Complex-Cycle-Consistent Diffusion Model for Monaural Speech Enhancement

Yi Li[1], **Yang Sun**[2], **Plamen P Angelov**[1]

[1]School of Computing and Communications, Lancaster University, UK
[2]Big Data Institute, University of Oxford, UK

Abstract

In this paper, we present a novel diffusion model-based monaural speech enhancement method. Our approach incorporates the separate estimation of speech spectra's magnitude and phase in two diffusion networks. Throughout the diffusion process, noise clips from real-world noise interferences are added gradually to the clean speech spectra and a noise-aware reverse process is proposed to learn how to generate both clean speech spectra and noise spectra. Furthermore, to fully leverage the intrinsic relationship between magnitude and phase, we introduce a complex-cycle-consistent (CCC) mechanism that uses the estimated magnitude to map the phase, and vice versa. We implement this algorithm within a phase-aware speech enhancement diffusion model (SEDM). We conduct extensive experiments on public datasets to demonstrate the effectiveness of our method, highlighting the significant benefits of exploiting the intrinsic relationship between phase and magnitude information to enhance speech. The comparison to conventional diffusion models demonstrates the superiority of SEDM.

Introduction

In real-world acoustic environments, speech signals are inevitably contaminated by background noise, which can significantly deteriorate speech quality and intelligibility. The primary goal of speech enhancement techniques is to separate the target speech signal from the background noise. Consequently, speech enhancement plays a pivotal role in various speech processing systems, including assisted living, teleconferencing, and automatic speech recognition (ASR) (Zhu et al. 2023; Yang, Pandey, and Wang 2023). Monaural speech enhancement presents one of the most challenging scenarios in this field, as it deals with a single channel.

Traditional deep learning-based methods for solving the monaural speech enhancement problem have been extensively studied. Recently, the diffusion model has not only achieved significant success in the field of image processing (Rahman, J. M. J. Valanarasu, and Patel 2023; Croitoru et al. 2023) but has also been introduced into the field of speech enhancement with excellent results (Hu et al. 2023; Lu et al. 2022). However, these methods have two limitations.

Firstly, these methods typically operate in the time-frequency (T-F) domain, where they estimate the magnitude response while leaving the phase response unaltered, as seen in (Wang, Narayanan, and Wang 2014; Li et al. 2021b). However, Wang et al. introduced a novel approach by proposing a complex ideal ratio mask that allows for simultaneous enhancement of both magnitude and phase spectra, operating in the complex domain (Williamson, Wang, and Wang 2016). Recent studies, such as (Welker, Richter, and Gerkmann 2022a), have widely adopted complex spectra in monaural speech enhancement due to their efficient performance improvement through the utilization of phase information from speech signals. Nevertheless, in most cases, magnitude and phase spectra are simultaneously enhanced by one or two neural networks to generate the final estimated speech spectra, neglecting the intrinsic relationship between magnitude and phase, which has been shown to be beneficial for further improving speech enhancement performance, as indicated in (Shimauchlt et al. 2017).

Secondly, diffusion models for speech enhancement (Hu et al. 2023; Lu et al. 2022) exploit Gaussian noise to learn denoising noisy speech based on maximum entropy and statistical inference. However, some recent diffusion-based image denoising techniques replace the Gaussian noise to real-world noise. Wu et al. synthesise realistic noise with the environmental settings to better model noise distribution complexity (Wu et al. 2023). The experiments prove that using real-world noise can boost the performance rather than Gaussian noise. Nevertheless, there has been a lack of recent research attempting to substitute real-world noise for Gaussian noise in the filed of speech enhancement. This has sparked our interest in exploring the problem.

We present three contributions to address these limitations in this paper. Firstly, we propose a novel speech enhancement diffusion model (SEDM). In contrast to conventional diffusion model-based speech enhancement methods (Lu, Tsao, and Watanabe 2021), we replace the Gaussian noise in the forward process with real-world noise. In each embedding, a noise clip is randomly selected from different noise types, e.g., $dwashing$, $dliving$, and $pstation$ (Thiemann, Ito, and Vincent 2013). Secondly, we propose a noise-aware reverse process that facilitates learning to generate both clean speech and noise spectra. Thirdly, we investigate the speech enhancement performance with independent net-

