CDiffuSE:
Conditional Diffusion Probabilistic Model for Speech Enhancement

Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe,
Alexander Richard, Cheng Yu, Yu Tsao

Carnegie Mellon University
Language Technologies Institute

Meta

Introduction

- We formulate 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
- In our experiment, conditional diffusion model not only improve over the vanilla diffusion model, but also outperform other generative models.

1

Denoising Diffusion Probabilistic Model

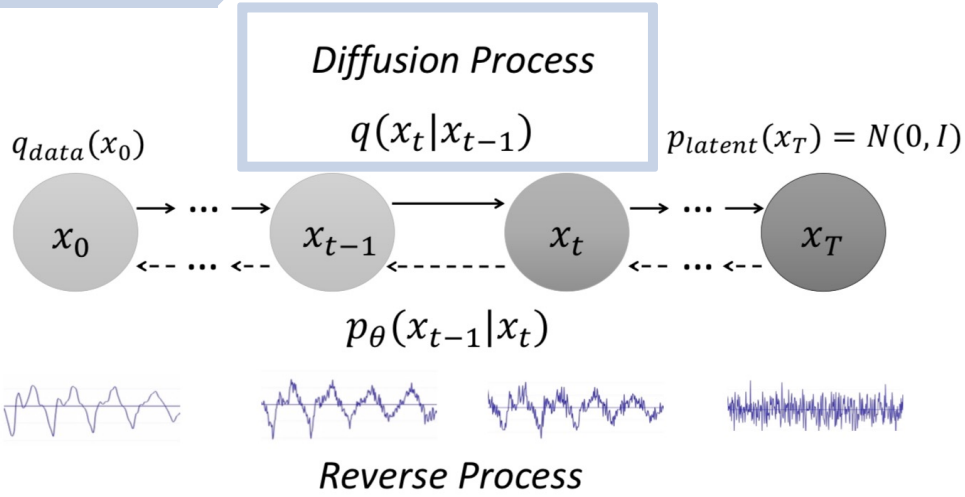
DDPM

Diffusion Probabilistic Model

Diffusion Markov process

Combines Gaussian noise into the clean speech x_0.

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



Diffusion Probabilistic Model

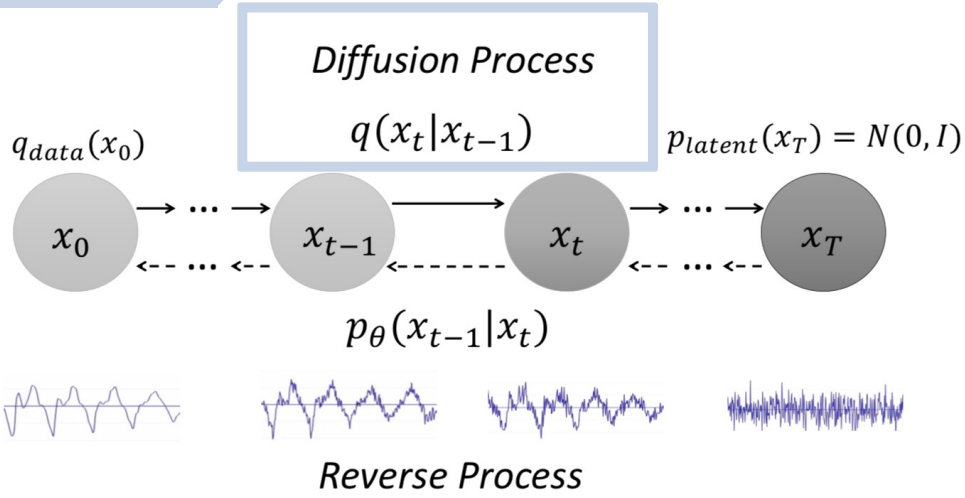
Diffusion Markov process

Combines Gaussian noise into the clean speech x_0.

