
DOSE: Diffusion Dropout with Adaptive Prior for Speech Enhancement

Wenxin Tai¹, Yue Lei¹, Fan Zhou^{1,2*}, Goce Trajcevski³, Ting Zhong^{1,2}
University of Electronic Science and Technology of China
Kashi Institute of Electronics and Information Industry
Iowa State University

Abstract

Speech enhancement (SE) aims to improve the intelligibility and quality of speech in the presence of non-stationary additive noise. Deterministic deep learning models have traditionally been used for SE, but recent studies have shown that generative approaches, such as denoising diffusion probabilistic models (DDPMs), can also be effective. However, incorporating condition information into DDPMs for SE remains a challenge. We propose a *model-agnostic* method called DOSE that employs two efficient condition-augmentation techniques to address this challenge, based on two key insights: (1) We force the model to prioritize the condition factor when generating samples by training it with dropout operation; (2) We inject the condition information into the sampling process by providing an informative adaptive prior. Experiments demonstrate that our approach yields substantial improvements in high-quality and stable speech generation, consistency with the condition factor, and inference efficiency. Codes are publicly available at <https://github.com/ICDM-UESTC/DOSE>.

1 Introduction

Speech enhancement (SE) aims to improve the intelligibility and quality of speech, particularly in scenarios where degradation is caused by non-stationary additive noise. It has significant practical implications in various fields such as telecommunications [1], medicine [2], and entertainment [3]. Modern deep learning models are often used to learn a deterministic mapping from noisy to clean speech. While deterministic models have long been regarded as more powerful in the field of SE, recent advancements in generative models [4, 5] have significantly closed this gap.

One such generative approach is based on using denoising diffusion probabilistic models (DDPMs) [6, 7], which have been shown to effectively synthesize natural-sounding speech. Several diffusion enhancement models have been developed [4, 5, 8], which try to learn a probability distribution over the data and then generate clean speech conditioned on the noisy input. A key challenge in using diffusion enhancement models is how to effectively incorporate condition information into learning and generating faithful speech [9, 10, 8]. Previous works address this issue through designing specific condition-injecting strategies [4, 9, 11] or devising complex network architectures [10, 5].

We conduct a thorough examination to understand the limitation of diffusion-based SE methods and find that diffusion enhancement models are susceptible to *condition collapse*, where the primary cause of inconsistent generation is the *non-dominant position of the condition factor*. We thus introduce a new paradigm to effectively incorporate condition information into the diffusion enhancement models. Specifically, we propose a Diffusion-drOpout Speech Enhancement method (DOSE), which is a model-agnostic SE method (Figure 1) that employs two efficient condition-augmentation tech-

*Corresponding author: fan.zhou@uestc.edu.cn

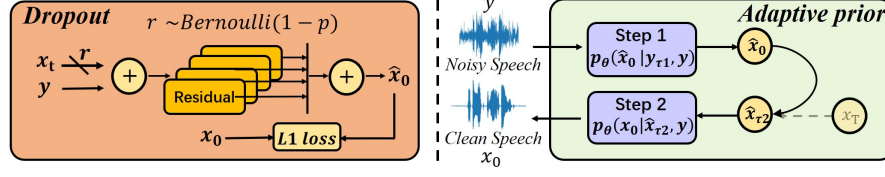


Figure 1: An illustration of the proposed DOSE. DOSE consists two primary procedures: (1) training a condition diffusion model using dropout operation, and (2) generating speech using a conditional diffusion model equipped with the adaptive prior.

niques: (1) During training, we randomly drop out intermediate-generated samples. This dropout mechanism guides the model’s attention toward the condition factors; (2) Instead of letting the model generate samples from scratch (Gaussian distribution), we employ an adaptive prior derived from the conditional factor to generate samples. Experiments on benchmark datasets demonstrate that our method surpasses recent diffusion enhancement models in terms of both accuracy and efficiency. Additionally, DOSE produces more natural-sounding speech and exhibits stronger generalization capabilities compared to deterministic mapping-based methods using the same network architecture.

2 Related works

There are two main categories of diffusion-based SE methods: (1) designing specific condition-injecting strategies [4, 9, 11, 5], or (2) generating speech with an auxiliary condition optimizer[12, 10, 8]. The first category considers noisy speech in the diffusion (or reverse) process, either by linearly interpolating between clean and noisy speech along the process [4, 9], or by defining such a transformation within the drift term of a stochastic differential equation (SDE) [11, 5].

Works from the second category rely on an auxiliary condition optimizer – a generator (diffusion model) synthesizes clean speech and a condition optimizer informs what to generate [12, 10, 8]. Both the generator and condition optimizer have the ability to denoise, with the latter undertaking the core part. Given the challenges in leveraging condition information [8], diffusion-based SE methods within this category often necessitate specific network architecture design to guarantee the participation of condition factors.

In a paradigm sense, our method is quite similar but different to the second branch – unlike previous approaches that require additional auxiliary networks, DOSE is an end-to-end diffusion-based SE method. In addition, DOSE is model-agnostic that does not need any specific network design to guarantee consistency between the generated sample and its corresponding condition factor.

3 Preliminaries

We now provide a brief introduction to the diffusion probabilistic model (diffusion models, for short), the definition of speech enhancement, and the condition collapse problem.

3.1 Diffusion models

A diffusion model [13, 6] consists of a forward (or, diffusion) process and a reverse process. Given a data point x_0 with probability distribution $p(x_0)$, the forward process gradually destroys its data structure by repeated application of the following Markov diffusion kernel:

$$p(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}), \quad t \in \{1, 2, \dots, T\}, \quad (1)$$

where β_1, \dots, β_T is a pre-defined noise variance schedule. With enough diffusion step T , $p(x_T)$ converges to the unit spherical Gaussian distribution. Based on the Markov chain, the marginal distribution at arbitrary timestep t has the following analytical form:

$$p(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad t \in \{1, 2, \dots, T\}, \quad (2)$$

where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$.

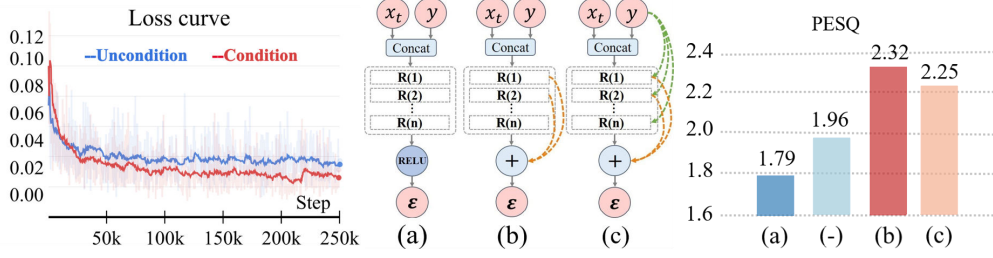


Figure 2: Investigation of the condition collapse problem. From left to right: (1) comparison of loss curves between unconditional and conditional diffusion models; (2) three variants; (3) PESQ performance of different variants, (-) represent the unprocessed speech.

As for the reverse process, it aims to learn a transition kernel from \mathbf{x}_t to \mathbf{x}_{t-1} , which is defined as the following Gaussian distribution [6]:

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t), \boldsymbol{\Sigma}(\mathbf{x}_t, t)), \quad (3)$$

where θ is the learnable parameter and $\boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_{\theta}(\mathbf{x}_t, t))$ denotes the mean of \mathbf{x}_{t-1} , which is obtained by subtracting the estimated Gaussian noise $\boldsymbol{\epsilon}_{\theta}(\mathbf{x}_t, t)$ in the \mathbf{x}_t . With such a learned transition kernel, one can approximate the data distribution $p(\mathbf{x}_0)$ via:

$$p_{\theta}(\mathbf{x}_0) = \int p_{\theta}(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) d\mathbf{x}_{1:T}, \quad (4)$$

where $p_{\theta}(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$.

3.2 Problem formulation

Speech enhancement refers to methods that try to reduce distortions, make speech sounds more pleasant, and improve intelligibility. In real environments, the monaural noisy speech \mathbf{y} in the time domain can be modeled as:

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \quad (5)$$

where \mathbf{x} and \mathbf{n} denote clean and noise signals, respectively. For human perception, the primary goal of speech enhancement is to extract \mathbf{x} from \mathbf{y} . Mapping-based speech enhancement methods directly optimize $p_{\theta}(\mathbf{x}|\mathbf{y})$, while diffusion enhancement methods generate clean samples through a Markov process $p_{\theta}(\mathbf{x}_{0:T-1}|\mathbf{x}_{1:T}, \mathbf{y})$.

3.3 Condition Collapse in diffusion enhancement models

The *condition collapse* problem in speech enhancement was first proposed in [8] and it refers to the limited involvement of the condition factor during conditional diffusion training, resulting in inconsistencies between the generated speech and its condition factor.

In this work, we argue that the condition factor \mathbf{y} indeed participates and helps the intermediate-generated sample \mathbf{x}_t approximate $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$. Our assertion is supported by the experiment depicted in the left part of Figure 2 – the diffusion model equipped with the condition factor exhibits a lower loss curve compared to the unconditional one². To better understand the condition collapse phenomenon, we devise two variants that explicitly modify the mutual information between the condition factor and the model’s output (Figure 2 (middle)). We use skip connections to add the condition factor to multiple layers, forcing the likelihood of maintaining a strong connection between the condition factor and output features. Since the dependence of the output on any hidden state in the hierarchy becomes weaker as one moves further away from the output in that hierarchy (cf. [14]), using skip connections can explicitly enhance connections between the generated sample and condition factor.

²We use DiffWave [7] as basic architecture and use the same experimental settings as [4, 9] – the only difference being the change in the way of condition-injecting since most speech enhancement methods will directly use noisy speech as the condition factor, rather than Mel-spectrogram.