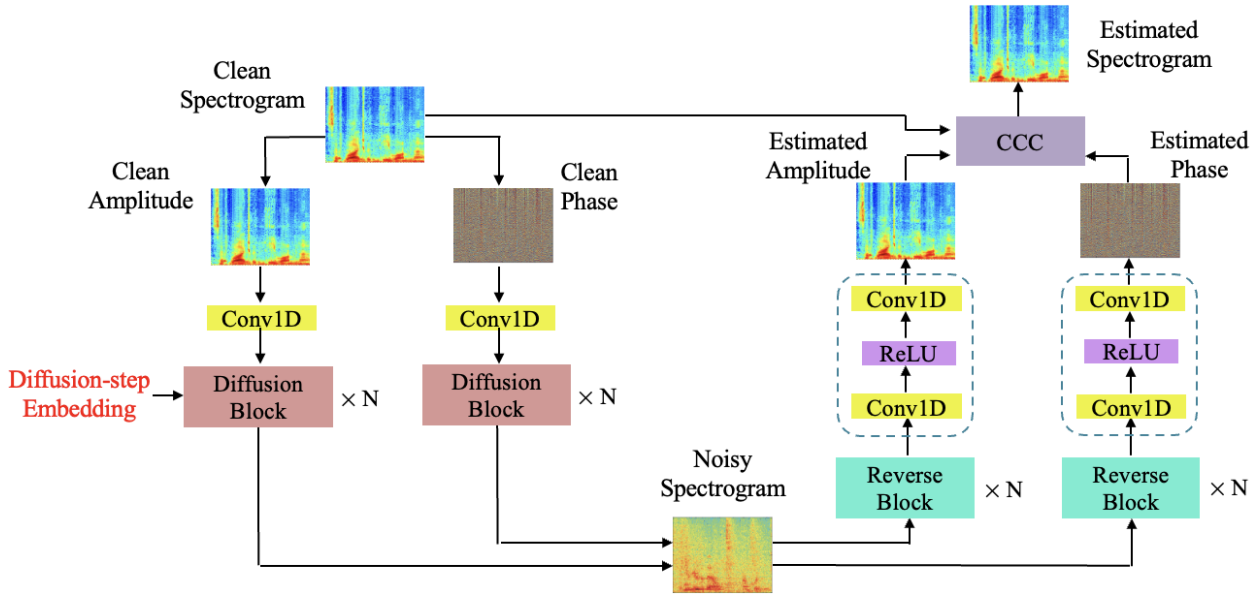


Figure 1: The overall architecture of SEDM consisted of a diffusion process (left) and a noise-aware reverse process (right).

works. In encoders of SEDM, we independently estimate the noise distributions in the magnitude and phase of the mixture spectrogram. Subsequently, the decoders generate the estimated magnitude and phase of the target speech spectrogram. The relationship between magnitude and phase, as shown in (Shimauchlt et al. 2017), is beneficial for further enhancing speech enhancement performance. Therefore, we calculate the commutative losses between the magnitude and phase features to further improve estimation accuracy.

Related Works

Phase-Aware Speech Enhancement

In conventional speech enhancement studies, the desired speech signal is typically reconstructed using the magnitude of the estimated speech signal and the phase of the noisy mixture. However, due to the reliance on noisy phase information, the speech enhancement performance may be compromised. Therefore, recent studies have started to incorporate the phase information of the desired speech signal. Polar coordinates (i.e. magnitude and phase) are commonly used when enhancing the STFT of noisy speech, as defined in (1):

$$S_{t,f} = |S_{t,f}|\, e^{i\theta_{S_{t,f}}} \tag{1}$$

where $|S_{t,f}|$ represents the magnitude response and $\theta_{S_{t,f}}$ represents the phase response of the short-time Fourier transform (STFT) at time t and frequency f. Each T-F unit in the STFT representation is a complex number with real and imaginary components.

Diffusion Model

Diffusion models, initially proposed in (Sohl-Dickstein et al. 2015), have demonstrated strong generative capabilities. A typical diffusion probabilistic model comprises two key processes: a forward/diffusion process and a reverse process. In the forward process, the model transforms clean input data into an isotropic Gaussian distribution by introducing Gaussian noise to the original signal at each step. Conversely, in the reverse process, the diffusion probabilistic model predicts a noise signal and subtracts this predicted noise signal from the noisy input to recover the clean signal. As the first diffusion model-based speech enhancement work, Lu et al. introduced the concept of a supportive reverse process in their work (Lu, Tsao, and Watanabe 2021), where noisy speech is added at each time-step to the predicted speech signal.

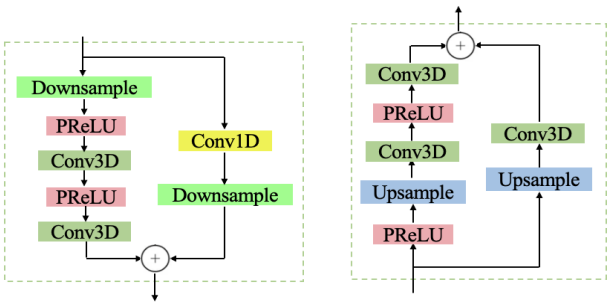
Phase-Aware Diffusion Models

The speech enhancement algorithm aims to estimate the speech signal $s(t)$ from noisy speech signal $y(t)$. To achieve that, we design a diffusion model-based training pipeline as presented in Figure 1. The proposed SEDM consists of a diffusion network (left side of Figure 1) and a noise-aware reverse network (left side of Figure 1) for a diffusion process and a reverse diffusion, respectively. In Figure. 1, the input and output are spectra, as we omit Short-Time Fourier Transform (STFT) for speech signals.

Diffusion Process

We disassemble STFT of clean speech signals into magnitude and phase components as the input for the diffusion network. The diffusion network consists of a stack of N diffusion blocks, each with a residual channel size of C, followed by a 1×1 convolutional layer. A single diffusion block is presented in Figure 2 (a).

Similar to U-net models (Zhao and Wang 2020; Li et al. 2023), we replace pooling operations in each diffusion block with downsampling operators to complement a typical contracting network with successive layers. Additionally, we



(a) Diffusion Block (b) Reverse Block

Figure 2: The proposed diffusion block and reverse block.

employ Parametric Rectified Linear Unit (PReLU) activation functions (He et al. 2015), which allow for flexibility in the range of values in the feature map, encompassing both negative and non-negative values. This flexibility contributes to more accurate estimation of the speech source. Furthermore, we incorporate residual connections (He et al. 2016) between the input of each diffusion block and the output of the final 3×1 convolutional layer. This addition aids in achieving rapid convergence, particularly when training a deep diffusion network with a large value of N, by mitigating issues related to vanishing gradients.

