cot\alpha - jut\csc\alpha} \quad (2)$$

In this paper, the proposed method uses an order range of $\alpha \in [0, \frac{\pi}{2}]$, and we utilize a order $p \in [0,1]$ to control the size of the α so that $\frac{p\pi}{2} = \alpha \in [0, \frac{\pi}{2}]$. FFT is a special form of FrFT when $p = 1$ i.e. $\alpha = \pi$. For the characterization of $\mathcal{K}_\alpha(u,t)$ in FrFT, IFrFT can be obtained by transforming F_α by $\mathcal{F}^{-\alpha}$ again, and the definition is given below

$$x(t) = \mathcal{F}^{-\alpha}\{F_\alpha(u)\}(t) = \int_R F_\alpha(u)\mathcal{K}_\alpha^*(u,t)du \quad (3)$$

Analogous to the STFT, the mathematical form of STFrFT is given by

$$STFrFT_{\alpha}(t,u) = \int_{-\infty}^{\infty} x(\tau) g(\tau - t) K_{\alpha}(\tau, u) d\tau, \quad (4)$$

where $g(\tau)$ denotes the chosen window function centered around the origin, e.g., Hamming Window or Hanning Window.

As can be visualized in Figure 3, u is able to dynamically display the entire time-frequency space by changing the size of alpha by adjusting the order p, which is STFT when $p = \frac{\pi}{2}$.

In practice, it is also necessary to interpolate discrete signals to ensure accurate computation, and in this paper, sinc interpolation is used to recover continuous signals from discrete samples. It has the unique property of being the "perfect" interpolation kernel, as it perfectly reconstructs a bandlimited signal from its samples in the context of the Shannon-Nyquist theorem. The sinc function is given by:

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \quad (5)$$

The following section describes the process of how to select an STFrFT spectrum corresponding to the optimal order from the candidate views as an auxiliary view with the STFT to form the multi-view input.

3.2 Order Selection

Selecting the optimal fractional order in SE algorithms is crucial because it allows for a precise balance between time and frequency domain representation, enhancing the signal's clarity and intelligibility. By adjusting the fractional order, the algorithm can better suppress noise while preserving important speech features such as pitch and formants, thus improving the SNR and overall speech quality. The optimal fractional order adapts to different noise types, environmental conditions, and speech signal characteristics, ensuring both effective noise reduction and the preservation of natural speech. This selection also ensures computational efficiency by avoiding unnecessary complexity, leading to a more robust and effective enhancement process. In order to select the optimal order, we devised a method of selection using the spectral centroid. The calculating spectral centroid is defined as follows:

$$C = \frac{\sum_{k=1}^{N} f_k \cdot |X(f_k)|}{\sum_{k=1}^{N} |X(f_k)|}, \quad (6)$$

where C is the spectral centroid, N is the number of discrete frequency points in the spectrum, f_k is the $k-th$ frequency point, $X(f_k)$ is the complex spectrum of the signal at the frequency point f_k (usually the result of the Fourier transform), and $|X(f_k)|$ is its modulus, which represents the amplitude at that frequency point.

We calculated the spectral centers $C^T_{p=0}, \cdots C^T_{p=0.9}$ for all candidate STFrFTs time lengths of T.

$$C'^T_p = C^T_p - \frac{\sum_{i=0}^{M} C^T_{p=i}}{P}, \quad (7)$$

Algorithm 1 Compute the spectral centroid

Input: STFrFT spectrum $x^{T\times F}_{p=0}$, $x^{T\times F}_{p=0.1}$,$\cdots$,$x^{T\times F}_{p=0.9}$, $size = T \times F$
Output: $cent$

```
Initialize set cent.
for i = 0 to 0.9 do
    Initialize set centroids, freq = FFTfrequencies
    for t = 1 to T do
        Initialize centroid, mag = |x^{t×F}_{α=i}|
        centroid = sum(freq × mag) / sum(mag)
        centroids ← centroid
    end for
    cent ← sum(centroids)
end for
return cent
```

where M is the number of order and C'^T_p is the distance from C^T_p to average spectral centroid. Finally, based on the C'^T_p of each order, the order p corresponding to the maximum C'^T_{max}, average C'^T_{avg} and minimum C'^T_{min} are obtained.

