Shallow Diffusion for Fast Speech Enhancement (Student Abstract)

Yue Lei[1], Bin Chen[1], Wenxin Tai[1,2*], Ting Zhong[1,2], Fan Zhou[1,2]

[1]University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China
[2]Kash Institute of Electronics and Information Industry, Kashgar 844000, China
{leiyue828, binchen4110}@gmail.com, wxtai@std.uestc.edu.cn, {zhongting, fan.zhou}@uestc.edu.cn

Abstract

Recently, the field of Speech Enhancement has witnessed the success of diffusion-based generative models. However, these diffusion-based methods used to take multiple iterations to generate high-quality samples, leading to high computational costs and inefficiency. In this paper, we propose SDFEN (**S**hallow **D**iffusion for **F**ast sp**E**ech e**N**hancement), a novel approach for addressing the inefficiency problem while enhancing the quality of generated samples by reducing the iterative steps in the reverse process of diffusion method. Specifically, we introduce the shallow diffusion strategy initiating the reverse process with an adaptive time step to accelerate inference. In addition, a dedicated noisy predictor is further proposed to guide the adaptive selection of time step. Experiment results demonstrate the superiority of the proposed SDFEN in effectiveness and efficiency.

Introduction

Speech enhancement (SE) plays an important role in many speech-related tasks, e.g., speech recognition and speech synthesis, aiming to improve the perceptual quality of speech signals in the presence of non-stationary background noise. Some studies attempt to introduce the diffusion model in SEs, achieving great success (Lu et al. 2022; Tai et al. 2023). However, a profound challenge persists within these diffusion-enhanced models: *high computation overhead* due the need for multiple time-step iterations in diffusion models to attain high-quality speech samples significantly hinder their practical deployment in real-world scenarios.

To address the problem of *high computation overhead*, we present our solution, SDFEN, for fast Speech enhancement. Specifically, inspired by (Liu et al. 2021) in audio synthesis, we introduce the shallow diffusion strategy, initiating the inference process with a intermediate timestep t rather than generating sample with a full trajectory.

To further reduce the steps in reverse process, we propose a novel noise predictor that aims to predict the proper prior with sufficient knowledge at time step t in the above shallow diffusion process, enabling the restoration of clean speech. An overview of proposed SDFEN is in Figure 1.

*Corresponding author: wxtai@std.uestc.edu.cn

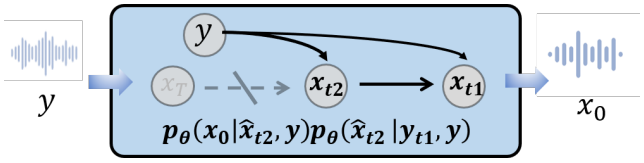

Figure 1: Overview of the proposed SDFEN architecture. Specifically, we generate the speech in only two steps. In the first step, a prior condition is generated from the noisy speech y, which is then applied to the diffusion model to obtain the estimated speech x_{t_2}. In the second step, we reuse the generated speech x_{t_2} to generate better estimates.

Methodology

Problem Definition. Speech enhancement aims to transfer noisy-reverberant speech to clean speech. Fomally, a noisy signal $\boldsymbol{y}$ in time domain can be expressed as $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$ where $\boldsymbol{x}$ and $\boldsymbol{n}$ denote clean and noise signal, respectively. The goal of speech-enhanced diffusion is to enhance intelligibility and quality by extracting $\boldsymbol{x}$ from $\boldsymbol{y}$:

$$p_\theta(\boldsymbol{x}_0|\boldsymbol{y}) = \int \underbrace{p(\boldsymbol{x}_T)}_{\text{Prior}} \prod_{t=1}^{T} \underbrace{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})}_{\text{Posterior}} d\boldsymbol{x}_{1:T}, \quad (1)$$

where $\boldsymbol{x}_T$ is sampled from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
Shallow diffusion. Recent studies (Liu et al. 2021) show that the sample generated at a relatively large time step t contains a large amount of noise. Inspired by this, we introduce the shallow diffusion that generate speech from an intermediate time step τ:

$$p_\theta(\boldsymbol{x}_0|\boldsymbol{y}) = \int \underbrace{p(\boldsymbol{x}_\tau|\boldsymbol{y})}_{\text{Prior}} \prod_{t=1}^{\tau} \underbrace{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})}_{\text{Posterior}} d\boldsymbol{x}_{1:t}, \quad (2)$$

where $p(\boldsymbol{x}_\tau|\boldsymbol{y})$ denotes a conditional prior at time step τ (we implement it via standard forward diffusion (Ho, Jain, and Abbeel 2020)):

$$\hat{\boldsymbol{x}}_\tau \sim p(\boldsymbol{x}_\tau|\boldsymbol{y}) := \sqrt{\bar{\alpha}_t}\hat{\boldsymbol{x}}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}. \quad (3)$$

In this way, the computational costs in the reverse process can been significantly reduced (from T time step to t time step), while having (almost) no significant influence in denoising performance. In practice, we devise a noise predictor $p_\psi(\boldsymbol{x}_\tau|\boldsymbol{y})$ to generate a better estimation.