In addition, each diffusion block includes a skip connection (He et al. 2016) to the next block to maintain the desired information within the signal. Within the diffusion network, each diffusion block generates feature maps at specific resolutions, which are then scaled to produce a latent representation of the noisy speech feature with multiple resolutions. During the training of the diffusion network, the optimal weighted combinations of these multi-resolution spectra are learned in relation to the target, which is the original speech feature representation.

Within the forward process $q\left(Y_1, \cdots, Y_N \mid Y_0\right)$ of each diffusion block, we embed either the magnitude or phase of a noise clip's spectrogram into the input speech spectrogram Y_0 as:

$$q\left(Y_1, \cdots, Y_N \mid Y_0\right) = \prod_{n=1}^{N} q\left(Y_n \mid Y_{n-1}\right) \tag{2}$$

where n is the index of diffusion block. Different from conventional diffusion models, we replace the Gaussian noise by real-world noises during the diffusion process. To achieve that, we divide four-minute noise sequences among all noise types from the DEMAND dataset (Thiemann, Ito, and Vincent 2013) into clips to align with the length of the input speech signals. We randomly select one clip I_n and progressively add to a clean speech spectrogram as a stochastic differential equation (SDE) (Rogers and Williams 2000):

$$dY_n = \mu\left(Y_n, n\right) dn + \sigma\left(Y_n, n\right) dI_n \tag{3}$$

where $\mu(Y_n, n)$ is the drift term representing the deterministic component and $\sigma(Y_n, n)$ is the diffusion term representing the stochastic (random) component. The final diffusion block generates the latent representation of the noisy speech spectra.

Reverse Process

Different from conventional diffusion models, we propose a noise-aware reverse process that initiates the sampling process from the noisy speech spectrogram. At each reverse block, we estimate both the clean speech and noise spectra, all while minimizing the introduction of additional noise signals. A single reverse block is presented in Figure 2 (b).

In the proposed noise-aware reverse process, we aim to generate each Y_{m-1} from the previous step Y_m. To achieve that, we define each reverse step with two trainable parameters γ_m and θ_m as:

$$Y_{m-1} = \frac{1}{\gamma_m}\left(Y_m - \frac{\theta_m}{1-\gamma_m} I_m\right) + \sigma_m \tag{4}$$

where σ_m is the variance of the estimated speech spectra distribution, which can be calculated as:

$$\sigma_m = \frac{1-\bar{\gamma}_{m-1}}{1-\bar{\gamma}_m}\theta_m \tag{5}$$

Again, similarly to the U-net models, reverse blocks utilize upsampling operators to restore the original spectrogram size. The final estimated magnitude and phase of the target speech spectrogram are derived from the last reverse blocks. Subsequently, the proposed CCC block further enhances the estimation accuracy of both magnitude and phase using the clean spectrogram.

Finally, the phase is reconstructed by re-wrapping the estimated unwrapped phase of the speech signal. This phase, along with the recovered speech magnitude, is used in the speech recovery module to reconstruct the estimated speech signal. During the test stage, the diffusion network is disregarded, and the reverse network is employed to enhance the noisy speech signals.

Complex-Cycle-Consistent Learning

After estimating the magnitude and phase of speech sources from the reverse blocks, they are input into the proposed Complex-Cycle-Consistent block (CCC) along with the clean magnitude and phase as shown in Figure 3. The block consists of two long short-term memory (LSTM) networks, with their parameters denoted as θ_A and θ_P. We denote the estimated magnitude and phase of speech spectra as $\mathbf{S}_A$ and $\mathbf{S}_P$, respectively.

As the input, the magnitude and phase of the estimated speech are fed into the CCC module with the clean magnitude and phase. First, the magnitude loss is estimated with (3). We define $P \rightarrow A$ as the mapping of the spectra from the phase to the magnitude. We refer to $\mathbf{S}_{P\rightarrow A}$ as the new phase reconstruction from the last cycle. Then, the loss between the magnitude of clean speech spectra and $\mathbf{S}_{P\rightarrow A}$ is computed with the L2 norm of the error as:

$$\mathcal{L}_{\mathbf{S}_{P\rightarrow A}} = \|\mathbf{S}_A - \mathbf{S}_{P\rightarrow A}\|_2^2 \tag{6}$$

In the training stage, the loss term $\mathcal{L}_{\mathbf{S}_{P\rightarrow A}}$ is relatively large, in contrast to the loss $\mathcal{L}_{\mathbf{S}_A}$. Therefore, we use a weight λ_1 to attenuate $\mathcal{L}_{\mathbf{S}_{P\rightarrow A}}$. The combined magnitude loss can be obtained as:

$$\mathcal{L}_{\mathbf{S}'_A} = \mathcal{L}_{\mathbf{S}_A} + \lambda_1 \cdot \mathcal{L}_{\mathbf{S}_{P\rightarrow A}} \tag{7}$$