After selecting the STFrFT spectrum corresponding to the completed optimal fractional order, we connect it to the STFT spectrum in the channel dimension as a multiview input, and the advantages of this to construct a multiview will be discussed later.

3.3 Multi-view Construction

The most direct way to utilize the multi-view idea for SE is to utilize the combination of time domain and frequency domain. However, the time-domain and frequency-domain features are heterogeneous, requiring fusion through feature mapping that projects both into the same high-dimensional space. This introduces a non-consistency factor between the views, making the fusion dependent on the quality of the feature mapping and the design of the loss function [Xue *et al.*, 2024; Zheng *et al.*, 2024]. In multi-view fusion, keeping consistency between views helps to describe the consistent target more accurately [Son *et al.*, 2024; Wu *et al.*, 2025], and if the consistency between two views is low, it often can broaden the feature space to some extent, but it does not help much in feature extraction of the target itself. The STFT is a special case of the STFrFT, differing only in the order p of the transform kernel, while remaining consistent in the time dimension. This consistency allows both to simultaneously characterize the distributions of speech and noise within the same time frame, enabling them to be input into the network as a whole and mitigating, to some extent, the conflicting nature of the viewpoints.

In deep learning, especially in SE networks, the instability of spectral resolution affect the performance of the SE model, and retaining fixed input features helps to maintain stability during the training process [Shi *et al.*, 2023]. STFrFT dynamically adjusts the degree of time-frequency localization of the signal by introducing a fractional parameter p, which allows STFrFT to provide more flexible time-frequency resolution theoretically, but also increases the computational complexity. Selecting different fractional order parameters will have

different effects on the time-frequency localization of the signal, which may lead to fluctuations in the time-frequency resolution, especially when the selected fractional order is large or small. The instability of the spectral resolution may affect the performance of the enhanced model. STFT as a fixed mode involved in training can provide stable time-frequency representation, thus yes the model can be maintained to a certain extent stable training, the introduction of the optimal short-time fractional view can be utilized to exploit the complementary nature of the multiview network.

3.4 Multi-view Fusion SE Framework

In this paper, we design the Multi-view Fusion SE Framework (MFSE) architecture and the overall architecture is shown in Figure 2, and we follow [Sun *et al.*, 2024; Chen *et al.*, 2024] using a UNet architecture. The encoder consists of three convolutional layers, two downsampling layers, and three Down-Up Blocks. The first convolution fuses multi-view information and expands it to a specified dimension, while the next two convolutions use depth-separable convolutions to integrate downsampled information for better compatibility with the Down-Up Block module. The decoder includes a UNet upsampling module, a mask deconstructor, and a complex domain deconstructor. The UNet upsampling module mirrors the encoder, with three upsampling layers and three Down-Up Blocks. The UNet encoder and decoder parts of the corresponding layers utilize the fusion module for fusing the conduction information flow. The mask deconstructor is used to decode the features into a frequency domain mask, which is used to identify the noise components in the frequency domain and mask them out, and then combine them with the phase information to transform them into the real and imaginary parts of the frequency domain, where $MASK$ is the output of the mask deconstructor and Y_m and Y_p represent the amplitude spectrum and phase of the input main view STFT, respectively.

$$\begin{aligned}\hat{Y}_r &= (MASK \times Y_m)\cos Y_p \\ \hat{Y}_i &= (MASK \times Y_m)\sin Y_p\end{aligned} \tag{8}$$

The complex domain deconstructor is used to decode the features into real and imaginary parts in the frequency domain and finally the final output is weighted and fused with the real and imaginary parts converted by the mask deconstructor as shown by

$$\begin{aligned}\hat{Y}_r'' &= a\hat{Y}_r + (1-a))\hat{Y}_r' \\ \hat{Y}_i'' &= a\hat{Y}_i + (1-a)\hat{Y}_i'\end{aligned} \tag{9}$$

Through experimentation, we finally set a to 0.75.

