Revisiting Denoising Diffusion Probabilistic Models for Speech Enhancement: Condition Collapse, Efficiency and Refinement

Wenxin Tai[1], **Fan Zhou**[1], **Goce Trajcevski**[2], **Ting Zhong**[1]*

[1] University of Electronic Science and Technology of China
[2] Iowa State University
amperetai@gmail.com, fan.zhou@uestc.edu.cn, gocet25@iastate.edu, zhongting@uestc.edu.cn

Abstract

Recent literature has shown that denoising diffusion probabilistic models (DDPMs) can be used to synthesize high-fidelity samples with a competitive (or sometimes better) quality than previous state-of-the-art approaches. However, few attempts have been made to apply DDPM for the speech enhancement task. The reported performance of the existing works is relatively poor and significantly inferior to other generative methods. In this work, we first reveal the difficulties in applying existing diffusion models to the field of speech enhancement. Then we introduce DR-DiffuSE, a simple and effective framework for speech enhancement using conditional diffusion models. We present three strategies (two in diffusion training and one in reverse sampling) to tackle the condition collapse and guarantee the sufficient use of condition information. For efficiency, we introduce the fast sampling technique to reduce the sampling process into several steps and exploit a refinement network to calibrate the defective speech. Our proposed method achieves state-of-the-art performance to the GAN-based model and shows a significant improvement over existing DDPM-based algorithms.

Introduction

Background interference can contaminate and vastly degrade the speech quality for human listeners. Speech enhancement (SE) aims to improve the hearing comfort by separating human voice and noise signals. Traditional SE methods have been extensively studied for a few decades (Boll 1979; Gerkmann and Hendriks 2011; Hu et al. 2013) and have achieved significant accomplishments. However, they often fail when faced with non-stationary scenarios. Recently, this issue has been relatively well alleviated with approaches based on the deep learning (DL) paradigms (Zhao et al. 2016; Tai et al. 2021a).

Most of the DL-based SEs focus on minimizing the overall difference between the target speech and its denoised estimate and, usually, L_p-norm distance training objectives are adopted for network training (Li et al. 2021; Tai et al. 2021b). However, these discriminative models have no notion of what realistic examples typically look like and how the features used for discrimination are combined with other features that define an object (Geirhos et al. 2020; Phan et al. 2020). In addition, since models are trained on a finite set of training data, they usually cannot generalize to unseen situations, e.g., different noise types, reverberation, and different signal-to-noise ratios (SNRs) (Welker, Richter, and Gerkmann 2022; Yu et al. 2021).

Conditional generative models have recently shown promises to overcome many limitations of their discriminative counterparts (Song et al. 2018; Geirhos et al. 2020; Fu et al. 2019, 2021). This property burgeoned another branch of SE – instead of learning a direct mapping from noisy to clean speech, likelihood-based generative SE models prefer to learn the distribution of clean speech and treat SE as a conditional generation problem. A few generative SE models have been proposed, such as generative adversarial networks (GANs)-based (Soni, Shah, and Patil 2018; Zhang et al. 2020; Routray and Mao 2022), variational autoencoders (VAEs)-based (Pariente, Deleforge, and Vincent 2019; Fang et al. 2021), and flow-based models (Maiti and Mandel 2020; Strauss and Edler 2021). Compared to discriminative approaches, generative models are more robust to unseen scenarios and are expected to produce more natural speech, as they optimize the predictive performance and learn the input distribution jointly (Pascual, Bonafonte, and Serra 2017; Welker, Richter, and Gerkmann 2022).

Denoising diffusion probabilistic models (DDPMs) (Ho, Jain, and Abbeel 2020) have emerged as a class of generative models, demonstrating their capability to achieve significant performance improvements in both image and audio synthesis. Specifically, in diffusion models, a fixed forward process gradually perturbs the data in a step-wise fashion towards fully random noise. A parametrized reverse process is learned to perform iterative denoising and to generate samples from the noise. Compared to other generative models, DDPMs demonstrate surprisingly high sample quality and have easier training patterns while having (almost) no restrictions on model architecture (Sohl-Dickstein et al. 2015). Recently, some researchers have tried to apply DDPMs to SE (Zhang, Jayasuriya, and Berisha 2021; Lu, Tsao, and Watanabe 2021; Lu et al. 2022). Although better generalizations have been made, the performance of DDPM-based methods is still lower than that of other generative SE models. Therefore, an important question is: *what hinders the performance of DDPM in the field of SE?*

*Corresponding author.

Following are the **major challenges**:

C1: *Weak condition influence*. Many works try to apply conditional information as an additional input of the model, hoping that the model will automatically pay attention to conditional information (Saharia et al. 2021; Li et al. 2022; Whang et al. 2022). However, by looking at the optimization objective of the diffusion model, we find that the training process seems to be independent of the conditional factors. As a result, the semantic correspondence between the conditioning noisy speech and the synthesized clean speech can sometimes get lost.

C2: *Inefficient/Inaccurate inference procedure*. While DDPMs inherently are step-wise iterative learning models, a guarantee of high sample quality typically comes at the cost of hundreds to thousands of denoising steps. When reducing the sampling steps to accelerate inference, however, quality degradation due to perceivable background noise has been observed (Dhariwal and Nichol 2021; Kong et al. 2021).

Present work & Contributions. We investigate the drawbacks of exploiting diffusion models for SE, and propose DR-DiffuSE – a denoising and refining model – to tackle SE with the following specifics.

S1: We design an auxiliary conditional generation network to generate reliable condition representations. We also propose a conditional DDPM to generate speech conditioned on the previous generated reliable condition. To better utilize the condition information and prevent the model from the condition collapse problem, we present a novel dual-path parallel network architecture that can provide fine-grained condition guidance for the diffusion model. Moreover, we provide additional non-parameterized guidance during the reverse process by interpolating diffusing condition features with the intermediate output of DDPM.

S2: We introduce a fast sampling technique that can reduce the inference process from hundreds or thousands steps into several (e.g., 6) steps, significantly improving the speech generation efficiency. To compensate for the quality degeneration, we design a refinement network to refine the defective speech generated from the conditional DDPM. By sharing parameters between the refinement network and the condition generation network, DR-DiffuSE is much more robust and generalizable in complex scenarios.