As shown in Figure 2 (right), an increase in mutual information (connections) leads to a significant improvement in the consistency between the generated sample and the condition factor ($a \rightarrow b$). However, it requires a meticulously designed model to guarantee its effectiveness ($b \rightarrow c$). While previous studies [5, 10, 8] focus on explicitly enhancing the consistency between the output speech and condition factor through specific network architecture design, we explore the possibility of a solution independent of the model architecture. This would broaden the applicability of our method, as it enables slight modifications to existing deterministic mapping-based models to transform them into diffusion enhancement models.

4 Methodology

Considering the diffusion model provides a transition function from \mathbf{x}_t to \mathbf{x}_{t-1} , typical condition generation process is represented as:

$$p_{\theta}(\mathbf{x}_0|\mathbf{y}) = \int \underbrace{p(\mathbf{x}_T)}_{\text{Prior}} \prod_{t=1}^T \underbrace{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})}_{\text{Condition}} d\mathbf{x}_{1:T}, \quad \mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}). \quad (6)$$

Our experiments above indicate that $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})$ will easily collapse to $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$, resulting in the condition generation process degenerating into a vanilla unconditional process:

$$\int p_{\theta}(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}) d\mathbf{x}_{1:T} \Rightarrow \int p_{\theta}(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) d\mathbf{x}_{1:T}. \quad (7)$$

As a result, facilitating automatic learning of the joint distribution for both clean and noisy speech samples does not work well for the speech enhancement task.

4.1 Condition augmentation I: Adaptive Prior

Let's revisit Eq. (6): since we cannot easily inject the condition factor into the condition term, how about the prior term? For example, we can modify the condition generation process as:

$$p_{\theta}(\mathbf{x}_0|\mathbf{y}) = \int \underbrace{p(\mathbf{x}_{\tau}|\mathbf{y})}_{\text{Conditional}} \prod_{t=1}^{\tau} \underbrace{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}_{\text{Unconditional}} d\mathbf{x}_{1:\tau}, \quad (8)$$

where $p(\mathbf{x}_{\tau}|\mathbf{y})$ is formulated as $p(\mathbf{x}_{\tau}|\mathbf{y}) = \mathcal{N}(\mathbf{x}_{\tau}; \sqrt{\bar{\alpha}_{\tau}}\mathbf{y}, (1 - \bar{\alpha}_{\tau})\mathbf{I})$. The following propositions verify the feasibility of our proposal.

Proposition 1. *For any $\xi > 0$ such that $0 < \xi < M$ for some finite positive value M , there exists a positive value $\tau \in \{0, \dots, T\}$ that satisfies:*

$$D_{KL}(p(\mathbf{x}_t|\mathbf{x})\|p(\mathbf{x}_t|\mathbf{y})) \leq \xi, \quad \forall \tau \leq t \leq T, \quad (9)$$

where $p(\mathbf{x}_t|\mathbf{c}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{c}, (1 - \bar{\alpha}_t)\mathbf{I})$.

Remark 1. *This proposition indicates that, given a tolerable margin of error ξ and a well-trained diffusion model, we can always find a suitable τ such that we are able to recover the clean speech \mathbf{x} from its noisy one \mathbf{y} using Eq. (8).*

While **Proposition 1** allows us to generate clean speech \mathbf{x} given the noisy speech \mathbf{y} using Eq. (8), it does not guarantee that our model will achieve successful recovery with a high probability.

Proposition 2. *Let \mathbf{x} be the clean sample, \mathbf{y} be it's corresponding noisy one, and \mathbf{x}' be any neighbor from the neighbor set $\mathcal{S}(\mathbf{x})$. Then diffusion enhancement models can recover \mathbf{x} with a high probability if the following inequality is satisfied:*

$$\log \left(\frac{p(\mathbf{x})}{p(\mathbf{x}')} \right) > \frac{1}{2\sigma_t^2} (\|\mathbf{x} - \mathbf{y}\|_2^2 - \|\mathbf{x}' - \mathbf{y}\|_2^2), \quad \forall \mathbf{x}' \in \mathcal{S}(\mathbf{x}), \quad (10)$$

where $\sigma_t^2 = \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}$.

Remark 2. *Assuming that the condition factor \mathbf{y} is closer to \mathbf{x} than to \mathbf{x}' , we obtain a non-positive right-hand side (RHS). For a given \mathbf{x} , the left-hand side (LHS) value is fixed, and to ensure the inequality always holds, a smaller σ_t^2 is preferred.*

As shown in Figure 3, σ_t^2 will increase as the timestep t increases. Thus, according to **Proposition 2**, we should choose a small τ for Eq. (8) to maximize the probability of successfully recovering the clean speech from the noisy one. However, constrained by **Proposition 1**, τ cannot be too small. In other words, the clean speech distribution $p(\mathbf{x}_\tau)$ and the noisy speech distribution $p(\mathbf{y}_\tau)$ will get closer over the forward diffusion process, and the gap $|\mathbf{n}| = |\mathbf{y} - \mathbf{x}|$ between the noisy speech and the clean one will indeed be “washed out” by the increasingly added noise. Since the original semantic information will also be removed if τ is too large, there should be a trade-off when we set τ for the diffusion enhancement model.

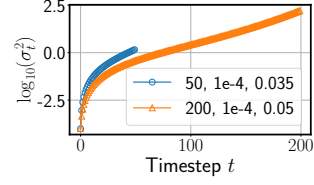


Figure 3: The change curves of $\log_{10} \sigma_t^2$. Elements in legend are T, β_1, β_T respectively.

Condition optimizer. We find that both propositions are correlated with the condition factor \mathbf{y} . If we can reduce the gap between the condition factor and clean speech, we can choose a smaller τ , effectively increasing the likelihood of recovering clean speech. One simple idea is to employ a neural network f_ψ to optimize the condition factor, as demonstrated in [15]. Accordingly, we can rewrite Eq. (8) as:

$$p_{\theta, \psi}(\mathbf{x}_0 | \mathbf{y}) = \int p_\psi(\mathbf{x}_\tau | \mathbf{y}) \prod_{t=1}^{\tau} p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) d\mathbf{x}_{1:\tau}, \quad (11)$$

where $p_\psi(\mathbf{x}_\tau | \mathbf{y}) = \mathcal{N}(\mathbf{x}_\tau; \sqrt{\bar{\alpha}_\tau} f_\psi(\mathbf{y}), (1 - \bar{\alpha}_\tau) \mathbf{I})$.

In practice, we should also consider failure cases of the condition optimizer in complex scenarios, especially the issue of excessive suppression that has been reported in recent literature [16, 17, 18]. To mitigate this issue, we use $0.5\mathbf{c} + 0.5\mathbf{y}$ (like a simple residual layer) as a mild version of the condition factor:

$$p_\psi(\mathbf{x}_\tau | \mathbf{y}) = \mathcal{N}(\mathbf{x}_\tau; 0.5\sqrt{\bar{\alpha}_\tau} (f_\psi(\mathbf{y}) + \mathbf{y}), (1 - \bar{\alpha}_\tau) \mathbf{I}). \quad (12)$$

We call $p_\psi(\mathbf{x}_\tau | \mathbf{y})$ the adaptive prior as it varies with different noisy samples \mathbf{y} .

4.2 Condition augmentation II: Diffusion Dropout

Aside from changing the prior $p(\mathbf{x}_T)$ to conditional prior $p_\psi(\mathbf{x}_\tau | \mathbf{y})$, we also optimize the condition term $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{y})$. Instead of designing specific condition-injecting strategies [4, 9, 11] or devising complicated network architecture [8, 10, 5], we attempt to “do subtraction” by discarding some shared (intermediate-generated samples & condition factor) and important (target-related) information from intermediate-generated samples. Naturally, if we discard some information from \mathbf{x}_t , then the diffusion enhancement model is forced to use the condition factor \mathbf{y} to recover the speech. Taking a further step, we can even discard the entire \mathbf{x}_t , as the condition factor \mathbf{y} alone is sufficient for recovering the clean speech \mathbf{x}_0 (this is what deterministic models do). To this end, we define a neural network $f_\theta(d(\mathbf{x}_t, p), \mathbf{y}, t)$ to approximate $p(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{y}) = \mathcal{N}(\mathbf{x}_{t-1}; f_\theta(d(\mathbf{x}_t, p), \mathbf{y}, t), \Sigma(\mathbf{x}_t, t)), \quad (13)$$

where $d(\mathbf{x}_t, p)$ is the dropout operation:

$$d(\mathbf{x}_t, p) = \begin{cases} \mathbf{x}_t & \text{if } r = 1 \\ \epsilon & \text{if } r = 0 \end{cases}, \quad r \sim \text{Bernoulli}(1 - p). \quad (14)$$

4.3 DOSE training

Ho et al. [6] and much of the following work choose to parameterize the denoising model through directly predicting ϵ with a neural network $\epsilon_\theta(\mathbf{x}_t, t)$, which implicitly sets:

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta). \quad (15)$$

In this case, the training loss is also usually defined as the mean squared error in the ϵ -space $\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2$. Although this standard specification works well for training an unconditional diffusion model, it is not suited for DOSE – for two reasons.