Down-Up Block, the main feature extraction module of MFSE, consists of three parts: (i) TF-DenseBlock (ii) TF-ConvBlock (iii) Low-High Frequency Fusion (LHFF) based on multi-granularity attention block. The details of the three modules are shown in Figure 2(c).

TF-DenseBlock uses depth-separable expansion convolution combined with the residual structure of DenseNet, where larger sensory fields tend to give better results in speech tasks. Under the premise of ensuring lightweight, TF-DenseBlock performs feature extraction and fusion of features at different scales, expanding the sensory ability of the extractor on the features.

TF-ConvBlock utilizes depth-separable convolution for feature extraction in the time and frequency dimensions respectively after TF-DenseBlock extracts multiscale features, keeping the extractor's ability in the time and frequency dimensions dynamically balanced to improve the stability and performance of model training.

The multi-granularity attention block leverages multi-granularity features to perform stagewise fusion of symmetric UNet features at the same stage. It employs channel and spatial attention mechanisms for coarse-grained feature extraction, followed by a 7×7 convolution and sigmoid activation to compute fine-grained attention for each spectral feature. The symmetric features are then weighted and fused. To address the neglect of the correlation between low-frequency and high-frequency information in previous SE methods, we designed the LHFF module. The LHFF module divides the spectral features into four equal parts along the frequency axis. It passes the low-frequency features in sequence with other frequency bands through a multi-granularity attention block and then splices them in the frequency dimension. LHFF progressively maps from coarse-grained to fine-grained features, using the well-recovered low-frequency information to guide the optimization of high-frequency components, which are strongly affected by noise.

3.5 Loss Function

The loss function is used in this paper as a weighted fusion of the magnitude spectrum with three weighted fusions of the complex spectrum and the time-domain loss. The magnitude spectrum is defined as follows:

$$\mathcal{L}_{\text{Mag.}} = E_{Y_m,\hat{Y}_m''}\left[\left\|Y_m - \hat{Y}_m''\right\|^2\right], \tag{10}$$

where $\hat{Y}_m'' = \sqrt{\hat{Y}_r''^2 + \hat{Y}_i''^2}$. In order to more accurately portray the degree of spectral reduction, we optimize using the complex spectral loss, which is defined by

$$\begin{aligned}\mathcal{L}_{\text{RI}} = {} & E_{Y_r,\hat{Y}_r''}\left[\left\|Y_r - \hat{Y}_r''\right\|^2\right] \\ & + E_{Y_i,\hat{Y}_i''}\left[\left\|Y_i - \hat{Y}_i''\right\|^2\right]\end{aligned} \tag{11}$$

Utilizing time loss enables a greater focus on the metric SSNR is given by

$$\mathcal{L}_{\text{Time}} = E_{x,\hat{x}}\left[\|x - \hat{x}\|_1\right] \tag{12}$$

where $\hat{x} = ISTFT(\hat{Y}_r'', \hat{Y}_i'')$, x is clean speech lable. The final loss is formulated as follows:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{\text{Mag.}} + (1-\alpha)\mathcal{L}_{\text{RI}} + \beta\mathcal{L}_{\text{Time}} \tag{13}$$

where we assign the value of α to 0.9 and β to 0.2.

4 Experiments

In order to verify the noise reduction performance and generalization ability of the proposed method, we conducted a series of experiments.

