CONDITIONAL DIFFUSION PROBABILISTIC MODEL FOR SPEECH ENHANCEMENT

Yen-Ju Lu[1,3]*, Zhong-Qiu Wang*[1]*, Shinji Watanabe*[1]*, Alexander Richard*[2]*, Cheng Yu*[3]*, and Yu Tsao*[3]

[1]Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Reality Labs Research, Pittsburgh PA, USA
[3]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

Speech enhancement is a critical component of many user-oriented audio applications, yet current systems still suffer from distorted and unnatural outputs. While generative models have shown strong potential in speech synthesis, they are still lagging behind in speech enhancement. This work leverages recent advances in diffusion probabilistic models, and proposes a novel speech enhancement algorithm that incorporates characteristics of the observed noisy speech signal into the diffusion and reverse processes. More specifically, we propose a generalized formulation of the diffusion probabilistic model named conditional diffusion probabilistic model that, in its reverse process, can adapt to non-Gaussian real noises in the estimated speech signal. In our experiments, we demonstrate strong performance of the proposed approach compared to representative generative models, and investigate the generalization capability of our models to other datasets with noise characteristics unseen during training.

Index Terms— speech enhancement, diffusion probabilistic model, generative model, deep learning

1. INTRODUCTION

Speech enhancement, a key element for immersive audio experiences in telecommunication as well as a crucial front-end processor for robust speech recognition [1, 2], assistive hearing [3], and robust speaker recognition [4, 5], is a challenging and still unsolved problem in audio processing. Riding on the advance of deep learning, considerable progress has been made in the past decade [6, 7]. Deep learning based approaches can be roughly divided into two categories, based on the criteria used to estimate the transformation function from noisy-reverberant speech to clean speech. The first category trains discriminative models to minimize the difference between enhanced and clean speech, where the difference can be a point-wise L_p-norm distance [8], or can be computed based on a perceptual metric [9, 10]. The second category considers the distribution of the clean speech signals to form the objective function. Well-known examples along this direction include generative adversarial networks (GANs) [11, 12], Bayesian wavenet [13], variational autoencoders [14], and flow-based models [15]. While the best performing approaches typically fall into the first category [16, 6], they usually introduce unpleasant speech distortion and phonetic inaccuracies to the enhanced speech [17, 18, 19]. Generative approaches that aim to match the distribution of speech signals rather than regressive approaches optimizing a point-wise loss hold the promise to produce more natural sounding speech, although they are currently lagging behind regressive approaches and require more research to unfold their potential.

This work investigates diffusion probabilistic models [20], a class of generative models that have shown outstanding performance in image generation [21, 22] and audio synthesis [23, 24, 25], for speech enhancement. Diffusion probabilistic models convert clean input data to an isotropic Gaussian distribution in a step-by-step diffusion process and, in a reverse process, gradually restore the clean input by predicting and removing the noise introduced in each step of the diffusion process. These models, in their vanilla formulation, assume isotropic Gaussian noise in each step of the diffusion process as well as the reverse process. However, in realistic conditions, the noise characteristics are usually non-Gaussian, which violates the model assumption when directly combining the noisy speech signal in the sampling process. We address this problem by formulating a generalized conditional diffusion probabilistic model that incorporates the observed noisy data into the model. We derive the corresponding conditional diffusion and reverse processes as well as the evidence lower bound (ELBO) optimization criterion [21], and show that the resulting model is a generalization of the original diffusion probabilistic model. In our experiment, we will demonstrate that our formulation can not only improve over the vanilla diffusion probabilistic model, but also outperform other generative models.

2. DIFFUSION PROBABILISTIC MODEL

A T-step diffusion model [21] consists of two processes: the *diffusion* process with steps $t \in \{0, 1, \cdots, T\}$ and the *reverse* process $t \in \{T, T-1, \cdots, 0\}$. We start with a brief summary of the vanilla diffusion probabilistic model, *i.e.*, we revisit the original diffusion- and reverse process.

