An Association-based Fusion Method for Speech Enhancement

Shijie Wang[1] , **Qian Guo**[2] , **Lu Chen**[1] , **Liang Du**[1] , **Zikun Jin**[1] , **Zhian Yuan**[1] , **Xinyan Liang**[1*]
[1]Institute of Big Data Science and Industry, Key Laboratory of Evolutionary Science Intelligence of Shanxi Province, Shanxi University, Taiyuan 030006, China
[2] Shanxi Key Laboratory of Big Data Analysis and Parallel Computing, School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
wshijie0@163.com, czguoqian@163.com, chenlu@sxu.edu.cn, duliang@sxu.edu.cn,
jinzikun@163.com, jczhianyuan@163.com, liangxinyan48@163.com

Abstract

Deep learning-based speech enhancement (SE) methods predominantly draw upon two architectural frameworks: generative adversarial networks and diffusion models. In the realm of SE, capturing the local and global relations between signal frames is crucial for the success of these methods. These frameworks typically employ a UNet as their foundational backbone, integrating Long Short-Term Memory (LSTM) networks or attention mechanisms within the UNet to effectively model both local and global signal relations. However, the coupled relation modeling way may not fully harness the potential of these relations. In this paper, we propose a novel, decoupled Association-based Fusion Speech Enhancement method (AFSE). AFSE first constructs a graph that encapsulates the association between each time window of the speech signal, and then models the global relations between frames by fusing the features of these time windows in a manner akin to graph neural networks. Furthermore, AFSE leverages a UNet with dilated convolutions to model the local relations, enabling the network to maintain a high-resolution representation while benefiting from a wider receptive field. Experimental results demonstrate that the AFSE method significantly improves performance in speech enhancement tasks, validating the effectiveness and superiority of our approach. The code is available at https://github.com/jie019/AFSE_IJCAI2025.

1 Introduction

Speech enhancement (SE) aims to improve the quality of a speech signal by reducing noise and other distortions, thereby making it more suitable for further analysis or utilization [O'Shaughnessy, 2024]. SE has become a critical process in various applications including communications, hearing aids, and medical applications [Chhetri *et al.*, 2023].

Recent advancements in deep learning (DL) have brought significant improvements to SE. DL-based approaches leverage the power of neural networks to model the complex relations between noisy and clean speech signals, showing remarkable success in various scenarios [Tai *et al.*, 2021]. These methods typically focus on minimizing the overall difference between the target speech and its denoised speech using Lp-norm distance [Li *et al.*, 2021]. However, these discriminative models inherently lack the ability to capture the underlying structure of speech signals and their variability in real-world environments [Phan *et al.*, 2020]. Specifically, discriminative models primarily focus on feature mapping between noisy and clean signals, ignoring the higher-order relations within the signal where such relations are critical. As a result, they often struggle to generalize to unseen noise conditions or complex acoustic scenarios.

Generative models have been introduced to improve SE by learning the distribution of clean speech signals and generating plausible denoised outputs [Welker *et al.*, 2022]. The type of methods are mainly based on two architectural frameworks: generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs). It is well known that capturing the local and global relations between signal frames is crucial for the success of these methods. These frameworks typically employ a UNet as their foundational backbone, integrating Long Short-Term Memory (LSTM) networks or attention mechanisms within the backbone to effectively model both local and global signal relations. However, the coupled relation modeling way may not fully harness the potential of these relations.

Association is often used to characterize the relation strength between multiple variables in data. Its ability for relation modeling has been proved in multi-modal classification [Liang *et al.*, 2022], representation learning [Liang *et al.*, 2025]. The integration of association-based methods with advanced DL algorithms has the potential to significantly improve the performance of signal enhancement systems. In this paper, we propose a decoupled relation modeling method, called Association-based Fusion Speech Enhancement (AFSE) method. Specifically, AFSE first constructs a graph that encapsulates the association between each time window of the speech signal, and then models the global relations between frames by fusing the features of these time windows in a manner akin to graph neural networks. Furthermore, AFSE leverages a UNet with dilated convolutions to model the local relations, enabling the network to maintain a

[*] Corresponding author

high-resolution representation while benefiting from a wider receptive field. The contributions of our work are as follows:

- We propose an effective method for modeling global relation between frames by combining the association with graph neural network.
- An association-based fusion speech enhancement (AFSE) method is proposed, which achieves a better usage of relation between signal frames in an decoupled relation modeling strategy.
- The extensive comparison experiments on two public datasets show that AFSE achieves competitive performance with fewer model parameters compared to the state-of-the-art speech enhancement methods.

2 Related Work

2.1 Speech Enhancement (SE)

Recently, SE methods driven by deep learning are mainly based on the generative adversarial networks (GANs) and diffusion models.

GANs-based SE: GAN is a representative generative method and has been adopted in various domains. In the field of SE, GANs have received widespread attention in recent years and have achieved remarkable results. SEGAN [Pascual *et al.*, 2017] is one of the earliest GAN models applied to speech enhancement. It improves the clarity of speech by using binary labels (such as real or fake) in the discriminator to distinguish enhanced speech from original speech. Subsequent work, such as FSEGAN [Donahue *et al.*, 2018], further optimized this structure and proposed an improved generative model for speech enhancement tasks. However, these early methods rely on discrete labels for training and may not be able to directly optimize speech quality evaluation indicators related to human perception. To this end, MetricGAN [Fu *et al.*, 2019] introduced continuous labels based on evaluation indicators and improved the training method of GAN. MetricGAN-OKD [Shin *et al.*, 2023] further optimizes this process and achieves multi-indicator optimization through online knowledge distillation, which improves the effect of speech enhancement.