Comprehensive experiments conducted on two benchmark datasets show that DR-DiffuSE achieves the comparable and even better performance compared to the state-of-the-art generative SE models, as well as significant improvement over existing DDPM-based algorithms under both seen and unseen conditions.

Preliminaries & Motivations

We now describe the SE task, the basic background of DDPMs, and the challenges *w.r.t.* the use of condition information and inference efficiency.

Speech Enhancement (SE) refers to methods that try to reduce distortions, make speech sounds more pleasant, reduce listening effort and improve intelligibility. In real environments, the received noisy signal $\boldsymbol{y}$ in the time domain can be modeled as: $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$ where $\boldsymbol{x}$ and $\boldsymbol{n}$ denote clean and noise signals, respectively. We assume the dataset is composed of M data pairs $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)|1 \leq i \leq M\}$, where $\boldsymbol{x}_i$ is the i-th ground truth clean speech, and $\boldsymbol{y}_i$ is the noisy one. For human perception, the primary goal of speech enhancement is to improve the intelligibility and quality via extracting $\boldsymbol{x}$ from $\boldsymbol{y}$.

Denoising Diffusion Probabilistic Model

DDPMs refer to a series of generative neural networks consisting of two processes: forward diffusion and parameterized reverse.

Forward Diffusion. A forward diffusion maps data to noise by gradually perturbing the input data through a stochastic process that starts from a data sample and iteratively generates noisier samples using a simple Gaussian diffusion kernel. Let $\boldsymbol{x}^0 \sim q_{\text{data}}$. The diffusion process is a Markovian noising process which gradually adds Gaussian noise to the data to produce noised samples $\boldsymbol{x}^1$ through $\boldsymbol{x}^T$:

$$q(\boldsymbol{x}^1, \cdots, \boldsymbol{x}^T|\boldsymbol{x}^0) = \prod_{t=1}^{T} q\left(\boldsymbol{x}^t \mid \boldsymbol{x}^{t-1}\right), \tag{1}$$

and each step of the noising process adds Gaussian noise according to some variance schedule given by β_t (a predefined, small positive constant):

$$q(\boldsymbol{x}^t \mid \boldsymbol{x}^{t-1}) = \mathcal{N}(\boldsymbol{x}^t; \sqrt{1-\beta_t}\boldsymbol{x}^{t-1}, \beta_t \boldsymbol{I}). \tag{2}$$

And, there is a closed form expression for $q(\boldsymbol{x}^t|\boldsymbol{x}^0)$:

$$q(\boldsymbol{x}^t|\boldsymbol{x}^0) = \mathcal{N}(\boldsymbol{x}^t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}^0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Therefore, when T is large enough, $\bar{\alpha}_t$ goes to 0, and $q(\boldsymbol{x}^T|\boldsymbol{x}^0)$ becomes the latent Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Reverse process. If we reverse the above diffusion process and sample from $q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t\right)$, we can reconstruct the true sample from a Guassian input $\boldsymbol{x}^T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$:

$$q(\boldsymbol{x}^0, \cdots, \boldsymbol{x}^T) = q(\boldsymbol{x}^T) \prod_{t=1}^{T} q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t\right). \tag{4}$$

Unfortunately, $q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t\right)$ is not easy to estimate because it needs to use the entire dataset. Thus, DDPMs propose to learn a model $p_{\boldsymbol{\theta}}$ to approximate the conditional probabilities as:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}^0, \cdots, \boldsymbol{x}^T) = p_{\boldsymbol{\theta}}(\boldsymbol{x}^T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t\right). \tag{5}$$

Note that if β_t is small enough, $q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t\right)$ will also be Gaussian. Thus, $p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t\right)$ can be represent as:

$$p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t\right) = \mathcal{N}(\boldsymbol{x}^{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}^t, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}^t, t)), \tag{6}$$

which is a parametrized reverse process that turns over the forward diffusion and performs iterative denoising. It represents data synthesis and is trained to generate data by converting random noise into realistic data. Both forward and reverse processes often use thousands of iterations (especially in the reverse process), making DDPM impractical for real-time SE systems.

Weak Condition Influence

One of the main reasons that the diffusion-based models (Lu et al. 2022; Zhang, Jayasuriya, and Berisha 2021) cannot achieve the effect of previous SE methods is the lack of exploiting the condition information. The most common way of utilizing condition information in the training process is to concatenate the condition representation $\boldsymbol{y}$ with the intermediate noisy spectrogram $\boldsymbol{x}^t$ along the channel dimension and serve them as the data input $(\boldsymbol{x}^t, \boldsymbol{y})$ (Li et al. 2022; Whang et al. 2022), *or* add the conditional information $\boldsymbol{y}$ into each block, which is similar to using the time step t in existing approaches (Kong et al. 2021; Lu, Tsao, and Watanabe 2021; Lu et al. 2022). However, *condition collapse* occurs in both two solutions.

Definition 1. ***Condition collapse.*** *Condition collapse refers to the problem where conditional DDPMs appear to ignore the condition input and produce the output arbitrarily.*

The failure to leverage conditional information is related to the training objective of the diffusion model. Reviewing the training process of the diffusion model, it requires to minimize the Kullback–Leibler (KL) divergence between $q(\boldsymbol{x}^{t-1}|\boldsymbol{x}^t, \boldsymbol{x}^0)$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}^{t-1}|\boldsymbol{x}^t)$. Since $\boldsymbol{x}^t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}^0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, we can derive that $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}^t, t)$ should equal to $\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)$ given $(\boldsymbol{x}^t, t)$. Using the condition factor $\boldsymbol{y}$ as extra-input to the function $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}^t, t)$ indicates that $\boldsymbol{y}$ has no obvious connections to the training target. Consequently, $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}^t, \boldsymbol{y}, t)$ may estimate $\boldsymbol{x}^{t-1}$ with the help of $(\boldsymbol{x}^t, \boldsymbol{y}, t)$, or just $(\boldsymbol{x}^t, t)$.

