A two-stage beamforming and diffusion-based refiner system for 3D speech enhancement

Feilong Chen[1], Wenmo Lin[1], Chengli Sun[1], Qiaosheng Guo[2]

Abstract

Speech enhancement in 3D reverberant environments is a challenging and significant problem for many downstream applications, such as speech recognition, speaker identification, and audio analysis. Existing deep neural network models have shown efficacy for 3D speech enhancement tasks, but they often introduce distortions or unnatural artifacts in the enhanced speech. In this work, we propose a novel two-stage refiner system that integrates a neural beamforming network and a diffusion model for robust 3D speech enhancement. The neural beamforming network performs spatial filtering to suppress the noise and reverberation, while the diffusion model leverages its generative capability to restore the missing or distorted speech components from the beamformed output. To the best of our knowledge, this is the first work that applies the diffusion model as a backend refiner to 3D speech enhancement. We investigate the effect of training the diffusion model with either enhanced speech or clean speech, and find that clean speech can better capture the prior knowledge of speech components and improve the speech recovery. We evaluate our proposed system on different datasets and beamformer architectures, and show that it achieves consistent improvements in metrics like WER and NISQA, indicating that the diffusion model has strong generalization ability and can serve as a backend refinement module for 3D speech enhancement, regardless of the front-end beamforming network. Our work demonstrates the effectiveness of integrating discriminative and generative models for robust 3D speech enhancement, and also opens up a new direction for applying generative diffusion models to 3D speech processing tasks, which can be used as a backend to various beamforming enhancement methods.

Keywords: Speech enhancement; 3D speech signal; Diffusion model; Beamforming; Multi-channel

✉ Feilong Chen
flchen@nchu.edu.cn

Wenmo Lin
linwm2022@163.com

Chengli Sun
sunchengli@nchu.edu.cn

Qiaosheng Guo
arvin@jushengtai.com

[1] School of Information and Engineering, Nanchang Hangkong University, Nanchang 330063, China

[2] Chaoyang jushengtai (Xinfeng) Technology Co Ltd, Ganzhou 341001, China

Code is available at https://github.com/flchenwhu/3D-SE-Diffusion

1 Introduction

Speech enhancement is a task that aims to restore clean and clear speech signals from noisy or reverberant speech [9]. Many downstream applications, such as speech recognition, speaker identification, and audio analysis, rely on high-quality speech signals as input. Deep learning-based speech enhancement methods have achieved superior performance over traditional methods in recent years [3, 12, 22]. Common methods include time-frequency (T-F) masking [35], spectral mapping [6], and direct processing of speech in the time domain [7]. These methods are discriminative, as they learn to map noisy speech to clean speech using a supervised objective function. However, some discriminative methods may introduce unpleasant speech distortion or unnatural artifacts in the enhanced speech, due to the mismatch between the objective function and the perceptual quality. To address this issue, generative methods use generative models to learn the properties of speech as prior knowledge, and then use this knowledge to enhance speech [34]. Classic generative speech enhancement models include variational autoencoders [13], generative adversarial networks [8], and others. Recently, the diffusion model, a novel generative model, has been applied to speech enhancement. It first uses a Markov chain to gradually convert speech into noise in the diffusion process, and then reverses this process in the denoising process to gradually generate target speech from noise [19, 20, 31]. Later, the score-based diffusion model for speech enhancement emerged, which incorporates stochastic differential equations (SDEs) into the diffusion model. This model has a backward SDE corresponding to each step of the forward SDE, and unlike the Markov chain, this backward process can be executed with a numerical solver [29, 34].

Neural beamforming, a deep learning method for speech enhancement based on neural networks, has been widely studied. Some methods firstly employ single-channel deep noise suppression network to produce a mask for each single-channel input, which are then used to derive the multi-channel spatial covariances of the noise signals for the minimum variance distortionless response (MVDR) beamformers [4, 10]. Another method, which is called all-deep-learning beamforming (ADL-MVDR), integrates mask estimation, spatial covariance calculation, and frame by frame beamforming into a single network [36]. These neural beamforming methods have achieved good results in multi-channel speech enhancement.

The two-stage model is a newly emerging architecture for speech enhancement, which consists of a two-level structure. This structure can effectively improve the speech quality and intelligibility by leveraging an intermediate prior that guides the subsequent optimization. The intermediate prior decomposes the original task into multiple subtasks, and enhances the interpretability of the model compared to the one-stage structure that only has one black box processing task. Therefore, it has attracted much attention. Some methods perform coarse filtering on the first stage, and then enter the second stage for finer filtering. The two-stage structure can be seen as a process of first enhancing and then refining the speech signals to obtain more accurate estimates of clean speech. For example, LeBlanc et al. proposed a two-stage enhancement method that uses deep neural networks for initial speech enhancement on the first stage, and a genetic algorithm based tuning optimization for adaptive multi-band spectral subtraction for further enhancement on the second stage [14]. Another method is to have a two-level structure targeting different goals, gradually achieving enhancement in sequence. As proposed by Nossier et al., the first stage uses the amplitude spectrum as the training target for denoising in the frequency domain, and the second stage performs further denoising and speech reconstruction in the time domain [25]. In addition, the diffusion model can also be applied to two-stage networks. Some researchers use the diffusion model as the backend of other speech enhancement methods to optimize and improve the enhanced speech, in order to obtain more pleasant and natural speech signals [27, 32].

3D speech enhancement is a challenging and worthwhile research topic that aims to produce realistic and high-quality speech signals. The task goal is to obtain clean single-channel speech from 3D multi-channel speech signals with noise and reverberation. 3D speech enhancement is a relatively underexplored domain, with most existing methods developed in association with the L3DAS Challenge in recent years. Notably, the top solutions in the

L3DAS22 Challenge employ beamforming techniques and two-stage architectures. ESPNET-SE, the first place solution, utilizes a two-stage structure that iteratively combines a neural network and a multi-frame multi-channel Wiener beamforming filter [18]. PCG-AIID, the third place solution, employs a two-stage structure of coarse filtering and spatial beamforming [16]. However, currently existing methods are all discriminative, and no one has yet applied generative methods to 3D speech enhancement. One reason is that generative methods require too many resources, especially for processing multi-channel signals. Another challenge is that generative methods need to handle the complex spatial information and acoustic characteristics of 3D speech signals. Given that the goal of 3D speech enhancement is to obtain clean single-channel speech from multi-channel noisy speech, using the generative model as the back-end will greatly reduce the difficulty of its application. Therefore, we can use generative model as the refiner for 3D speech enhancement.

To address these challenges, we propose a two-stage refiner system that combines a beamforming network and a diffusion generative model for 3D speech enhancement. The first stage uses a discriminative beamforming network to enhance the multi-channel speech and obtain a single-channel target speech with suppressed noise. This facilitates the second stage application of the generative diffusion model. The second stage uses a generative diffusion model to further refine the enhanced speech and recover the speech components that are lost or distorted during the beamforming process. The main contributions of this paper can be summarized as follows:

(1) A novel beamforming and diffusion-based refiner system are proposed for 3D speech enhancement that leverages the advantages of both neural beamforming network and generative diffusion model to obtain high-quality single-channel speech. The generative diffusion model utilizes its generation capabilities to restore the lost or distorted speech components during the beamforming enhancement process.