2.1. Diffusion Process

Given the clean speech data x_0, the diffusion process $q_{\text{data}}(x_0)$ of the first diffusion step ($t = 0$) is defined as the data distribution x_0 on $\mathbb{R}^L$, where L is the signal length in samples. For the the t-th diffusion step, we have a step-dependent variable $x_t \in \mathbb{R}^L$ with the same signal length L. The diffusion process from data x_0 to the variable x_T can be formulated based on a fixed Markov chain:

$$q(x_1, \cdots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

with a Gaussian model $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$, where β_t is a small positive constant. In other words, in each step a Gaussian noise is added to the previous sample x_{t-1}. According to the pre-defined schedule $\beta_1, \cdots, \beta_T$, the overall process gradually converts clean x_0 to a latent variable with an isotropic Gaussian distribution of $p_{\text{latent}}(x_T) = \mathcal{N}(0, I)$.

By substituting the Gaussian model of $q(x_t|x_{t-1})$ into Eq. (1) and by marginalizing $x_1, \ldots, x_{t-1}$, the sampling distribution of x_t

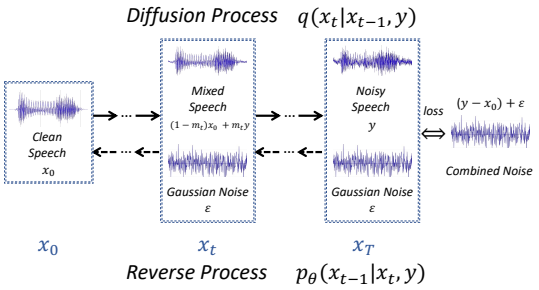

Fig. 1. Diffusion process (solid arrows) and reverse process (dashed arrows) of the proposed conditional diffusion probabilistic model.

can be derived as the following distribution conditioned on x_0:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

2.2. Reverse Process

The reverse process converts the latent variable $x_T \sim N(0, I)$ to x_0, also based on a Markov chain similar to Eq. (1):

$$p_\theta(x_0, \cdots, x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{3}$$

where $p_\theta(\cdot)$ is the distribution of the reverse process with learnable parameters θ. Unlike the diffusion process, the following marginal likelihood is intractable:

$$p_\theta(x_0) = \int p_\theta(x_0, \cdots, x_{T-1}|x_T) \cdot p_{\text{latent}}(x_T) dx_{1:T}. \tag{4}$$

Therefore, we use the ELBO to form an approximated objective function for model training. In [21], it is reported that minimizing the following equation leads to higher generation quality:

$$c + \sum_{t=1}^{T} \kappa_t \mathbb{E}_{x_0,\epsilon} \parallel \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \parallel_2^2, \tag{5}$$

with constants c and κ_t. Here ϵ_θ is the model trained to estimate the Gaussian noise ϵ in x_t. After optimizing Eq. (5), the corresponding reverse process equation becomes:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t I), \tag{6}$$

where the mean $\mu_\theta(x_t, t)$ is:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)). \tag{7}$$

The $\mu_\theta(x_t, t)$ predicts the mean of x_{t-1} distribution by removing the estimated Gaussian noise $\epsilon_\theta(x_t, t)$ in the x_t, and the variance is fixed to a constant $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

3. CONDITIONAL DIFFUSION PROBABILISTIC MODEL

The original diffusion process in Section 2.1 starts from the clean data $q_{\text{data}}(x_0)$ and adds Gaussian noise into the speech signal. In the proposed conditional diffusion probabilistic model, we incorporate the noisy data y into the diffusion process, as shown in Fig. 1.

3.1. Conditional Diffusion Processes

In the conditional diffusion process, we use an interpolation parameter m_t to combine the clean data x_0 and the noisy data y, the summation of x_0 and the real noise n, on the solid arrows in Fig. 1. Instead of starting from the Markov chain Gaussian model $q(x_t|x_{t-1})$ in the original diffusion process, we first define the following conditional diffusion process $q(x_t|x_0, y)$:

$$q_{\text{cdiff}}(x_t|x_0, y) = \mathcal{N}(x_t; (1-m_t)\sqrt{\bar{\alpha}_t}x_0 + m_t\sqrt{\bar{\alpha}_t}y, \delta_t I), \tag{8}$$

where δ_t is the variance. Unlike the original diffusion process $q(x_t|x_0)$ in Eq. (2), we assume that the Gaussian mean in Eq. (8) is represented as a linear interpolation between the clean data x_0 and the noisy data y with the interpolation ratio m_t. m_t starts from $m_0 = 0$ and is gradually increased to $m_T \approx 1$, turning the mean of x_t from the clean speech x_0 to noisy speech y as in Fig. 1.