Empirically, two cases may cause the condition collapse problem in the field of SE: (i) The condition representation is too noisy to provide reliable information for predicting the movement of the diffusion process, i.e., $0 \leftarrow \mathcal{I}(\boldsymbol{x}^{t-1}, \boldsymbol{y}) \leq \mathcal{I}(\boldsymbol{x}^{t-1}, (\boldsymbol{y}, t)) \ll \mathcal{I}(\boldsymbol{x}^{t-1}, (\boldsymbol{x}^t, t))$, where $\mathcal{I}$ denotes the mutual information. In this case, the model tends to completely ignore the condition information and estimate $\boldsymbol{x}^{t-1}$ with only $(\boldsymbol{x}^t, t)$; and (ii) Inappropriate architecture design restricts the utilization of condition information for DDPMs generation. Insensitive to condition information, previous DDPM-based SE methods (Lu, Tsao, and Watanabe 2021; Zhang, Jayasuriya, and Berisha 2021; Lu et al. 2022) forcibly add conditional information into each block. However, such an operation requires the conditional information to adapt to all latent manifold space of each block, which is hard, if not impractical, since the feature representation space of different blocks generally have different grains. The above inappropriate architecture design weakens the value of condition representation inadvertently, but leads to sub-optimal conditional generation quality.

Trade-offs between Efficiency and Quality

While DDPMs inherently are iterative-learning models, a guarantee of high sample quality typically comes at the cost of hundreds or thousands of denoising steps. It's extremely time-consuming compared to other generative methods such as GAN and VAE. This limitation restricts the application of DDPMs in many scenarios requiring real-time SE. To speed up the reverse sampling procedure, a few studies (Lu, Tsao, and Watanabe 2021; Lu et al. 2022) introduce the fast sampling technique (Kong et al. 2021), which accelerates the inference process by reducing the number of the reverse denoising steps. Although this operation can boost efficiency, it inevitably sacrifices the speech generation quality. Therefore, making compromise between efficiency and quality is mandatory in previous methods.

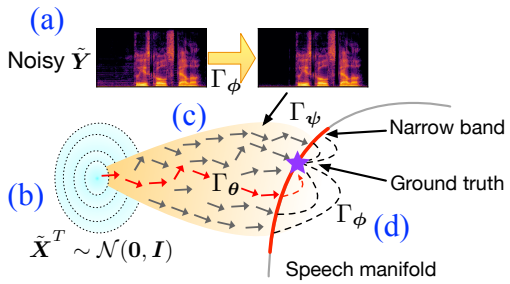

Figure 1: The workflow of DR-DiffuSE: (a) A condition generation network that generates reliable condition. (b) Generate $\tilde{\boldsymbol{X}}^T$ by sampling from a random Gaussian distribution. (c) Leverage the generated reliable condition representation to provide guidance for the diffusion model, encouraging the generative model to remove noise along the expected trace. Due to the stochastic property of DDPMs and the rough estimation caused by the fast sampling, the generated samples will fall in a narrow band of speech manifold. (d) Recalibration via the refinement network $\Gamma_{\boldsymbol{\phi}}$.

Methodology

We first reformulate SE as a conditional generation problem, and then present the details of data representation. We propose three strategies to address the condition collapse issue – two in diffusion training and one in reverse sampling – to guarantee the sufficient condition information utilization. To alleviate the inefficiency problem while maintaining the generation quality, we introduce a fast sampling technique to reduce the sampling process into several steps and exploit a refinement network to calibrate the defective speech. The idea of DR-DiffuSE is illustrated in Figure 1.

Formulation and Data Representation

We assume a dataset of input-output speech pairs, denoted as $\mathcal{D}=\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^M$, which represents samples drawn from an unknown conditional distribution $q(\boldsymbol{x}|\boldsymbol{y})$. We approach this problem by adapting the DDPM to conditional speech generation – the goal of our model is to reverse the Gaussian diffusion process by iteratively recovering signals from noise through a reverse Markov chain conditioned on $\boldsymbol{y}$. Specifically, the conditional DDPM generates a target clean speech $\boldsymbol{x}^0$ after T refinement steps. Starting with a pure noise speech $\boldsymbol{x}^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the model refines the speech through successive iterations $\boldsymbol{x}^{[0:T-1]}$ according to learned conditional transition distributions $p_{\boldsymbol{\theta}}(\boldsymbol{x}^{t-1}|\boldsymbol{x}^t, \boldsymbol{y})$. Thus,

the complete reverse diffusion process for generation is:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}^0, \cdots, \boldsymbol{x}^T|\boldsymbol{y}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}^T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t, \boldsymbol{y}\right). \quad (7)$$

Different from prior DDPM-based works (Lu, Tsao, and Watanabe 2021; Lu et al. 2022; Zhang, Jayasuriya, and Berisha 2021) that directly use time-domain speech waveform as input, we represent speech signals in complex-valued time-frequency domain via Short-time Fourier transform (STFT), as Fourier theory provides a feasible feature representation. We treat the clean speech $\boldsymbol{X}$ and the noisy speech $\boldsymbol{Y}$ as elements of $\mathbb{C}^{T\times F}$ – which can also be represented as $\mathbb{R}^{2\times T\times F}$ if we couple the magnitude and phase into Cartesian coordinates, constructing real and imaginary pairs.

Since the global distribution of STFT speech amplitudes is typically heavy-tailed, the information visible in untransformed spectrograms is dominated by only a small portion of bins. Inspired by recent works (Ju et al. 2022; Yu et al. 2022) that power compression can decrease the dynamic range of the spectrum and improve the significance of low-energy regions with more informative speech components, we apply an amplitude compression to adjust the energy distribution that may attain approximate normalization. The transformation and its associated inverse transformation is defined as:

$$\tilde{\boldsymbol{X}} = \sqrt{|\boldsymbol{X}|}e^{i\angle(\boldsymbol{X})} \quad \Leftrightarrow \quad \boldsymbol{X} = |\tilde{\boldsymbol{X}}|^2 e^{i\angle(\tilde{\boldsymbol{X}})}. \quad (8)$$

Speech Generation via Conditional DDPM

To prevent the diffusion model from the condition collapse, we put forward three complementary solutions: (i) generate the reliable condition via a condition generation network, (ii) design a dual-path parallel network for fine-grained condition guidance, and (iii) provide additional guidance during the reverse process by interpolating diffusing condition features with intermediate output of DDPM.