Methods	Pub/Year	Params	Input	PESQ ↑	CISG ↑	CBAK ↑	COVL ↑	SSNR ↑	STOI ↑
Noisy	-	-	-	1.97	3.3	2.44	2.63	1.68	0.91
SEGAN	arvix/2017	-	T	2.16	3.48	2.94	2.80	7.73	0.92
MMSEGAN	ICASSP/2018	-	T-F	2.53	3.80	3.12	3.14	-	0.93
MetricGAN	ICML/2019	-	T	2.86	3.99	3.18	3.42	-	-
DSEGAN	SPL/2020	-	T	2.39	3.46	3.11	2.90	8.72	0.932
MetricGAN+	Interspeech/2021	-	T-F	3.15	4.14	3.16	3.64	-	-
PSMGAN	CSL/2022	-	T-F	2.92	3.88	3.45	3.62	-	-
MetricGAN-OKDv2	ICML/2023	0.82M	T-F	3.12	4.17	3.13	3.64	-	-
PR-WaveGlow	ICASSP/2020	-	T-F	-	3.80	2.40	3.10	-	0.91
DEMUCS	Interspeech/2020	18.87M	T	2.93	4.22	3.25	3.52	-	-
TFT-Net	IJCAI/2020	-	T-T-F	2.75	3.93	3.44	3.34	-	-
SE-Flow	ICASSP/2021	-	T	2.28	3.70	3.03	2.97	7.93	-
CleanUNet	ICASSP/2022	39.77M	T	2.88	4.32	3.41	3.63	-	-
GaGNet	Appl Acoust/2022	5.94M	T-F	2.94	4.26	3.45	3.59	-	-
S4ND UNet	Interspeech/2023	0.75M	T-F	2.99	4.37	3.56	3.70	-	-
Dual-S4D	TASLP/2024	10.8M	T-F	2.55	3.94	3.00	3.23	-	0.934
DiffuSE	APSIPA ASC/2021	-	T	2.44	3.66	2.83	3.03	-	-
CDiffuSE	ICASSP/2022	-	T	2.52	3.72	2.91	3.10	5.28	0.914
DR-DifuSE	AAAI/2023	3.55M	T-F	3.09	<u>4.38</u>	<u>3.57</u>	<u>3.76</u>	<u>9.52</u>	<u>0.949</u>
DOSE	NeurIPS/2023	-	T-F	2.56	3.83	3.27	3.19	-	0.936
VPIDM	TASLP/2024	-	T-F	<u>3.16</u>	4.23	3.53	3.70	-	-
MFSE(Ours)	-	2.9M	T-F-F	**3.17**	**4.46**	**3.79**	**3.89**	**10.63**	**0.952**

Table 1: Comparing with state-of-the-art models on VoiceBank+Demand dataset for PESQ, CISG , CBAK, COVL , SSNR , STOI metrics, our approach achieves well results. "-" denotes the result is not provided in the original paper. "T" indicates that the input is time domain. "T-F" indicates that the input is frequency domain STFT, "T-T-F" indicates that the input contains both time and frequency domains, and "T-F-F" indicates that the input is frequency domain STFT and STFrFT.

4.1 Datasets

VoiceBank+DEMAND is widely used for SE benchmarking. It includes recordings from 30 speakers and 10 noise types, divided into training and test sets with 28 and 2 speakers, respectively. Training uses four signal-to-noise ratios (SNRs) of [0, 5, 10, 15] dB, while testing uses [2.5, 7.5, 12.5, 17.5] dB. The pristine audio is sourced from the VoiceBank corpus, consisting of 11,572 speech instances from 28 speakers in the training set, and 824 utterances from 2 unseen speakers in the test set. The dataset covers a wide range of noisy environments, including public spaces (cafeterias, restaurants, offices), domestic settings (kitchens, living rooms), and urban transport hubs (cars, metros, buses, and subway stations). For our experiments, all speech samples were down-sampled to 16 kHz.