(2) We demonstrate that the generative diffusion model can serve as a universal backend for various neural beamforming networks and improve their denoising performance. This indicates the excellent generalization ability of the generative diffusion model as a refinement backend.

(3) We show that training the generative diffusion model with clean speech can learn the prior knowledge of speech more effectively and achieve better speech restoration results than training with neural network enhanced data.

In summary, our proposed system can handle complex and noisy scenarios and is a powerful tool for 3D speech enhancement. The refiner system can generate clean single-channel speech with high fidelity and intelligibility. The generative diffusion model is a versatile tool for speech refinement that can generalize across different beamforming networks.

2 Background

2.1 3D speech beamforming filter

The 3D B format signal can be expressed as:

$$\boldsymbol{M_b} = [M_W(l, f), M_Y(l, f), M_Z(l, f), M_X(l, f)]^T \tag{1}$$

$M(l, f)$ represents frequency domain signal, l , f represents frame and frequency respectively, b represents B-format, W, Y, Z, X are B-format channel indexs. The method for estimating B-format single-channel clean voice $\hat{S}_*(l, f)$ is as follows:

$$\hat{S}_*(l, f) = \boldsymbol{W^T}(l, f)\boldsymbol{M_b}(l, f) \tag{2}$$

$$\boldsymbol{W_b(l.f)} = [W_W(l, f), W_Y(l, f), W_Z(l, f), W_X(l, f)]^T \tag{3}$$

T represents matrix transposition. $\boldsymbol{W}(l.f)$ can be regarded as a B-format beamforming filter [28], and can be estimated by neural network. Firstly, the noise signal is processed using short-time Fourier transform (STFT), and then the parameters of the beamforming filter are estimated using a neural network. Finally, the STFT noise signal is multiplied

by the filter parameter matrix, summed, and transformed back to the time domain using inverse short-time Fourier transform (ISTFT) to obtain a single-channel signal. The beamforming process in the frequency domain is shown in Figure 1. Firstly, the signal is input into the network, and a neural network is used to estimate the beamforming filter. Then, the original signal is multiplied by the filter to obtain the estimated single-channel clean speech signal.

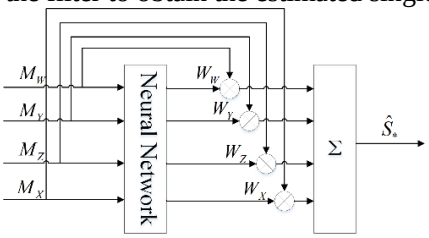

Fig. 1 Multi-channel 3D speech beamforming filter

2.2 Score-based generative diffusion model

The core idea of score-based speech enhancement diffusion model is to gradually increase the noise level during the forward diffusion process until the input speech is completely replaced by noise, and to train the score model for the reverse process by estimating the score $\nabla_{\mathrm{x}} \log p_{data}(x)$, the gradient of the log probability density with respect to the data. Once the score-based model is trained, an iterative program called Langevin dynamics can be used to extract samples from it [34].

The diffusion and reverse process described by SDE can be expressed as [1, 30]:

$$dx_t = f(x_t,t)dt + g(t)dw \tag{4}$$

$$dx_t = [f(x_t,t) - g(t)^2 \nabla_{x_t} \log p_t(x_t)]dt + g(t)d\bar{w} \tag{5}$$

where f is drift, g is diffusion, t is the expression of the progress degree of the forward process or the reverse process, w is the standard Wiener process, $\bar{w}$ is a standard Wiener process for time flowing in reverse, dt is the infinitesimal process step size, and in equation (4) dt is forward, in equation (5) is reverse. The score $\nabla_{x_t} \log p_t(x_t)$ is the logarithmic density at this process, which can be approximated by a process related score model $s_\theta(x_t,t)$ and solved by a predictor corrector for the reverse process.

They use speech $y = x_0 + n$ as input, where x_0 is clean speech, n is noise. The diffusion process is as follows:

$$dx_t = \gamma(y - x_t)dt + g(t)dw \tag{6}$$

$$g(t) = \sigma_{\min}(\sigma_{\max}/\sigma_{\min})^t \sqrt{2\log(\sigma_{\max}/\sigma_{\min})} \tag{7}$$

$\sigma_{\min}$ and $\sigma_{\max}$ parameterize the variance table of the added Gaussian noise, and γ is a constant that can be interpreted as a stiffness parameter for pulling x_0 to y as process t progresses. The state distribution of this process is called the perturbation kernel:

$$p_{0t}(x_t|x_0,y) = \mathcal{N}_{\mathbb{C}}(x_t; \mu_{x_0,y}(t), \sigma(t)^2\mathbf{I}) \tag{8}$$

where $\mathbf{I}$ is an identity matrix. Since each state of a forward process can be represented by its mean $\mu_{x_0,y}(t)$ and variance $\sigma(t)^2$, we can effectively sample x_t in process t, and the mean and variance can be expressed as [30]:

$$\mu_{x_0,y}(t) = e^{-\gamma t}x_0 + (1 - e^{-\gamma t})y \tag{9}$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2((\sigma_{\max}/\sigma_{\min})^{2t} - e^{-2\gamma t})\log(\sigma_{\max}/\sigma_{\min})}{\gamma + \log(\sigma_{\max}/\sigma_{\min})} \tag{10}$$

Because in the case of limited processes, it is not possible to reach $\mu_{x_0,y}(t)$ to y, they choose the stiffness parameter γ based on the condition of $\mathbb{E}[|\mu_{x_0,y}(1) - y|^2] < 10^{-3}$. They also apply amplitude transform and inverse

transform to all complex STFT coefficients c, to better adapt to the assumptions of scored-based diffusion models when representing speech signals in the complex unilateral STFT domain.

The training task is as in [10] and learn the parameter θ that minimizes the following:

$$\mathbb{E}_{t,x_0,x_t|x_0}[\| s_\theta(x_t,t,y)-\nabla_{x_t}\log p_{0t}(x_t \mid x_0,y)\|_2^2] \tag{11}$$

Due to $y=x_0+n$ enters (11) only as input to the model and as a regulating signal, the model is trained to estimate the Gaussian noise added in the forward process, and therefore this training task is considered purely generative [34]. Figure 2 illustrates the speech changes during the forward and reverse processes of the diffusion model. The forward process adds Gaussian white noise to the speech signal, gradually transforming it into a simple prior distribution. The reverse process reconstructs the previous signal step by step, eventually generating a clean speech signal.