Reliable condition information generation. Considering that $p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^{t-1}|\tilde{\boldsymbol{X}}^t, \tilde{\boldsymbol{Y}}, t)$ needs to remove complex noise from noisy speech $\tilde{\boldsymbol{Y}}$ and Gaussian noise from $\tilde{\boldsymbol{X}}^t$ simultaneously, we propose to decompose non-Gaussian and Gaussian noise estimates separately, towards mitigating the feature modeling pressure of DDPMs while reducing the risk that DDPMs tend to ignore condition information. Specifically, we exploit an auxiliary network $\Gamma_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Y}})$ to generate a better/reliable condition than noisy spectrogram $\tilde{\boldsymbol{Y}}$. To this end, we train $\Gamma_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Y}})$ to estimate the clean speech $\tilde{\boldsymbol{X}}^0$ to the greatest extent. The new form of the conditional DDPM becomes:

$$p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^{t-1}|\tilde{\boldsymbol{X}}^t, \Gamma_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Y}}), t). \quad (9)$$

Fine-grained condition representation. Unlike previous works that use condition information casually, we propose a more elegant way to provide fine-grained and matched conditional guidance for each block in DDPM. Given the reliable output from the condition generation network, we design an auxiliary condition delivery network $\Gamma_{\boldsymbol{\psi}}$ which has a similar architecture with that of the diffusion model $\Gamma_{\boldsymbol{\theta}}$. Due to the architecture similarity, we can construct a dual-path parallel network architecture and pass the same-level and fine-grained condition information from the auxiliary condition delivery network to the diffusion model. For simplicity, we use $\boldsymbol{\theta}$ to denote $\{\boldsymbol{\theta}, \boldsymbol{\psi}\}$ (parameters of the conditional DDPM) in the following content.

Additional guidance in reverse process. So far, to prevent the collapse problem, we have already presented reliable condition generation and well-defined information exchange strategies during training. These tricks, however, do not guarantee the resistance of the model to condition collapse. Here, we further explore a different approach by exploiting a classifier $p(\boldsymbol{C}|\tilde{\boldsymbol{X}}^t)$ to improve the diffusion generator and maintain the stability of model training. A recent work (Dhariwal and Nichol 2021) shows one way to achieve this goal, wherein a pre-trained diffusion model can be conditioned using the gradients of a classifier. In particular, considering that training an extra classifier is inefficient and inaccuracy, we present a non-parameterized difference minimizer instead, utilizing the gradient between the diffusion condition representation $\boldsymbol{C}^t$ and intermediate output $\tilde{\boldsymbol{X}}^t$ of diffusion model as the outside guidance. Suppose we have the mean $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_{\boldsymbol{\theta}}$ and variance $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ for each step t. and let $\boldsymbol{C}$ denote the output of the condition generation network. We can obtain $\tilde{\boldsymbol{X}}^t$ by sampling from:

$$\tilde{\boldsymbol{X}}^t = \mathcal{N}(\boldsymbol{\mu} + \zeta\boldsymbol{\Sigma}\nabla \log p_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}}(\boldsymbol{C}^t \mid \tilde{\boldsymbol{X}}^t), \boldsymbol{\Sigma}), \quad (10)$$

where ζ is a gradient scale and $\boldsymbol{C}^t$ denotes the diffusing result $\boldsymbol{C}^t = \mathcal{N}(\boldsymbol{C}^t; \sqrt{\bar{\alpha}_t}\boldsymbol{C}, (1-\bar{\alpha}_t)\boldsymbol{I})$ according to Eq.(3).

When we use MSE criterion as the difference minimizer and substitute the optimizing target from the maximization of $p_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}}(\boldsymbol{C}^t \mid \tilde{\boldsymbol{X}}^t)$ to the minimization of the difference between two variables, the gradient for $\tilde{\boldsymbol{X}}^t$ can be defined as $2(\tilde{\boldsymbol{X}}^t - \boldsymbol{C}^t)$. Applying the gradient into Eq.(10), we have:

$$\tilde{\boldsymbol{X}}^t = \mathcal{N}(\boldsymbol{\mu} - 2\zeta\boldsymbol{\Sigma}(\tilde{\boldsymbol{X}}^t - \boldsymbol{C}^t), \boldsymbol{\Sigma}). \quad (11)$$

Note that this gradient-based outsider guidance is equivalent to interpolate the estimated intermediate results from conditional DDPM with diffusing condition features. That is, we first generate a candidate $\tilde{\boldsymbol{X}}^t$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, we calculate the difference between $\tilde{\boldsymbol{X}}^t$ and $\boldsymbol{C}^t$, and further compute the gradient $2(\tilde{\boldsymbol{X}}^t - \boldsymbol{C}^t)$. Finally, we update the $\tilde{\boldsymbol{X}}^t$ to $(1-2\zeta\boldsymbol{\Sigma})\tilde{\boldsymbol{X}}^t + 2\zeta\boldsymbol{\Sigma}\boldsymbol{C}^t$.