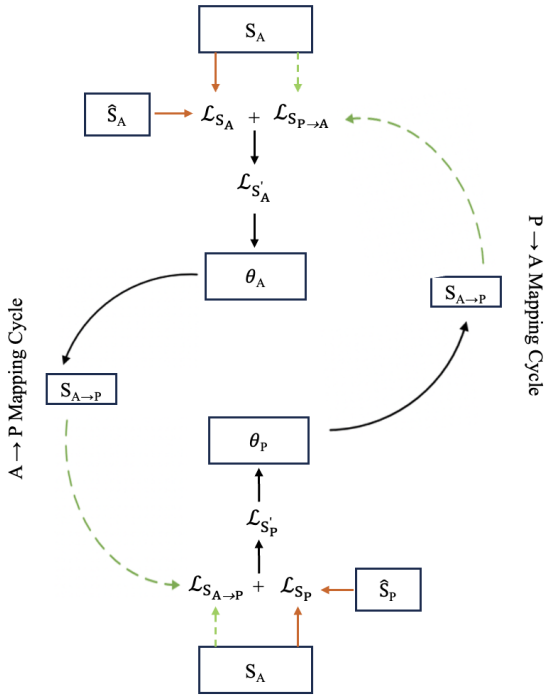


Figure 3: The proposed complex-cycle-consistent learning for speech (CCC) enhancement.

The combined magnitude loss $\mathcal{L}_{\mathbf{S}'_A}$ is applied to train θ_A which is used to map the spectra from the magnitude to the phase $\mathbf{S}_{A\to P}$ in the $A \to P$ mapping cycle. Similarly, the loss between the phase of the clean speech spectra and the reconstruction from the mapping $\mathbf{S}_{A\to P}$ is written as:

$$\mathcal{L}_{\mathbf{S}_{A\to P}} = \|\mathbf{S}_P - \mathbf{S}_{A\to P}\|_2^2 \tag{8}$$

Accordingly, the combined phase loss is presented as:

$$\mathcal{L}_{\mathbf{S}'_P} = \mathcal{L}_{\mathbf{S}_P} + \lambda_2 \cdot \mathcal{L}_{\mathbf{S}_{A\to P}} \tag{9}$$

where a weight λ_2 is a weight parameter. Then, θ_P is trained with the combined phase loss and yields a new mapping magnitude reconstructions as $\mathbf{S}_{P\to A}$ for the next epoch. Parameters θ_A and θ_P trained with the cycle-consistent learning approach and finally outputs the magnitude and phase of estimated speech spectra. The pseudocode of the proposed CCC module is summarized as Algorithm 1.

Algorithm 1: Proposed complex-cycle-consistent learning

Input: magnitude of the clean speech spectra $\mathbf{S}_A$, phase of the clean speech spectra $\mathbf{S}_P$, epoch $E = 1, 2, ..., E_{max}$
Output: Estimated speech $\hat{\mathbf{S}}_A$ and $\hat{\mathbf{S}}_P$
Initialize reverse network parameters θ_A and θ_P
while $E = 1$ **do**
    Estimate $\hat{\mathbf{S}}_A$ and $\hat{\mathbf{S}}_P$
    Calculate the losses: $\mathcal{L}_{\mathbf{S}_A}$ and $\mathcal{L}_{\mathbf{S}_P}$
end while
for $E = 2, ..., E_{max}$ **do**
    Run the mapping cycle $\mathbf{S}_{A\to P}$ by using θ_A
    Update $\mathcal{L}_{\mathbf{S}'_A}$ as (8)
    Run the mapping cycle $\mathbf{S}_{P\to A}$ by using θ_P
    Update $\mathcal{L}_{\mathbf{S}'_P}$ as (10)
    $\mathcal{L}_{\mathbf{S}_A} = \mathcal{L}_{\mathbf{S}'_A}$, $\mathcal{L}_{\mathbf{S}_P} = \mathcal{L}_{\mathbf{S}'_P}$
    Update θ_A, θ_P by minimizing $\mathcal{L}_{\mathbf{S}_A}$ and $\mathcal{L}_{\mathbf{S}_P}$
end for
Estimate $\hat{\mathbf{S}}_A$ and $\hat{\mathbf{S}}_P$ with trained θ_A and θ_P

Experiments

Datasets

We extensively perform experiments on several public speech datasets, including IEEE (IEEE Audio and Electroacoustics Group 1969), TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) (Garofolo et al. 1993), VOICE BANK (VCTK) (Veaux, Yamagishi, and King 2013), and Deep Noise Suppression (DNS) challenge (Reddy et al. 2021). To generate noisy speech signals in training and test, we randomly collect and use 10 of 15 noise types *psquare*, *dliving*, *dkitchen*, *nriver*, *tcar*, *dwashing*, *npark*, *omeeting*, *ohallway* and *pstation* from Diverse Environments Multichannel Acoustic Noise Database (DEMAND) (Thiemann, Ito, and Vincent 2013). Each noise interference has a unique case and lasts four minutes long, and it is divided into two clips with an equal length. One is used to match the lengths of the speech signals to generate training data in the diffusion process and the other is used to generate development and inference data.

Model Configuration

We set the number of diffusion blocks and channels as [N,C] ∈ [30,63],[40,128],[50,128] for small, medium, and large SEDM models (SEDM-S, SEDM-M, SEDM-L), respectively. The number of reverse blocks is equal to the number of diffusion blocks, i.e., $M = N$ .The kernel size of Bi-DilConv is 3, and the dilation is doubled at each layer within each block as [1, 2, 4, ..., 2^{n-1}]. Each LSTM in CCC consists of three hidden layers and 30 features in the hidden state. Further studies on model backbones are out of scope of this paper.

The proposed model is trained by using the Adam optimizer with a weight decay of 0.0001, a momentum of 0.9, and a batch size of 64. We train the networks for 200 epochs, where we warm-up the network in the first 20 epochs by without CCC losses. The initial learning rate is 0.03, and is multiplied by 0.1 at 120 and 160 epochs. All the experiments are run on Tesla V100 GPUs.