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

Diffusion step:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$



Diffusion Probabilistic Model

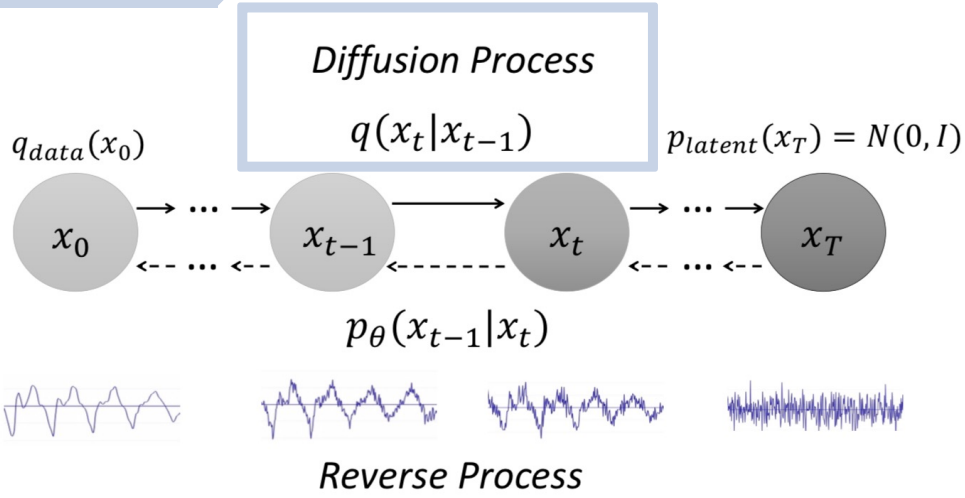
Diffusion Markov process

Combines Gaussian noise into the clean speech x_0.

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

Diffusion step:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$



Marginalized diffusion step:

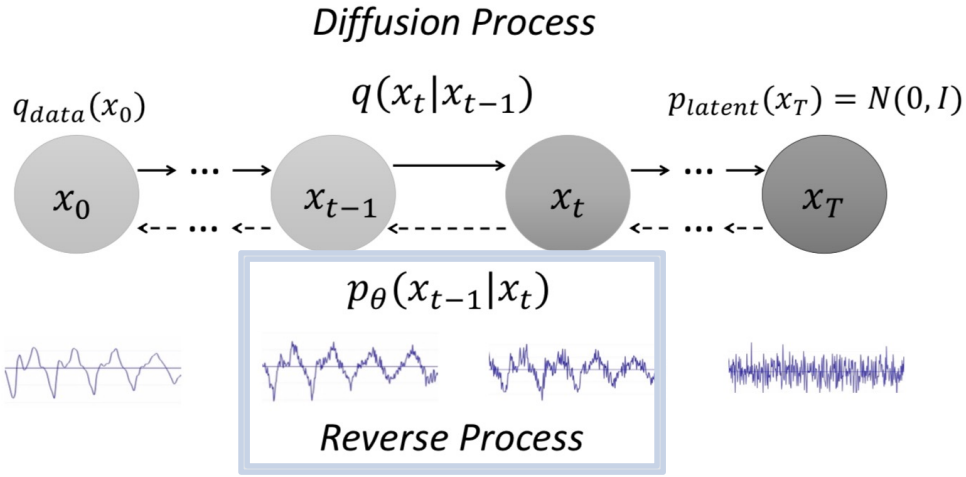
$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \quad \text{where } \alpha_t = 1-\beta_t \text{ and } \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s.$$

Diffusion Probabilistic Model

Reverse Markov process

Estimates the clean speech x_0 from Gaussian noise x_T.