Reverse diffusion process. The goal of training the reverse diffusion is to maximize the log-likelihood of the clean spectrogram conditioned on the paired noisy spectrogram $\mathbb{E}[\log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^0|\tilde{\boldsymbol{Y}})]$. Since directly optimizing the exact log-likelihood is intractable, we alternatively maximize its vari-

ational lower bound:

$$\begin{aligned}
\mathcal{L}_{\text{VLB}} &= \mathbb{E}_q \left[\log \frac{q(\tilde{\boldsymbol{X}}^1, \cdots, \tilde{\boldsymbol{X}}^T | \tilde{\boldsymbol{X}}^0)}{p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^0, \cdots, \tilde{\boldsymbol{X}}^T | \tilde{\boldsymbol{Y}})} \right] \\
&= \mathbb{E}_q \left[\log \frac{q(\tilde{\boldsymbol{X}}^{1:T} | \tilde{\boldsymbol{X}}^0)}{p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^{1:T} | \tilde{\boldsymbol{X}}^0, \tilde{\boldsymbol{Y}})} - \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^0 | \tilde{\boldsymbol{Y}}) \right] \\
&= D_{\text{KL}}(q(\tilde{\boldsymbol{X}}^{1:T} | \tilde{\boldsymbol{X}}^0) \| p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^{1:T} | \tilde{\boldsymbol{X}}^0, \tilde{\boldsymbol{Y}})) - \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^0 | \tilde{\boldsymbol{Y}}) \\
&\geq -\log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^0 | \tilde{\boldsymbol{Y}}), \qquad (12)
\end{aligned}$$

which can be adapted into the training objective $\mathcal{L}$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \mathbb{E}_q[-\log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^0 | \tilde{\boldsymbol{X}}^1, \Gamma_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Y}})) + \\
&\underbrace{\sum_{t=2}^{T} D_{\text{KL}}(q(\tilde{\boldsymbol{X}}^{t-1} | \tilde{\boldsymbol{X}}^t, \tilde{\boldsymbol{X}}^0) \| p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^{t-1} | \tilde{\boldsymbol{X}}^t, \Gamma_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Y}})))]}_{\text{Conditional DDPM training}} \\
&+ \underbrace{\mathbb{E}_q \left[D_{\text{KL}}(q(\tilde{\boldsymbol{X}}^0) \| p_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Y}}) \right]}_{\text{Condition generation training}}. \qquad (13)
\end{aligned}$$

For the conditional DDPM, the loss function can be defined as minimizing the KL divergence at each step. To simplify the training objective, (Ho, Jain, and Abbeel 2020) suggests to fix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ as $\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \boldsymbol{I}$ and only estimate the $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^t, \boldsymbol{C}, t)$ via the re-parameterization trick:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^t, \boldsymbol{C}, t) = \frac{1}{\sqrt{\alpha_t}}(\tilde{\boldsymbol{X}}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^t, \boldsymbol{C}, t)).$$

The goal is to train $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^t, \boldsymbol{C}, t)$ that can predict $\boldsymbol{\epsilon}$:

$$\mathcal{L}(\boldsymbol{\theta}) = E_{t \sim \mathcal{U}([T]), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}} \left(\tilde{\boldsymbol{X}}^t, \boldsymbol{C}, t \right) \right\|^2 \right],$$

where $\mathcal{U}([T])$ is the uniform distribution over $\{1, 2, \cdots, T\}$.

Speech generation. After training conditional DDPM, the network $\boldsymbol{\theta}$ learns to predict the noise $\boldsymbol{\epsilon}$ added to each step, which can be used to obtain $\tilde{\boldsymbol{X}}^{t-1}$ from $\tilde{\boldsymbol{X}}^t$ via:

$$\frac{1}{\sqrt{\alpha_t}}(\tilde{\boldsymbol{X}}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}^t, \boldsymbol{C}, t)) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t} \boldsymbol{\epsilon}.$$

Then we update $\tilde{\boldsymbol{X}}^{t-1}$ by interpolating conditional guidance from the condition generation network:

$$\tilde{\boldsymbol{X}}^{t-1} = (1 - 2\zeta\boldsymbol{\Sigma})\tilde{\boldsymbol{X}}^{t-1} + 2\zeta\boldsymbol{\Sigma}\boldsymbol{C}^{t-1}. \qquad (14)$$

Starting with a pure noise speech $\tilde{\boldsymbol{X}}^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the model iteratively refines the speech through successive iterations $\tilde{\boldsymbol{X}}^{[0:T-1]}$ according to Eq.(7).

Sampling Acceleration

Another issue with existing neural diffusion process is a large number of iterative steps required to reconstruct the target distribution during reverse sampling. In this work, we adapt the *fast sampling technique* (Kong et al. 2021) to enable more efficient inference.

Fast sampling technique. Given a trained model from above, (Kong et al. 2021) discovered that the most effective denoising steps in sampling occur near $t = 0$ and accordingly derived a fast sampling algorithm. The key idea is to break down the T-step reverse process into a S-step process with carefully designed variance schedule. Let $S \ll T$ be the number of steps in the reverse process and $\{\eta_s\}_{s=1}^S$ be the user defined variance schedule. We can derive γ_s and $\bar{\gamma}_s$:

$$\gamma_s = 1 - \eta_s, \bar{\gamma}_s = \prod_{i=1}^{s} \gamma_s, \qquad (15)$$

where γ_s and $\bar{\gamma}_s$ correspond to α_t and $\bar{\alpha}_t$ in original reverse sampling process. At s-th sampling step, we use $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot, s)$ to eliminate noise. The aligned s^{align} is defined as:

$$s^{\text{align}} = t + \frac{\sqrt{\bar{\alpha}_t} - \sqrt{\bar{\gamma}_s}}{\sqrt{\bar{\alpha}_t} - \sqrt{\bar{\alpha}_{t+1}}}, \qquad (16)$$

where s^{align} is a floating-point number, and is different from the interger diffusion-step in training. Finally, the parameterizations of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are defined as:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\text{fast}}(\cdot, s^{\text{align}}) = \frac{1}{\sqrt{\gamma_s}}(\tilde{\boldsymbol{X}}^s - \frac{\eta_s}{\sqrt{1-\bar{\gamma}_s}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot, s^{\text{align}})),$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\text{fast}}(\cdot, s^{\text{align}}) = \frac{1-\bar{\gamma}_{s-1}}{1-\bar{\gamma}_s}\eta_s \boldsymbol{I}, \qquad (17)$$

where $(\cdot, s^{\text{align}})$ denotes $(\tilde{\boldsymbol{X}}^s, \boldsymbol{C}, s^{\text{align}}))$. Therefore, the sampling efficiency is significantly improved by converting the reverse steps from T step to S step ($S \ll T$).