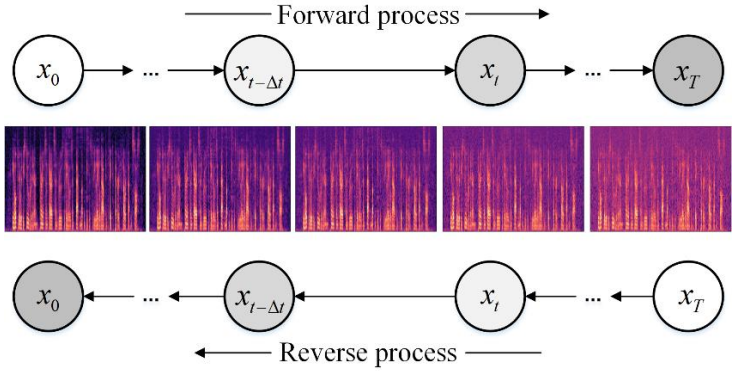

Fig. 2 Forward and Reverse process of diffusion model

3 proposed Methods

3.1 Dual microphone beamforming

Due to the fact that 3D speech simulates the sound heard by the human ears in a reverberant environment, two microphone arrays simulate the human ears for sound collection. We splice the B-format signals collected by two microphone arrays into an eight-channel signal as input, and its form is as follows: $[WA, YA, ZA, XA, WB, YB, ZB, XB]$, where A/B represents two microphone arrays. W, Y, Z, X is the channel index of microphone signals in the array.

So a complete input signal can be expressed as:

$$\begin{aligned}\boldsymbol{M}(l,f)=[&M_{WA}(l,f),M_{YA}(l,f),M_{ZA}(l,f),M_{XA}(l,f),\\&M_{WB}(l,f),M_{YB}(l,f),M_{ZB}(l,f),M_{XB}(l,f)]^T\end{aligned} \tag{12}$$

The corresponding beamforming filter is:

$$\begin{aligned}\boldsymbol{W}(l,f)=[&W_{WA}(l,f),W_{YA}(l,f),W_{ZA}(l,f),W_{XA}(l,f),\\&W_{WB}(l,f),W_{YB}(l,f),W_{ZB}(l,f),W_{XB}(l,f)]^T\end{aligned} \tag{13}$$

Then the beamforming filtering result is:

$$\hat{S}(l,f)=\boldsymbol{W}^{\mathrm{T}}(l,f)\boldsymbol{M}(l,f) \tag{14}$$

3.2 Score-based diffusion refiner

3.2.1 Refiner-Noisy

We use a score-based generative diffusion model as our refiner to improve the single-channel speech obtained from the enhanced neural beamforming network.

Unlike [34], we use beamformed speech $y = x_0 + n - l$ as input, where x_0 is clean speech, n is the residual noise after enhancement and l is the effective component lost during enhancement. Since $y = x_0 + n - l$ enters (11) only as input to the model and as a regulating signal, this training task is also considered purely generative. Algorithm 1 shows the training process of Refiner-Noisy.

Algorithm 1 Refiner-Noisy

For $i = \Delta t, 2\Delta t, ..., T$ do
  Sample $x_t \sim \mathcal{N}_{\mathbb{C}}(x_t; \mu_{x_0,y}(t), \sigma(t)^2\mathbf{I})$
  Compute $\mu_{x_0,y}(t)$ and $\sigma(t)^2$ using Eqs.(9) and (10)
  Take gradient step on
  $\mathbb{E}_{t,x_0,x_t|x_0}[|| s_\theta(x_t, t, y) - \nabla_{x_t} \log p_{0t}(x_t \mid x_0, y) ||_2^2]$
End for

Algorithm 2 Refiner-Clean

For $i = \Delta t, 2\Delta t, ..., T$ do
  Sample $x_t \sim \mathcal{N}_{\mathbb{C}}(x_t; x_0, \sigma(t)^2\mathbf{I})$
  Compute $\mu_{x_0,y}(t)$ and $\sigma(t)^2$ using Eqs. (10)
  Take gradient step on
  $\mathbb{E}_{t,x_0,x_t|x_0}[|| s_\theta(x_t, t, y) - \nabla_{x_t} \log p_{0t}(x_t \mid x_0, y) ||_2^2]$
End for

3.2.2 Refiner-Clean

In the previous section, we used $y = x_0 + n - l$ as input and mentioned that the diffusion model used was only trained to estimate the Gaussian noise added to the forward process. Our original intention was to use the score-based model as the backend of the beamformer to refine and repair speech, in order to improve the enhancement effect. The use of enhanced speech as training data has achieved results (see Chapter 3 for details). Since we are going to use it as a backend to repair the lost effective components, we think that adding noise and lost components to the input may affect the learning of the inherent properties of clean speech. Therefore, this time we use clean speech as the input to train the score-based diffusion model.

At this time, $y = x_0$, and equation (6) becomes:

$$dx_t = \gamma(x_0 - x_t)dt + g(t)dw \tag{15}$$

According to equation (9), $\mu_{x_0,y}(t) = x_0$, becomes a constant value, and the state distribution of the process (8) becomes simpler:

$$p_{0t}(x_t \mid x_0, y) = \mathcal{N}_{\mathbb{C}}(x_t; x_0, \sigma(t)^2\mathbf{I}) \tag{16}$$

At the same time, the impact of environmental noise and reverberation is completely removed, so that the task is closer to the generation task aimed at generating clean speech signals. Considering that in the inverse problem, the optimization formula for estimating the target image from noise is expressed as [23]:

$$\hat{x} = \arg\min_{x \in \mathbb{R}^n} f(x) \ \textit{with} \ f(x) = g(x) + h(x) \tag{17}$$

And it can be regarded as the maximum posterior probability probability estimator, when:

$$g(x) = -\log(p_{y|x}(x)) \ \textit{and} \ h(x) = -\log(p_x(x)) \tag{18}$$

where $p_{y|x}$ is the likelihood relating x to measurements y and p_x is the prior distribution. If $y = x_0 + n - l$ is used for training, there is not only the added Gaussian noise, but also the residual noise after enhancement and some lost effective components. This leads to a relatively small $p_{y|x}$, which leads to a larger $g(x)$, $f(x)$ and a smaller posterior probability. Using clean speech does not have these effects, which is more conducive to repairing more lost components in the reverse process and participating in multi-channel speech enhancement tasks as a refined repair backend. Algorithm 2 shows the training process of Refiner-Clean.

3.2.3 Refinement process

For refinement, the initial complex spectrum x_T is obtained by sampling from the prior distribution of $t = T$:

$$x_T \sim \mathcal{N}_{\mathbb{C}}(y, \sigma(T)^2\mathbf{I}) \tag{19}$$

Then we use the predictive corrector in [33] to perform the reverse process, and use a predictive correction combination of reverse diffusion sampling and annealed Langevin dynamics. We set the required parameter SNR of the corrector to 0.33 based on experience, and conducted a total of 50 iterations, with each iteration involving one correction.

3.3. Two-stage beamforming and diffusion-based refiner system

We first use pretrained SE (speech enhancement) module (beamforming stage) to perform noise and reverberation suppression on the 3D speech signals and outputs a single-channel enhanced signal. Then the diffusion model refiner stage further refines the enhanced signal and recovers the speech components that are lost or distorted during the beamforming stage. The overview of our system is shown in Figure 3.