Algorithm 1 DOSE Training

```

1: choose  $p$ 
2: repeat
3:    $\mathbf{x}_0 \sim p(\mathbf{x})$ 
4:    $t \sim \text{Uniform}(\{1, \dots, T\})$ 
5:    $\epsilon \sim \mathcal{N}(0, I)$ 
6:   Take gradient descent step on
7:    $\nabla_{\theta} \|\mathbf{x}_0 - \mathbf{x}_{\theta}(d(\mathbf{x}_t, p), \mathbf{y}, t)\|_2^2$ 
8: until converged

```

Algorithm 2 DOSE Sampling

```

1: choose  $\tau_1, \tau_2, \tau_2 < \tau_1 \leq T$ 
2: Step 1: Generate  $\hat{\mathbf{x}}_{\tau_2}$ 
3:    $\mathbf{y}_{\tau_1} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{\tau_1}}\mathbf{y}, (1 - \bar{\alpha}_{\tau_1})\mathbf{I})$ 
4:    $\hat{\mathbf{x}}_0 = f_{\theta}(\mathbf{y}_{\tau_1}, \mathbf{y}, \tau_1)$ 
5:    $\hat{\mathbf{x}}_{\tau_2} \sim \mathcal{N}(0.5\sqrt{\bar{\alpha}_{\tau_2}}(\hat{\mathbf{x}}_0 + \mathbf{y}), (1 - \bar{\alpha}_{\tau_2})\mathbf{I})$ 
6: Step 2: Generate  $\hat{\mathbf{x}}_0$ 
7:    $\hat{\mathbf{x}}_0 = f_{\theta}(\hat{\mathbf{x}}_{\tau_2}, \mathbf{y}, \tau_2)$ 
8: return  $\hat{\mathbf{x}}_0$ 

```

First, we cannot estimate ϵ without the help of \mathbf{x}_t because ϵ and \mathbf{y} are independent. Second, as discussed earlier, we want DOSE to start with a small timestep and we strive to make τ small. However, as τ approaches zero, small changes in \mathbf{x} -space have an increasingly amplified effect on the implied prediction in ϵ -space (Eq. (15)). In other words, the efforts made by diffusion enhancement models become so negligible that diffusion models lose their ability to calibrate the speech at small timesteps.

So, we need to ensure that the estimation of $\hat{\mathbf{x}}_0$ remains flexible as the timestep t gets smaller. Considering the equivalently of the ϵ -space loss $\|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, t)\|_2^2$ to a weighted reconstruction loss in \mathbf{x} -space $\frac{1}{\sigma_t^2} \|\mathbf{x}_0 - \mathbf{x}_{\theta}(\mathbf{x}_t, t)\|_2^2$, we can directly estimate the clean speech \mathbf{x}_0 at each timestep t :

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}), t \in \{1, \dots, T\}} [\|\mathbf{x}_0 - f_{\theta}(d(\mathbf{x}_t, p), \mathbf{y}, t)\|_2^2] \quad (16)$$

4.4 DOSE inference

After training, the ideal scenario is that $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})$ approximates $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ precisely, enabling us to generate clean speech using Eq. (6). However, when applied in practice, it is difficult to completely eliminate errors (both sample error and true error). If these errors are not effectively managed or corrected during the generation process, the quality of the generated samples may deteriorate, leading to artifacts, blurriness, etc [19, 20]. This issue is particularly pronounced when using diffusion models for fine-grained conditional generation tasks, as diffusion models require a large number of steps to generate samples, which will significantly reduce the consistency between the generated sample and its condition factor (see §5.3, Figure 6).

The adaptive prior (Sec 4.1) provides an opportunity to address the error accumulation issue. Specifically, we can select a suitable τ smaller than T , conditioned on an adaptive prior, and generate speech in fewer steps. We can extend Eq. 11 by transforming the unconditional diffusion enhancement model into a conditional one:

$$p_{\theta, \psi}(\mathbf{x}_0|\mathbf{y}) = \int p_{\psi}(\mathbf{x}_{\tau}|\mathbf{y}) \prod_{t=1}^{\tau} p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}) d\mathbf{x}_{1:\tau}, \quad (17)$$

and the number of sampling steps is reduced from T to $\tau + 1$.

Readers familiar with diffusion models may recall that the standard process repeatedly applies a “single-step” denoising operation $\mathbf{x}_{t-1} = \text{denoise}(\mathbf{x}_t; t)$ that aims to convert a noisy sample at some timestep t to a (slightly less) noisy sample at the previous timestep $t - 1$. In fact, each application of the one-step denoiser consists of two steps: (1) an estimation of the fully denoised sample \mathbf{x}_0 from the current timestep t , and (2) computing a (properly weighted, according to the diffusion model) average between this estimated denoised sample and the noisy sample at the previous timestep $t - 1$. Thus, instead of performing the entire τ -step diffusion process to denoise a sample, it is also possible to run *denoise* once and simply output the estimated sample in one shot [21]. Accordingly, Eq. (17) can be further rewritten as:

$$p_{\theta, \psi}(\mathbf{x}_0|\mathbf{y}) = \int p_{\psi}(\mathbf{x}_{\tau}|\mathbf{y}) p_{\theta}(\mathbf{x}_0|\mathbf{x}_{\tau}, \mathbf{y}) d\mathbf{x}_{\tau} \quad (18)$$

We can even achieve DOSE without the condition optimizer $f_{\psi}(\cdot)$ – using conditional diffusion enhancement model instead. For example, we can generate clean speech via:

$$p_{\theta}(\mathbf{x}_0|\mathbf{y}) = \int \int p_{\theta}(\hat{\mathbf{x}}_{\tau_2}|\mathbf{y}_{\tau_1}, \mathbf{y}) p_{\theta}(\mathbf{x}_0|\hat{\mathbf{x}}_{\tau_2}, \mathbf{y}) d\hat{\mathbf{x}}_{\tau_2} d\mathbf{y}_{\tau_1}, \quad (19)$$

where τ_1, τ_2 ($\tau_2 < \tau_1 \leq T$) are two pre-defined hyper-parameters. The motivation behind Eq. (19) is that, once we have trained a neural network $f_\theta(x_t, y, t)$ that can accurately estimate x_0 (Eq. (16)), according to the theoretical analysis in Sec 4.1, we can first choose a suitable value for τ_1 to ensure a relatively good approximation of x_0 :

$$\hat{x}_0 = f_\theta(y_{\tau_1}, y, \tau_1) \approx f_\theta(x_{\tau_1}, y, \tau_1) \quad (20)$$

In the second step, once we have obtained a good condition factor, we can choose a smaller timestep $\tau_2 < \tau_1$ to get a better estimation of x_0 than \hat{x}_0 generated in the first step.

Summary. DOSE has three important benefits: (1) By dropping x_t entirely, we make the condition factor y the “protagonist”, automatically enhancing the consistency between the generated sample and the condition factor. (2) By training the model with this modified training objective, DOSE can perform well not only on Gaussian noise ($x_t \rightarrow x_0$) but also on various types of non-Gaussian noise ($y \rightarrow x_0$). (3) DOSE is efficient (2 steps), faster than existing diffusion enhancement models.

5 Experiments

We compare DOSE with prevailing diffusion enhancement methods and deterministic mapping-based enhancement methods in §5.1. We conduct a counterfactual verification to understand the intrinsic mechanism of DOSE in §5.2. We show two visual cases of excessive suppression and error accumulation in §5.3. While providing a self-contained version of our main results, we note that we also have additional quantitative observations reported in the Appendices. Specifically, we compare DOSE with other baselines via subjective evaluation (Appendix A.2); We investigate the significance of the proposed adaptive prior and explain why we need to use a mild version of the condition factor (Appendix A.3); We examine the effect of our new training objective and demonstrate the necessity of using it (Appendix A.4); We explain why we use two steps in speech generation (Appendix A.5); We provide parameter sensitivity experiments (Appendix A.6); We show plenty of visual cases of excessive suppression and error accumulation (Appendix A.7 and A.8). To help readers better understand our research, we include a discussion subsection in Appendix A.9. Specifically, we: (1) Analyze the reasons behind the superior generalizability of diffusion enhancement models compared to deterministic mapping-based models (from the robust training perspective); (2) Explain why we use 0.5 in the mild version of the condition factor; (3) Discuss the broader impacts of speech enhancement methods.

Dataset and baselines. Following previous works [4, 9, 8], we use the VoiceBank-DEMAND dataset [22, 23] for performance evaluations. To investigate the generalization ability of models, we use CHiME-4 [24] as another test dataset following [9], i.e., the models are trained on VoiceBank-DEMAND and evaluated on CHiME-4. We compare our model with recent open-sourced diffusion enhancement models such as DiffuSE [4], CDiffuSE [9], SGMSE [11], SGMSE+ [5], and DR-DiffuSE [8]. Since the only difference between SGMSE+ and SGMSE is their network architecture, we compare our model with just one of them.

Evaluation metrics. We use the following metrics to evaluate SE performance: the perceptual evaluation of speech quality (PESQ) [25], short-time objective intelligibility (STOI) [26], segmental signal-to-noise ratio (SSNR), the mean opinion score (MOS) prediction of the speech signal distortion (CSIG) [27], the MOS prediction of the intrusiveness of background noise (CBAK) [27] and the MOS prediction of the overall effect (COVL) [27]. Besides these metrics, we also design two MOS metrics (MOS and Similarity MOS) for subjective evaluation.

Configurations. To ensure a fair comparison, we keep the model architecture exactly the same as that of the DiffWave model [7] for all methods³. DiffWave takes 50 steps with the linearly spaced training noise schedule $\beta_t \in [1 \times 10^{-4}, 0.035]$ [4]. We train all methods for 300,000 iterations using 1 NVIDIA RTX 3090 GPU with a batch size of 16 audios. We select the best values for τ_1 and τ_2 according to the performance on a validation dataset, a small subset (10%) extracted from the training data. More experiment settings can be found in Appendix A.10.

³Since the focus of our work is on studying the capabilities of diffusion dropout and adaptive prior for consistency enhancement, we use off-the-shelf architectures to avoid confounding our findings with model improvements. This decision (using DiffWave) rests on both its widely validated effectiveness and the minimal changes it required in the baseline experimental setup.

Table 1: Comparison of different diffusion enhancement methods.