Speech Refinement

Different from other generative models, diffusion models catch dynamic dependencies from noisy audio instead of clean ones, which introduces more variation information (i.e, noise levels) in addition to the spectrogram fluctuation (Huang et al. 2022). This stochastic property leads to the same condition that can generate distinct samples, which is fine and necessary for generative tasks. However, the SE task has a unique ground-truth objective. In addition, with limited receptive field patterns, a distinct degradation will happen when reducing the reverse iterations, e.g., when adopting the fast sampling technique (Kong et al. 2021; Whang et al. 2022). More importantly, since the performance of conditional DDPM is seriously influenced by the quality of condition signals, the limited generalization of the condition generation network under mismatched conditions will severely affect the SE performance in unseen scenarios.

To overcome above issues, we propose a refinement network to improve the flawed speech from the previous conditional DDPM. The basic idea is to reuse the condition generation network as the refinement network. Towards that goal, we supervise the refinement network $\Gamma_{\boldsymbol{\phi}}$ by minimizing the difference between the refined speech $\hat{\boldsymbol{X}}$ and ground truth

Method	Year	Domain	STOI(%)	PESQ	CSIG	CBAK	COVL	SSNR
Unprocessed	–	–	92.1	1.97	3.35	2.44	2.63	1.68
SEGAN	2017	Time	–	2.16	3.48	2.94	2.80	7.73
MMSEGAN	2018	Time	93.0	2.53	3.80	3.12	3.14	–
MetricGAN	2019	Time	–	2.86	3.99	3.18	3.42	–
DSEGAN	2020	Time	93.2	2.39	3.46	3.11	2.90	8.72
PR-WaveGlow	2020	Time-Frequency	91.0	–	3.80	2.40	3.10	–
M-CRGAN	2020	Time-Frequency	94.0	2.92	4.16	3.24	3.54	–
MetricGAN+	2021	Time-Frequency	–	**3.15**	4.14	3.16	<u>3.64</u>	–
SE-Flow	2021	Time	–	2.28	3.70	3.03	2.97	7.93
CycleGAN	2021	Time-Frequency	<u>94.3</u>	2.90	<u>4.24</u>	**3.57**	3.49	<u>9.33</u>
PSMGAN	2022	Time-Frequency	–	2.92	3.88	<u>3.45</u>	3.62	–
DiffuSE (Large, 6 steps)	2021	Time	–	2.44	3.66	2.83	3.03	–
CDiffuSE (Large, 6 steps)	2022	Time	91.4	2.52	3.72	2.91	3.10	5.28
DR-DiffuSE	–	Time-Frequency	**94.9**	<u>3.09</u>	**4.38**	**3.57**	**3.76**	**9.52**

Table 1: Comparison of different generative models on VoiceBank-DEMAND dataset.

clean speech $\tilde{\boldsymbol{X}}^0$. The loss function of the refinement network is calculated by the RI (real and imaginary) components and the magnitude of the estimated spectrum as:

$$\mathcal{L}^{\text{RI}}(\phi) = \left\| \hat{\boldsymbol{X}}_r - \tilde{\boldsymbol{X}}_r^0 \right\|_F^2 + \left\| \hat{\boldsymbol{X}}_i - \tilde{\boldsymbol{X}}_i^0 \right\|_F^2,$$

$$\mathcal{L}^{\text{Mag}}(\phi) = \left\| \sqrt{|\hat{\boldsymbol{X}}_r|^2 + |\hat{\boldsymbol{X}}_i|^2} - \sqrt{|\tilde{\boldsymbol{X}}_r^0|^2 + |\tilde{\boldsymbol{X}}_i^0|^2} \right\|_F^2,$$

$$\mathcal{L}(\phi) = \lambda_{\text{RI}} \mathcal{L}^{\text{RI}}(\phi) + \lambda_{\text{Mag}} \mathcal{L}^{\text{Mag}}(\phi), \tag{18}$$

where λ_{RI} and λ_{Mag} are the weighting hyper-parameters. With a well-trained refinement network, DR-DiffuSE is able to speed up the reverse sampling speed by dozens of times while maintaining or even improving the SE performance.

Experiments

We now describe the details of our experimental evaluation. Code is available at https://github.com/judiebig/DR-DiffuSE.

Network Design

Architecture. Both sub-networks use complex spectrum as input and have a similar network topology – the standard encoder-decoder UNet architecture. Specifically, for condition generation network, instead of using a regular convolutional layer as the basic unit for encoder and decoder, five BiConvGLUs (bi-convolutional gated linear units) (Tai et al. 2021b) are utilized to consecutively compress the spectral features while capturing the local spectral-temporal patterns. In the bottleneck layer, we use stacked temporal convolution modules (S-TCM) proposed in (Li et al. 2021) to capture the long-range time dependencies. Compared to condition generation network Γ_{ϕ}, the architecture of conditional DDPM $\Gamma_{\boldsymbol{\theta}}$ has a little difference – we insert the representation of diffusion time step and fine-grained condition information into each block. ***Encoder***: Within each block, the kernel size is set to (2, 3) in the time and frequency axis except (2, 5) in the first block. We fix the channel dimension of all blocks to 64, and the stride is set to (1, 2). The stride we set means that while the frequency size is gradually halved, the time size remains unchanged to meet the real-time requirement. After each convolution, InstanceNorm and PReLU layers are followed to accelerate training. ***Bottleneck***: We use 3 groups of S-TCMs as the sequential module, each of which stacks 6 S-TCM units with dilation rate exponentially increasing to obtain a large temporal receptive field, i.e., (1, 2, 4, 8, 16, 32). In this way, the network can leverage the relation among different temporal scales to boost the speech information recovery. ***Decoder***: The decoder has a symmetrical architecture with *Encoder*, where the compressed features are gradually interpolated and restored to the original size. Namely, the decoder is composed of five BiConvGLU (Transposed convolution, "ConvTranspose2d") with fixed channel dimension 64, and the stride is set to (1, 2) with kernel size (2, 3). In particular, the kernel size of the last BiConvGLU is set to (2, 5). We use the skip connection strategy to mitigate the information loss.