DNS Challenge corpus consists of over 500 hours of clean audio data recorded by 2150 distinct speakers, alongside more than 180 hours of diverse noise recordings. We synthesize 30s long clips by augmenting clean speech utterances and noise. Four signal-to-noise level bands were used to generate the generalization performance test samples, namely [-5-0dB,0-5dB, 5-10dB, 10-15dB] to establish 10,000 speech samples each, and after that, 100 random samples of noisy each were used to evaluate the model's ability to generalize.

We follow [Défossez *et al.*, 2020; Lu *et al.*, 2021; Lu *et al.*, 2022; Sun *et al.*, 2024; Guo *et al.*, 2024b] using Perceptual Evaluation of Speech Quality (PSEQ), prediction of the signal distortion (CSIG), prediction of the background intrusiveness (CBAK), prediction of the overall speech quality (COVL), segmental signal-to-noisy ratio (SSNR) and Short-Time Objective Intelligibility (STOI) these indicators to evaluate our methodology.

Order select	PESQ	CISG	CBAK	COVL	SSNR	STOI
SC-min	3.15	4.43	3.78	3.86	10.50	0.951
SC-avg	3.17	4.46	3.79	3.89	10.63	0.952
SC-max	3.16	4.45	3.77	3.88	10.44	0.952

Table 2: Evaluation metrics of different orders under the VoiceBank+Demand dataset.

4.2 Trainning Setup

Throughout the training process, the speech data was uniformly partitioned into 51,040 points, using an FFT size of 510, a window length of 510, a hop length of 160, and a sampling rate of 16kHz. The training configuration included a batch size of 4, a learning rate of 5e-4, with learning rate decay occurring every 30 epochs and a decay coefficient of 0.5. The model architecture used a base channel size of 64. All models were trained for 120 epochs using the AdamW optimizer. Training was performed on a single 46GB NVIDIA L40 GPU.

Model	Params	PESQ	CISG	CBAK	COVL
MFSE	2.9M	3.17	4.46	3.79	3.89
Single-Spectrum	2.9M	3.05	4.38	3.72	3.79
w/o TF-DenseBlock	2.9M	3.10	4.41	3.74	3.83
w/o TF-ConvBlock	2.2M	3.04	4.41	3.71	3.78
w/o LHfreqFusion	2.3M	3.12	4.42	3.75	3.80

Table 3: Conduct ablation experiments on major design options

4.3 Performance on VoiceBank+Demand

Compared with other comparative methods on the Voice Bank+Demand dataset, the algorithm proposed in this paper has a good performance on all the metrics, as shown in Table 1, where it can be seen that the MFSE improves on CISG, CBAK, COVL and SSNR by 5%-11.7% over the second place metrics in the table.

In order to analyze the effect of the optimal order on the model performance, we also performed the experiments in Table 2 according to from the max spectral centroid (SC-max), the min spectral centroid (SC-min), and the average spectral centroid (SC-avg), and the experiments found that the SC-avg chosen to have the best metrics of the order voice enhancement.

4.4 Generalizability Performance on DNS Challenge

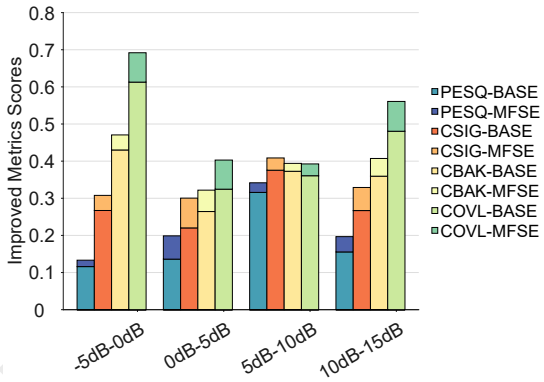

Figure 4: Training on VoiceBank+Demand dataset followed by generalizability test on DNS challenge dataset.