Method	Year	Efficiency	Dataset	STOI(%) \uparrow	PESQ \uparrow	CSIG \uparrow	CBAK \uparrow	COVL \uparrow
Unprocessed	—	—		92.1	1.97	3.35	2.44	2.63
DiffWave	2021	1 step (dis)		93.3	2.51	3.72	3.27	3.11
DiffuSE	2021	6 steps	VB	93.5 ± 0.20 ± 0.05	2.39 ± 0.12 ± 0.01	3.71 ± 0.01 ± 0.01	3.04 ± 0.23 ± 0.01	3.03 ± 0.08 ± 0.01
CDiffuSE	2022	6 steps		93.7 ± 0.40 ± 0.05	2.43 ± 0.08 ± 0.01	3.77 ± 0.05 ± 0.01	3.09 ± 0.18 ± 0.01	3.09 ± 0.02 ± 0.01
SGMSE	2022	50 steps		93.3 ± 0.00 ± 0.08	2.34 ± 0.17 ± 0.01	3.69 ± 0.03 ± 0.01	2.90 ± 0.37 ± 0.01	3.00 ± 0.11 ± 0.01
DR-DiffuSE	2023	6 steps		92.9 ± 0.04 ± 0.06	2.50 ± 0.01 ± 0.02	3.68 ± 0.04 ± 0.02	3.27 ± 0.00 ± 0.02	3.08 ± 0.03 ± 0.02
DOSE	—	2 steps		93.6 ± 0.30 ± 0.05	2.56 ± 0.05 ± 0.01	3.83 ± 0.11 ± 0.01	3.27 ± 0.00 ± 0.01	3.19 ± 0.08 ± 0.01
Unprocessed	—	—		71.5	1.21	2.18	1.97	1.62
DiffWave	2021	1 step (dis)		72.3	1.22	2.21	1.95	1.63
DiffuSE	2021	6 steps	CHIME-4	83.7 ± 11.4 ± 0.05	1.59 ± 0.36 ± 0.01	2.91 ± 0.70 ± 0.01	2.19 ± 0.24 ± 0.01	2.19 ± 0.56 ± 0.01
CDiffuSE	2022	6 steps		82.8 ± 10.5 ± 0.05	1.58 ± 0.36 ± 0.01	2.88 ± 0.67 ± 0.01	2.15 ± 0.20 ± 0.01	2.18 ± 0.55 ± 0.01
SGMSE	2022	50 steps		84.5 ± 12.2 ± 0.05	1.57 ± 0.34 ± 0.02	2.92 ± 0.71 ± 0.01	2.18 ± 0.23 ± 0.02	2.18 ± 0.55 ± 0.01
DR-DiffuSE	2023	6 steps		77.6 ± 5.30 ± 0.06	1.29 ± 0.07 ± 0.04	2.40 ± 0.19 ± 0.02	2.04 ± 0.09 ± 0.01	1.78 ± 0.15 ± 0.01
DOSE	—	2 steps		86.6 ± 14.3 ± 0.05	1.52 ± 0.30 ± 0.01	2.71 ± 0.50 ± 0.01	2.15 ± 0.20 ± 0.01	2.06 ± 0.43 ± 0.01

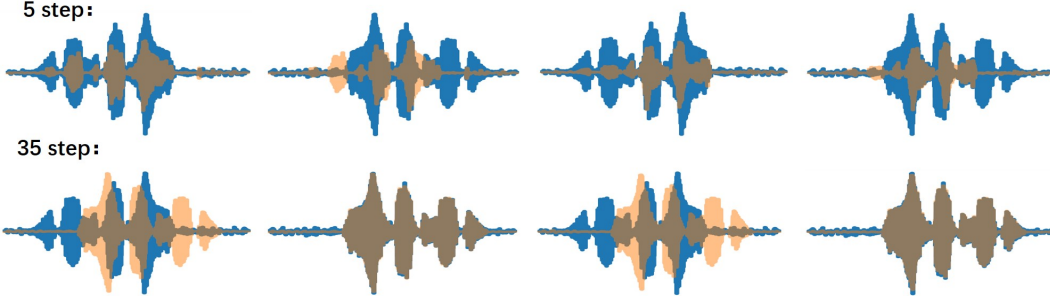


Figure 4: Counterfactual visualization. The first two columns are associated with a dropout probability of 0.1, while the last two columns are associated with a dropout probability of 0.9. In each row, the blue waveforms in the first and third columns are the counterfactual samples, and the blue waveforms in the second and fourth columns are the normal samples. The orange waveforms are generated samples from the model.

5.1 Performance comparison

We compare our method with previous diffusion enhancement methods and summarize our experimental results in Table 1. We observe that: (1) Diffusion enhancement methods have better generalizability than deterministic methods. (2) Methods with specific condition-injecting strategies, such as DiffuSE, CDiffuSE, and SGMSE, have strong generalization but perform slightly worse than deterministic mapping-based methods in matched scenarios. (3) Method (DR-DiffuSE) with auxiliary condition optimizer, performs better in matched scenarios and shows a slight improvement in mismatched scenarios. (4) Our method performs well in both matched and mismatched scenarios and is on par with state-of-the-art diffusion enhancement models while requiring fewer steps.

5.2 Counterfactual verification

We perform a counterfactual verification to gain insights into the underlying mechanism of DOSE. To verify whether dropout can increase the “discourse power” of the conditional factor, we keep the condition factor y fixed and reverse the intermediate-generated speech at a specific step ($reverse(x_t)$). This reversed intermediate-generated speech is called a counterfactual sample. Notably, if the final generated speech is more similar to the condition factor than the counterfactual speech, we can conclude that the condition factor plays a dominant role in the generation process. Otherwise, we can say that the condition factor is less influential.

As shown in Figure 4, we compare the performance of two models with different dropout probabilities (0.1 vs. 0.9). We have two findings here: (1) A higher dropout probability encourages the model to prioritize the condition factor even with a small timestep t . (2) When timestep t is large, DOSE ef-

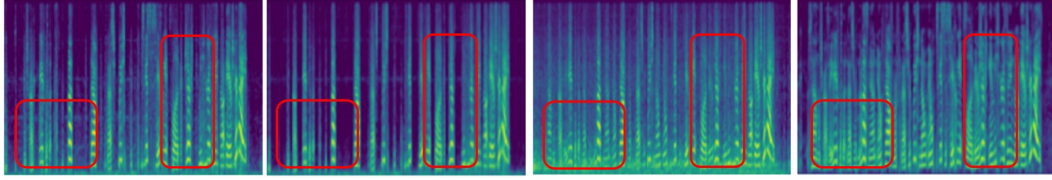


Figure 5: Excessive suppression visualization (unconditional diffusion enhancement model on CHIME-4). From left to right: (1) DiffWave (dis); (2) adaptive prior with the estimated condition; (3) adaptive prior with the mild condition; (4) clean speech.

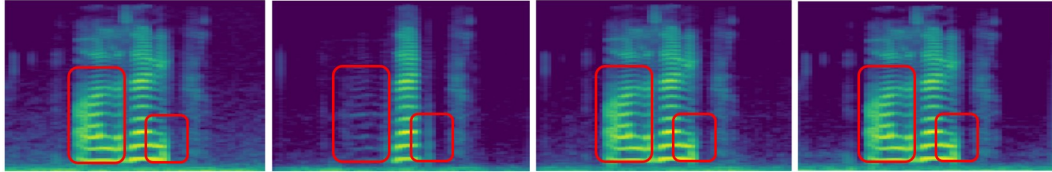


Figure 6: Error accumulation visualization (VB, DOSE). From left to right: (1) noisy speech; (2) full (50) steps; (3) 2 steps; (4) clean.

effectively captures condition information, ensuring the model’s robustness to noise and maintaining consistency in the early stages of inference.

5.3 Excessive suppression & error accumulation

We provide a visual case of excessive suppression in Figure 5 and a visual case of error accumulation in Figure 6. From Figure 5, we can see that: (1) The deterministic model fails in mismatched scenarios and generates samples that lose speech details; (2) The diffusion enhancement model generate defective speech when directly using the estimated speech as the condition factor; (3) The diffusion enhancement model equipped with a mild version of the condition factor can recover clean speech effectively. From Figure 6, we notice that: (1) Full-step generation can remove noise and generate natural-sounding speech. However, it can’t guarantee the consistency between the generated speech and condition factor; (2) Two-step speech generation with adaptive prior can promise consistency and high quality simultaneously.

6 Conclusions

In this work, we present a new approach DOSE that effectively incorporates condition information into diffusion models for speech enhancement. DOSE uses two efficient condition-augmentation techniques to address the condition collapse problem. Comprehensive experiments on benchmark datasets demonstrate the efficiency and effectiveness of our method.

In our method, there are two groups of hyper-parameters: the dropout probability p for the dropout operation and two timesteps τ_1, τ_2 for the adaptive prior. These parameters are critical to model performance. For example, if the dropout probability is set too high, the diffusion enhancement model will rely solely on the condition factor to estimate the speech. Then our diffusion enhancement model will degenerate into a deterministic model, losing its generalizability. We also need to make a trade-off when choosing the timestep τ (especially τ_1): On one hand, a large τ is needed to reduce the gap between the clean speech and condition factor. On the other hand, the original semantic information will also be removed if τ is set too large.

In practice, it is necessary to evaluate the model on a subset of data and then empirically set the hyperparameters. These manually defined hyper-parameters are selected based on the Empirical Risk Minimization (ERM) principle and may not be optimal for every individual sample. Thus, an important direction for future research is to develop methods that can adaptively choose hyper-parameters for different samples. It is also expected that the model can adaptively select appropriate coefficients when forming a mild version of the conditioning factor.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (Grant No.62176043 and No.62072077), Natural Science Foundation of Sichuan Province (Grant No.2022NSFSC0505), Kashgar Science and Technology Bureau (Grant No.KS2023025), and National Science Foundation SWIFT (Grant No.2030249).

References

- [1] Steven L Gay and Jacob Benesty. *Acoustic signal processing for telecommunication*, volume 551. Springer Science & Business Media, 2012.
- [2] Tim Van den Bogaert, Simon Doclo, Jan Wouters, and Marc Moonen. Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids. *The Journal of the Acoustical Society of America*, 125(1):360–371, 2009.
- [3] Ivan Tashev. Recent advances in human-machine interfaces for gaming and entertainment. *International journal of information technologies and security*, 3(3):69–76, 2011.
- [4] Yen-Ju Lu, Yu Tsao, and Shinji Watanabe. A study on speech enhancement based on diffusion probabilistic model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 659–666. IEEE, 2021.
- [5] Julius Richter, Simon Welker, Jean-Marie Lemerrier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *arXiv preprint arXiv:2208.05830*, 2022.
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [7] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.
- [8] Wenxin Tai, Fan Zhou, Goce Trajcevski, and Ting Zhong. Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement. In *AAAI*, 2023.
- [9] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7402–7406. IEEE, 2022.
- [10] Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*, 2022.
- [11] Simon Welker, Julius Richter, and Timo Gerkmann. Speech enhancement with score-based generative models in the complex STFT domain. In *Proc. Interspeech 2022*, pages 2928–2932, 2022. doi: 10.21437/Interspeech.2022-10653.
- [12] Jianwei Zhang, Suren Jayasuriya, and Visar Berisha. Restoring degraded speech via a modified diffusion model. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 2753–2757. International Speech Communication Association, 2021.
- [13] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [14] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- [15] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022.