Time encoding: As the parameters of the neural network are shared across time, we employ sinusoidal position embeddings to encode time step t. At time step t, we follow (Ho, Jain, and Abbeel 2020) to embed the step index into an 128-dimensional positional encoding vector $\boldsymbol{e}$:

$$\begin{aligned} \boldsymbol{e}_t = \Big[&\sin\left(10^{\frac{0\times4}{63}} t\right), \cdots, \sin\left(10^{\frac{63\times4}{63}} t\right) \\ &\cos\left(10^{\frac{0\times4}{63}} t\right), \cdots, \cos\left(10^{\frac{63\times4}{63}} t\right) \Big] \end{aligned} \tag{19}$$

When applying the fast sampling technique, we generate the embedding of s^{align} by linearly interpolating with its surrounding neighbors. Suppose the neighbors are t and $t+1$, we have:

$$\boldsymbol{e}_{s^{\text{align}}} = \boldsymbol{e}_t + (s^{align} - t)(\boldsymbol{e}_{t+1} - \boldsymbol{e}_t) \tag{20}$$

Condition injecting: Given the generated reliable condition representation $\boldsymbol{C}$, we first add it as the additional input – concatenating with the diffusion noisy spectrogram $\tilde{\boldsymbol{X}}^t$. Then we present a parallel condition delivery strategy that

injects the same-level fine-grained condition feature from each block of the condition delivery network to each block of the diffusion model. After the processing of each BiConvGLU block, we first use a convolution layer with (1, 1) kernel to transform the condition feature into a new representation space. Then, we inject condition information by simply adding it to the output of BiConvGLU. Finally, a convolution layer with (1, 1) kernel is applied to aggregate information.

Datasets, Protocols and Evaluation Metrics

We use the VoiceBank-DEMAND dataset (Veaux, Yamagishi, and King 2013; Thiemann, Ito, and Vincent 2013) for performance evaluations. To investigate the generalization ability of models, we use CHiME-4 (Vincent et al. 2017) as another test dataset following (Lu et al. 2022), i.e., the models are trained on VoiceBank-DEMAND and evaluated on CHiME-4. We use the following metrics to evaluate SE performance: the perceptual evaluation of speech quality (PESQ) (Rix et al. 2001), short-time objective intelligibility (STOI) (Taal et al. 2010), segmental signal-to-noise ratio (SSNR), the mean opinion score (MOS) prediction of the speech signal distortion (CSIG) (Hu and Loizou 2007), the MOS prediction of the intrusiveness of background noise (CBAK) (Hu and Loizou 2007) and the MOS prediction of the overall effect (COVL) (Hu and Loizou 2007).

Comparison with Advanced Baselines

The overall performance of DR-DiffuSE and the baselines is shown in Table 1, yielding the following observations:

Our DR-DiffuSE consistently outperforms the baselines by a significant margin, except for the PESQ score that MetricGAN+ surpasses DR-DiffuSE by 0.06. This is probably because that MetricGAN+ designed a specific metric-based loss function that is strongly positive with PESQ scores (it is also the origin of their method name). Besides, since the generative model in GAN is exactly the same as a standard discriminative SE model, the only difference is an extra model that can provide additional signals that are hard to design loss function, e.g., metric scores like PESQ and STOI. Therefore, their performance is often comparable to or better than classic discriminative methods. However, GAN models are known for potentially unstable training due to their adversarial training nature.

Likelihood-based generative models like DiffuSE and CDiffuSE typically perform poorly and cannot meet the requirement of the SE task. This phenomenon is partially related to their optimization target – diffusion models are not trained to directly minimize the difference between the generated audio and the reference audio, but to de-noise intermediate noisy samples iteratively to recover the clean speech and estimate a surrogate variational bound (Huang et al. 2022). Besides, improper condition information usage also restricts their performance. To compensate for the loss of quality caused by the optimization target of DDPM and fast sampling technique, DR-DiffuSE exploits the refinement training to calibrate the defective speech. Furthermore, DR-DiffuSE consists of three schemes to guarantee the full utilization of condition information. As a result, the above designs ensure that DR-DiffuSE performs significantly better than previous works.

Method	PESQ	CSIG	CBAK	COVL
Unprocessed	1.27	2.61	1.93	1.88
WaveCRN	1.43	2.53	2.03	1.91
Demucs	1.38	2.50	2.08	1.88
Conv-TasNet	1.63	1.70	1.82	1.54
CDiffuSE (Large)	1.66	2.98	2.19	2.27
D-DiffuSE (200 steps)	1.72	3.00	2.26	2.36
D-DiffuSE+R (200 steps)	<u>1.81</u>	<u>3.04</u>	<u>2.47</u>	<u>2.38</u>
DR-DiffuSE	**1.85**	**3.06**	**2.61**	**2.41**

Table 2: Trained on Voicebank, tested on CHIME-4.

Generalization and Efficiency Considerations

Compared to discriminative counterparts, generative models are more robust to unseen noisy and are expected to produce more natural speech, as they jointly optimize the predictive performance and data distribution. Here, we investigate the generalization capabilities of DR-DiffuSE against other L_p-loss-based discriminative approaches and recent DDPM-based SE methods. As shown in Table 2, given a domain shift in the testing data, regression-based approaches (Demucs (Defossez, Synnaeve, and Adi 2020), Conv-TasNet (Luo and Mesgarani 2019), and WaveCRN (Hsieh et al. 2020)) suffer from a significant performance drop. DDPM-based generative methods are much more resilient against such shifts in signal characteristics. Compared to CDiffuSE, DR-DiffuSE achieves significant improvement, showing that it is particularly resistant to shifts in noise characteristics of the speech data. Furthermore, D-DiffuSE+R (apply refinement training but directly generate samples from the conditional DDPM) reaps better performance than D-DiffuSE, demonstrating the benefit of retraining condition generation network: preventing the condition generation network fail in unseen scenarios and thus improving the generalization of the conditional DDPM.

Conclusion

This work first revealed the difficulties in applying diffusion models to the field of speech enhancement. We introduced DR-DiffuSE, a simple and effective framework for speech enhancement using conditional diffusion models. To address the condition collapse problem, we presented three approaches (two in diffusion training and one in reverse sampling) to guarantee the sufficient use of condition information. Besides, we introduced a fast sampling technique to reduce the sampling process into several steps and proposed a refinement network to calibrate the defective speech. In the future, we will focus on few-shot learning conditions by exploiting DR-DiffuSE to extract knowledge from unlabeled data, and generating semi-labeled data to alleviate the problem of poor SE performance caused by limited examples.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No.62072077 and No.62176043), Natural Science Foundation of Sichuan Province (Grant No. 2022NSFSC0505), and National Science Foundation SWIFT (Grant No.2030249). We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. Special thanks to Yen-Ju Lu for his kind help!