Diffusion-based SE: In the field of SE, DDPMs [Ho *et al.*, 2020] as a class of generative model have made significant progress in recent years. The diffusion model gradually adds noise to the data through a fixed forward process, and iteratively de-noises through a parameterized reverse process to generate samples from the noise. Compared to other generative models, diffusion models demonstrate excellent sample quality and a simpler training process, while placing few restrictions on the model architecture. The application of specific diffusion models to SE includes a variety of methods. For example, DiffuSE [Lu *et al.*, 2021] is an SE method based on a diffusion model that generates high-quality speech samples by step-by-step de-noising. CDiffuSE [Lu *et al.*, 2022] further optimizes this process by combining conditional generation techniques to improve model generalization and enhancement. In addition, DR-DiffuSE [Tai *et al.*, 2023b] significantly improves noise reduction and voice quality by introducing fast sampling technology and refining networks.

In summary, GANs often suffer from training instability and mode collapse. DDPMs are computationally expensive due to the iterative denoising steps. Furthermore, they model the local and global relationships of data in a coupled manner, which may not fully capture the intrinsic relationships within the data. By adopting a decoupled modeling strategy and explicitly modeling the relationships between global frames, we can better capture the intrinsic relationships within the data.

2.2 Graph Neural Networks (GNNs)

We utilize the message-passing mechanism of GNNs to achieve global frame fusion after obtaining the global association between frames. GNNs directly operate on the graph structure and aggregate information via a message-passing mechanism [Zhou *et al.*, 2020].

The k-th layer of the GNN message-passing scheme is defined as:

$$h_v^{(k)} = \text{Cat}\left(h_v^{(k-1)}, \text{Agg}\left(\left\{h_u^{(k-1)}, e_{uv} \mid u \in N_v\right\}\right)\right) \quad (1)$$

where Cat$(\cdot)$, Agg$(\cdot)$ and N_v denote the concatenation, aggregate functions, and immediate neighbors of node v, respectively; $h_v^{(k)}$ is the representation vector of node v in the k-th layer, and e_{uv} is the edge vector between nodes u and v.

3 The Proposed AFSE

The whole framework of AFSE is shown in Fig. 1, which consists of (1) frame representation learning, (2) global frame fusion with association between frames, (3) local frame fusion with dilated convolution, (4) loss function.

3.1 Frame Representation Learning

To improve the effectiveness of modeling relationships, we first extract feature representations for each frame. This way enhances the ability to capture detailed characteristics. Specifically, the mixture signal in the time domain is first transformed into the time-frequency (T-F) domain using the Short-Time Fourier Transform (STFT), and its formulation in T-F domain is written as:

$$X_{k,l} = Y_{k,l} + N_{k,l} \quad (2)$$

where $X \in \mathbb{R}^{2\times L\times K}$, $Y \in \mathbb{R}^{2\times L\times K}$, and $N \in \mathbb{R}^{2\times L\times K}$ denote the complex-valued noisy signal, clean signal, and noise, respectively. 2 is the number of channels, L is the number of time steps, and K is the number of frequency bins. $k \in \{1, \cdots, K\}$ and $l \in \{1, \cdots, L\}$ is the index of frequency bin and time, respectively.

Since the global distribution of speech amplitudes obtained with STFT is typically heavy-tailed, the information visible in untransformed spectrograms is dominated by only a small portion of bins. Inspired by recent results from [Ju *et al.*, 2022] that power compression can decrease the dynamic range of the spectrum and improve the significance of low-energy regions with more informative speech components, we apply an amplitude compression to adjust the energy distribution to achieve approximate normalization. The transformation [Tai *et al.*, 2023b] is defined as follows:

$$\tilde{X} = \sqrt{|X|}e^{i\angle X}, \quad \tilde{X} \in \mathbb{R}^{2\times L\times K} \quad (3)$$