Given the interpolation formulation in Eq. (8), we can derive $q_{\text{cdiff}}(x_t|x_0) = \int q_{\text{cdiff}}(x_t|x_0, y)p_y(y|x_0)dy$ by marginalizing y in the multiplication of $q_{\text{cdiff}}(x_t|x_0, y)$ and $p_y(y|x_0)$ with the special case where $n \sim \mathcal{N}(0, I)$. Then, $q_{\text{cdiff}}(x_t|x_0)$ becomes equivalent to the original diffusion process $q(x_t|x_0)$ in Eq. (2) when

$$\delta_t = (1-\bar{\alpha}_t) - m_t^2\bar{\alpha}_t. \tag{9}$$

This analytical result indicates that our model is a generalization of the original diffusion probabilistic model. In our previous study [25], we investigated directly utilizing noisy signal in the reverse process; the idea is found to work well experimentally, but there lacks a theoretical justification. In Sec. 3.2, we will propose a conditional reverse process that is theoretically sound. To further research the effect of incorporating noisy signal in the diffusion model, in Sec. 3.3, we will set δ_t according to Eq. (9) so that the conditional diffusion process becomes a generalized version of the original diffusion process[1].

3.2. Conditional Reverse Processes

In the conditional reverse process, we start from x_T, noisy speech signal y with variance δ_T, according to Eq. (8) with $m_T = 1$:

$$p_{\text{cdiff}}(x_T|y) = \mathcal{N}(x_T, \sqrt{\bar{\alpha}_T}y, \delta_T I). \tag{10}$$

Based on the Markov chain, similar to Eq. (6), the conditional reverse process on the dashed arrows in Figure 1 aims to predict x_{t-1} based on x_t and y:

$$p_{\text{cdiff}}(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y, t), \tilde{\delta}_t I), \tag{11}$$

where the $\mu_\theta(x_t, y, t)$ is the estimated mean of the conditional reverse process. The concrete form of the variance $\tilde{\delta}_t$ is introduced later. In contrast to the vanilla diffusion model, we further condition the diffusion model on y. Therefore, similar to Eq. (7), the mean $\mu_\theta(x_t, y, t)$ in each reverse step is a linear combination of x_t, y, and estimated noise ϵ with weights c_{xt}, c_{yt} and $c_{\epsilon t}$,

$$\mu_\theta(x_t, y, t) = c_{xt}x_t + c_{yt}y - c_{\epsilon t}\epsilon_\theta(x_t, y, t), \tag{12}$$

where the $\epsilon_\theta(x_t, y, t)$ is the model to estimate the Gaussian and non-Gaussian noise combination. The coefficients c_{xt}, c_{yt}, and $c_{\epsilon t}$ can be derived from the ELBO optimization criterion, see Section 3.3.

3.3. Coefficient Estimation by Optimizing ELBO

By modifying the derivations in [21], we obtain the ELBO condition for the conditional diffusion process to optimize the likelihood:

$$\begin{aligned} ELBO = -\mathbb{E}_q \Big(& D_{\text{KL}}(q_{\text{cdiff}}(x_T|x_0, y)||p_{\text{latent}}(x_T|y)) \\ & + \sum_{t=2}^{T} D_{\text{KL}}(q_{\text{cdiff}}(x_{t-1}|x_t, x_0, y)||p_\theta(x_{t-1}|x_t, y)) \\ & - \log p_\theta(x_0|x_1, y) \Big). \end{aligned} \tag{13}$$

[1]It is difficult to derive $q_{\text{cdiff}}(x_t|x_{t-1})$ for satisfying the original diffusion process if we first define $q_{\text{cdiff}}(x_t|x_{t-1}, y)$, because the distribution of y depends on x_0 as $y = x_0 + n$.