After the model training was completed, to test the difference in generalization ability using the single view model (BASE) and the multi-view model (MFSE), we tested it using the DNS challenge dataset. We evaluated the model trained using the BASE model and the model trained using SC-avg on four different signal-to-noise ratio distributions and showed the improvement in various metrics as shown in the figure 4, which suggests that the approach using multiple views has better generalization ability.

4.5 Ablation Study

We conducted ablation experiments by separately using only the STFT instead of the multi-view method (Single-Spectrum), without the TF-DenseBlock (w/o TF-DenseBlock), without the TF-ConvBlock (w/o TF-ConvBlock), and without the LHFF (w/o LHFF). The experimental results after ablation are shown in Table 3. The analysis of the ablation experiments yielded that the number of model parameters remained basically unchanged after the addition of STFrFT as the multiview input, and PESQ, CISG, CBAK, and COVL were improved by 0.12, 0.07, 0.07, and 0.1, respectively, which verified the validity of the multiview input based on STFrFT as SE at the experimental level. When the TF-DenseBlock is omitted, the number of parameters remains largely unaffected, but all metrics show a degradation, with PESQ experiencing the most significant decline. When the TF-ConvBlock is not used, both PESQ and COVL metrics deteriorate more noticeably. The absence of LHFF leads to an even greater reduction in the PESQ and COVL scores.

We visualized the spectra obtained from the ablation experiments, as shown in Figure 5. In the figure, the yellow boxed region in "w/o TF-ConvBlock" shows poorer recovery compared to MFSE, the green boxed region in "w/o TF-DenseBlock" is less well-recovered than MFSE, and the blue boxed region in "w/o LHFF" also exhibits worse recovery than MFSE.

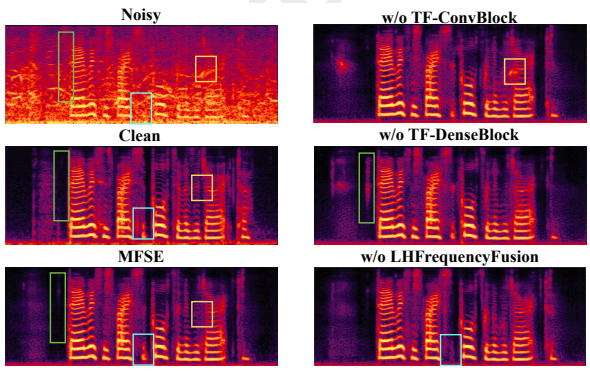

Figure 5: Visualization of ablation experiment results. The green, blue and yellow boxes are used to mark the significantly different locations of the enhanced speech after the ablation experiment.

5 Conclusion

We propose a multi-view SE approach based on short-time fractional Fourier transform. Specifically, it is divided into two parts: multi-view construction and enhanced network. The former uses STFrFT and STFT to construct multi-view input to expand the representation space for speech and noise separation, while the latter uses the UNet architecture to combine the correlation between the low-frequency and high-frequency information of speech to reconstruct speech. Experiments show that our proposed method performs well in the benchmark test VoiceBank+Demand and has some generalization ability under multiple signal-to-noise ratios in another benchmark test DNS challenge.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. T2495251, 62306171, 62406218, 62472267), the Science and Technology Major Project of Shanxi (No. 202201020101006), and Fundamental Research Program of Shanxi Province (Nos. 202303021211023, 202303021221075, 202203021222183).

References

[Chen *et al.*, 2024] Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, 33:1002–1015, 2024.

[Dang *et al.*, 2023] Feng Dang, Qi Hu, Pengyuan Zhang, and Yonghong Yan. Forknet: simultaneous time and time-frequency domain modeling for speech enhancement. *arXiv preprint arXiv:2305.08292*, 2023.

[Défossez *et al.*, 2020] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, pages 3291–3295, Shanghai, China, October 2020. ISCA.

[Fu *et al.*, 2021] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. Metricgan+: An improved version of metricgan for speech enhancement. In *Interspeech*, pages 201–205, Brno, Czechia, August 2021. ISCA.