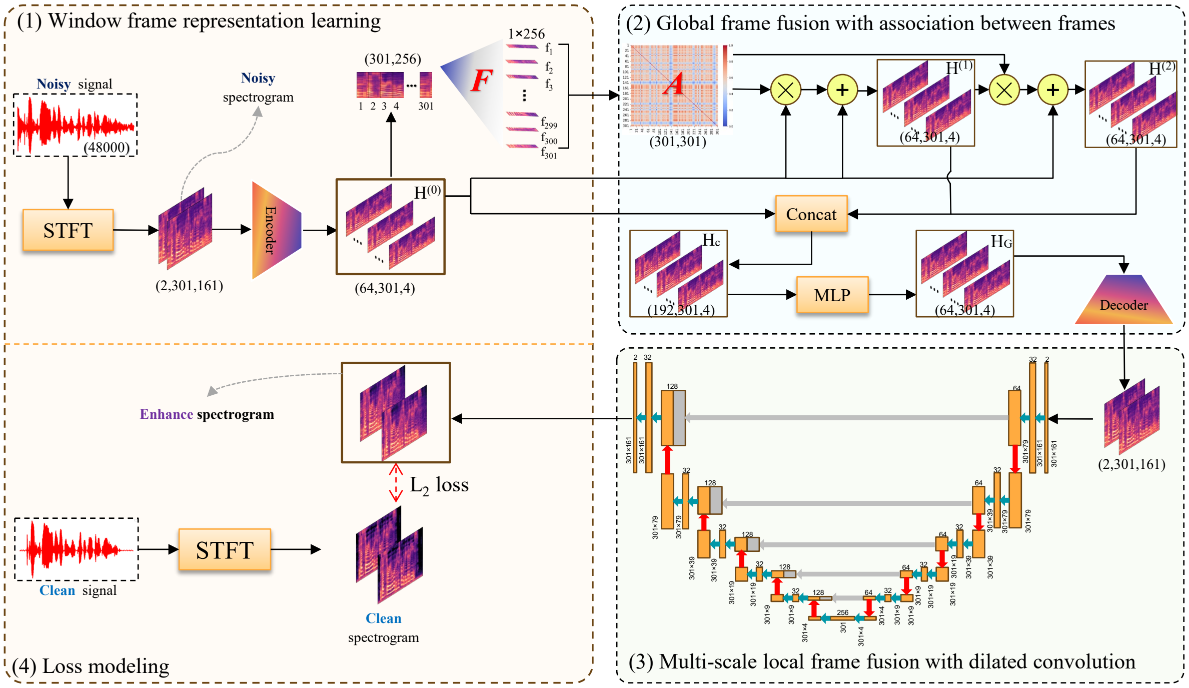

Figure 1: The whole framework of AFSE

where $\angle X$ denotes the phase, i denotes imaginary number, and e is a natural constant. Next, the noisy spectrum $\tilde{X}$ is passed into an encoder Enc to obtain an intermediate feature representation H:

$$H = \text{Enc}(\tilde{X}), \quad H \in \mathbb{R}^{C\times L\times M} \tag{4}$$

where C is the number of channels and M denotes the encoded feature dimension. To model the relationship between frames, we convert the features H into features F:

$$F = Reshape(H, (L, C \times M)), \quad F \in \mathbb{R}^{L\times(C\times M)} \tag{5}$$

where $Reshape(H, size)$ denotes the shape of tensor H is changed $size$. F contains L frame feature vectors, where each f_t is defined as:

$$f_t \in \mathbb{R}^{(C\times M)}, \quad t \in \{1, 2, \ldots, L\} \tag{6}$$

3.2 Global Frame Fusion with Association

To enhance the signal, we model the relationships within the signal across different time windows. The association-based method leverages these relationships to improve the signal quality by reducing noise while preserving the essential characteristics of the original signal. This section mainly introduces how to perform global frame fusion with association in a GNN-like work manner. To this end, we model the association between frames as graph.

Modeling Association Graph Between Frames. We get the representation of each frame by the frame representation learning, and we use association analysis to get the relationship graph between frames. Specifically, we construct a dense graph for each signal, where each frame is represented as a node, and any two nodes are connected with an edge based on their association. Thus, the graph can be formulated as $G(V, E)$, where V is the set of vertices and E is the set of edges. Here, $H^{(0)} = Reshape(H, (L, C, M)) \in \mathbb{R}^{L\times C\times M}$ is the embeddings denoting the nodes in the graph, $A \in \mathbb{R}^{L\times L}$ be the adjacency matrix of the graph $G(V, E)$. Each of its elements represents an edge which is defined as:

$$A_{ij} = 1 - \frac{\|f_i - f_j\|_2}{\max\limits_{i,j} \|f_i - f_j\|_2} \tag{7}$$

where a greater value means a stronger correlation, and vice versa means a weaker correlation. The constructed graph has a global perception and can mine the relationship between frames because A_{ij} measures the connection weight between any two frames in the signals. Therefore, we named the graph $G(V, E)$ as the global graph.

Fusion with Association. After constructing this densely connected graph, we update the node embedding by correlation fusion in a residual way, formulated as:

$$H^{(k)} = \alpha H^{(k-1)} + (1-\alpha)(\tilde{A}H^{(k-1)}) \tag{8}$$

where $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the normalized adjacency matrix. D denote the diagonal degree matrix, where $D_{ii} = \sum_{j=1}^{n} A_{ij}$. α is the mixing coefficient, and $H^{(k)}$ denotes the output embedding of the k-th graph convolution layer.

Next we combine embeddings from different layers to capture multi-level information. After k graph convolution layers, the concatenated embeddings are:

$$H_c = \text{Cat}(H^{(0)}, H^{(1)}, \ldots, H^{(k)}) \tag{9}$$

where Cat denotes the tensor concatenation operation. Then we transforms the concatenated embedding to produce the final output, incorporating non-linear relationships. The final output is obtained through a Multi-Layer Perceptron (MLP):

$$H_G = \sigma\left(\mathrm{MLP}(H_c)\right) \tag{10}$$

where σ is an activation function. Finally, we restores the transformed output to its original dimensions through a decoder Dec for subsequent local relation fusion.

$$X_G = \mathrm{Dec}(H_G), \quad X_G \in \mathbb{R}^{2\times L\times K} \tag{11}$$

3.3 Multi-scale Local Frame Fusion

In this section, we focus on the fusion of local frames at multiple scales using a specialized module that can leverage the relation among different temporal scales to boost the speech recovery. The model uses the complex spectrum as input and follows a standard encoder-decoder U-Net architecture.