Algorithm 1 Training

for $i = 1, 2, \cdots, N_{\text{iter}}$ **do**
  Sample $(x_0, y) \sim q_{\text{data}}$, $\epsilon \sim \mathcal{N}(0, I)$, and
  $t \sim$Uniform$(\{1, \cdots, T\})$
  $x_t = ((1 - m_t)\sqrt{\bar{\alpha}_t}x_0 + m_t\sqrt{\bar{\alpha}_t}y) + \sqrt{\delta_t}\epsilon$
  Take gradient step on
  $\nabla_\theta \parallel \frac{1}{\sqrt{1-\bar{\alpha}_t}}(m_t\sqrt{\bar{\alpha}_t}(y - x_0) + \sqrt{\delta_t}\epsilon) - \epsilon_\theta(x_t, y, t) \parallel_2^2$
  according to Eq. (21)
end for

Algorithm 2 Sampling

Sample $x_T \sim \mathcal{N}(x_T, \sqrt{\bar{\alpha}_T}y, \delta_T I)$,
for $t = T, T-1, \cdots, 1$ **do**
  Compute c_{xt}, c_{yt} and $c_{\epsilon t}$ using Eqs. (18), (19), and (20)
  Sample $x_{t-1} \sim p_{\text{cdiff}}(x_{t-1}|x_t, y) =$
  $\mathcal{N}(x_{t-1}; c_{xt}x_t + c_{yt}y - c_{\epsilon t}\epsilon_\theta(x_t, y, t), \tilde{\delta}_t I)$
end for
return x_0

To optimize Eq. (13), we first need the distribution $q_{\text{cdiff}}(x_t|x_{t-1}, y)$. Generally, the diffusion process define $q_{\text{cdiff}}(x_t|x_{t-1}, y)$ first and derive $q_{\text{cdiff}}(x_t|x_0, y)$ by marginalizing $x_0, \cdots, x_{t-1}$. Instead, we first design the interpolation form in Eq. (8) as mentioned in Sec. 3.1. Therefore, we compare the coefficients of the marginalized result and Eq. (8) to compute the coefficients of $q_{\text{cdiff}}(x_t|x_{t-1}, y)$ as:

$$q_{\text{cdiff}}(x_t|x_{t-1}, y) = \mathcal{N}\Big(x_t; \frac{1 - m_t}{1 - m_{t-1}}\sqrt{\alpha}_t x_{t-1} + \Big(m_t - \frac{1 - m_t}{1 - m_{t-1}}m_{t-1}\Big)\sqrt{\bar{\alpha}}_t y, \delta_{t|t-1}I\Big), \quad (14)$$

where the $\delta_{t|t-1}$ is also calculated by δ_t to satisfy Eq. (9) as:

$$\delta_{t|t-1} = \delta_t - \left(\frac{1 - m_t}{1 - m_{t-1}}\right)^2 \alpha_t \delta_{t-1}. \quad (15)$$

Then, by combining Eqs. (8) and (14), $q_{\text{cdiff}}(x_{t-1}|x_t, x_0, y)$ can be derived through Bayes' theorem and the Markov chain property:

$$q_{\text{cdiff}}(x_{t-1}|x_t, x_0, y) = \mathcal{N}\Big(x_{t-1}; \frac{1 - m_t}{1 - m_{t-1}}\frac{\delta_{t-1}}{\delta_t}\sqrt{\alpha}_t x_t + (1 - m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}\sqrt{\bar{\alpha}}_{t-1}x_0 + \Big(m_{t-1}\delta_t - \frac{m_t(1 - m_t)}{1 - m_{t-1}}\alpha_t\delta_{t-1}\Big)\frac{\sqrt{\bar{\alpha}_{t-1}}}{\delta_t}y, \tilde{\delta}_t I\Big), \quad (16)$$

where $\tilde{\delta}_t$, the variance term of $q_{\text{cdiff}}(x_{t-1}|x_t, x_0, y)$, is

$$\tilde{\delta}_t = \frac{\delta_{t|t-1} * \delta_t}{\delta_{t-1}}. \quad (17)$$

