A Study on Speech Enhancement Based on Diffusion Probabilistic Model

Yen-Ju Lu, Yu Tsao, Shinji Watanabe

Outline

- Introduction
- Diffusion Model
- Experiments
- Conclusion

Introduction

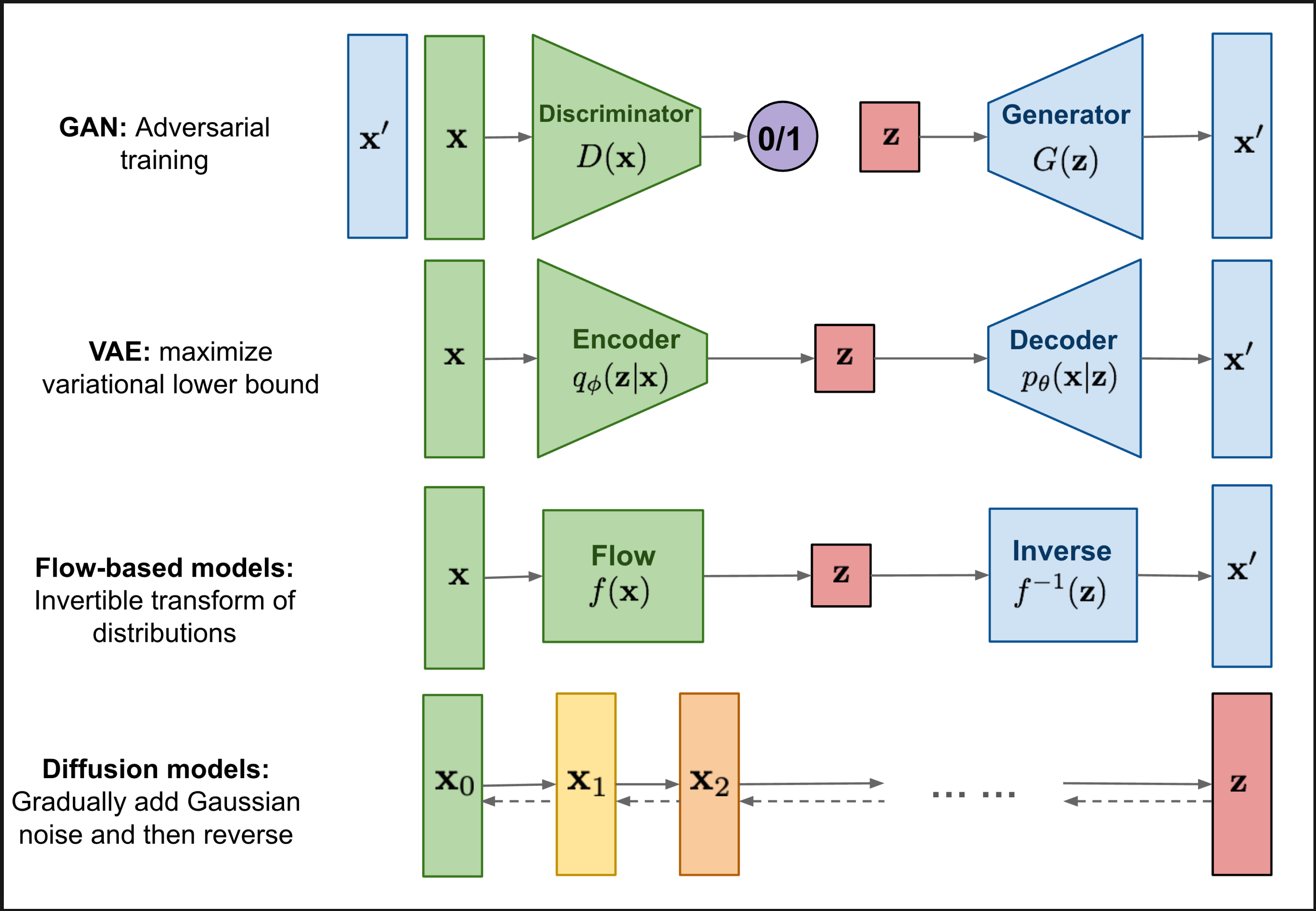
Diffusion Model is a novel generation method that has achieved good results in both image and speech generation tasks.

Based on this, DiffWave has become the state of the art in speech synthesis with only a few parameters.