References

Boll, S. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2): 113–120.

Defossez, A.; Synnaeve, G.; and Adi, Y. 2020. Real time speech enhancement in the waveform domain. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 8780–8794.

Fang, H.; Carbajal, G.; Wermter, S.; and Gerkmann, T. 2021. Variational autoencoder for speech enhancement with a noise-aware encoder. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 676–680. IEEE.

Fu, S.-W.; Liao, C.-F.; Tsao, Y.; and Lin, S.-D. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning (ICML)*, 2031–2041. PMLR.

Fu, S.-W.; Yu, C.; Hsieh, T.-A.; Plantinga, P.; Ravanelli, M.; Lu, X.; and Tsao, Y. 2021. Metricgan+: An improved version of metricgan for speech enhancement. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Gerkmann, T.; and Hendriks, R. C. 2011. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4): 1383–1393.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 6840–6851.

Hsieh, T.-A.; Wang, H.-M.; Lu, X.; and Tsao, Y. 2020. Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement. *IEEE Signal Processing Letters*, 27: 2149–2153.

Hu, X.; Wang, S.; Zheng, C.; and Li, X. 2013. A cepstrum-based preprocessing and postprocessing for speech enhancement in adverse environments. *Applied acoustics*, 74(12): 1458–1462.

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

Huang, R.; Lam, M. W.; Wang, J.; Su, D.; Yu, D.; Ren, Y.; and Zhao, Z. 2022. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 4157–4163.

Ju, Y.; Rao, W.; Yan, X.; Fu, Y.; Lv, S.; Cheng, L.; Wang, Y.; Xie, L.; and Shang, S. 2022. TEA-PSE: Tencent-ethereal-audio-lab personalized speech enhancement system for ICASSP 2022 DNS CHALLENGE. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 9291–9295. IEEE.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*.

Li, A.; Liu, W.; Zheng, C.; Fan, C.; and Li, X. 2021. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1829–1843.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*.

Lu, Y.-J.; Tsao, Y.; and Watanabe, S. 2021. A study on speech enhancement based on diffusion probabilistic model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 659–666. IEEE.

Lu, Y.-J.; Wang, Z.-Q.; Watanabe, S.; Richard, A.; Yu, C.; and Tsao, Y. 2022. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7402–7406. IEEE.

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

Maiti, S.; and Mandel, M. I. 2020. Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 206–210. IEEE.

Pariente, M.; Deleforge, A.; and Vincent, E. 2019. A Statistically Principled and Computationally Efficient Approach to Speech Enhancement Using Variational Autoencoders. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Pascual, S.; Bonafonte, A.; and Serra, J. 2017. SEGAN: Speech enhancement generative adversarial network. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Phan, H.; McLoughlin, I. V.; Pham, L.; Chén, O. Y.; Koch, P.; De Vos, M.; and Mertins, A. 2020. Improving GANs for

speech enhancement. *IEEE Signal Processing Letters*, 27: 1700–1704.

Rix, A. W.; Beerends, J. G.; Hollier, M. P.; and Hekstra, A. P. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 749–752. IEEE.

Routray, S.; and Mao, Q. 2022. Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network. *Computer Speech Language*, 71: 101270.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2021. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2256–2265. PMLR.

Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; and Kushman, N. 2018. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

Soni, M. H.; Shah, N.; and Patil, H. A. 2018. Time-frequency masking-based speech enhancement using generative adversarial network. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5039–5043. IEEE.

Strauss, M.; and Edler, B. 2021. A flow-based neural network for time domain speech enhancement. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 5754–5758. IEEE.

Taal, C. H.; Hendriks, R. C.; Heusdens, R.; and Jensen, J. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 4214–4217. IEEE.

Tai, W.; Lan, T.; Wang, Q.; and Liu, Q. 2021a. IDANet: An Information Distillation and Aggregation Network for Speech Enhancement. *IEEE Signal Processing Letters*, 28: 1998–2002.

Tai, W.; Li, J.; Wang, Y.; Lan, T.; and Liu, Q. 2021b. Foster Strengths and Circumvent Weaknesses: a Speech Enhancement Framework with Two-branch Collaborative Learning. *arXiv preprint arXiv:2110.05713*.

Thiemann, J.; Ito, N.; and Vincent, E. 2013. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19, 035081. Acoustical Society of America.

Veaux, C.; Yamagishi, J.; and King, S. 2013. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *International Conference Oriental COCOSDA*, 1–4. IEEE.

Vincent, E.; Watanabe, S.; Nugraha, A. A.; Barker, J.; and Marxer, R. 2017. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech Language*, 46: 535–557.

Welker, S.; Richter, J.; and Gerkmann, T. 2022. Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via stochastic refinement. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, 16293–16303.

Yu, G.; Li, A.; Zheng, C.; Guo, Y.; Wang, Y.; and Wang, H. 2022. Dual-branch Attention-In-Attention Transformer for single-channel speech enhancement. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 7847–7851. IEEE.

Yu, G.; Wang, Y.; Wang, H.; Zhang, Q.; and Zheng, C. 2021. A two-stage complex network using cycle-consistent generative adversarial networks for speech enhancement. *Speech Communication*, 134: 42–54.

Zhang, J.; Jayasuriya, S.; and Berisha, V. 2021. Restoring degraded speech via a modified diffusion model. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2753–2757.

Zhang, Z.; Deng, C.; Shen, Y.; Williamson, D. S.; Sha, Y.; Zhang, Y.; Song, H.; and Li, X. 2020. On loss functions and recurrency training for GAN-based speech enhancement systems. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Zhao, Y.; Wang, D.; Merks, I.; and Zhang, T. 2016. DNN-based enhancement of noisy and reverberant speech. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 6525–6529. IEEE.