To optimize the KL divergence term in Eq. (13), $\tilde{\delta}_t$ is also used as the variance of $p_{\text{cdiff}}(x_{t-1}|x_t, y)$ in Eq. (11) to match $q_{\text{cdiff}}(x_{t-1}|x_t, x_0, y)$, and the coefficients c_{xt}, c_{yt}, $c_{\epsilon t}$ in Eq. (12) are then be derived as:

$$c_{xt} = \frac{1 - m_t}{1 - m_{t-1}}\frac{\delta_{t-1}}{\delta_t}\sqrt{\alpha}_t + (1 - m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}\frac{1}{\sqrt{\alpha}_t}, \quad (18)$$

$$c_{yt} = (m_{t-1}\delta_t - \frac{m_t(1 - m_t)}{1 - m_{t-1}}\alpha_t\delta_{t-1})\frac{\sqrt{\bar{\alpha}_{t-1}}}{\delta_t}, \quad (19)$$

$$c_{\epsilon t} = (1 - m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}\frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\alpha_t}}. \quad (20)$$

Now, given the explicit form of all distributions in Eq. (13), the ELBO to be optimized simplifies to

$$c' + \sum_{t=1}^{T} \kappa'_t \mathbb{E}_{x_0,\epsilon,y} \parallel (\frac{m_t\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}}(y - x_0) + \frac{\sqrt{\delta_t}}{\sqrt{1 - \bar{\alpha}_t}}\epsilon) - \epsilon_\theta(x_t, y, t) \parallel_2^2 \quad (21)$$

with constants c' and κ'_t, and the ϵ is the Gaussian noise in x_t. Because we have the interpolation form of x_t with the coefficient m_t in Eq. (8), the optimization target in Eq. (21) keeps the simple form in training. Comparing to Eq. (5), the $\epsilon_\theta(x_t, y, t)$ in the conditional diffusion model estimates both Gaussian noise ϵ and non-Gaussian noise $y - x_0$ in x_t. Therefore, the proportion of $y - x_0$ and ϵ coefficients is the same as y and the standard deviation in Eq. (8).

3.4. CDiffuSE Training and Sampling Algorithm

In the conditional reverse process, according to Eq. (11) and (21), $\epsilon_\theta(x_t, y, t)$ computes the combined noise, which is then deducted from the combination of x_t and y to obtain cleaned data x_{t-1}. Finally, iterative application of this process over all T steps yields the clean signal x_0. The overall diffusion and reverse process of the conditional diffusion probabilistic models are described in Algorithms 1 and 2. When the interpolation weight m_t of the real noise is set to 0, the optimization target in Eq. (21) and the reverse process in (11) becomes (5) and (6) as in the original diffusion probabilistic models.

In our previous study [25], a supportive reverse process was proposed as a less theoretically rigorous implementation to carry out the reverse process from the noisy speech (rather than isotropic Gaussian noise in the original reverse process) without changing the diffusion process. In our proposed CDiffuSE model, we remove the assumption that the real noise in y follows the Gaussian distribution and avoid the mismatch issue between the diffusion and reverse process.

4. EXPERIMENTS

In this section, we evaluate the performance of our approach against other generative speech enhancement models and we show generalization capabilities under conditions where state of the art approaches such as Demucs [16] collapse. The samples of the CDiffuSE-enhanced signals can be found online[2].

4.1. Experimental Setup

Dataset: we evaluate the CDiffuSE model on the VoiceBank-DEMAND dataset [26]. The dataset consists of 30 speakers from the VoiceBank corpus [27], which is further divided into a training set and a testing set with 28 and 2 speakers, respectively. The training utterances are artificially contaminated with eight real-recorded noise samples from the DEMAND database [28] and two artificially generated noise samples at 0, 5, 10, and 15 dB SNR levels, amounting to 11,572 utterances. The testing utterances are mixed with different noise samples at 2.5, 7.5, 12.5, and 17.5 dB SNR levels, amounting to 824 utterances in total. We consider perceptual evaluation of speech quality (PESQ) [29], prediction of the signal distortion (CSIG), background intrusiveness (CBAK), and overall speech quality (COVL) [30] as the evaluation metrics.