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



Diffusion Probabilistic Model

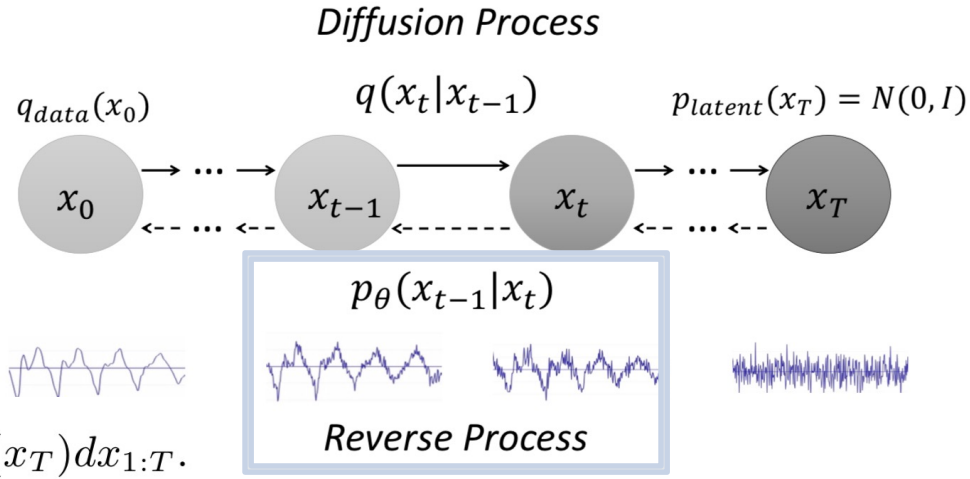
Reverse Markov process

Estimates the clean speech x_0 from Gaussian noise x_T.

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

Intractable marginal likelihood:

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



Diffusion Probabilistic Model

Reverse Markov process

Estimates the clean speech x_0 from Gaussian noise x_T.

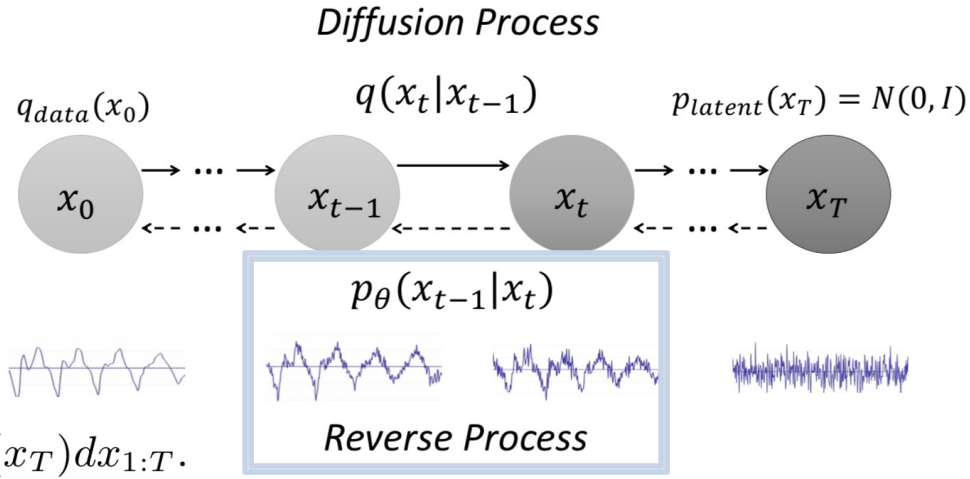
$$p_\theta(x_0, \cdots, x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t),$$

Intractable marginal likelihood:

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

ELBO as Optimization target:

$$-\mathbb{E}_q \left(\text{KL}\left(q(x_T|x_0) \| p_{\text{latent}}(x_T)\right) + \sum_{t=2}^{T} \text{KL}\left(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t)\right) - \log p_\theta(x_0|x_1) \right)$$



Diffusion Probabilistic Model

Diffusion Process (Training)

Training neural network to reduce the loss from ELBO

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

Neural Network

Model ε_θ takes x_t and t as input to estimate noise ε in x_t.

c and κ_t: constants

Diffusion Probabilistic Model

Diffusion Process (Training)

Training neural network to reduce the loss from ELBO

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

Neural Network

Model ε_θ takes x_t and t as input to estimate noise ε in x_t.

c and κ_t: constants

Reverse Process (Sampling)

Predict μ_θ, mean of the previous distribution x_{t-1} by removing ε_θ from the x_t sample.