Encoder. Instead of regular convolutional layers, the encoder and decoder employ five BiConvGLU (Bi-directional Convolutional Gated Linear Units) layers. These BiConvGLUs are designed to capture local spectral-temporal patterns while compressing the spectral features. The use of BiConvGLU enhances the model's capacity to handle complex features in both time and frequency domains.

Each convolutional block within the model uses a kernel size of (2, 3) along the time and frequency axes, except for the first block, which uses a kernel size of (2, 5). The stride is set to (1, 2), meaning that while the frequency size is gradually halved, the time size remains unchanged to ensure real-time processing capability. After each convolution, the model applies Instance Normalization (InstanceNorm) and a Parametric ReLU (PReLU) activation function, which help to stabilize and accelerate the training process.

Bottleneck Layer. In the bottleneck layer, which serves as the transition between the encoder and decoder, the model incorporates Stacked Temporal Convolution Modules (S-TCMs) as proposed by [Li *et al.*, 2021]. These S-TCMs are crucial for capturing long-range dependencies in the time domain, which is essential for tasks that require understanding temporal dynamics, such as speech enhancement.

The model integrates three groups of S-TCMs as the sequential module, each group stacks six S-TCM units with exponentially increasing dilation rates (1, 2, 4, 8, 16, 32). This design allows the network to cover a large temporal receptive field, which is vital for effectively leveraging relations across different temporal scales. This approach significantly boosts the network's ability to recover detailed speech information across various time scales.

Algorithm 1 AFSE Training Procedure

Input: the number of samples M, total epochs N, loss functions Q, hops K, activation function σ, concatenation operation Cat, mixing coefficient α
Parameter: **W**: a trainable weight matrix , **U**: Unet, **Enc**: Encoder, **Dec**: Decoder

```
for epoch = 1 to Total epochs $N$ do
    for $m$ = 1 to $M$ do
        Sample a pair of noisy and clean speeches $(X, Y)$
        $H^0 = \mathbf{Enc}(X)$
        The adjacency matrix A is obtained by Eq. 7
        $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$
        $H_c = H^0$
        for $k$ = 1 to $K$ do
            $H^{(k)} = (1-\alpha)\tilde{A}H^{(k-1)} + \alpha H^{(k-1)}$
            $H_c = Cat(H_c, H^k)$
        end for
        $X_G = \mathbf{Dec}(\sigma(H_c\mathbf{W}))$
        $\hat{X} = \mathbf{U}(X_G)$
        $loss = Q(\hat{X}, Y)$
        Update the learnable parameters W, U, Enc, Dec
    end for
end for
```

Decoder. The decoder mirrors the encoder's structure but operates in reverse. It gradually restores the compressed features to their original size using five BiConvGLU layers with a fixed channel dimension of 64. The stride remains to be set to (1, 2), and the kernel size matches that of the encoder blocks, except for the last BiConvGLU, which uses a kernel size of (2, 5). The model also incorporates skip connections between corresponding encoder and decoder layers, which help to mitigate information loss and enhance the quality of the reconstructed output.

$$\hat{X} = \mathrm{U}(X_G), \hat{X} \in \mathbb{R}^{2\times L\times K} \tag{12}$$

where U represents Unet. $\hat{X}$ is the enhanced spectrum. Finally, we use the Inverse Short-Time Fourier Transform (ISTFT) to convert the enhanced speech from the time-frequency domain back into the time domain.

The whole algorithm is shown in Algorithm 1.

3.4 Loss Function

The loss function used to train the network aims to minimize the difference between the RI (real and imaginary) of the estimated spectrum and the target spectrum, as well as the magnitude of the estimated spectrum and the target spectrum.

RI components loss is defined as:

$$L_{\mathrm{RI}}(\phi) = \|\hat{X}_r - Y_r\|_F^2 + \|\hat{X}_i - Y_i\|_F^2 \tag{13}$$

This term measures the Frobenius norm of the difference between the RI $\hat{X}_r$, $\hat{X}_i$ components of the estimated spectrum and the target spectrum Y_r and Y_i.

Magnitude loss is defined as:

$$L_{\mathrm{Mag}}(\phi) = \left\| \sqrt{|\hat{X}_r|^2 + |\hat{X}_i|^2} - \sqrt{|Y_r|^2 + |Y_i|^2} \right\|_F^2 \tag{14}$$

This term measures the difference in magnitude between the estimated and target spectra. By focusing on the magnitude, this loss helps the model to better capture the overall energy distribution of the signal.