- [16] Andong Li, Chengshi Zheng, Cunhang Fan, Renhua Peng, and Xiaodong Li. A recursive network with dynamic attention for monaural speech enhancement. In *Proc. Interspeech 2020*, pages 2422–2426, 2020. doi: 10.21437/Interspeech.2020-1513. URL <http://dx.doi.org/10.21437/Interspeech.2020-1513>.
- [17] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [18] Santiago López-Tapia and Nicolás Pérez de la Blanca. Fast and robust cascade model for multiple degradation single image super-resolution. *IEEE Transactions on Image Processing*, 30:4747–4759, 2021.
- [19] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.
- [20] Jean-Marie Lemerrier, Julius Richter, Simon Welker, and Timo Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *arXiv preprint arXiv:2212.11851*, 2022.
- [21] Nicholas Carlini, Florian Tramèr, J Zico Kolter, et al. (certified!!) adversarial robustness for free! In *International Conference on Learning Representations*, 2023.
- [22] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *International Conference Oriental COCOSA*, pages 1–4. IEEE, 2013.
- [23] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19, page 035081. Acoustical Society of America, 2013.
- [24] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech Language*, 46:535–557, 2017.
- [25] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pages 749–752. IEEE, 2001.
- [26] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pages 4214–4217. IEEE, 2010.
- [27] Yi Hu and Philippos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.
- [28] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022.

A Appendix

In the supplemental material,

- **A.1:** We provide the proofs for Propositions 1 and 2.
- **A.2:** We compare DOSE with other baselines via subjective evaluation.
- **A.3:** We investigate the significance of the proposed adaptive prior and explain why we need to use a mild version of the condition factor.
- **A.4:** We examine the effect of our new training objective and demonstrate the necessity of using it.
- **A.5:** We explain why we use two steps in speech generation.
- **A.6:** We provide parameter sensitivity experiments.
- **A.7:** We present several visual cases of excessive suppression.
- **A.8:** We present several visual cases of error accumulation.
- **A.9:** We analyze the reasons behind the superior generalizability of diffusion enhancement models compared to deterministic mapping-based models (from the robust training perspective), explain why we use 0.5 in the mild version of the condition factor, and discuss the broader impacts of speech enhancement methods.
- **A.10:** We include more information about speech processing and basic architecture.

A.1 Mathematical proofs

We now present the proofs of the Propositions stated in the main text.

Proposition 1. (cf. §4.1): *For any $\xi > 0$ such that $0 < \xi < M$ for some finite positive value M , there exists a positive value $\tau \in \{0, \dots, T\}$ that satisfies:*

$$D_{KL}(p(\mathbf{x}_t|\mathbf{x})||p(\mathbf{x}_t|\mathbf{y})) \leq \xi, \quad \forall \tau \leq t \leq T, \quad (9)$$

where $p(\mathbf{x}_t|\mathbf{c}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{c}, (1 - \bar{\alpha}_t)\mathbf{I})$.

Proof. Given two Gaussian distributions P, Q defined over a vector space \mathbb{R}^d , the KL divergence of multivariate Gaussian distributions is defined as follows:

$$D_{KL}(P||Q) = \frac{1}{2} \left(\text{Tr}(\Sigma_2^{-1}\Sigma_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - d + \ln \left(\frac{\det \Sigma_2}{\det \Sigma_1} \right) \right). \quad (21)$$

Here, $\boldsymbol{\mu}_1 \in \mathbb{R}^d$ and $\Sigma_1 \in \mathbb{R}^{d \times d}$ are the mean and covariance matrix of distribution P , and $\boldsymbol{\mu}_2 \in \mathbb{R}^d$ and $\Sigma_2 \in \mathbb{R}^{d \times d}$ are the mean and covariance matrix of distribution Q . d is the dimensionality of the vectors (i.e., the number of dimensions in the vector space), and Tr denotes the trace operator.

Note that when the two Gaussian distributions have diagonal covariance matrices (i.e., when the different dimensions are independent), the above formula simplifies to the sum of the KL divergences of each univariate Gaussian distribution. Thus, given two Gaussian distributions $p(\mathbf{x}_t|\mathbf{x})$, $p(\mathbf{x}_t|\mathbf{y})$ and Eq. (5), the KL divergence between these two distributions can be calculated as follows:

$$D_{KL}(p(\mathbf{x}_t|\mathbf{x})||p(\mathbf{x}_t|\mathbf{y})) = \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t} \|\mathbf{y} - \mathbf{x}\|_2^2 = \frac{1}{\sigma_t^2} \|\mathbf{n}\|_2^2. \quad (22)$$

According to the definition of the diffusion model and Figure 3, $D_{KL}(p(\mathbf{x}_t|\mathbf{x})||p(\mathbf{x}_t|\mathbf{y}))$ is a monotonically decreasing function. Ideally, for a bounded error \mathbf{n} with (almost) infinite timestep T , we have:

$$\lim_{t \rightarrow 0} D_{KL}(p(\mathbf{x}_t|\mathbf{x})||p(\mathbf{x}_t|\mathbf{y})) = +\infty, \quad \lim_{t \rightarrow T} D_{KL}(p(\mathbf{x}_t|\mathbf{x})||p(\mathbf{x}_t|\mathbf{y})) = 0. \quad (23)$$

According to Bolzano’s theorem, there exists at least one point τ in the interval $\{0, \dots, T\}$ such that $D_{KL}(p(\mathbf{x}_\tau|\mathbf{x})||p(\mathbf{x}_\tau|\mathbf{y})) = \xi$. Then, Eq. (9) holds for $\tau \leq t \leq T$. \square

Testing audio


0:00 / 0:05

▶ 🔊 ⋮

Insrtuction:

How nature(i.e. human-sounding) is this recording ?

Please focus on examining the audio **quality and naturalness**, and ignore the differences of style (**timbre,emotion and prosody**)

What's your rating  of voice

Select an rating for your favorite audio

☒ Excellent-Completely natural speech -5

☐ 4.5

☐ Good-Mostly natural speech -4

☐ 3.5

☐ Fair-Equally natural and unnatural speech -3

☐ 2.5

☐ Poor-Mostly unnatural speech -2

☐ 1.5

☐ Bad-Completely unnatural speech -1

You selected Excellent-Completely natural speech -5

Figure 7: Screenshot of MOS test.

Proposition 2. (cf. §4.1): *Let \mathbf{x} be the clean sample, \mathbf{y} be it's corresponding noisy one, and \mathbf{x}' be any neighbor from the neighbor set $\mathcal{S}(\mathbf{x})$. Then diffusion enhancement models can recover \mathbf{x} with a high probability if the following inequality is satisfied:*

$$\log \left(\frac{p(\mathbf{x})}{p(\mathbf{x}')} \right) > \frac{1}{2\sigma_t^2} (\|\mathbf{x} - \mathbf{y}\|_2^2 - \|\mathbf{x}' - \mathbf{y}\|_2^2), \quad \forall \mathbf{x}' \in \mathcal{S}(\mathbf{x}), \quad (10)$$

where $\sigma_t^2 = \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}$ is the variance of the Gaussian noise added at timestep t in the forward diffusion process.

Proof. The main idea is to prove that any point \mathbf{x}' quite similar but different to the ground-true speech \mathbf{x} should have a lower density than \mathbf{x} in the conditional distribution so that the diffusion enhancement models can recover \mathbf{x} with a high probability. In other words, we should have:

$$p(\mathbf{x}_0 = \mathbf{x} | \mathbf{x}_t = \mathbf{y}_t) > p(\mathbf{x}_0 = \mathbf{x}' | \mathbf{x}_t = \mathbf{y}_t) \quad (24)$$

According to Bayes' theorem, we have:

$$\begin{aligned} p(\mathbf{x}_0 = \mathbf{x} | \mathbf{x}_t = \mathbf{y}_t) &= \frac{p(\mathbf{x}_0 = \mathbf{x}, \mathbf{x}_t = \mathbf{y}_t)}{p(\mathbf{x}_t = \mathbf{y}_t)} \\ &= p(\mathbf{x}_0 = \mathbf{x}) \cdot \frac{p(\mathbf{x}_t = \mathbf{y}_t | \mathbf{x}_0 = \mathbf{x})}{p(\mathbf{x}_t = \mathbf{y}_t)}. \end{aligned} \quad (25)$$

Applying Eq. (25) to Eq. (24), we obtain:

$$\begin{aligned} p(\mathbf{x}) \cdot \frac{1}{\sqrt{(2\pi\sigma_t^2)^d}} \exp \frac{-\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma_t^2} &> p(\mathbf{x}') \cdot \frac{1}{\sqrt{(2\pi\sigma_t^2)^d}} \exp \frac{-\|\mathbf{x}' - \mathbf{y}\|_2^2}{2\sigma_t^2} \\ \Leftrightarrow \log \left(\frac{p(\mathbf{x})}{p(\mathbf{x}')} \right) &> \frac{1}{2\sigma_t^2} (\|\mathbf{x} - \mathbf{y}\|_2^2 - \|\mathbf{x}' - \mathbf{y}\|_2^2), \quad \forall \mathbf{x}' \in \mathcal{S}(\mathbf{x}), \end{aligned} \quad (26)$$

and the proof is now complete. \square

A.2 Subjective evaluation

We conduct two types of Mean Opinion Score (MOS) tests to verify the quality of synthesized audio through human evaluation.