Moreover, all the speech utterances are resampled to 16 kHz. They are converted to spectrogram using fast Fourier transform (FFT), with a window of 512 samples (32ms) with an overlap of 256 samples (16ms) between the neighboring windows. Since the input and the output of the proposed method and baselines are both magnitude spectrogram and the dimension of single axis is set to 257. A linear processing layer is stacked when splitting the feature map to convert the spectrogram to feature vectors of 512 dimensions.

	Configuration			IEEE			TIMIT		
Method	N	C	K	STOI (%)	PESQ	fwSNRseg (dB)	STOI (%)	PESQ	fwSNRseg (dB)
Unprocessed	-	-	-	42.3	1.52	3.11	41.5	1.44	3.04
DCTCRN	7	256	5	73.4	2.36	12.88	78.5	2.45	13.24
RemixIT	64	512	21	74.8	2.42	13.13	79.3	2.56	13.97
FRCRN	6	128	7	76.5	2.50	13.87	80.2	2.59	14.43
CMGAN	16	64	8	75.2	2.47	12.98	79.6	2.55	13.78
SCP-GAN	-	-	-	77.3	2.66	14.04	81.5	2.77	15.00
SEDM-S	30	63	3	79.0	2.68	13.97	81.2	2.76	14.85
SEDM-M	40	128	3	79.7	2.73	14.22	81.7	2.81	15.18
SEDM-L	50	128	3	**80.2**	**2.75**	**14.30**	**81.9**	**2.83**	**15.29**

Table 1: Speech enhancement performance comparisons on the **IEEE** and **TM** datasets. The number of residual blocks, channels and kernel is denoted as N, C, and K, respectively.

Competitors

In this work, we compare the proposed method with six state-of-the-art models DCTCRN (Li et al. 2021a), RemixIT (Tzinis et al. 2022), FRCRN (Zhao, Nguyen, and Ma 2021), CMGAN (Cao, Abdulatif, and Yang 2022), and SCP-GAN (Zadorozhnyy and Q. Ye 2022), which reach state-of-the-art benchmarks in the DNS challenge and VCTK + DEMAND datasets. It is highlighted that we reproduce these models with the same experimental setting, e.g., training data and reverberations, as the proposed method for fair comparison.

Results

In this section, we firstly evaluate the speech enhancement performance of SEDM family and compare to state-of-the-art benchmarks on commonly used datasets, i.e., IEEE, TIMIT, VCTK, and DNS challenge. Then, we compare the proposed models to other diffusion models in the literature. Finally, we provide some visualizations and ablation study to further confirm the effectiveness of contributions.

Evaluations on the IEEE and TIMIT Datasets

The first experiment is conducted on IEEE and TIMIT (IEEE Audio and Electroacoustics Group 1969; Garofolo et al. 1993). In the training and development stages, 600 recordings from 60 speakers and 60 recordings from 6 speakers are randomly selected in each dataset, respectively. To evaluate and compare the quality of the enhanced speech with various methods, we use the short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and frequency-weighted segmental signal-to-noise ratio (fwSNRseg) as performance measures on the IEEE and TIMIT datasets. The STOI and the PESQ are bounded in the range of [0, 1] and [-0.5, 4.5], respectively (Hu and Loizou 2008). The fwSNRseg is estimated by computing the segmental signal-to-noise ratios (SNRs) in each spectral band and summing the weighed SNRs from all bands (Liu, Ma, and Chen 2017) in the range of [-10, 35] dB.

Table 1 shows the averaged speech enhancement performance of the proposed method as compared to state-of-the-art models using the IEEE and the TIMIT datasets, with three SNR levels (-5, 0, 5 dB) and ten noise interferences in . Each result is the average of 360 noisy speech signals (120 clean speech signals $\times$3 SNR levels). From Table 1, it can be observed that in all the evaluated models, SEDM-L offers the best effectiveness.

Evaluations on the VCTK and DNS Challenge Datasets

We perform extensive experiments to evaluate whether SEDM family can achieve a good speech enhancement performance over the VCTK dataset. We randomly generate 11572 noisy mixtures with 10 background noises at one of 4 SNR levels (15, 10, 5, and 0 dB) in the training stage. The test set with 2 speakers, unseen during training, consists of a total of 20 different noise conditions: 5 types of noise sourced from the DEMAND dataset at one of 4 SNRs each (17.5, 12.5, 7.5, and 2.5 dB). This yields 824 test items, with approximately 20 different sentences in each condition per test speaker. To evaluate and compare the quality of the enhanced speech with various methods, we use mean opinion score (MOS) predictor of signal distortion (CSIG), MOS predictor of background intrusiveness (CBAK), MOS predictor of overall speech quality (COVL) to map the enhancement between [1, 5] (Hu and Loizou 2008). Furthermore, similar to (Macartney and Weyde 2018; Deng et al. 2020), PESQ and segmental signal-to-noise ratio (SSNR) are used as well. Table 2 shows the averaged speech enhancement results on the VCTK dataset (Veaux, Yamagishi, and King 2013).