As mentioned earlier, we consider two types of refiners: Refiner-Noisy and Refiner-Clean. Refiner-Noisy is trained with the enhanced slightly noisy speech, as shown by the orange dashed line in Figure 3. Refiner-Clean is trained with the clean speech, as shown by the red dashed line in Figure 3. The two refiners are obtained during the training phase and can serve as universal backends for different beamformers to refine the speech.

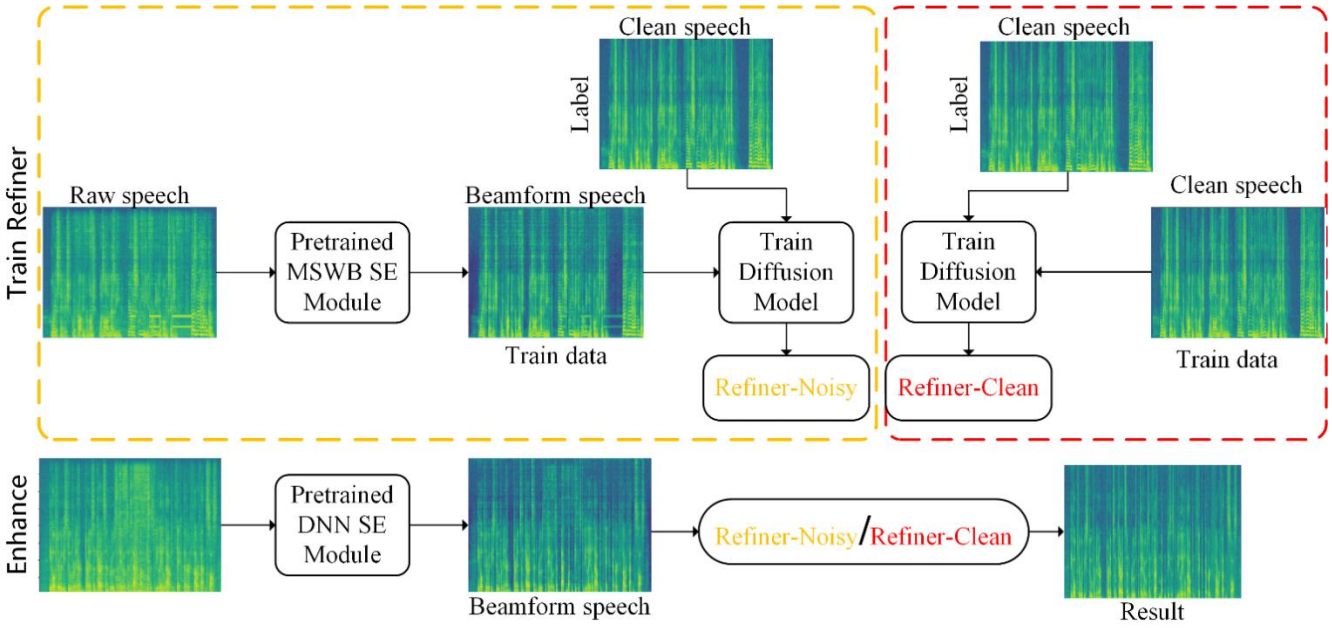

Fig. 3 Two-stage Beamforming and diffusion-based Refiner system

The specific implementation is as follows:

Training phase: we first train the neural beamforming network using the MSWB SE module to obtain a beamformer, which we choose because it is the best model among our models for 3D speech enhancement; then we use it to enhance the speech that has not participated in the beamformer training; the enhanced single-channel speech is used as the training data for the score-based diffusion model to train a Refiner-Noisy; the Refiner-Clean is directly trained with clean speech.

Enhancement phase: “Pretrained DNN SE Module” can be any beamforming network; we use the beamforming network to enhance test speech; then we refine the enhanced single-channel speech with one of our refiners.

4 System Setup

4.1 Neural network architecture

4.1.1 Neural beamforming network

Multiple-input Multiple-output UNet Beamforming (MMUB) [28] uses multi-channel U-Net to estimate beamforming filters, and multiplies them with multi-channel complex spectrograms, then sums the filtered multi-channel signals to obtain the estimation of the target signal.

Filter-and-Sum Network-Transform-Average-Concatenate (FaSNet-TAC) [21] is a time-domain beamforming model that adds a TAC module to a single stage filter and sum network. The TAC module enables the system to make global decisions using information from all microphones, while the filter and sum network estimates beamforming filters through model based methods.

Multiple-input Single-output WNet Beamforming (MSWB) [17] constructed a two-stage U-Net network using the method of first-stage coarse wave and second-stage fine filtering, and the beamforming enhancement process of the second stage will be adaptively adjusted based on the effect of the previous stage and the information in the original signal to improve the final enhancement effect. Compared to the above two methods, it achieves the best performance, and this network is used as the pre-trained SE module during the training procedure of our Refine-Beamformer system. Its network architecture is shown in Figure 4.