Reference audio

▶ 0:00 / 0:05

Testing audio

▶ 0:00 / 0:05

Insruction:
How similar is this recording to the reference audio?

Please focus on the similarity of the style (**specker identity,emotion and prosody**) to the reference, and ignore the **audio quality**

What's your favorite audio:
choose one audio
☒ Reference Audio
☐ Testing Audio
You selected Reference Audio

What's your rating of voice
Select an rating for your favorite audio
☒ Excellent-Completely natural speech -5
☐ 4.5
☐ Good-Mostly natural speech -4
☐ 3.5
☐ Fair-Equally natural and unnatural speech -3
☐ 2.5
☐ Poor-Mostly unnatural speech -2
☐ 1.5
☐ Bad-Completely unnatural speech -1
You selected Excellent-Completely natural speech -5

Figure 8: Screenshot of Similarity MOS test.

Table 2: MOS tests under different scenarios.

Method	Scenarios	MOS↑	Similarity MOS↑	Scenarios	MOS↑	Similarity MOS↑
Unprocessed		3.60 \pm 0.31	3.33 \pm 0.34		3.30 \pm 0.30	3.10 \pm 0.32
DiffWave (dis)		3.80 \pm 0.21	3.75 \pm 0.42		2.10 \pm 1.20	1.00 \pm 2.10
DiffuSE		3.40 \pm 0.20	4.17 \pm 0.26		3.00 \pm 0.27	3.33 \pm 0.21
CDiffuSE	Matched	3.85 \pm 0.25	4.12 \pm 0.31	Mismatched	2.55 \pm 0.25	3.42 \pm 0.32
SGMSE		3.65 \pm 0.05	3.95 \pm 0.62		2.83 \pm 0.47	3.41 \pm 0.31
DR-DiffuSE		3.80 \pm 0.20	3.84 \pm 0.51		2.48 \pm 0.33	1.45 \pm 1.65
DOSE		4.05 \pm 0.45	4.35 \pm 1.02		3.48 \pm 0.18	3.17 \pm 0.07
		\pm 0.29	\pm 0.21		\pm 0.26	\pm 0.23

Naturalness. For audio quality evaluation, we conduct the MOS (mean opinion score) tests and explicitly instruct the raters to “focus on examining the audio quality and naturalness, and ignore the differences of style (timbre, emotion, and prosody)”. The testers present and rate the samples, and each tester is asked to evaluate the subjective naturalness on a 1-5 Likert scale.

Consistency. For audio consistency evaluation, we explicitly instruct the raters to “focus on the similarity of the speech (content, timbre, emotion, and prosody) to the reference, and ignore the differences of audio quality”. This is slightly different from the original definition of SMOS for speech synthesis. In the SMOS (similarity mean opinion score) tests, we pair each synthesized utterance with a ground truth utterance to evaluate how well the synthesized speech matches that of the target speaker. The testers present and rate the samples, and each tester is asked to evaluate the subjective consistency on a 1-5 Likert scale.

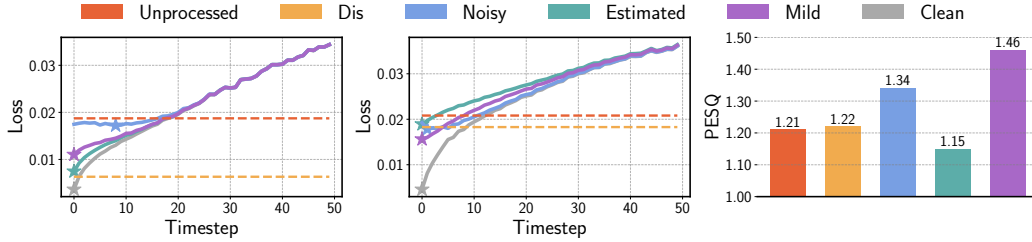


Figure 9: Performance of the unconditional diffusion enhancement model with adaptive prior. From left to right: (1) each step loss on matched VB; (2) each step loss on mismatched CHIME-4; (3) PESQ comparison for different priors on mismatched CHIME-4.

Our subjective evaluation tests are crowd-sourced and conducted by 15 volunteers. The screenshots of instructions for testers have been shown in Figure 7 and Figure 8. We paid \$10 to participants hourly and totally spent about \$300 on participant compensation.

The MOS results with the 95% confidence interval are shown in Table 2. we observe that: (1) Our method surpasses all baselines, demonstrating the strong ability of the proposed framework in synthesizing natural speech; (2) Our model can synthesize consistent speech to the golden speech, which is aligned with our motivation for algorithm design. It’s also exciting to see that DOSE yields similar scores to methods with specific condition-injecting strategies (i.e., DiffuSE, CDiffuSE, and SGMSE) on Similarity MOS.

A.3 Adaptive prior analysis

We now investigate the significance of the proposed adaptive prior (§4.1) and show why we need to use a mild version (Eq. (12)) of the condition factor.

We design three variants to investigate the effect of different condition optimizer settings on denoising performance. These variants are: (a) applying adaptive prior with the noisy speech; (b) applying adaptive prior with the estimated one (from the deterministic model); (c) applying adaptive prior with the mild condition (Eq. (12)). To control variables, we use an unsupervised diffusion model with adaptive priors. We conduct experiments on the matched VB dataset and mismatched CHIME-4 dataset respectively, and our results are shown in Figure 9.

As shown in Figure 9 (left), we plot the one-step loss on matched VB and obtain the following observations: (1) The trend of loss curves is in line with our analysis in §4.1 that we need to find a trade-off timestep for better performance. (2) Equipping the unconditional diffusion enhancement model with the adaptive prior technique has a certain denoising ability but is inferior to its counterpart discriminative model in matched scenarios. We attribute the second phenomenon to the limited denoising capacity of the unconditional diffusion enhancement models (cf. [9, 21]).

Although the performance of the unconditional diffusion enhancement model equipped with adaptive prior is mediocre in matched scenarios, as illustrated in Figure 9 (mid), it exhibits greater stability than discriminative models in mismatched scenarios. To verify the influence of different priors, we compare their PESQ on mismatched CHIME-4, shown in Figure 9 (right). We see that: (1) The deterministic model fails in mismatched scenarios and generates samples that are even worse than the unprocessed ones; (2) The diffusion enhancement model has strong generalizability and performs significantly better than the deterministic model; (3) The diffusion enhancement model loses its capability when using the estimated speech as the condition factor; and (4) Although the estimated speech is worse than the unprocessed one, the diffusion enhancement model equipped with a mild version of the condition factor achieves the best performance. This implies that the estimated speech can provide additional complementary guidance to the diffusion model, and the model can adaptively “separate the wheat from the chaff”. Thus, using a mild version of the condition factor is important and necessary.

In summary, our research shows the strong generalizability of the diffusion enhancement model. However, we also find that the unconditional diffusion enhancement model has mediocre performance. This suggests that relying solely on the adaptive prior technique is not sufficient, further emphasiz-

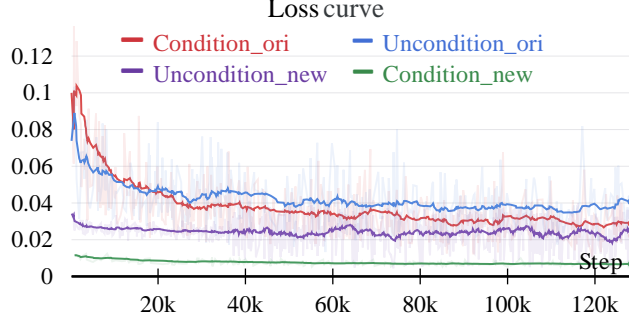


Figure 11: Investigation of the training objective. we compare the ϵ -space and x -space loss curves.

ing the importance of the diffusion dropout operation – training conditional diffusion enhancement models in a supervised manner.

A.4 Training objective investigation

We now examine the effect of our new training objective and demonstrate the necessity of using it.

Let’s recall the sampling process of DOSE. In the first step, we generate a relatively good estimation of the clean speech. In the second step, we use DOSE with a small timestep to generate a better one. We plot the relationship between $\Delta\epsilon$ and Δx_0 in Figure 10. Specifically, we fix the Δx_0 as a constant and use Eq. (15) to calculate the corresponding $\Delta\epsilon$. This experiment aims to show how effort for calibration in x -space is equivalent to that in ϵ -space. From Figure 10 we see that, as t approaches zero, small changes in x -space amplify the implied prediction in ϵ -space. There is a nearly 100-fold difference between the values of $\Delta\epsilon$ and Δx_0 . This implies that the efforts of the diffusion enhancement model at small timesteps become negligible, causing diffusion models to lose their ability to recover natural-sounding speech from the defective speech estimated in the small timestep.

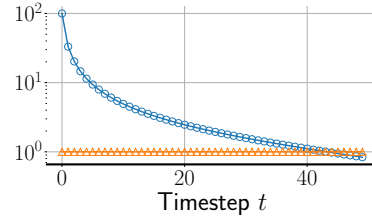


Figure 10: The relationship between $\Delta x_0 = |\hat{x}_0 - x_0|$ (orange) and $\Delta\epsilon = |\hat{\epsilon} - \epsilon|$ (blue).

We also plot the training loss curves in Figure 11. After substituting the training objective from ϵ -space to x -space, we observe that the loss change becomes more significant. This demonstrates more effective participation of conditioning factors in diffusion training: training diffusion enhancement model with this new objective allows for easier and more effective exploitation of conditioning factors.

A.5 Complexity analysis & why use two-steps generation?

In this subsection, we explain why we use two steps to generate speech.

Why not one-step speech generation? According to §4.4, we can generate speech in one shot using a trained conditional diffusion enhancement model to reduce error accumulation [21]. This one-step speech generation provides an appealing performance (Table 3). However, if we consider the diffusion model training as a multi-task paradigm, the denoising task at a smaller timestep t is easier than that at a larger timestep [21]. Correspondingly, the primary estimation error occurs at the large timestep area and directly estimating clean speech at a large timestep will result in sub-optimal performance [28]. Meanwhile, we can’t choose a small timestep t as it will lead to the mismatched problem discussed in §4.1.