[Guo *et al.*, 2024a] Qian Guo, Xinyan Liang, Yuhua Qian, Zhihua Cui, and Jie Wen. A progressive skip reasoning fusion method for multi-modal classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 429—437, New York, NY, October 2024. Association for Computing Machinery.

[Guo *et al.*, 2024b] Zilu Guo, Qing Wang, Jun Du, Jia Pan, Qing-Feng Liu, and Chin-Hui Lee. A variance-preserving interpolation approach for diffusion models with applications to single channel speech enhancement and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3025–3038, 2024.

[Hu *et al.*, 2020] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. In *Interspeech*, pages 2472–2476, Shanghai, China, October 2020. ISCA.

[Kong *et al.*, 2022] Zhifeng Kong, Wei Ping, Ambrish Dantrey, and Bryan Catanzaro. Speech denoising in the waveform domain with self-attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7867–7871, Singapore, May 2022. IEEE.

[Li *et al.*, 2022] Andong Li, Chengshi Zheng, Lu Zhang, and Xiaodong Li. Glance and gaze: A collaborative learning framework for single-channel speech enhancement. *Applied Acoustics*, 187:108499, 2022.

[Liang *et al.*, 2022] Xinyan Liang, Yuhua Qian, Qian Guo, Honghong Cheng, and Jiye Liang. AF: An association-based fusion method for multi-modal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9236–9254, 2022.

[Liang *et al.*, 2024] Xinyan Liang, Pinhan Fu, Qian Guo, Keyin Zheng, and Yuhua Qian. DC-NAS: Divide-and-conquer neural architecture search for multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13754–13762, Philadelphia, Pennsylvania, March 2024. Association for the Advancement of Artificial Intelligence.

[Liang *et al.*, 2025] Xinyan Liang, Pinhan Fu, Yuhua Qian, Qian Guo, and Guoqing Liu. Trusted multi-view classification via evolutionary multi-view fusion. In *Proceedings of the 13th International Conference on Learning Representations*, pages 1–14, Singapore, April 2025.

[Lin *et al.*, 2024] Xin Lin, Yang Zhang, and Shiyuan Wang. Mixed t-domain and tf-domain magnitude and phase representations for gan-based speech enhancement. *Scientific Reports*, 14(1):17698, 2024.

[Lu *et al.*, 2021] Yen-Ju Lu, Yu Tsao, and Shinji Watanabe. A study on speech enhancement based on diffusion probabilistic model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 659–666, Tokyo, Japan, December 2021. IEEE.

[Lu *et al.*, 2022] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7402–7406, Singapore, May 2022. IEEE.

[Luo and Mesgarani, 2019] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.

[Maiti and Mandel, 2020] Soumi Maiti and Michael I Mandel. Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 206–210, Barcelona, Spanish, May 2020. IEEE.

[Nareddula *et al.*, 2021] Santhan Kumar Reddy Nareddula, Subrahmanyam Gorthi, and Rama Krishna Sai Subrahmanyam Gorthi. Fusion-net: Time-frequency information fusion y-network for speech enhancement. In *Interspeech*, pages 3360–3364, Brno, Czechia, August 2021. ISCA.

[P-J *et al.*, 2023] K P-J, CH Yang, Sm Siniscalchi, et al. A multi-dimensional deep structured state space approach to speech enhancement using small-footprint models. In *Interspeech*, pages 2453–2457, Dublin, Ireland, August 2023. ISCA.

[Pascual *et al.*, 2017] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative

adversarial network. In *Interspeech*, pages 3642–3646, Stockholm, Sweden, August 2017. ISCA.

[Phan *et al.*, 2020] Huy Phan, Ian V McLoughlin, Lam Pham, Oliver Y Chén, Philipp Koch, Maarten De Vos, and Alfred Mertins. Improving gans for speech enhancement. *IEEE Signal Processing Letters*, 27:1700–1704, 2020.