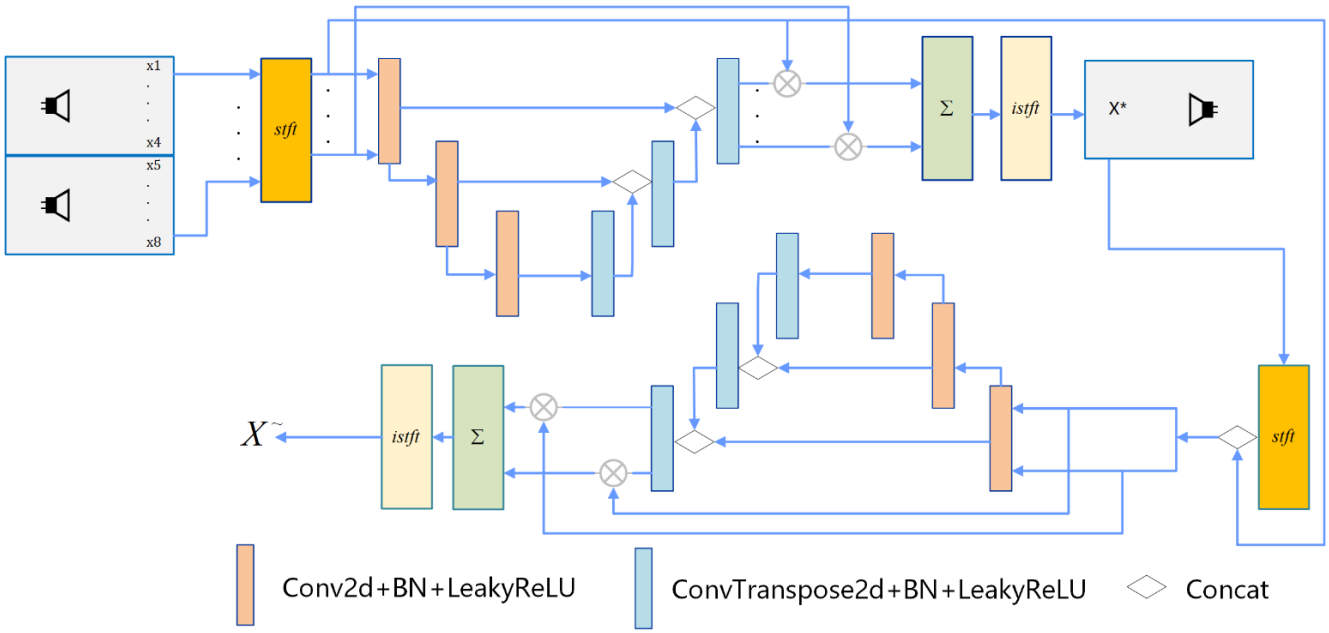

Fig. 4 MSWB SE Module

4.1.2 Refiner Model

In this work, we adopt the score-based diffusion model as a refiner to achieve 3D speech enhancement. The score-based diffusion model leverages the diffusion model to restore speech from noise [34], and has shown remarkable performance in speech enhancement. We use a deep complex U-Net with one complex-valued output channel for estimating the score and two complex-valued input channels for the training task and the score estimation, and its encoder/decoder pair is shown in Figure.5. This network incorporates time-embedding layers

into all encoder and decoder blocks, providing the DNN with information about the time-step. We also use a combination of reverse diffusion sampling and annealed Langevin dynamics to gradually enhance the speech [34].

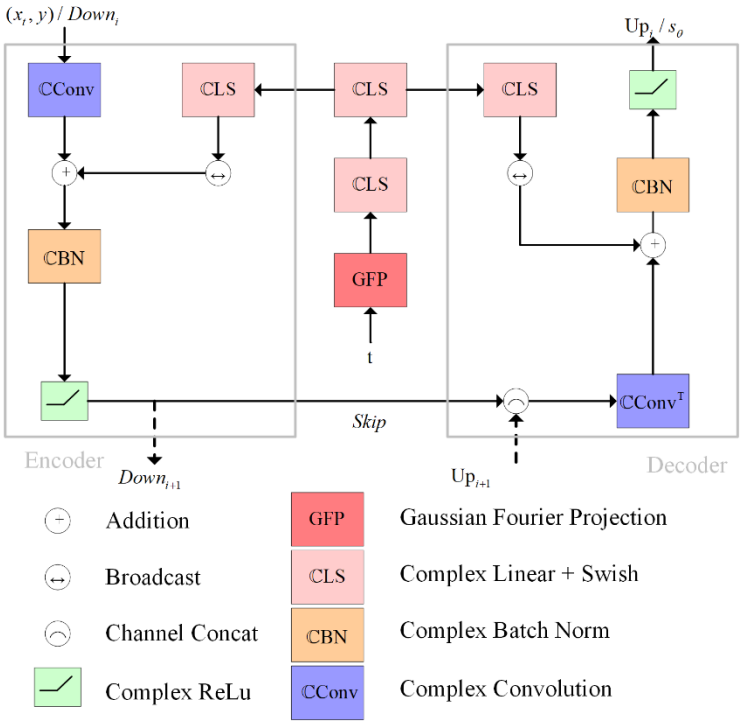

Fig. 5 Score Model Encoder/Decoder Pair Framework [34]

4.2 Dataset

We use the 3D audio dataset provided by L3DAS22 Challenge (https://www.l3das.com/editions.html) for model training. The multi-channel signal is simulated by convolving single-channel clean speech and multi-channel impulse response (IR) signals, and then adding background noise that is also convolved with another impulse response signal.

Clean speech comes from the LibriSpeech corpus [26], and noise signals come from FSD50K [5].

The signal capture is performed in a large office measuring 6 meters (length), 5 meters (width), and 3 meters (height). Two Ambonics microphone arrays (MicA, MicB) are placed in the center of the room. Each microphone array has 4 microphones and generates 4 channel signals. The distance between the A microphone array and the B microphone array is 20cm, which is close to the distance between the human ears. Both microphone arrays are located at a height of 1.3 meters, which is close to the ear height of a sitting person. The position of the source signal is randomly selected from 252 different positions.

4.3 Training procedure and parameters

We used L3DAS22 Task1 train100 from the L3DAS22 challenge dataset as the training set for the beamformer, and trained a batch size 3 model on NVIDIA GTX1660 GPU, and updated the parameters using the Adam gradient optimizer. The initial learning rate is 0.001, and it will decay at a rate of halving each time when the validation loss does not decrease within 2 rounds. The minimum learning rate is set to 0.0000001, and the training will stop and enter the testing phase when the validation loss does not decrease within 5 rounds. We used the single-channel dataset enhanced by beamforming from the L3DAS22 Task1 train360 as the training set for the Refiner-Noisy, and

the clean speech from L3DAS22 Task1 train360 as the training set for the Refiner-Clean. We trained a batch size 16 model on NVIDIA Tesla P100 GPU with a learning rate of 10^{-4}. We parameterize SDE (9) as follows: $\gamma = 1.5$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.5$, $t_{\varepsilon} = 0.03$. We track the exponential moving average of DNN weights with a decay of 0.999 for sampling. Both beamformer and refiner are supervised trained, with clean single-channel speech corresponding to the input signal as label, and 70% of the dataset is used for training, while the remaining 30% is used as validation set.

4.4 Loss function

The selection of an appropriate loss function is critical for training deep learning models. In this work, we evaluated two candidate loss functions - mean absolute error (MAE) and mean squared error (MSE) - for training the refiner system. We adopted a two-stage framework consisting of a neural beamforming network for initial enhancement followed by a score-based diffusion model refiner for further improvement.

For the beamformer, we utilized MAE as the loss function based on previous findings [28] that showed its superior performance over other losses.

For the diffusion model refiner, we compared MAE and MSE losses by training on a dataset enhanced by the MSWB (Figure 2). Models were trained with a batch size of 4 on an NVIDIA GTX1660 GPU. We monitored the scale invariant signal distortion ratio (SI-SDR) [2] and extended short-term objective intelligibility (ESTOI) [15] on the validation set during training. As shown in Table 1, the diffusion model trained with the MAE loss achieved substantially higher ESTOI (0.79 vs 0.60) and SI-SDR (9.05 vs 7.19) compared to MSE. The highest values achieved for each evaluation metric during validation of the refiner model training are highlighted in bold in Table 1.

Table 1 Comparison of different loss function of refiners

Loss	ESTOI ↑	SI-SDR ↑
MAE	**0.79**	**9.05**
MSE	0.60	7.19

This suggests that applying the MAE loss function consistently for both the beamforming and diffusion-based refiner stages leads to better performance. The reason is that changing the loss function may alter the training objective, thus compromising the enhancement effect of the first stage. By contrast, using the same loss function ensures the alignment of the two models' objectives, facilitating more effective refinement based on the initial beamformed output. Therefore, we conclude that the MAE loss function is more suitable than MSE for training the full pipeline proposed in this work.