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

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

θ:the parameters of the neural network.
$\tilde{\beta}_t$: the variance of $q(x_{t-1}|x_t, x_0)$

2

Conditional Diffusion Probabilistic Model

CDPM

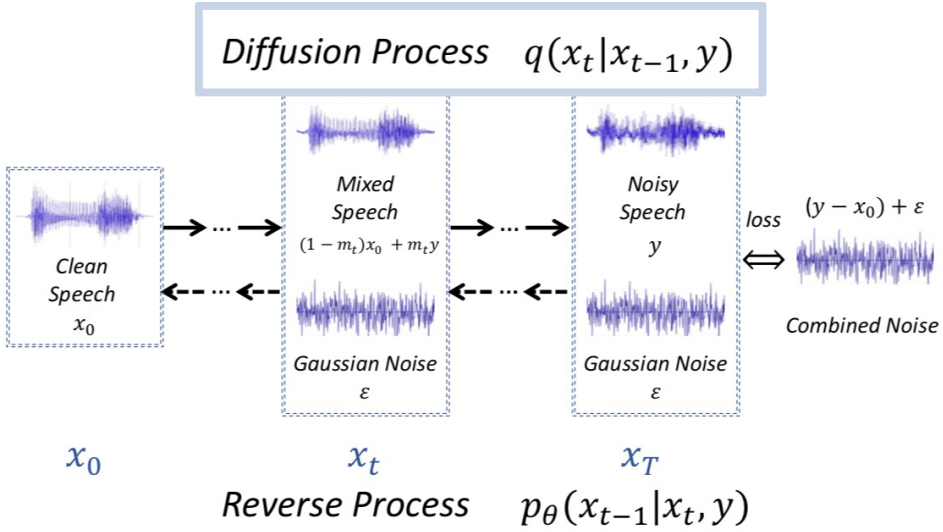
Conditional Diffusion Probabilistic Model

- Vanilla diffusion probabilistic model generates clean speech signal x_0 from Gaussian noise ε.
- Generalized conditional diffusion probabilistic model incorporates the observed noisy data into the model.
- Generate speech signal x_0 from the mixture of Gaussian noise ε and noisy speech y.

Conditional Diffusion Probabilistic Model

Conditional Diffusion Process

Starts from clean speech x_0 and gradually becomes a combination of noisy speech y and Gaussian noise ε.



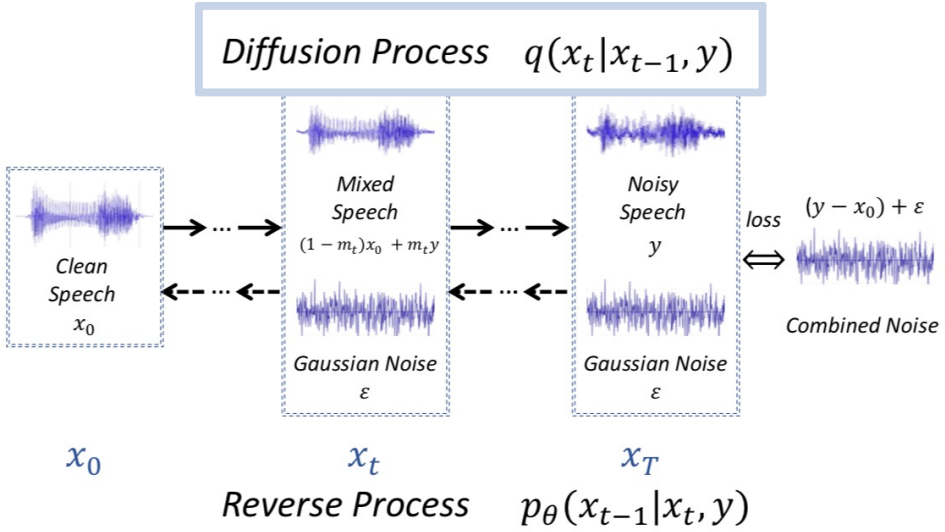
Conditional Diffusion Probabilistic Model

Conditional Diffusion Process

Starts from clean speech x_0 and gradually becomes a combination of noisy speech y and Gaussian noise ε.

$$q_{\text{cdiff}}(x_t|x_0, \text{y}) =$$

$$\mathcal{N}\left(x_t; (1-m_t)\sqrt{\bar{\alpha}_t}x_0 + m_t\sqrt{\bar{\alpha}_t}y, \delta_t I\right),$$