Model Architecture and Training: we implement CDiffuSE based on the same model architecture and the same pre-training strategy with clean Mel-filterbank conditioner as that of DiffuSE reported in [25]. We investigate two systems, namely Base and Large CDiffuSE, which respectively take 50 and 200 diffusion steps. The linearly spaced training noise schedule is reduced to $\beta_t \in [1 \times 10^{-4}, 0.035]$ for Base CDiffuSE, and to $\beta_t \in [1 \times 10^{-4}, 0.0095]$ for Large CDiffuSE. The interpolation parameter m_t in Section 3.1 is set to $m_t =$

[2]https://github.com/neillu23/CDiffuSE

$\sqrt{(1-\bar{\alpha}_t)/\sqrt{\bar{\alpha}_t}}$ which satisfies the $m_0 = 0$ and $m_t \approx 1$ requirement. We train both Base and Large CDiffuSE models for 300,000 iterations, based on an early stopping scheme. The batch size is set to 16 for Base CDiffuSE and to 15 for Large CDiffuSE. The fast sampling scheme [23] is used in the reverse processes with the inference schedule $\gamma_t = [0.0001, 0.001, 0.01, 0.05, 0.2, 0.35]$ for both Base CDiffuSE and Large CDiffuSE. The proposed CDiffuSE model performs enhancement in the time domain. After the reverse process is completed, the enhanced waveform further combine the original noisy signal with the ratio 0.2 to recover the high frequency speech in the final enhanced waveform, as suggested in [16, 31].

4.2. Evaluation results

4.2.1. Results on VoiceBank-DEMAND

In Table 1, we report the results of CDiffuSE and DiffuSE from [25]. As expected, the large models for DiffuSE and CDiffuSE both outperform the smaller base models. Moreover, CDiffuSE shows improved performance over the diffusion probabilistic model baseline DiffuSE. Note that the key to success here is that CDiffuSE has had direct access to the noisy data while learning the reverse diffusion process, allowing it to actively compensate for the noise characteristics in the input signals. Being able to leverage noise from the input signal, our approach improves on all the metrics, confirming that the theoretically sound CDiffuSE leads to improved results in practice. We additionally compare CDiffuSE to other time-domain generative models, namely SEGAN [11], SE-Flow [15], and improved deep SEGAN (DSEGAN) [32]. CDiffuSE outperforms its competitors on all metrics - with the exception of CBAK - and achieves a particularly significant improvement in PESQ, see Table 2.

Table 1. Results of DiffuSE and CDiffuSE on VoiceBank.

Method	PESQ	CSIG	CBAK	COVL
Unprocessed	1.97	3.35	2.44	2.63
DiffuSE (Base) [25]	2.41	3.61	2.81	2.99
CDiffuSE (Base)	**2.44**	**3.66**	**2.83**	**3.03**
DiffuSE (Large) [25]	2.43	3.63	2.81	3.01
CDiffuSE (Large)	**2.52**	**3.72**	**2.91**	**3.10**

Table 2. Performance comparison of CDiffuSE and time-domain generative models on VoiceBank.

Method	PESQ	CSIG	CBAK	COVL
Unprocessed	1.97	3.35	2.44	2.63
SEGAN [11]	2.16	3.48	2.94	2.80
DSEGAN [32]	2.39	3.46	**3.11**	2.90
SE-Flow [15]	2.28	3.70	3.03	2.97
CDiffuSE (Base)	**2.44**	3.66	2.83	**3.03**
CDiffuSE (Large)	**2.52**	**3.72**	2.91	**3.10**

Table 3. Comparison of CDiffuSE and discriminative models.

(a) Trained and tested on VoiceBank (**matched** condition).

Method	PESQ	CSIG	CBAK	COVL
Unprocessed	1.97	3.35	2.44	2.63
WaveCRN [33]	2.63	3.95	3.06	3.29
Demucs* [16]	2.65/3.07	**3.99**/4.31	**3.33**/3.40	**3.32**/3.63
Conv-TasNet [34]	**2.84**	2.33	2.62	2.51
CDiffuSE (Large)	2.52	3.72	2.91	3.10

(b) Trained on VoiceBank, tested on CHiME-4 (**mismatched** condition).

Method	PESQ	CSIG	CBAK	COVL
Unprocessed	1.27	2.61	1.93	1.88
WaveCRN [33]	1.43	2.53	2.03	1.91
Demucs* [16]	1.38	2.50	2.08	1.88
Conv-TasNet [34]	1.63	1.70	1.82	1.54
CDiffuSE (Large)	**1.66**	**2.98**	**2.19**	**2.27**

*We directly used the default setup from https://github.com/facebookresearh/denoiser to test performance, and we also copied the Demucs results in [16] to the right-side of "/" in Table 3(a).