4.5 Evaluation

The evaluation metric for 3D speech enhancement is a combination of short-term objective intelligibility (STOI) and word error rate (WER). We conducted pre-training on the 960h Librispeech corpus and computed the WER using the Wav2Vec speech recognition model [11]. The final metric is defined as:

$$Metric = (STOI + (1 - WER)) / 2 \tag{20}$$

STOI is in the range of 0-1, and the larger the value, the better. WER value is within the range of 0-1, and the smaller the value, the better. Therefore, the final measurement is in the range of 0-1, and the higher the value, the better.

In addition, we use non-intrusive speech quality assessment (NISQA) to evaluate the naturalness of enhanced speech at different stages [24]. NISQA is a perceptual quality measure that reflects how natural and pleasant the speech sounds to human listeners. It can capture the distortions and artifacts introduced by speech enhancement methods, such as spectral distortion, musical noise, and phase distortion. We use NISQA to compare the quality of speech enhanced by beamforming and diffusion-based refiner system. The higher the NISQA, the better.

5 Result and Analysis

5.1 Evaluation results and analysis

To validate the enhancement and refinement capability of our proposed two-stage beamforming and diffusion-based refiner system, we evaluated the beamformer and refiner system both independently using the L3DAS22 Task1 dev dataset. The overall evaluation results are shown in Table 2.

Table 2 shows the comparison of several configurations: beamformers alone (FaSNet-TAC, MMUB, MSWB 4/8ch), beamformers followed by a diffusion-based refiner trained on speech data enhanced by MSWB 8ch beamforming (Refiner-Noisy), and beamformers followed by a diffusion-based refiner trained on clean speech (Refiner-Clean). MSWB (8ch/4ch) denotes the network with single microphone array/dual microphone array inputs, while FaSNet-TAC and MMUB only have dual microphone array inputs. The arrow next to the metric indicates whether higher (↑), or lower (↓) values are better.

The results indicate that the refiner system (both Refiner-Noisy and Refiner-Clean) improved the metrics of Metric, WER, and NISQA scores for all beamformer front-ends, except for STOI. This demonstrates their robust enhancement and refinement capability for 3D speech with noise. It also shows that the diffusion model as the back end of speech enhancement has effective generation characteristics and strong generalization performance. This generalization performance allows the diffusion model to be used as a backend refiner and combined with various neural beamforming methods to form a more robust speech enhancement system.

Table 2 Evaluation results of the system and front-end on Dev

Method	STOI ↑	WER ↓	Metric ↑	NISQA ↑
FaSNet-TAC[21]	**0.777**	0.465	0.656	2.315
+Refiner-Noisy	0.774	0.436	0.669	2.382
+Refiner-Clean	**0.777**	**0.427**	**0.675**	**2.507**
MMUB[28]	**0.9**	0.207	0.847	2.835
+Refiner-Noisy	0.894	**0.177**	0.858	2.887
+Refiner-Clean	0.898	0.178	**0.86**	**2.990**
MSWB4ch[17]	**0.905**	0.191	0.857	3.064
+Refiner-Noisy	0.899	**0.161**	0.869	3.184
+Refiner-Clean	0.903	0.162	**0.871**	**3.271**
MSWB8ch[17]	**0.925**	0.158	0.884	3.241
+Refiner-Noisy	0.918	**0.128**	0.895	3.306
+Refiner-Clean	0.923	0.13	**0.896**	**3.396**

We hypothesize that the diffusion model can capture the intrinsic properties of speech signals and use them to refine the enhanced speech, regardless of the training conditions. This leads to an interesting question: will Refiner-Clean, which is trained on clean speech, have better performance than Refiner-Noisy, which is trained on enhanced speech by neural beamforming methods?

Remarkably, the experimental data shows that Refiner-Clean achieved slightly better performance than Refiner-Noisy, with consistent gains in the two important metrics of Meric and NISQA across different beamformers. This suggests that training with clean speech can help the refiner learn the prior knowledge of effective speech better, and thus repair the speech more clearly. One possible explanation for this finding is that training with enhanced signals may affect the refiner's judgment of effective speech due to the addition of noisy speech components and loss of speech components. However, the influence is very limited, because the noise within the signal does not spread throughout all time periods of the signal, and only exists in some time periods. In contrast, training with clean speech signals does not have such an impact, and the refiner can better leverage the diffusion process to restore the lost components.

A possible drawback of our proposed refiner system is that the STOI after refinement is slightly lower than before. The reason is that noise is added during the diffusion process, and the reverse process uses the predicted mean to repair effective speech. While repairing more effective speech, it also leads to a slight decrease in STOI caused by the residual Gaussian random noise in the signal. Furthermore, we observe that training the refiner with enhanced speech (Refiner-Noisy) leads to the lowest STOI among the three configurations, which confirms this point. This is because the enhanced speech has higher noise content and loss of speech components, which may impair the refiner's ability to learn the speech prior.

On the other hand, the improvement in WER performance indicates that the refined speech has fewer word errors, such as substitutions, deletions, and insertions, which effectively reduces the speech recognition errors. This demonstrates the ability of diffusion models to repair speech components that are lost or distorted. Moreover, the improvement in NISQA directly reflects the enhanced speech quality and naturalness of the refined speech. This confirms that using diffusion models as beamforming backend refiners can effectively improve the 3D speech enhancement performance and robustness.

It should be noted that the main focus of this paper is to investigate the performance and generalization of the refiner system based on the diffusion model in refining 3D speech enhanced by beamforming methods, rather than to compete with related SOTA models.

To further validate the effectiveness of our proposed refiner system, we also evaluated our best performing model series on a blind test set of L3DAS22 Task1 test dataset. As shown in Table 3, the results are consistent with those on the dev dataset, indicating that our proposed refiner system can achieve reliable and robust speech enhancement and refinement.

Table 3 Evaluation results of our best system series on a blind test set of L3DAS22 Task1 dataset

Method	STOI ↑	WER ↓	Metric ↑	NISQA ↑
MSWB8ch[17]	**0.923**	0.136	0.894	3.181
+Refiner-Noisy	0.916	**0.113**	0.902	3.234
+Refiner-Clean	0.921	**0.113**	**0.904**	**3.324**

To further demonstrate the generalization and effectiveness of our proposed method across different datasets, we also conducted experiments on the L3DAS21 Task1 dev set, which has different noise and reverberation characteristics than the L3DAS22 Task1 dataset. We observed that the NISQA metric was consistently improved by our refiner for all three beamformers, as shown in Figure 6. This indicates that our method can effectively enhance the speech quality across different acoustic conditions.