The total loss is a weighted sum of the RI components loss and the magnitude loss, and is defined as:

$$L(\phi) = \lambda_{\mathrm{RI}} L_{\mathrm{RI}}(\phi) + \lambda_{\mathrm{Mag}} L_{\mathrm{Mag}}(\phi) \tag{15}$$

Methods	Pub./Year	Do.	G/L	Params↓	CSIG↑	CBAK↑	COVL↑	PESQ↑	SSNR↑	STOI(%)↑
Unprocessed	-	-	-	-	3.35	2.44	2.63	1.97	1.68	92.1
DEMUCS	Interspeech/2020	T	G&L	18.87M	4.22	3.25	3.52	2.93	-	-
CleanUNet	ICASSP/2022	T	G&L	39.77M	4.32	3.41	3.63	2.88	-	-
GaGNet	Appl Acoust/2022	T-F	L	5.94M	4.26	3.45	3.59	2.94	-	-
TaylorSENet	IJCAI/2022	T-F	L	5.40M	4.31	3.08	3.64	2.90	2.40	**95.3**
S4ND U-Net	Interspeech/2023	T-F	G	**0.75M**	4.37	3.56	3.70	2.99	-	-
Dual-S4D	TASLP/2024	T-F	G&L	10.8M	3.94	3.00	3.23	2.55	-	93.4
DiffuSE	APSIPA ASC/2021	T	L	-	3.66	2.83	3.03	2.44	-	-
CDiffuSE	ICASSP/2022	T	L	-	3.72	2.91	3.10	2.52	5.28	91.4
DR-DiffuSE	AAAI/2023	T-F	L	3.55M	4.38	3.57	3.76	3.09	9.52	94.9
DOSE	NeurIPS/2023	T-F	L	-	3.83	3.27	3.19	2.56	-	93.6
VPIDM	TASLP/2024	T-F	L	-	4.23	3.53	3.70	3.16	-	-
SEGAN	Interspeech/2017	T	L	-	3.48	2.94	2.80	2.16	7.73	-
MMSEGAN	ICASSP/2018	T-F	L	-	3.80	3.12	3.14	2.53	-	93.0
MetricGAN	ICML/2019	T	G	-	3.99	3.18	3.42	2.86	-	-
MetricGAN+	Interspeech/2021	T-F	G	-	4.14	3.16	3.64	3.15	-	-
DSEGAN	SPL/2020	T	L	-	3.46	3.11	2.90	2.39	8.72	93.2
MGAN-OKDv2	ICML/2023	T-F	G	0.82M	4.17	3.13	3.64	3.12	-	-
AFSE(Ours)	-	T-F	G&L	2.09M	**4.44**	**3.66**	**3.85**	**3.18**	**10.17**	95.0

Table 1: Comparison results on VoiceBank-DEMAND, where ↑ and ↓ denote that the larger/smaller the value is, the better the performance is, respectively.

where λ_{RI} and λ_{Mag} are weighting hyper-parameters and are set to 0.5 and 0.5 with empirical trials, respectively. This combined loss ensures that the network learns to accurately predict both the phase and magnitude of the signal, leading to more effective speech enhancement.

4 Experiment

To evaluate the proposed AFSE, we use two datasets: VoiceBank-DEMAND (VBD) and DNS Challenge.

Training Configuration. We sample all the utterances at 16 kHz. The window size is set as 20 ms, with 50% overlap between adjacent frames. 320-point FFT is utilized, leading to 161-D in the feature axis. The model is trained on Pytorch platform with a NVIDIA RTX 4090 GPU. We use the Adam optimizer with a batch size of 8 to train the proposed model, and the learning rate is initialized as $1e^{-3}$. Moreover, we train the model by 60 epochs for two datasets.

Evaluation Metrics. We use the following metrics to evaluate SE performance: narrow-band (NB) [Rix *et al.*, 2001] and wide-band (WB) perceptual evaluation of speech quality (PESQ) [Rec, 2005] for speech quality, short-time objective intelligibility (STOI) [Taal *et al.*, 2010] for intelligibility, segmental signal-to-noise ratio(SSNR), scale-invariant signal-to-noise ratio (SISNR) [Roux *et al.*, 2019] and the mean opinion score (MOS) prediction of the speech signal distortion (CSIG) [Hu and Loizou, 2008] for speech distortion, the MOS prediction of the intrusiveness of background noise (CBAK) [Hu and Loizou, 2008] for noise intrusion, the MOS prediction of the overall effect (COVL) [Hu and Loizou, 2008] for overall signal quality.

4.1 VoiceBank-DEMAND

The VoiceBank-DEMAND [Valentini-Botinhao *et al.*, 2016] consists of 30 speakers, with 28 used for the training/validation dataset and the remaining two for the test dataset. The training/validation and test datasets contain 11,572 utterances with four signal-to-noise ratio (SNR) (15, 10, 5, and 0 dB) levels and 824 utterances with four SNR (17.5, 12.5, 7.5, and 2.5 dB) levels, respectively. In the dataset, we compare our model with seventeen state-of-the-art methods which are classified into following three categories:

- GAN-based methods: SEGAN [Pascual *et al.*, 2017], MMSEGAN [Soni *et al.*, 2018], MetricGAN [Fu *et al.*, 2019], MetricGAN+ [Fu *et al.*, 2021], DSEGAN [Phan *et al.*, 2020] and MGAN-OKDv2 [Shin *et al.*, 2023].
- Diffusion-based methods: DiffuSE [Lu *et al.*, 2021], CDiffuSE [Lu *et al.*, 2022], DR-DiffuSE [Tai *et al.*, 2023b], DOSE [Tai *et al.*, 2023a] and VPIDM [Guo *et al.*, 2024].
- Other methods: DEMUCS [Défossez *et al.*, 2020], CleanUNet [Kong *et al.*, 2022], GaGNet [Li *et al.*, 2022b], TaylorSENet [Li *et al.*, 2022a], S4ND U-Net [Ku *et al.*, 2023] and Dual-S4D [Sun *et al.*, 2024].