Since the conditional diffusion model can learn vital information from both the intermediate-generated and noisy speech – the estimated speech can provide complementary guidance to the diffusion model (Appendix A.3) – allowing us to further improve the result by generating speech with multiple steps.

Table 3: Ablation study for two-step speech generation.

Method	Scenarios	PESQ↑	STOI(%)↑	Scenarios	PESQ↑	STOI(%)↑
Unprocessed		1.97	92.1		1.21	71.5
DOSE (fixed 1 step)	Matched	2.47 ^{+0.50} _{±0.01}	93.0 ^{+0.90} _{±0.05}	Mismatched	1.38 ^{+0.17} _{±0.01}	82.8 ^{+11.3} _{±0.05}
DOSE (handpicked 1 step)		2.50 ^{+0.53} _{±0.01}	93.4 ^{+1.30} _{±0.05}		1.51 ^{+0.31} _{±0.01}	86.4 ^{+15.7} _{±0.05}
DOSE (fixed 2 steps)		2.48 ^{+0.51} _{±0.01}	93.1 ^{+1.00} _{±0.05}		1.44 ^{+0.23} _{±0.01}	83.6 ^{+12.1} _{±0.05}
DOSE (handpicked 2 steps)		2.56 ^{+0.59} _{±0.01}	93.6 ^{+1.50} _{±0.05}		1.52 ^{+0.32} _{±0.01}	86.6 ^{+15.1} _{±0.05}

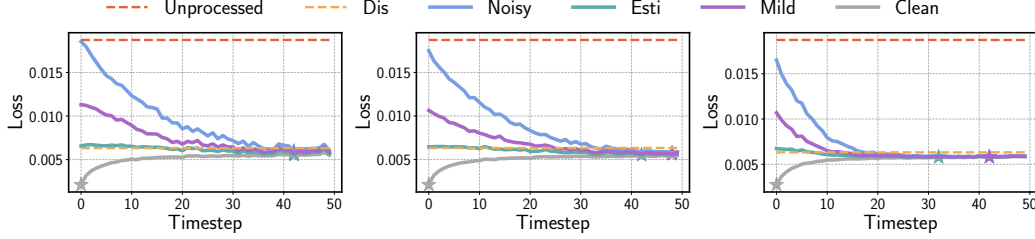


Figure 12: Performance of the conditional diffusion enhancement model with different dropout probability in VB. From left to right: (1) without dropout; (2) dropout 50% ; (3) dropout 90%.

Hyper-parameter selection. Basically, we need to set optimal hyper-parameters by evaluating the performance with a small batch of data. Suppose we choose the number of sampling steps as K ($K < T$), and the amount of test data as N , then the computational complexity of the grid search is $\mathcal{O}(NT!/(T-K)!)$. Since T is always set as a large number in the diffusion model’s setup, the complexity introduced by choosing a large K is often unmanageable.

Trivial solution. We have a simple alternative solution: defining the hyper-parameters empirically (e.g., equal intervals [7]). We present speech quality comparisons between empirically defined (fixed) and handpicked hyper-parameters in Table 3. As shown, the conditional diffusion enhancement model with handpicked hyper-parameters performs better than that with empirically defined hyper-parameters.

Although fixed hyper-parameters have inferior performance compared to handpicked ones, they still show appealing performance compared to prevailing diffusion enhancement baselines. Therefore, in situations where we can’t evaluate the model in advance, we can use empirically defined hyper-parameters instead.

A.6 Parameter sensitivity

There are two groups of hyper-parameters that are critical to model performance: the dropout probability p for model training and two timesteps τ_1, τ_2 for model inference. In this subsection, we conduct parameter sensitivity experiments to investigate how these hyper-parameters affect the model’s performance.

Dropout probability p . We vary the dropout probability p , setting it to $\{0.0, 0.5, 0.9, 1.0\}$, and plot the one-step loss on matched VB (Figure 12) and mismatched CHIME-4 (Figure 13) respectively. Please note that the figure with $p = 1$ has been excluded since it degenerates to a deterministic mapping-based model and the performance remains unchanged across all timesteps.

From Figure 12, we see that: (1) The proposed conditional diffusion enhancement model works. Compared to the deterministic mapping-based model, our model performs slightly better in matched scenarios and significantly better in mismatched scenarios; (2) When intermediate-generated speech \mathbf{x}_t is unreliable (large t), the condition factor \mathbf{y} plays a dominant role. (3) As the dropout probability increases, the model focuses more on the condition factor, while when the dropout probability is small, the loss curve oscillates when t gets large.

From Figure 13, We find that: when using a higher dropout probability (such as $p = 0.5$ and $p = 0.9$), the model can generally achieve better generalizability. Note that if the dropout probability is set too high, the diffusion enhancement model will rely solely on the condition factor to estimate the clean

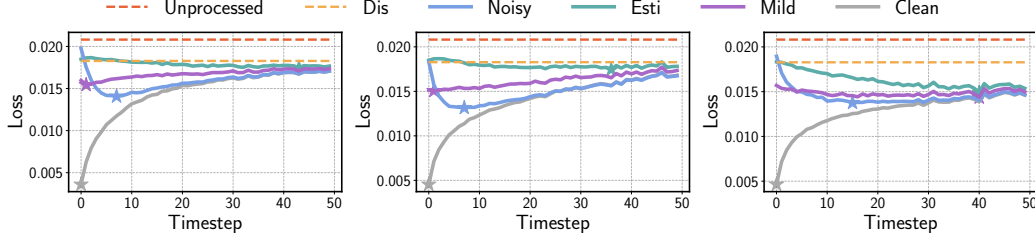


Figure 13: Performance of the conditional diffusion enhancement model with different dropout probability in CHIME-4. From left to right: (1) without dropout; (2) dropout 50% ; (3) dropout 90%.

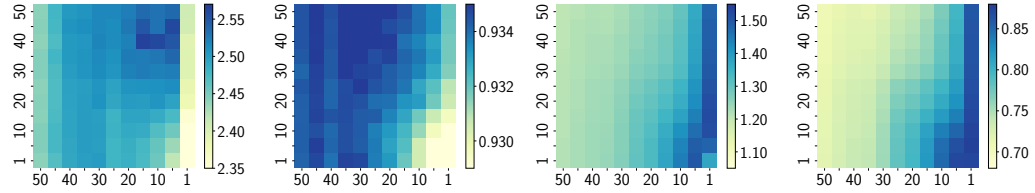


Figure 14: Performance of two-step speech generation. From left to right: (1) PESQ on matched VB dataset; (2) STOI on matched VB dataset; (3) PESQ on mismatched CHIME-4 dataset; (4) STOI on mismatched CHIME-4 dataset.

speech – and the diffusion enhancement model will degenerate into a deterministic model, losing its generalizability.

Timesteps τ_1 and τ_2 . Considering the computational complexity of the “one-step” grid search, we search optimal hyperparameters with a slightly larger step. Specifically, we select both optimal τ_1 and τ_2 from the predefined set $\{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. We show the PESQ and STOI performance of different combinations in Figure 14. We can observe that $\tau_1 > \tau_2$ (excluding STOI on matched VB dataset, and the difference is negligible: $0.930 \sim 0.934$) will lead to better performance, which is in line with our analysis (§4.4).

A.7 Visual case of excessive suppression

We present several visual cases of excessive suppression.

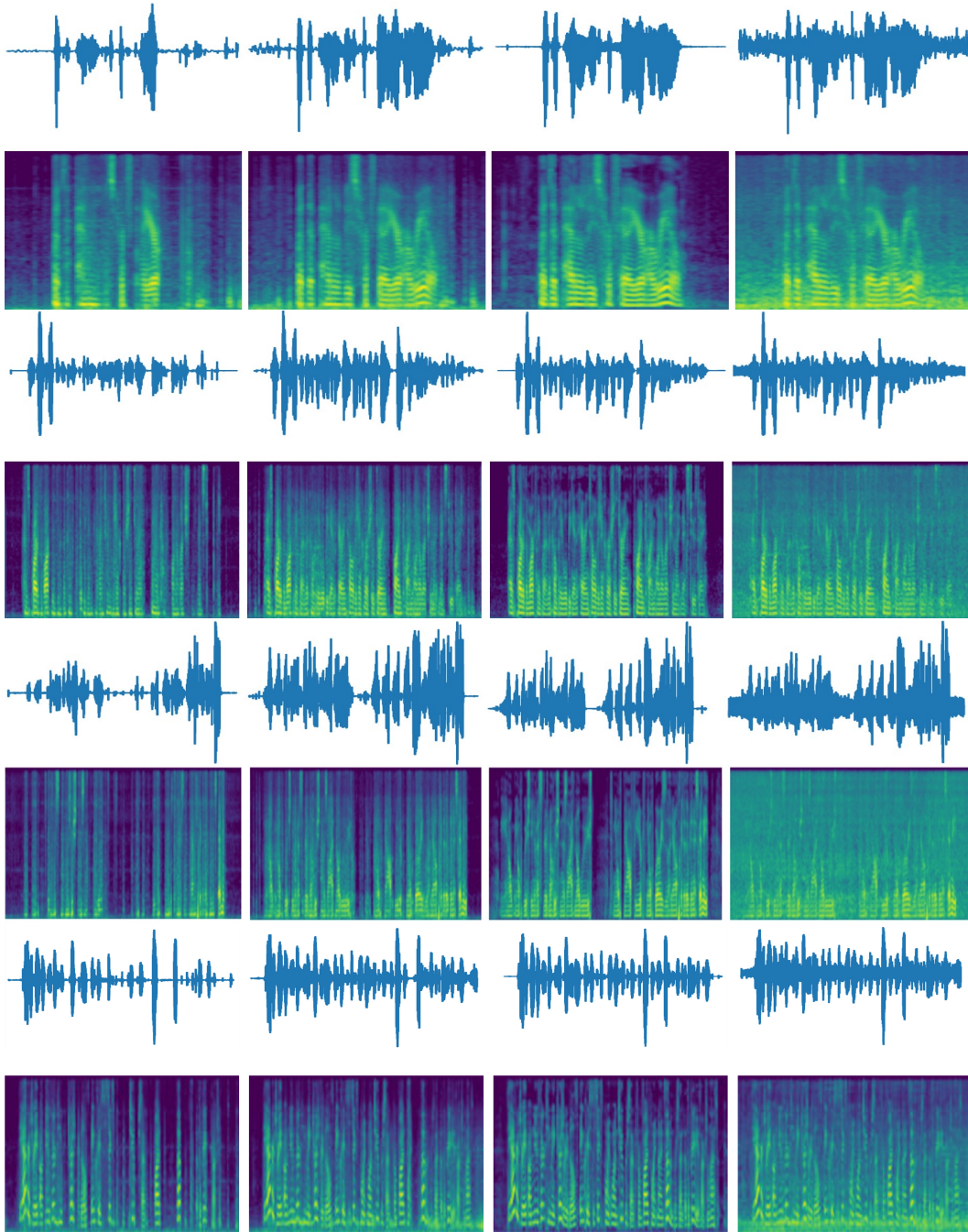


Figure 15: Excessive suppression visualization (CHIME-4, DOSE). From left to right: (1) estimated condition; (2) 2 steps; (3) clean; (4) noisy speech.