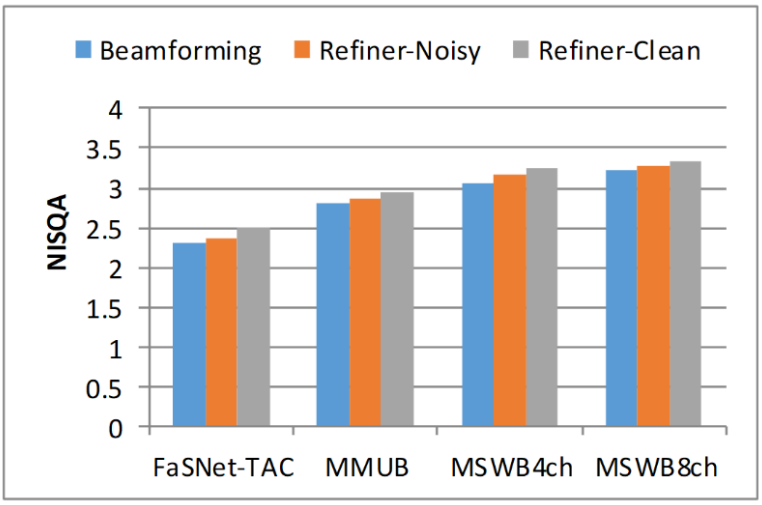

Fig. 6 NISQA on L3DAS21 dev set

We only used the NISQA metric to evaluate on the L3DAS21 dev set, because NISQA is a perceptual quality measure that reflects how natural and pleasant the speech sounds to human listeners. It can capture the distortions and artifacts introduced by speech enhancement methods, such as spectral distortion, musical noise, and phase distortion. Therefore, NISQA is more suitable for evaluating the refinement effect of the diffusion model as a backend, which is the main focus of this paper.

5.2 Mel spectrogram analysis

We also analyze the mel spectrograms processed by the system and front-end. Due to the close similarity of the spectrogram results of the two refiners, we only choose MSWB8ch as the beamformer and the corresponding Refiner-Clean as the Refine system to compare mel spectrogram of clean speech, as shown in Figure 7.

We randomly selected two speech signals for mel spectrogram analysis. Firstly, from an overall perspective, the mel spectra of clean speech are both the brightest and clearest, while the spectra only formed by beamforming are the dimmest. In contrast, the spectra refined by diffusion models are relatively bright. Secondly, we analyze from the part of spectrogram and find that in the spectrogram of speech (a), the refined spectrogram has more repaired components, as shown in the initial part of the timeline circled in the yellow oval dashed line. Moreover, in the spectrogram of speech (b), the refined spectrogram has more repaired components at high frequencies than the beamforming only spectrogram. These observations undoubtedly validate the feasibility and effectiveness of the diffusion model as a refined backend for repairing speech.

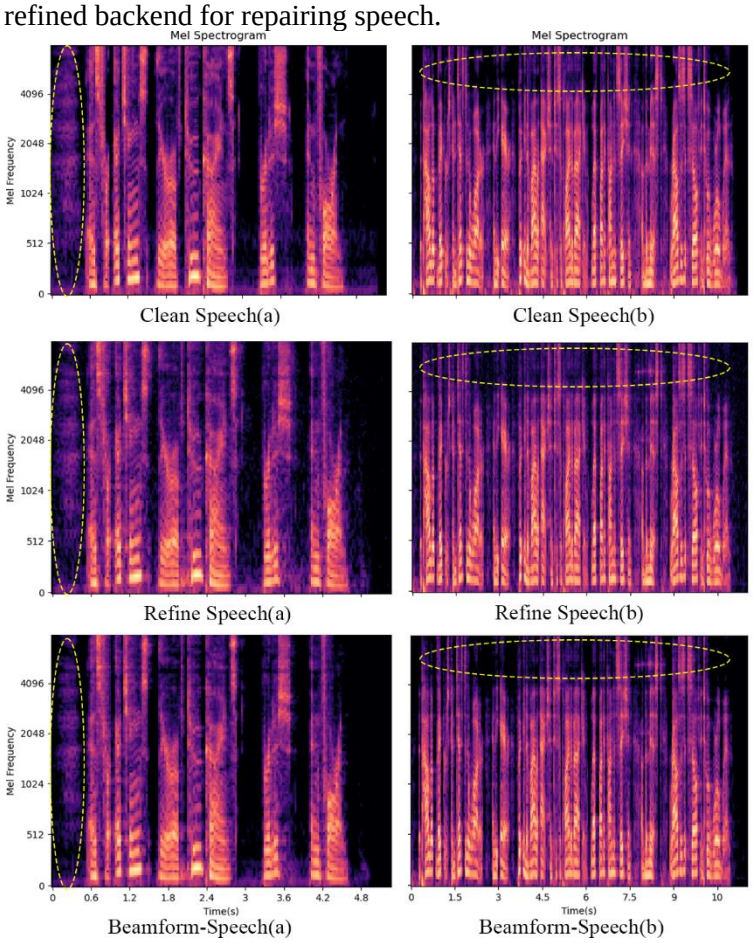

Fig. 7 Mel Spectrogram

However, we also noticed that there are some small components in the refined spectrogram that do not exist in the clean speech, which also confirms the argument in section 4.1 that there is residual Gaussian random noise after the reverse process generates speech from noise. Nevertheless, these components are much fewer than those that effectively repaired, and the improvement in metric also indicates this. This corresponds precisely to the slight sacrifice of STOI mentioned in section 5.1.

5.3 Discussion

Unlike the one-stage model that directly predicts the clean speech from the noisy speech, our two-stage model decomposes the speech enhancement task into two subtasks: noise suppression and refinement. Our two-stage method is based on the idea of integrating a neural beamforming network as the front-end and a score-based diffusion model as the backend for the multi-channel 3D speech enhancement task. This method leverages the advantages of the discriminative beamforming network, such as spatial filtering and noise reduction, and the powerful generation capabilities of the diffusion model, such as speech restoration and refinement. This method not only achieves good results in terms of relevant metrics, but also achieves consistent results across different datasets and beamforming networks, while avoiding the difficulty of directly processing multi-channel speech by the diffusion model.

Except for the multi-channel 3D speech enhancement task of the L3DAS competition, the idea of our two-stage method also has the potential to be applied to other tasks such as multi-channel speech separation, speaker recognition, etc. For example, the beamforming network can be replaced by a speech separation network to obtain the source signals, and the diffusion model can be conditioned on the speaker identity to generate the desired speech [29]. This kind of task generality will provide more possibilities for the researchers.

In terms of model method generality, our proposed method achieved consistent results across different datasets and beamforming methods. The authors can adjust the first stage beamforming network or the second stage diffusion model according to the specific objective.

Of course, although our two-stage method achieves good results and generalization ability, and has good performance in terms of Metric, WER, and NISQA metrics, its performance on the STOI metric has decreased. STOI is a short-term objective intelligibility measure that reflects how intelligible the speech is to human listeners. STOI is mainly related to the signal-to-noise ratio (SNR) and the speech intelligibility index (SII) of the speech. The reason why our method based on diffusion as the back-end causes STOI to decrease is that the diffusion process adds noise to the speech, and the denoising process uses the predicted mean to restore the speech. While restoring more speech components, it also leads to a slight decrease in STOI caused by the residual Gaussian random noise in the signal. One possible way to improve the STOI performance of the refiner system is to use a more sophisticated noise model in the diffusion process. Another possible way is to use a more accurate estimator than the predicted mean in the denoising process. It is left for future work to explore how to improve the second stage (backend) diffusion model to address the STOI degradation problem.