We report two types of metrics including model parameters and enhancing performance the latter one contains CSIG, CBAK, COVL, PESQ, SSNR, STOI. The experiment results are reported in Table 1, where "T" and "F" denote time domain and spectrum domain, respectively; "G" and "L" denote global relation and local relation, respectively. From Table 1, the following conclusions can be drawn.

- AFSE achieves the best results in terms of all enhancing performance except for STOI. For model parameters, the S4ND U-Net ranks the first, followed by AFSE. However, S4ND U-Net ranks only the third, third, fourth, fifth in CSIG, CBAK, COVL, PESQ, respectively.

Methods	Pub./Year	Do.	G/L	Params↓	PESQ-WB↑	PESQ-NB↑	STOI(%)↑	SISNR(dB)↑
Noisy	-	-	-	-	1.58	2.45	91.52	9.07
NSNet	ICASSP/2020	T-F	G	-	2.15	2.87	94.47	15.61
DTLN	Interspeech/2020	T-F	G&L	-	-	3.04	94.76	16.34
DCCRN	Interspeech/2020	T-F	G&L	3.67M	-	3.27	-	-
PoCoNet	Interspeech/2020	T-F	G&L	-	2.75	-	-	-
FullSubNet	ICASSP/2021	T-F	G&L	5.64M	2.78	3.31	96.11	17.29
CTS-Net	TASLP/2021	T-F	L	4.35M	2.94	3.42	96.66	17.99
GaGNet(3000h)	Appl Acoust/2022	T-F	L	5.94M	3.17	3.56	97.13	18.91
FAF-Net(500h)	AAAI/2022	T-F	L	-	2.93	3.44	96.37	-
TaylorSENet(3000h)	IJCAI/2022	T-F	L	5.40M	3.22	3.59	97.36	19.15
AFSE(500h)	-	T-F	G&L	**2.09M**	3.23	3.67	97.34	18.63
AFSE(800h)	-	T-F	G&L	**2.09M**	3.27	3.69	97.40	18.95
AFSE(3000h)	-	T-F	G&L	**2.09M**	**3.29**	**3.71**	**97.50**	**19.22**

Table 2: Comparison with state-of-the-art methods on DNS dataset, where $A(t)$ denotes the algorithm A is trained using the data of t hours.

- On the one hand, compared with DR-DiffuSE, our method has achieved comprehensive advantages. The reason may be that the core units of DR-DiffuSE is convolution, cannot model the global relation between frames. On the other hand, the enhanced performance of S4ND U-Net is inferior to our method, possibly because of the lack of local relationship modeling. This indicates that both of the local and global relation are important for SE.
- Compared to DEMUCS and CleanUNet, which model both local and global relationships in a coulped manner, our superior performance may be attributed to the decoupled fusion of global and local relationships.

In summary, our method combines high performance with a low number of parameters. It achieves excellent results in speech quality and noise suppression. By maintaining a lower parameter count while delivering superior performance, our method is suited to practical applications where both quality and resource constraints are critical.

4.2 DNS Challenge

The Interspeech 2020 DNS challenge dataset [Reddy *et al.*, 2020] is a large speech enhancement dataset. The clean speechs are collected from Librivox and totally includes 500 hours utterances from 2150 speakers. The noise clips are from Audioset and Freesound, including 60000 noise clips with 150 classes. Following [Zheng *et al.*, 2021], we synthesize 500 hours noisy clips with SNR levels of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB for training. For evaluation, we use another 150 noisy clips from the test set without reverberation. The testing SNR levels are randomly distributed in the range from 0 dB to 20 dB.

For this dataset, we compare against nine state-of-the-art methods: LSTM based methods FullSubNet [Hao *et al.*, 2021] and DTLN [Westhausen and Meyer, 2020], U-Net based methods DCCRN [Hu *et al.*, 2020] and PocoNet [Isik *et al.*, 2020]. NSNet [Xia *et al.*, 2020] utilizes weighted speech distortion losses, and CTS-Net [Li *et al.*, 2021] is a two-stage complex spectral mapping method. FAF-Net [Yue *et al.*, 2022] leverages a reference-based feature alignment and fusion strategy to enhance speech quality. GaGNet [Li *et al.*, 2022b] utilizes a dual-path glance-and-gaze structure for collaborative spectrum estimation. TaylorSENet [Li *et al.*, 2022a] employs a Taylor-unfolding framework for decoupled magnitude and residual modeling. The results of NSNet are quoted from FullSubNet, and others are directly quoted from their respective papers.

To verify the superiority of the proposed SE system in more complex acoustic scenarios, we present the results on the Interspeech 2020 DNS-Challenge corpus, as shown in Table 2. The evaluation metrics include Params, WB-PESQ, NB-PESQ, STOI, and SISNR. From the results in Table 2, we can draw the following conclusions: (1) When the training data is 3000 hours, AFSE achieves the best results in terms of all evaluation metrics. With the same amount of training data, our performance performs better than TaylorSENet that ranks the first place among the comparison methods, but the number of parameters of AFSE is less than half of ones of TaylorSENet; (2) The gradual improvement in performance with increased training set size (500h, 800h, 3000h) indicates that AFSE benefits more from larger datasets, further showing the scalability and effectiveness of our AFSE; (3) Compared with the local fusion method such as FAF-Net, TaylorSENet, AFSE achieves better overall performance, possibly due to its more comprehensive integration of global and local information, reinforcing the importance of this dual-fusion strategy.