Accordingly, Eq. (2) can be rewritten as:

$$p_\theta(\boldsymbol{x}_0|\boldsymbol{y}) = \int p_\psi(\boldsymbol{x}_\tau|\boldsymbol{y}) \prod_{t=1}^{\tau} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y) d\boldsymbol{x}_{1:t}. \quad (4)$$

To further accelerate the inference process, we substitute the second term $\prod_{t=1}^{\tau} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)$ with one-step generation, so that we can further reduce the sampling steps from entire τ steps to 2 steps. And the updated inference process can be represented as:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{y}) = \int \int p_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_2}|\boldsymbol{y}_{t_1}, \boldsymbol{y}) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\hat{\boldsymbol{x}}_{t_2}, \boldsymbol{y}) d\hat{\boldsymbol{x}}_{t_2} d\boldsymbol{y}_{t_1}$$

where t_1, t_2 are two pre-defined hyper-parameters. And once we choose the appropriate parameters, the $\boldsymbol{x}_0$ can be obtained in only two steps.

Training Objective. (Lu, Tsao, and Watanabe 2021) and following studies parameterize the denoising model by predicting $\boldsymbol{\epsilon}$ using a neural network $\epsilon_\theta(\boldsymbol{x}_t, t)$ to achieve $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \tilde{\beta}_t I)$ which implicitly establishes:

$$\hat{\boldsymbol{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta \right).$$

As discussed earlier, we want SDFEN to start with a small timestep and we strive to make t small. However, as t approaches zero, small changes in $\mathbf{x}$-space have an increasingly amplified effect on the implied prediction in $\boldsymbol{\epsilon}$-space. In other words, the efforts made by the diffusion enhancement model at small time step become so negligible that diffusion models lose their ability to recover natural-sounding speech from defective speech.

To this end, we turn to predict in $\mathbf{x}$-space by reparameterizing the training target in $\boldsymbol{\epsilon}$-space loss, and finally obtain a new training target as:

$$\mathcal{L} = \mathbb{E}_{x_0 \sim p(\boldsymbol{x}), t \in [1,T]} \left[\| \boldsymbol{x}_0 - f_\theta \left(\boldsymbol{x}_t, \boldsymbol{y}, t \right) \|_2^2 \right]. \quad (5)$$

Experiments

Dataset and Baselines. We use the VoiceBank-DEMAND dataset (Veaux, Yamagishi, and King 2013) for performance evaluations. We compare SDFEN with the following diffusion enhancement baselines: DiffuSE (Lu, Tsao, and Watanabe 2021), CDiffuSE (Lu et al. 2022), SGMSE (Welker, Richter, and Gerkmann 2022), and DR-DiffuSE (Tai et al. 2023) on the VoiceBank-DEMAND dataset (Veaux, Yamagishi, and King 2013).

Evaluation metrics. We use the following metrics to evaluate SE performance: short-time objective intelligibility (STOI), the perceptual evaluation of speech quality (PESQ), the mean opinion score (MOS) prediction of the speech signal distortion (CSIG), and the MOS prediction of the intrusiveness of background noise (CBAK). In addition, we show the number of steps in the reverse process to assess the efficiency of models.

Implementation Details. We implement our method using DiffWave architecutre (Kong et al. 2021), the same architecture as all baselines. DiffWave takes 50 steps with the linearly spaced training noise schedule $\beta_t \in [1 \times 10^{-4}, 0.035]$. All methods are trained for 300k iterations on `RTX 3090` with a batch size of 16 audios. We determine optimal values for t_1 and t_2 by evaluating their performance on a validation dataset, extracted as a subset from the training data.

Method	Step	STOI	PESQ	CSIG	CBAK
*	–	92.1	1.97	3.35	2.44
DiffuSE	6	93.5	2.39	3.71	3.04
CDiffuSE	6	**93.7**	2.43	3.77	3.09
SGMSE	50	93.3	2.34	3.69	2.90
DR-DiffuSE	6	92.9	2.50	3.68	3.27
SDFEN	**2**	93.4	**2.55**	**3.81**	**3.27**

Table 1: Performance Comparison. We represent the number of steps in the reverse process as the efficiency of models. * denotes the unprocessed model. The best results are in bold font.

Performance Comparison. Table 1 summarizes the experimental results and we have the following observations: (1) SGMSE with 50 steps yields subpar results compared to that generated with fewer steps. This phenomenon is consistent with our hypothesis that generating samples starting from a relatively noisy step has no significant benefit for generating faithful speech. (2) Compared with all baselines, SDFEN achieves better generation results with only two steps, showing the strong efficiency and effectiveness of the shallow diffusion.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No.62176043 and No.62072077)

References

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *ICLR*.

Liu, J.; Li, C.; Ren, Y.; Chen, F.; and Zhao, Z. 2021. DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. In *AAAI*.

Lu, Y.-J.; Tsao, Y.; and Watanabe, S. 2021. A study on speech enhancement based on diffusion probabilistic model. In *2021 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*, 7402–7406. IEEE.

Tai, W.; Zhou, F.; Trajcevski, G.; and Zhong, T. 2023. Revisiting Denoising Diffusion Probabilistic Models for Speech Enhancement: Condition Collapse, Efficiency and Refinement. In *AAAI*.

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.

Welker, S.; Richter, J.; and Gerkmann, T. 2022. Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain. In *Proc. Interspeech 2022*, 2928–2932.