Method	PESQ	CSIG	CBAK	COVL	SSNR
Unprocessed	1.97	3.35	2.44	2.63	1.7
DCTCRN	3.30	3.69	3.90	4.53	10.1
RemixIT	3.38	3.85	3.99	4.68	10.2
FRCRN	3.43	3.92	4.20	4.71	11.6
CMGAN	3.41	3.94	4.12	4.63	11.1
SCP-GAN	3.52	3.97	**4.25**	4.75	10.8
SEDM-S	3.46	3.88	4.07	4.58	11.6
SEDM-M	3.50	3.94	4.15	4.71	11.7
SEDM-L	**3.59**	**4.06**	4.22	**4.89**	**11.8**

Table 2: Speech enhancement performance comparison on **VCTK**.

From this table, we can see that the proposed method outperforms the state-of-the-art methods in terms of all perfor-

mance measures. The proposed SEDM-L is 0.96 higher than SCP-GAN (11.8 vs. 10.8, SSNR).

The proposed method is further evaluated on the DNS challenge benchmark and compared with the state-of-the-art methods. The clean speech set includes over 500 hours of clips from 2150 speakers and the noise set includes over 180 hours of clips from 150 classes in the DNS challenge (Reddy et al. 2021). In the training stage, 75% of the clean speeches are mixed with the background noise but without reverberation at a random SNR in between -5 and 20 dB as (Hao et al. 2021). In the test stage, 150 noisy clips are randomly selected from the blind test dataset without reverberations. In these experiments, the averaged STOI (%), wide-band PESQ (WP), narrow-band PESQ (NP), and scale-invariant source-to-distortion ratio (SI-SDR) (dB) performances are presented in Table 3.

Method	WP	NP	STOI (%)	SI-SDR
Unprocessed	1.56	2.45	91.2	9.0
DCTCRN	2.82	3.17	94.6	10.8
RemixIT	**2.95**	3.33	97.1	**19.7**
FRCRN	2.65	3.23	96.1	11.1
CMGAN	2.54	3.10	94.1	10.6
SCP-GAN	2.84	3.25	95.2	10.9
SEDM-S	2.88	3.31	96.6	12.6
SEDM-M	2.91	3.35	97.2	13.0
SEDM-L	2.93	**3.42**	**97.4**	13.2

Table 3: Speech enhancement performance comparison on the **DNS challenge** dataset without reverberations.

We observe that the proposed SEDM model shows competitive performance compared to the state-of-the-art model, Remix, on the DNS challenge. Specifically, SEDM-L outperforms the state-of-the-art models in terms of NP and STOI metrics.

Comparison to other Diffusion Models

We further investigate the effectiveness of our diffusion model against state-of-the-art diffusion models in the literature, including denoising diffusion probabilistic model (DDPM) (Hu et al. 2020), diffusion probabilistic model-based speech enhancement (DiffuSE) (Lu, Tsao, and Watanabe 2021), noise-aware speech enhancement (NASE) (Hu et al. 2023), score-based generative models speech enhancement (SGMSE) (Welker, Richter, and Gerkmann 2022b), conditional diffusion probabilistic model for speech enhancement (CDiffuSE) (Lu et al. 2022), neural audio upsampling model (NU-Wave) (Han and Lee 2022). The evaluation results on VCTK (Veaux, Yamagishi, and King 2013) + DEMAND (Thiemann, Ito, and Vincent 2013) are reported in Table 4, and the experimental setting is the same as Table 2.

We observe that the proposed SEDM-L achieves the best performance, outperforming the second-best NASE (Hu et al. 2023) by 0.58 and 0.04 over PESQ and ESTOI, respectively.

Models	PESQ	ESTOI	SI-SDR
DDPM	2.28	0.64	8.5
NU-Wave	2.33	0.67	9.0
DiffuSE	2.41	0.72	10.9
CDiffuSE	2.58	0.79	12.4
SGMSE	2.93	0.87	17.3
NASE	3.01	0.87	17.6
SEDM-L	**3.59**	**0.91**	**19.4**

Table 4: Speech enhancement performance comparison to other diffusion models on the VCTK dataset.

Ablation Study

In this experiment, we investigate the effectiveness of each contribution. A ResNet152 (He et al. 2016) is used when the diffusion model shows ✗. Although the CCC mechanism is based on phase-aware spectrogram, we use a single diffusion model to generate speech spectra as presented in Figure 4 for the combination of (diffusion model: ✓, phase-aware: ✗, and CCC: ✓).