These results furthermore validate the effectiveness and superiority of our approach, showcasing its potential to set a new benchmark in the field of speech enhancement.

4.3 Further Analysis

In the part, we further analysis our model from six aspects: each module effectiveness, hyper-parameter sensibility, impact of different SNR values and effect of different lengths.

Ablation Study. This subsection is to investigate the effect of different components in AFSE via ablation experiments. The components include frame representation, Graph and Unet. In the ablation study, we compare AFSE with three its degeneration models: **Case 1**: Remove the frame representation module; **Case 2**: Remove the Graph module; and **Case 3**: Remove the Unet module.

The results are shown in Table 3. From the results, we

Methods	PESQ-WB	PESQ-NB	STOI(%)	SISNR(dB)	CSIG	CBAK	COVL
AFSE	**3.23**	**3.67**	**97.34**	**18.63**	**3.02**	**3.91**	**3.15**
Case 1	3.14	3.62	97.01	18.49	2.93	3.87	3.06
Case 2	3.14	3.62	97.02	18.50	2.94	3.87	3.06
Case 3	2.61	3.28	95.25	17.17	2.53	3.50	2.59

Table 3: Ablation results on DNS.

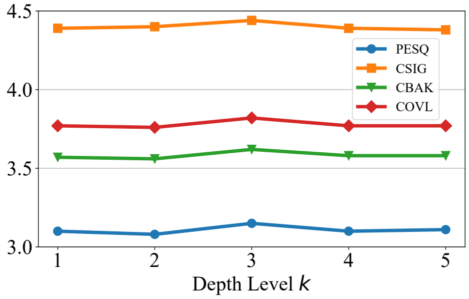

Figure 2: Sensitivity analysis of hyper-parameters k on VBD.

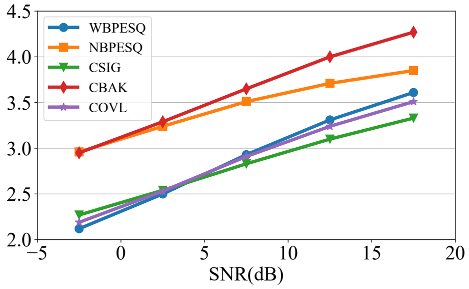

Figure 3: Enhancement results with the inputs setting to different SNR levels on DNS.

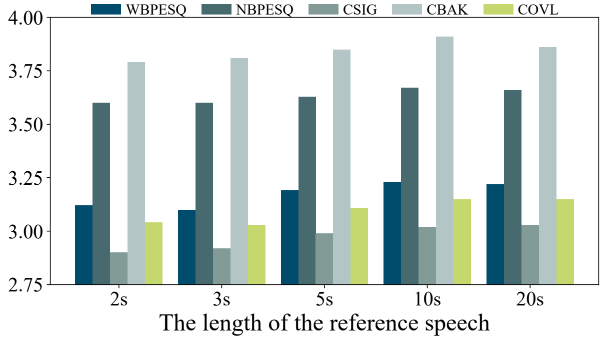

Figure 4: The speech enhancement results using references with different time lengths on DNS.

can observe that (1) Modeling the global relationships based on representations is better than that based on raw features; (2) Removing the graph structure leads to a performance decline, highlighting the importance of learning global associations between frames in speech enhancement. (3) Based on the results of removing the Unet module, it can be concluded that learning local correlation significantly contributes to improving model performance. In summary, each component in AFSE plays a key role.

Hyper-Parameters. We conducted an in-depth analysis of key hyper-parameters within our model. Fig. 2 visually demonstrates the model's performance variations across distinct k orders, ranging from 1 to 5. Our assessment encompasses the VoiceBank-DEMAND, evaluating the PESQ, CSIG, CBAK, COVL as the metric. From Fig. 2, we draw the following key observations: In general, opting for a relatively smaller hop order, such as 3, yields improved performance.

Analysis on Different SNR Values. We evaluate the performance of the model at different SNR levels on the DNS dataset. Since the speeches in different groups are different, we further generate a new DNS test set by adding noise with different SNR values to all the clean speeches in the test set. We classify its noise level (SNR ranging from -5 to 20 dB) into five groups and give the average WBPESQ, NBPESQ, CSIG, CBAK, COVL result for each group. The results in Fig. 3 are produced on the new DNS test set. This is evident from the upward trends across all evaluation metrics (WB-PESQ, NB-PESQ, CSIG, CBAK, COVL) that the performance of the AFSE improves as the SNR increases.

Analysis on Different Lengths. We evaluate the enhancement results using references with different lengths. The result is shown in Fig. 4. The result indicates that there is a noticeable improvement in all metrics as the reference length increases from 3s to 10s. This suggests that longer reference speech provides more contextual information, enabling better enhancement. However, the performance gains seem to plateau after a certain length (10s to 20s), indicating that beyond a certain point, increasing the reference length does not significantly boost enhancement quality. Considering the trade off between computing complexity and performance, we utilize the reference with 10s in our experiments.