4.2.2. Results on CHiME-4

Generative models typically aim to fit the distribution of the training samples instead of optimizing a point-wise L_p-loss. This property has made them state of the art in applications like text-to-speech and vocoding [35, 36] and also makes them more robust against domain shifts in the input data.

In this section, we investigate this property of our proposed CDiffuSE. We compare the generalization abilities of our approach to other, L_p-loss based approaches and demonstrate that our approach is particularly resistant towards shifts in noise characteristics of the speech data. The models in this section are trained on VoiceBank-DEMAND and evaluated on the simulated test data of CHiME-4 [37]. The CHiME-4 simulated test data is created based on real-recorded noises from four real-world environments (including street, pedestrian areas, cafeteria and bus) based on four speakers. We use the signals from the fifth microphone for evaluation.

As mentioned previously and as Table 3(a) shows, generative speech enhancement models are still lagging behind the performance of their regressive counterparts. A model from the latter category trained on VoiceBank and evaluated on the VoiceBank test set performs far better than most generative methods. Particularly, Demucs [16] and Conv-TasNet [34] outperform our CDiffuSE, which was the strongest generative model in Table 2.

Given a domain shift in test data, however, regression based approaches such as Demucs, Conv-TasNet, and WaveCRN suffer from a significant drop in performance, see Table 3(b). Different signal characteristics between the VoiceBank training data and the CHiME-4 test set suffice to let the evaluation scores fall drastically, in some cases even below the scores of unprocessed data. Our proposed CDiffuSE, on the contrary, proves to be much more resilient against such shifts in signal characteristics. While the scores on the CHiME-4 test set are lower than the VoiceBank scores, CDiffuSE degrades to a much smaller degree than its regressive competitors, leaving it with the best scores on the CHiME-4 test data and demonstrating its high robustness to variation in noise characteristics.

5. CONCLUSION

We proposed CDiffuSE, a conditional diffusion probabilistic model that can explore noise characteristics from the noisy input signal explicitly and thereby adapts better to non-Gaussian noise statistics in real-world speech enhancement problems. We showed that our model is a strict generalization of the original diffusion probabilistic model and achieves state of the art results compared to other generative speech enhancement approaches. In contrast to non-generative approaches, our method exposes great generalization capabilities to speech data with noise characteristics not observed in the training data. We were able to show that CDiffuSE maintains strong performance when regression-based approaches such as Demucs and Conv-TasNet collapse.

6. REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[2] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, 2020.

[3] E. W. Healy, J. L. Vasko, and D. Wang, "The optimal threshold for removing noise from speech is similar across normal and impaired hearing—a time-frequency masking study," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. EL581–EL586, 2019.

[4] J. H.L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[5] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *arXiv preprint arXiv:1709.01703*, 2017.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech 2013*.

[8] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[9] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[10] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML 2019*.

[11] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[12] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP 2018*.

[13] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian Wavenet," in *Proc. Interspeech 2017*.

[14] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *Proc. ICASSP 2020*.

[15] M. Strauss and B. Edler, "A flow-based neural network for time domain speech enhancement," in *Proc. ICASSP 2021*.

[16] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[17] P. Wang, K. Tan *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.

[18] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *Proc. ICASSP 2018*.

[19] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Snr-based progressive learning of deep neural network for speech enhancement." in *Proc. Interspeech 2016*.

[20] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. ICML 2015*.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arXiv:2006.11239*, 2020.

[22] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *arXiv preprint arXiv:2102.09672*, 2021.

[23] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[24] S. Liu, Y. Cao, D. Su, and H. Meng, "Diffsvc: A diffusion probabilistic model for singing voice conversion," *arXiv preprint arXiv:2105.13871*, 2021.

[25] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *arXiv preprint arXiv:2107.11876*, 2021.

[26] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[27] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. CASLRE 2013*.

[28] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013, p. 035081.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," vol. 2, 2001, pp. 749–752.

[30] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[31] M. Abd El-Fattah, M. I. Dessouky, S. Diab, and F. Abd El-Samie, "Speech enhancement using an adaptive wiener filtering approach," *Progress In Electromagnetics Research M*, vol. 4, pp. 167–184, 2008.

[32] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving gans for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

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

[34] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[35] A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[36] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[37] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.