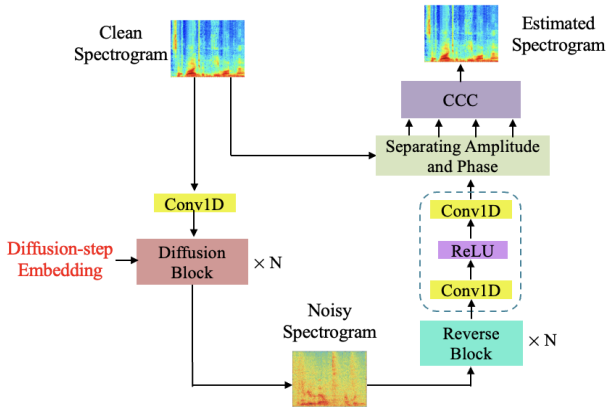


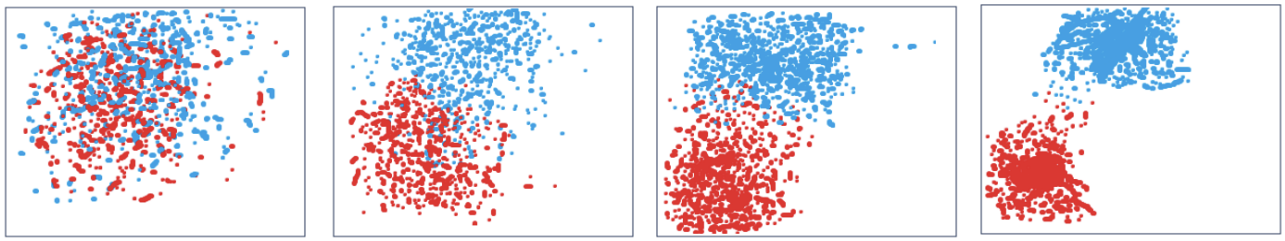
Figure 4: The pipeline of (diffusion model: ✓, phase-aware: ✗, and CCC: ✓). The clean speech spectra and the corresponding reconstruction are only converted into magnitude and phase components before CCC module.

The models are trained and tested on the IEEE dataset as in Section . Ablation study results are showed in Table 5.

Ablation Settings			PESQ
Diffusion Model	Phase-Aware	CCC	
✗	✗	✗	2.21
✓	✗	✗	2.43
✗	✓	✗	2.33
✗	✓	✓	2.60
✓	✓	✗	2.55
✓	✓	✓	**2.75**

Table 5: Ablation study of the three contributions in the proposed method.

Initially, we evaluate the effectiveness of diffusion model, which plays a pivotal role in learning desired features from



(a) N = 1 (b) N = 10 (c) N = 30 (d) N = 50

Figure 5: IEEE dataset visualization with diffusion t-SNE for different numbers of embeddings N.

noisy speech spectra. Diffusion model demonstrates a remarkable performance improvement from an initial PESQ of 2.21 to 2.43. This improvement can be attributed to the ability to effectively capture and model the temporal dynamics and dependencies present in speech signals.

Moreover, speech enhancement experiences a relatively slight improvement by exploiting phase information (i.e., 2.21→2.33). In the baselines, the speech signal is reconstructed by using the noisy phase and the estimated magnitude, which causes a phase loss between the clean speech signal and the corresponding reconstruction. However, the proposed phase-aware method utilizes θ_A and θ_P to estimate the phase of the target speech signal and noisy mixture, respectively, and thus improves the accuracy of estimation.

The final experiment in the ablation study involves the addition of CCC. As demonstrated in the appendix, the potential association between the magnitude and phase plays an important role in improving speech enhancement performance. With the proposed CCC mechanism, the magnitude and phase are estimated with the updated reconstruction of noisy speech features, which are better preserved in the estimated features.

Visualization of Learned Representation

As qualitative analysis, Figure 6 presents the t-distributed stochastic neighbour embedding (t-SNE) visualisation of the proposed model against different numbers of embeddings N using the SEDM family with C=128 and K=3 on IEEE.

Figure 6 shows the t-SNE visualisation using different perplexity settings. For small values of N, we observe that the feature embeddings are not quite separable for separation of clean speech (blue) and noise interference (red). For large perplexity values, the features representation from the SEDM-L is better separated. These t-SNE visualization results demonstrate that proposed methods are able to better learn discriminative feature representations with 50 embedding steps.

Noise Embeddings

In this section, we compare the proposed model trained with real-world noises to same backbones with Gaussian noise. Moreover, we evaluate the proposed model over both seen and unseen noise types in the test stage. The experimental setting is similar to Section , but we generate the test data using the remaining 5 out of 15 noise types for the unseen noise type scenario. Figure 5 presents the results.

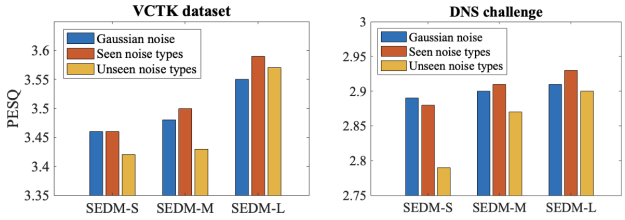


Figure 6: Speech enhancement performance over VCTK dataset (left) and DNS challenge (right). The blue bars indicate models trained with Gaussian noise and evaluated with 5 of 15 noise types from DEMAND. Both red and yellow bars show models trained with real-world noise, but evaluated with seen and unseen noise types, respectively.

Figure 5 shows speech enhancement performance of SEDM models against seen and unseen noise interferences in the test stage. We can observe that: (1) SEDM models trained with real-world noises suffer a performance degradation with unseen noise interferences due to noise domain mismatch. (2) SEDM-L demonstrates greater robustness compared to competitors across all noise interferences. (3) SEDM models trained with real-world noises are initially inferior to models trained with Gaussian noise in shallower networks. However, as the network depth increases, the proposed real-world noise-based models become more competitive, and in some cases, even surpass models trained with Gaussian noise.

Conclusions

In this paper, we have presented a diffusion model-based method to address the monaural speech enhancement problem. Different from the previous speech enhancement methods that ignore the intrinsic relationship between magnitude and phase information, our method estimated both the magnitude and phase information of the desired speech signal. In addition, the proposed complex-cycle-consistent mechanism provided mappings between the magnitude and phase to update the combined losses and further refined the estimation accuracy. The experimental results showed that the proposed method outperforms the state-of-the-art speech enhancement approaches over different public datasets. Our ablation experiments confirmed that real-world noise can, to a certain extent, serve as a substitute for Gaussian noise.