6 Conclusion

This paper proposes a novel two-stage refiner system for 3D speech enhancement, which combines a discriminative beamforming network and a generative diffusion model. The proposed system first applies a beamforming network to reduce noise and obtain initial enhanced signals from the multi-channel inputs. Then, it employs a generative diffusion model to refine the beamformed output by restoring speech components that are missing or distorted in the previous stage. Experimental results show that the diffusion-based backend effectively improves speech quality and intelligibility compared to stand-alone beamforming methods for 3D speech

enhancement, as quantified by metrics like WER and NISQA. Moreover, the refiner system generalizes well across different beamformer architectures and training data types. Our work highlights the potential of integrating discriminative and generative models for robust 3D speech enhancement, and also opens up new possibilities for applying generative diffusion models to multidimensional speech processing tasks, which are usually challenging due to the high dimensionality and complexity of multi-channel data. The generative diffusion model can leverage its strong generative capabilities to restore lost speech detail in both spatial and spectral domains. Future work can explore adopting generative diffusion models as an effective backend for other speech processing tasks, such as speech recognition, and speaker verification. We believe that our results strengthen the basis for leveraging the complementary strengths of discriminative and generative approaches for advanced 3D multi-channel speech systems.

Acknowledgements This work was supported partly by Jiangxi Province Degree and Postgraduate Education Teaching Reform Project (No. JXYJG-2023-134), Nanchang Hangkong University PhD Foundation (No. EA201904283) and Nanchang Hangkong University Graduate Foundation (No. YC2022-044).

Data availability All L3DAS challenge datasets used in this study are publicly available in https://www.l3das.com/editions.html .

References

[1] Anderson, B. D.: Reverse-time diffusion equation models: Stochastic Processes and their Applications 12(3): 313-326 (1982)
[2] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33: 12449-12460 (2020)
[3] Choi, H.-S., Park, S., Lee, J. H., et al.: Real-time denoising and dereverberation wtih tiny recurrent u-net. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2021)
[4] Erdogan, H., Hershey, J. R., Watanabe, S., et al.: Improved mvdr beamforming using single-channel mask prediction networks. In Interspeech, pp. 1981-1985 (2016)
[5] Fonseca, E., Favory, X., Pons, J., et al.: Fsd50k: an open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30: 829-852 (2021)
[6] Fu, S. W., Hu, T. Y., Tsao, Y., Lu, X.: Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In 2017 IEEE 27th international workshop on machine learning for signal processing (MLSP). pp. 1-6 (2017).
[7] Fu, S.-W., Tsao, Y., Lu, X., Kawai, H.: Raw waveform-based speech enhancement by fully convolutional networks. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE (2017)
[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial networks. Communications of the ACM 63(11): 139-144 (2020)
[9] Hendriks, R. C., Gerkmann, T., Jensen, J.: DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art. Synthesis Lectures on Speech and Audio Processing 9(1): 1-80 (2013)
[10] Heymann, J., Drude, L., Haeb-Umbach, R.: Neural network based spectral mask estimation for acoustic beamforming. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 196-200 (2016).
[11] Jensen, J., Taal, C. H.: An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(11): 2009-2022 (2016)
[12] Jiang, W., Sun, C., Chen, F., et al.: Low complexity speech enhancement network based on frame-level Swin transformer. Electronics, 12(6), 1330 (2023)
[13] Kingma, D. P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
[14] LeBlanc, R., Selouani, S. A.: A Two-Stage Deep Neuroevolutionary Technique for Self-Adaptive Speech Enhancement. IEEE Access 10: 5083-5102 (2022)
[15] Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J. R.: SDR–half-baked or well done?. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 626-630 (2019).
[16] Li, J., Zhu, Y., Luo, D., et al.: The PCG-AIID System for L3DAS22 Challenge: MIMO and MISO Convolutional Recurrent Network for Multi Channel Speech Enhancement and Speech Recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 9211-9215 (2022)
[17] Lin,W.,Chen,F.,Sun,C., Zhu, Z.: 3D Speech Enhancement Algorithm for Two-Stage U-Net Beamforming Network[J]. Computer Engineering and Applications, 59(22): 128-135 (2023).
[18] Lu, Y. J., Cornell, S., Chang, X., et al.: Towards low-distortion multi-channel speech enhancement: The ESPNET-SE submission to the L3DAS22 challenge. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 9201-9205 (2022)
[19] Lu, Y.-J., Tsao, Y., Watanabe, S.: A study on speech enhancement based on diffusion probabilistic model. 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE (2021)

[20] Lu, Y.-J., Wang, Z.-Q., Watanabe, S., et al.: Conditional diffusion probabilistic model for speech enhancement. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2022)
[21] Luo, Y., Chen, Z., Mesgarani, N., Yoshioka, T.: End-to-end microphone permutation and number invariant multi-channel speech separation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6394-6398 (2020)
[22] Luo, Y., Mesgarani, N.: 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)
[23] Mittag, G., Naderi, B., Chehadi, A., Möller, S.: Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. arXiv preprint arXiv:2104.09494 (2021)
[24] Mittag, G., Möller, S.: Deep learning based assessment of synthetic speech naturalness. arXiv preprint arXiv:2104.11673 (2021)
[25] Nossier, S. A., Wall, J., Moniri, M., et al.: Two-stage deep learning approach for speech enhancement and reconstruction in the frequency and time domains. In 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1-10 (2022).
[26] Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206-5210 (2015).
[27] Qiu, Z., Fu, M., Yu, Y., et al.: SRTNet: Time Domain Speech Enhancement Via Stochastic Refinement. arXiv preprint arXiv:2210.16805 (2022)
[28] Ren, X., Chen, L., Zheng, X., et al.: A neural beamforming network for b-format 3d speech enhancement and recognition. In 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1-6 (2021).
[29] Richter, J., Welker, S., Lemercier, J. M., et al.: Speech enhancement and dereverberation with diffusion-based generative models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2023).
[30] Särkkä, S., Solin, A.: Applied stochastic differential equations (Vol. 10). Cambridge University Press (2019)
[31] Serrà, J., Pascual, S., Pons, J., et al.: Universal speech enhancement with score-based diffusion. arXiv preprint arXiv:2206.03065 (2022)
[32] Sawata, R., Murata, N., Takida, Y., et al.: A versatile diffusion-based generative refiner for speech enhancement. arXiv preprint arXiv:2210.17287 (2022).
[33] Song, Y., Sohl-Dickstein, J., Kingma, D. P., et al.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
[34] Welker, S., Richter, J., Gerkmann, T.: Speech enhancement with score-based generative models in the complex STFT domain. arXiv preprint arXiv:2203.17004 (2022)
[35] Williamson, D. S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. IEEE/ACM transactions on audio, speech, and language processing 24(3): 483-492 (2015)
[36] Zhang, Z., Xu, Y., Yu, M., et al.: ADL-MVDR: All deep learning MVDR beamformer for target speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6089-6093 (2021).