[Routray and Mao, 2022] Sidheswar Routray and Qirong Mao. Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network. *Computer Speech & Language*, 71:101270, 2022.

[Shi *et al.*, 2023] Hao Shi, Masato Mimura, Longbiao Wang, Jianwu Dang, and Tatsuya Kawahara. Time-domain speech enhancement assisted by multi-resolution frequency encoder and decoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes island, Greece, June 2023. IEEE.

[Shin *et al.*, 2023] Wooseok Shin, Byung Hoon Lee, Jin Sob Kim, Hyun Joon Park, and Sung Won Han. MetricGAN-OKD: Multi-metric optimization of MetricGAN via online knowledge distillation for speech enhancement. In *International Conference on Machine Learning*, pages 31521–31538, Seoul, South Korea, July 2023. PMLR.

[Son *et al.*, 2024] Hosung Son, Min-jung Shin, Minji Cho, Joonsoo Kim, Kug-jin Yun, and Suk-Ju Kang. Cmvde: Consistent multi-view video depth estimation via geometric-temporal coupling approach. *IEEE Transactions on Multimedia*, 26:9710–9721, 2024.

[Soni *et al.*, 2018] Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5039–5043, Calgary, Canada, April 2018. IEEE.

[Strauss and Edler, 2021] Martin Strauss and Bernd Edler. A flow-based neural network for time domain speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758, Toronto, Canada, June 2021. IEEE.

[Sun *et al.*, 2024] Linhui Sun, Shuo Yuan, Aifei Gong, Lei Ye, and Eng Siong Chng. Dual-branch modeling based on state-space model for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1457–1467, 2024.

[Tai *et al.*, 2023a] Wenxin Tai, Yue Lei, Fan Zhou, Goce Trajcevski, and Ting Zhong. Dose: Diffusion dropout with adaptive prior for speech enhancement. In *Neural Information Processing Systems*, pages 40272–40293, New Orleans, Louisiana, December 2023. Curran Associates, Inc.

[Tai *et al.*, 2023b] Wenxin Tai, Fan Zhou, Goce Trajcevski, and Ting Zhong. Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13627–13635, Washington, DC, June 2023. AAAI.

[Tang *et al.*, 2021] 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*, pages 3816–3822, Yokohama, Japan, January 2021. International Joint Conferences on Artificial Intelligence.

[Tian *et al.*, 2024] Ye Tian, Zhe Wang, Jianguo Sun, and Liguo Zhang. Time-frequency domain fusion enhancement for audio super-resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2879–2887, Melbourne, Australia, October 2024. Association for Computing Machinery.

[Wang *et al.*, 2024] Lan Wang, Haitao Zhang, Youli Qiu, Yanji Jiang, Hao Dong, and Pengfei Guo. Improved speech separation via dual-domain joint encoder in time-domain networks. In *2024 International Conference on Electronic Engineering and Information Systems (EEISS)*, pages 233–239, Changsha, China, January 2024. IEEE.

[Wu *et al.*, 2025] Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 55–71, Milano, Italy, September 2025. Springer.

[Xue *et al.*, 2024] Yuanliang Xue, Guodong Jin, Tao Shen, Lining Tan, Nian Wang, Jing Gao, and Lianfeng Wang. Consistent representation mining for multi-drone single object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10845–10859, 2024.

[Yu *et al.*, 2020] Cheng Yu, Kuo-Hsuan Hung, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung. Time-domain multimodal bone/air conducted speech enhancement. *IEEE Signal Processing Letters*, 27:1035–1039, 2020.

[Zheng *et al.*, 2024] Qun Zheng, Xihong Yang, Siwei Wang, Xinru An, and Qi Liu. Asymmetric double-winged multi-view clustering network for exploring diverse and consistent information. *Neural Networks*, 179:106563, 2024.