Conditional Diffusion Probabilistic Model

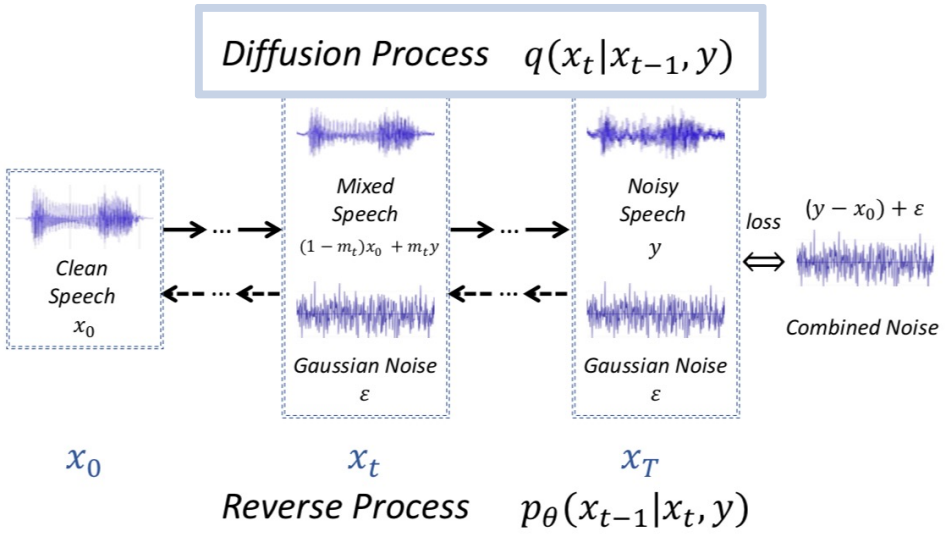
Conditional Diffusion Process

Starts from clean speech x_0 and gradually becomes a combination of noisy speech y and Gaussian noise ε.

$$q_{\text{cdiff}}(x_t|x_0, \text{y}) =$$

$$\mathcal{N}\left(x_t; (1-m_t)\sqrt{\bar{\alpha}_t}x_0 + m_t\sqrt{\bar{\alpha}_t}y, \delta_t I\right),$$

Interpolation



δ_t: the variance of x_t given x_0 and y.

m_t starts from $m_0 = 0$ and is gradually increased to $m_T \approx 1$

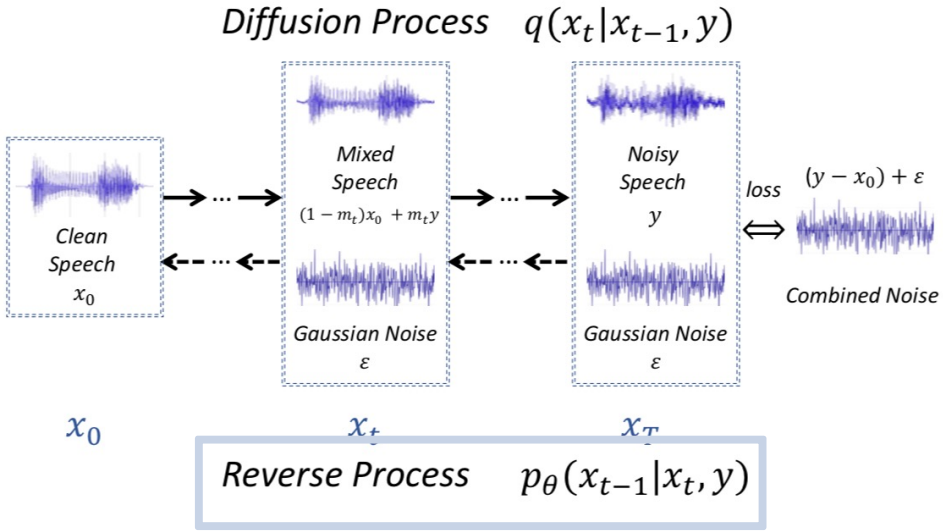
x_t starts from x_0 (clean speech) and gradually turns into $\mathcal{N}(x_T; \sqrt{\bar{\alpha}_T}y, \delta_T I)$