Acknowledgements

This work is supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617.

References

Cao, R.; Abdulatif, S.; and Yang, B. 2022. CMGAN: Conformer-based metric GAN for speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 10850–10869.

Deng, F.; Jiang, T.; Wang, X. R.; Zhang, C.; and Li, Y. 2020. NAAGN: noise-aware attention-gated network for speech enhancement. *Interspeech*.

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; Pallett, D. S.; and Dahlgren, N. L. 1993. TIMIT acoustic phonetic continuous speech corpus CD-ROM. *Linguistic Data Consortium*.

Han, S.; and Lee, J. 2022. NU-Wave 2: a general neural audio upsampling model for various sampling rates. *Interspeech*.

Hao, X.; Su, X. D.; Horaud, R.; and Li, X. F. 2021. FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, Y.; Chen, C.; Li, R.; Zhu, Q.; and Chng, E. S. 2020. Denoising diffusion probabilistic models. *Processdings of Neural Information Processing Systems (NeurIPS)*.

Hu, Y.; Chen, C.; Li, R.; Zhu, Q.; and Chng, E. S. 2023. Noise-aware speech enhancement using diffusion probabilistic model. *arXiv preprint arXiv:2307.08029*.

Hu, Y.; and Loizou, P. C. 2008. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1): 229–238.

IEEE Audio and Electroacoustics Group. 1969. IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio, Speech, and Language Processing*, AE-17(3): 225–246.

Li, Q.; Gao, F.; Guan, H.; and Ma, K. 2021a. Real-time monaural speech enhancement with short-time discrete cosine transform. *arXiv preprint arXiv: 2102.04629*.

Li, Y.; Sun, Y.; Horoshenkov, K.; and Naqvi, S. M. 2021b. Domain adaptation and autoencoder based unsupervised speech enhancement. *IEEE Transactions on Artificial Intelligence*, 3(1): 43 – 52.

Li, Y.; Sun, Y.; Wang, W.; and Naqvi, S. M. 2023. U-shaped Transformer with frequency-band aware attention for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1511–1521.

Liu, Z. X.; Ma, H. T.; and Chen, F. 2017. A new data-driven band-weighting function for predicting the intelligibility of noise-suppressed speech. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Lu, Y.-J.; Tsao, Y.; and Watanabe, S. 2021. A study on speech enhancement based on diffusion probabilistic model. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Lu, Y.-J.; Wang, Z.-Q.; Watanabe, S.; Richard, A.; Yu, C.; and Tsao, Y. 2022. Conditional diffusion probabilistic model for speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Macartney, C.; and Weyde, T. 2018. Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*.

Rahman, A.; J. M. J. Valanarasu, I. H.; and Patel, V. M. 2023. Ambiguous medical image segmentation using diffusion models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Reddy, C.; Dubey, H.; Koishida, K.; Nair, A.; Gopal, V.; Cutler, R.; Braun, S.; Gamper, H.; Aichner, R.; and Srinivasan, S. 2021. Interspeech 2021 deep noise suppression challenge. *Interspeech*.

Rogers, L. C. G.; and Williams, D. 2000. Diffusions, Markov processes and martingales,. *Cambridge University Press*.

Shimauchlt, S.; Kudo, S.; Koizumli, Y.; and Furuva, K. 2017. On relationships between amplitude and phase of short-time fourier transform. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning (ICML)*.

Thiemann, J.; Ito, N.; and Vincent, E. 2013. The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5): 3591 – 3591.

Tzinis, E.; Adi, Y.; Ithapu, V. K.; Xu, B.; Smaragdis, P.; and Kumar, A. 2022. RemixIT: continual self-training of speech enhancement models via bootstrapped remixing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1329–1341.

Veaux, C.; Yamagishi, J.; and King, S. 2013. The voice bank corpus: design, collection and data analysis of a large regional accent speech database. *IEEE Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*.

Wang, Y.; Narayanan, A.; and Wang, D. 2014. On training targets for supervised speech separation. *IEEE/ACM*

Transactions on Audio, Speech, and Language Processing, 22(12): 1849–1858.

Welker, S.; Richter, J.; and Gerkmann, T. 2022a. Speech enhancement with score-based generative models in the complex STFT domain. *Interspeech*.

Welker, S.; Richter, J.; and Gerkmann, T. 2022b. Speech enhancement with score-based generative models in the complex STFT domain. *Interspeech*.

Williamson, D. S.; Wang, Y.; and Wang, D. 2016. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3): 483 – 492.

Wu, Q.; Han, M.; Jiang, T.; Fan, H.; Zeng, B.; and Liu, S. 2023. Realistic noise synthesis with diffusion models. *arXiv preprint arXiv: 2305.14022*.

Yang, Y.; Pandey, A.; and Wang, D. 2023. Time-domain speech enhancement for robust automatic speech recognition. *Interspeech*.

Zadorozhnyy, V.; and Q. Ye, K. K. 2022. SCP-GAN: self-correcting discriminator optimization for training consistency preserving metric GAN on speech enhancement tasks. *arXiv preprint arXiv: 2210.14474*.

Zhao, S.; Nguyen, T. H.; and Ma, B. 2021. Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhao, Y.; and Wang, D. L. 2020. Noisy-reverberant Speech Enhancement Using DenseUNet with Time-frequency Attention. *Interspeech*.

Zhu, Q.-S.; Zhang, J.; Zhang, Z.-Q.; and Dai, L.-R. 2023. A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1927–1939.