5 Conclusions

In this paper, we have proposed an efficient association-based fusion method for speech signal enhancement that integrates global and local fusion techniques. This method efficiently captures both broad contextual relationships and detailed local patterns, achieving a better enhanced speech signal quality. Extensive experiments have demonstrated the superiority of this dual-fusion approach in improving speech enhancement performance. Our results suggest that combining global and local fusion strategies can be particularly effective in addressing the challenges of complex auditory environments. In the future, we plan to explore more refined strategies for balancing and optimizing these fusion techniques to further advance the field of speech enhancement.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 62306171, T2495251, 62406218, T2495253, 62373233, 62376146), the Science and Technology Major Project of Shanxi (No. 202201020101006), and Fundamental Research Program of Shanxi Province (No. 202203021222183).

References

[Chhetri *et al.*, 2023] Siddharth Chhetri, Manjusha Sanjeev Joshi, Chaitanya Vijaykumar Mahamuni, Repana Naga Sangeetha, and Tushar Roy. Speech enhancement: A survey of approaches and applications. In *International Conference on Edge Computing and Applications*, pages 848–856, 2023.

[Donahue *et al.*, 2018] Chris Donahue, Bo Li, and Rohit Prabhavalkar. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5024–5028, 2018.

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

[Fu *et al.*, 2019] Szuwei Fu, Chienfeng Liao, Yu Tsao, and Shoude Lin. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2031–2041, 2019.

[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 2021*, pages 201–205, 2021.

[Guo *et al.*, 2024] 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.

[Hao *et al.*, 2021] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP*, pages 6633–6637, 2021.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[Hu and Loizou, 2008] 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, 2008.

[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 2020*, pages 2472–2476, 2020.

[Isik *et al.*, 2020] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy. Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss. In *Interspeech 2020*, pages 2487–2491, 2020.

[Ju *et al.*, 2022] Yukai Ju, Wei Rao, Xiaopeng Yan, Yihui Fu, Shubo Lv, Luyao Cheng, Yannan Wang, Lei Xie, and Shidong Shang. TEA-PSE: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2022 dns challenge. In *ICASSP*, pages 9291–9295, 2022.

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

[Ku *et al.*, 2023] Pin-Jui Ku, Chao-Han Huck Yang, Sabato Siniscalchi, and Chin-Hui Lee. A multi-dimensional deep structured state space approach to speech enhancement using small-footprint models. In *Interspeech*, pages 2453–2457, 2023.

[Li *et al.*, 2021] Andong Li, Wenzhe Liu, Chengshi Zheng, Cunhang Fan, and Xiaodong Li. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1829–1843, 2021.

[Li *et al.*, 2022a] Andong Li, Shan You, Guochen Yu, Chengshi Zheng, and Xiaodong Li. Taylor, can you hear me now? A taylor-unfolding framework for monaural speech enhancement. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4193–4200, 2022.

[Li *et al.*, 2022b] 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.*, 2025] Xinyan Liang, Yuhua Qian, Qian Guo, and Keyin Zheng. A data representation method using distance correlation. *Frontiers of Computer Science*, 19(1):191303, 2025.

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

[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*, pages 7402–7406, 2022.

[O'Shaughnessy, 2024] Douglas O'Shaughnessy. Speech enhancement—a review of modern methods. *IEEE Transactions on Human-Machine Systems*, 54(1):110–120, 2024.

[Pascual *et al.*, 2017] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network. In *Interspeech*, pages 3642–3646, 2017.

[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.

[Rec, 2005] ITUT Rec. P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, CH–Geneva*, 2005.

[Reddy *et al.*, 2020] Chandan K.A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. The interspeech 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results. In *Interspeech*, pages 2492–2496, 2020.

[Rix *et al.*, 2001] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 749–752 vol.2, 2001.

[Roux *et al.*, 2019] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 626–630, 2019.

[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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 31521–31538, 2023.

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

[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.

[Taal *et al.*, 2010] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

[Tai *et al.*, 2021] Wenxin Tai, Tian Lan, Qianhui Wang, and Qiao Liu. Idanet: An information distillation and aggregation network for speech enhancement. *IEEE Signal Processing Letters*, 28:1998–2002, 2021.

[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 *Advances in Neural Information Processing Systems*, volume 36, pages 40272–40293, 2023.

[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*, volume 37, pages 13627–13635, 2023.

[Valentini-Botinhao *et al.*, 2016] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pages 146–152, 2016.

[Welker *et al.*, 2022] Simon Welker, Julius Richter, and Timo Gerkmann. Speech enhancement with score-based generative models in the complex STFT domain. In *Interspeech*, pages 2928–2932, 2022.

[Westhausen and Meyer, 2020] Nils L. Westhausen and Bernd T. Meyer. Dual-signal transformation lstm network for real-time noise suppression. In *Interspeech 2020*, pages 2477–2481, 2020.

[Xia *et al.*, 2020] Yangyang Xia, Sebastian Braun, Chandan K. A. Reddy, Harishchandra Dubey, Ross Cutler, and Ivan J. Tashev. Weighted speech distortion losses for neural-network-based real-time speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 871–875, 2020.

[Yue *et al.*, 2022] Huanjing Yue, Wenxin Duo, Xiulian Peng, and Jingyu Yang. Reference-based speech enhancement via feature alignment and fusion network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11648–11656, 2022.

[Zheng *et al.*, 2021] Chengyu Zheng, Xiulian Peng, Yuan Zhang, Sriram Srinivasan, and Yan Lu. Interactive speech and noise modeling for speech enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14549–14557, 2021.

[Zhou *et al.*, 2020] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.