Conditional Diffusion Probabilistic Model

Conditional Reverse Process

Estimates x_0 from x_T, noisy speech y with variance δ_T:

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



Conditional Diffusion Probabilistic Model

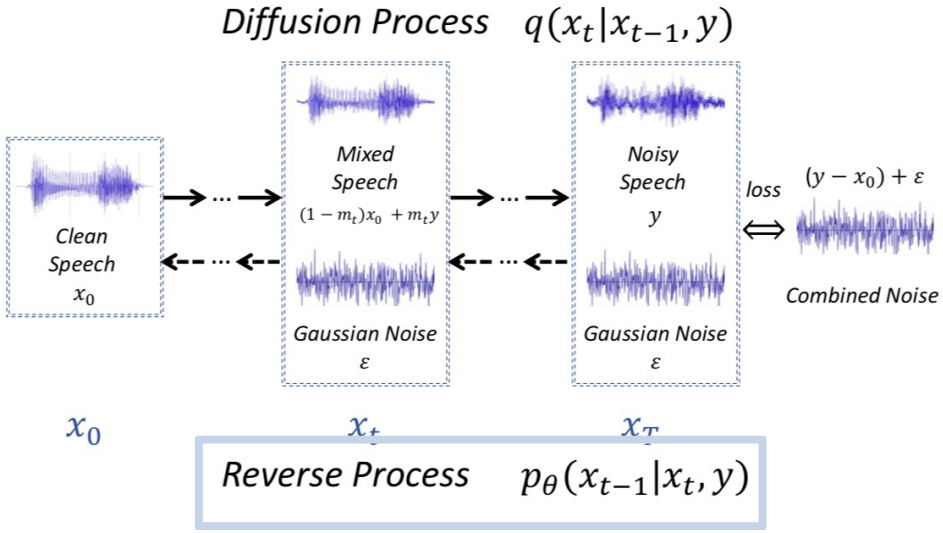
Conditional Reverse Process

Estimates x_0 from x_T, noisy speech y with variance δ_T:

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

Estimates the mean of x_{t-1} from x_t, noisy speech y, and predicted noise ε_θ.

$$p_{\text{cdiff}}(x_{t-1}|x_t, \text{y}) = \mathcal{N}(x_{t-1}; c_{xt}x_t + c_{yt}y - c_{\varepsilon t}\varepsilon_\theta(x_t, y, t), \tilde{\delta}_t I)$$



Conditional Diffusion Probabilistic Model

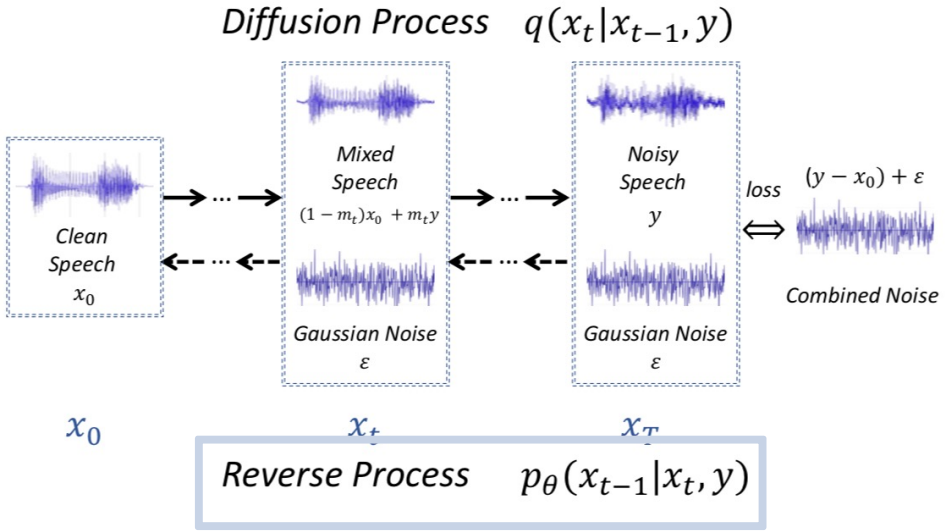
Conditional Reverse Process

Estimates x_0 from x_T, noisy speech y with variance δ_T:

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

Estimates the mean of x_{t-1} from x_t, noisy speech y, and predicted noise ε_θ.

$$p_{\text{cdiff}}(x_{t-1}|x_t, \text{y}) = \mathcal{N}(x_{t-1}; c_{xt}x_t + c_{yt}y - c_{\epsilon t}\varepsilon_\theta(x_t, y, t), \tilde{\delta}_t I)$$



To find the c_{xt}, c_{yt}, and $c_{\epsilon t}$, which bring to the lowest ELBO for the conditional diffusion probabilistic model.

Conditional Diffusion Probabilistic Model

Optimization

To optimize the likelihood of $p_\theta(x_0)$

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

$p_\theta(x_{t-1}|x_t,y)$: Sampling Process

$q_{\mathrm{cdiff}}(x_{t-1}|x_t,x_0,y)$: Diffusion Process (opposite direction)

Conditional Diffusion Probabilistic Model

Optimization

To optimize the likelihood of $p_\theta(x_0)$

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

$p_\theta(x_{t-1}|x_t,y)$: Sampling Process

$q_{\mathrm{cdiff}}(x_{t-1}|x_t,x_0,y)$: Diffusion Process (opposite direction)

Bayes' theorem:

$$q_{\mathrm{cdiff}}(x_{t-1}|x_t,x_0,y) = \frac{q_{\mathrm{cdiff}}(x_t|x_{t-1},x_0,y)\times q_{\mathrm{cdiff}}(x_{t-1}|x_0,y)}{q_{\mathrm{cdiff}}(x_t|x_0,y)}$$

Conditional Diffusion Probabilistic Model

Conditional Diffusion Process

Training neural network to reduce the loss from ELBO

Difference between signals **Gaussian noise**

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

Model ε_θ estimates the combination of the signal difference $(y - x_0)$ and Gaussian noise ε in x_t mixture.

Conditional Diffusion Probabilistic Model

Conditional Diffusion Process

Training neural network to reduce the loss from ELBO

Difference between signals **Gaussian noise**

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

Model ε_θ estimates the combination of the signal difference $(y - x_0)$ and Gaussian noise ε in x_t mixture.

c' and κ_t': constants

$\widetilde{\delta_t}$: the variance of $q_{\text{cdiff}}(x_{t-1}|x_t, x_0, y)$

Conditional Reverse Process

Predict the mean of previous distribution x_{t-1} by combining y and removing ε_θ from the x_t sample.

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

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

where

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

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

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

3

Experiment Results

Matched and Mismatched Condition

Experiment Results

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**

- CDiffuSE shows improved performance on all the metrics over the diffusion probabilistic model baseline DiffuSE.

Experiment Results

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**

- CDiffuSE shows improved performance on all the metrics over the diffusion probabilistic model baseline DiffuSE.
- CDiffuSE outperforms its competitors on all metrics with the exception of CBAK and achieves a particularly significant improvement in PESQ

Experiment Results

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/facebook-researh/denoiser to test performance, and we also copied the Demucs results in [16] to the right-side of "/" in Table 3(a).

- Generative speech enhancement models are still lagging behind the performance of their regressive counterparts.
- Given a domain shift in test data, regression based approaches suffer from a significant drop in performance.

Experiment Results

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/facebook-researh/denoiser to test performance, and we also copied the Demucs results in [16] to the right-side of "/" in Table 3(a).

- Generative speech enhancement models are still lagging behind the performance of their regressive counterparts.
- Given a domain shift in test data, regression based approaches suffer from a significant drop in performance.
- CDiffuSE degrades to a much smaller degree than its regressive competitors, demonstrates its high robustness to variation in noise characteristics.

Conclusions

- We proposed conditional diffusion probabilistic model that can explore noise characteristics from the noisy input signal explicitly 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.
- Our method has great generalization capabilities to speech data with noise characteristics not observed in the training data.

GitHub: https://github.com/neillu23/CDiffuSE