A.8 Visual case of error accumulation

We present several visual cases of error accumulation.

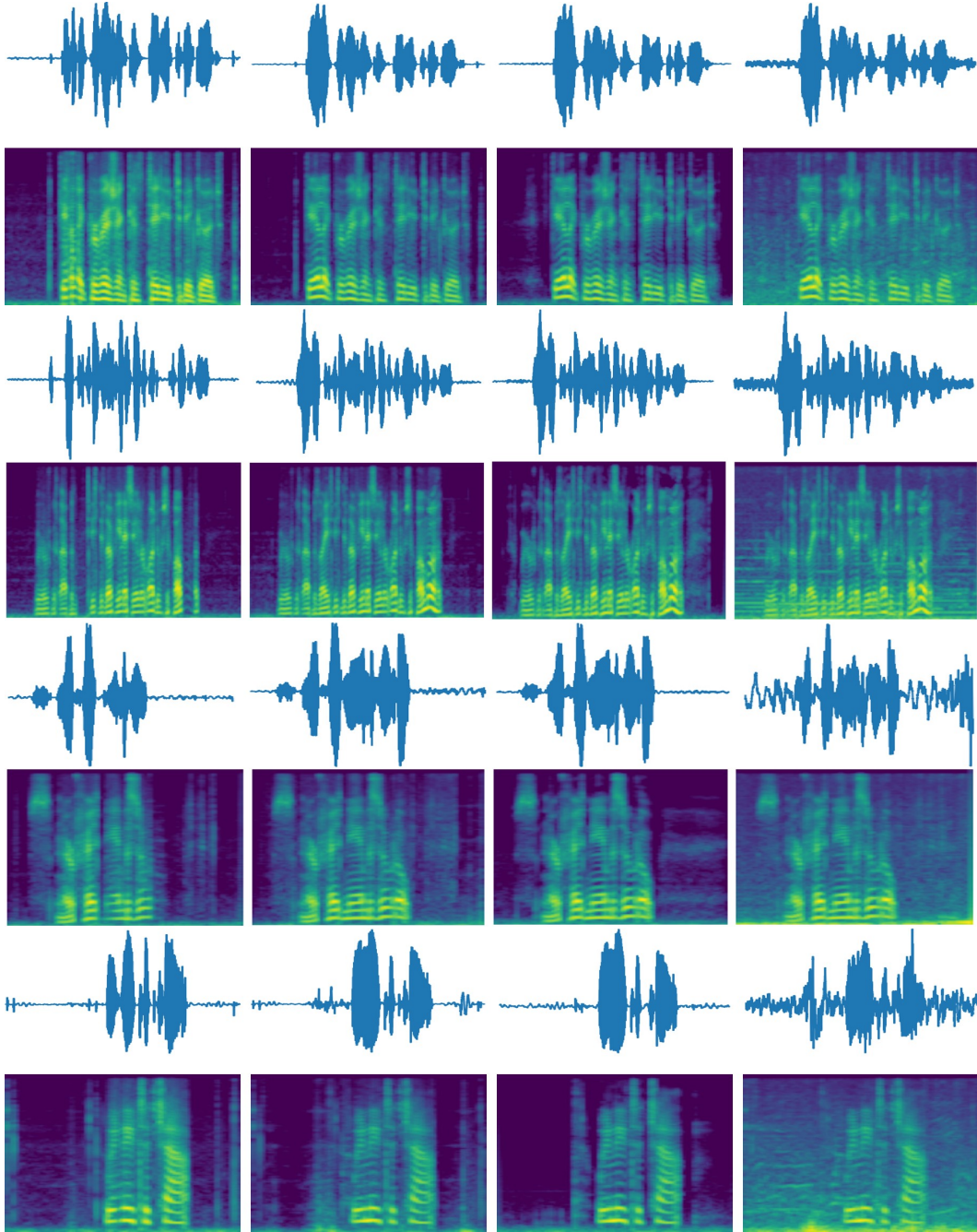


Figure 16: Error accumulation visualization (VB, DOSE). From left to right: (1) full (50) steps; (2) 2 steps; (3) clean; (4) noisy speech.

A.9 Discussion

To help readers better understand our approach, we analyze the reasons behind the better generalizability of diffusion enhancement models compared to deterministic mapping-based models (from the robust training perspective), explain why we use 0.5 in the mild version of the condition factor, and discuss the broader impacts of speech enhancement methods.

A.9.1 Generalization analysis

Given that full-step diffusion enhancement models have better generalizability than deterministic models, we now delve into the following question: if we are just using diffusion models as one-step (or two-step) denoisers, what accounts for their enhanced performance compared to deterministic mapping-based models?

To answer this question, we need to point out that training a diffusion model can be considered equivalent to training a multi-task paradigm. This involves training a model with shared parameters on multiple levels of Gaussian noise concurrently. Recent research [21] has demonstrated that the full training process of diffusion models leads to significantly improved one-shot denoising capabilities, which are more generalizable compared to previous works that trained standalone denoisers on a single noise level. Please refer to [21] (§5.2) for more details.

A.9.2 Why use 0.5 in the mild version of the condition factor? (reviewer boQC)

Employing an equal weight provides stability, yet the performance during instances with low SNR would be compromised (albeit still superior to direct utilization of y as a condition factor, unless the condition optimizer falters). One prospective solution involves introducing an additional adaptive strategy, i.e., $c = \alpha f_{\theta}(y) + (1 - \alpha)y$. We can design an adaptive *alpha* predictor and hope it can output α based on the quality of raw condition and samples from the condition optimizer. For example, when the condition optimizer produces lower-quality samples, giving more weight to the original condition factor would make sense. Conversely, if the raw condition factor has a low SNR, emphasizing the generated counterpart could be more effective. However, implementing this idea practically is intricate.

Given our strong reliance on diffusion-enhanced models to enhance generalization, any new adaptive strategy must be generalizable. For instance, training an adaptive alpha predictor on the (seen) VB-dataset (high SNR & consistent condition optimizer performance) could lead the model to consistently output higher α values for fusion. Unfortunately, this auxiliary model might not effectively adapt to variations when evaluating the mismatched (unseen) CHIME-4 dataset (low SNR & potential condition optimizer challenges). To this end, we might need other techniques such as data augmentation and adversarial training to improve its generalizability and robustness. This creates a dilemma: harnessing the speech diffusion model for overarching speech noise reduction generalization while simultaneously necessitating a pre-established generalized model to facilitate its implementation. So far, despite our efforts to train a strong alpha predictor, progress has been limited (the alpha predictor is still not generalizable, and the new system has no significant performance improvement over DOSE).

A.9.3 Broader impacts

As speech enhancement technology continues to advance and become more prevalent, it's important to consider its broader impacts.

Positive impacts. The impact of speech enhancement technology on real-life situations, particularly for individuals with hearing impairments, cannot be overstated. Hearing aids have long been the primary solution for those with hearing loss, but they are not always effective in noisy environments or for certain types of hearing loss. Speech enhancement technology can greatly improve speech intelligibility and communication for hearing aid users. For example, some hearing aids have AI-powered speech enhancement that boosts speech quality.

In addition to the benefits for individuals with hearing impairments, speech enhancement technology also has significant implications for various applications. In transportation, clearer and more intelligible speech can improve communication between pilots and air traffic control, leading to safer and more efficient air travel. In industry, speech enhancement can improve communication on

noisy factory floors, leading to increased productivity and safety. In educational settings, speech enhancement can improve student comprehension and engagement during lectures and presentations.

Negative impacts. While speech enhancement technology has the potential to greatly improve communication and speech intelligibility, one potential concern is that speech enhancement could modify the semantic content of speech, potentially misleading listeners. Thus, it’s important for developers of speech enhancement technology to consider this potential negative effect and work towards creating trustworthy systems.

A.10 Experimental details

Speech preprocessing. We process the speech waveform at a 16 kHz sampling rate. To maintain dimensionality consistency within mini-batches, we pad each utterance to 2 seconds (32000 points) using a zero-padding technique.

Basic architecture. To make a fair comparison, we use DiffWave [7] as the basic architecture following [4, 9] – the only difference being the change in the way of condition-injecting since most speech enhancement methods will directly use noisy speech as the condition factor, rather than Mel-spectrogram. We concatenate the condition factor with the intermediate-generated sample along the channel dimension as the model’s input. Specifically, the network is composed of 30 residual layers with residual channels 128. We use a bidirectional dilated convolution (Bi-DilConv) with kernel size 3 in each layer. We sum the skip connections from all residual layers. The total number of trainable parameters is 2.31M, slightly smaller than naive DiffWave (2.64M). Please refer to [7] and our code for more details.