Introduction

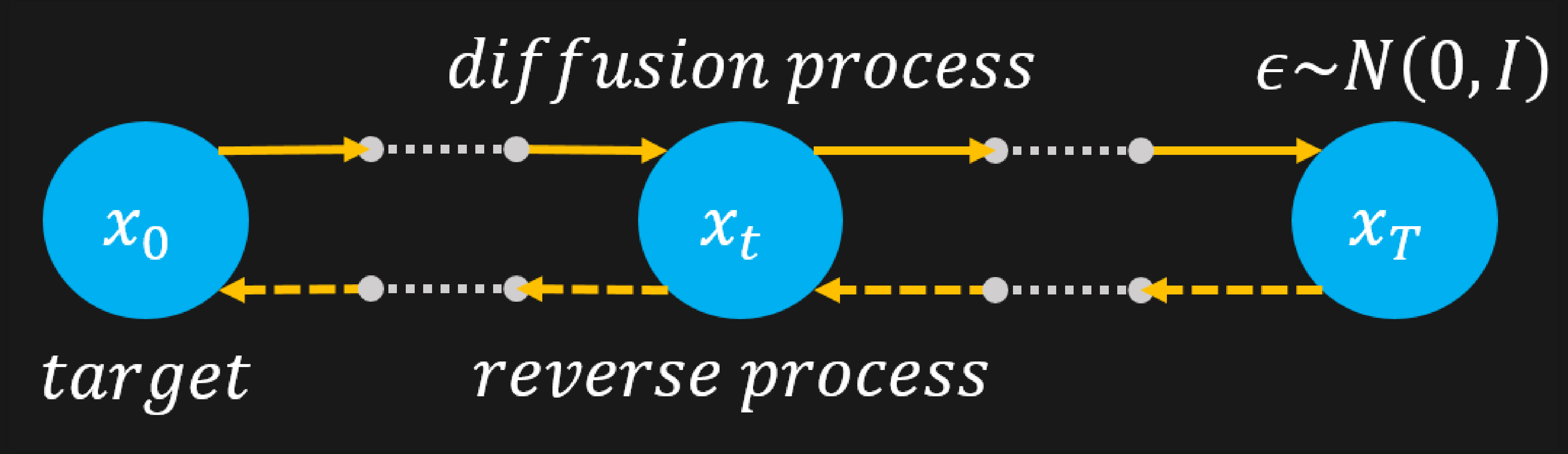
This paper attempts to apply DiffWave to Speech Enhancement and proposes a Supportive Reverse Process (SRP) specifically designed for this task to replace the original Reverse Process (RP).

Diffusion Model


GAN: Adversarial training
x'
x
Discriminator
D(x)
0/1
z
Generator
G(z)
x'
VAE: maximize variational lower bound
x
Encoder
qϕ(z|x)
z
Decoder
pθ(x|z)
x'
Flow-based models: Invertible transform of distributions
x
Flow
f(x)
z
Inverse
f⁻¹(z)
x'
Diffusion models: Gradually add Gaussian noise and then reverse
x0
x1
x2
z

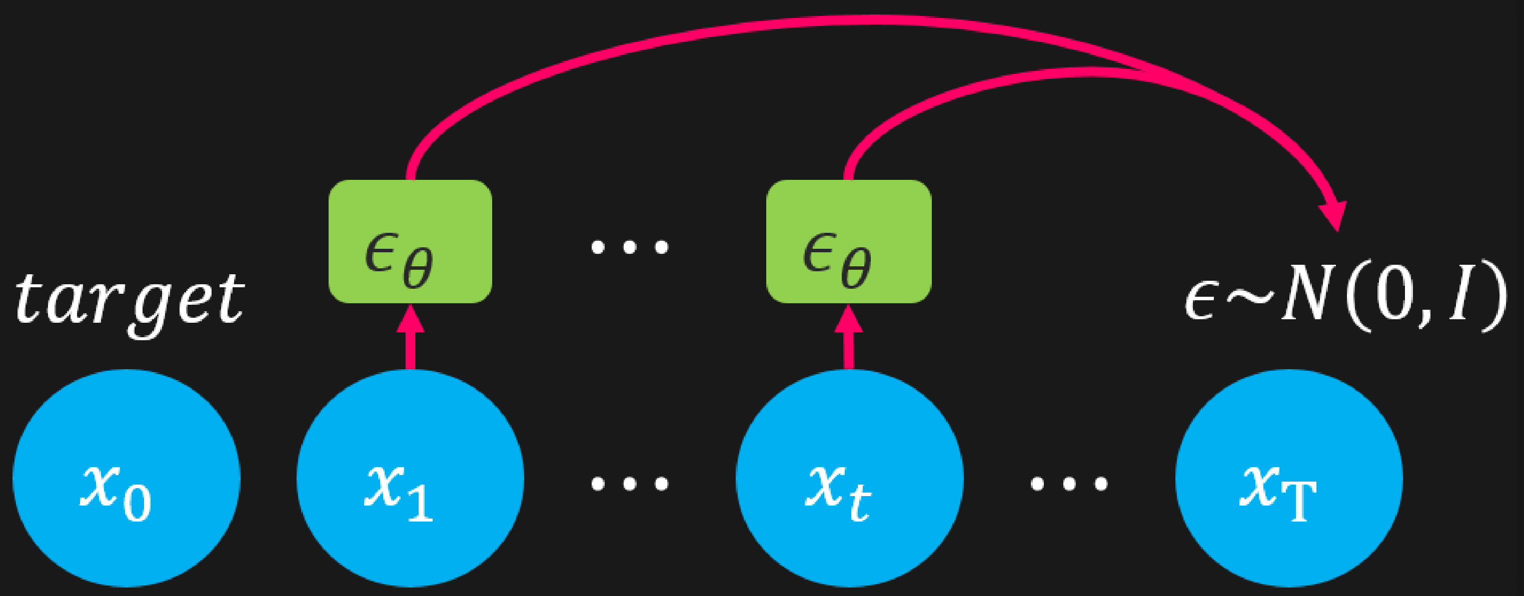
Params Equation

Param	Eq
β_t	$\{\beta_t\}_{t=1}^{T}$
α_t	$1 - \beta_t$
$\bar{\alpha}_t$	$\prod_{s=1}^{t} \alpha_s$
σ_t^2	$\frac{(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \beta_t$
γ_t	$\frac{\sigma_t}{\sqrt{\bar{\alpha}_{t-1}}}$



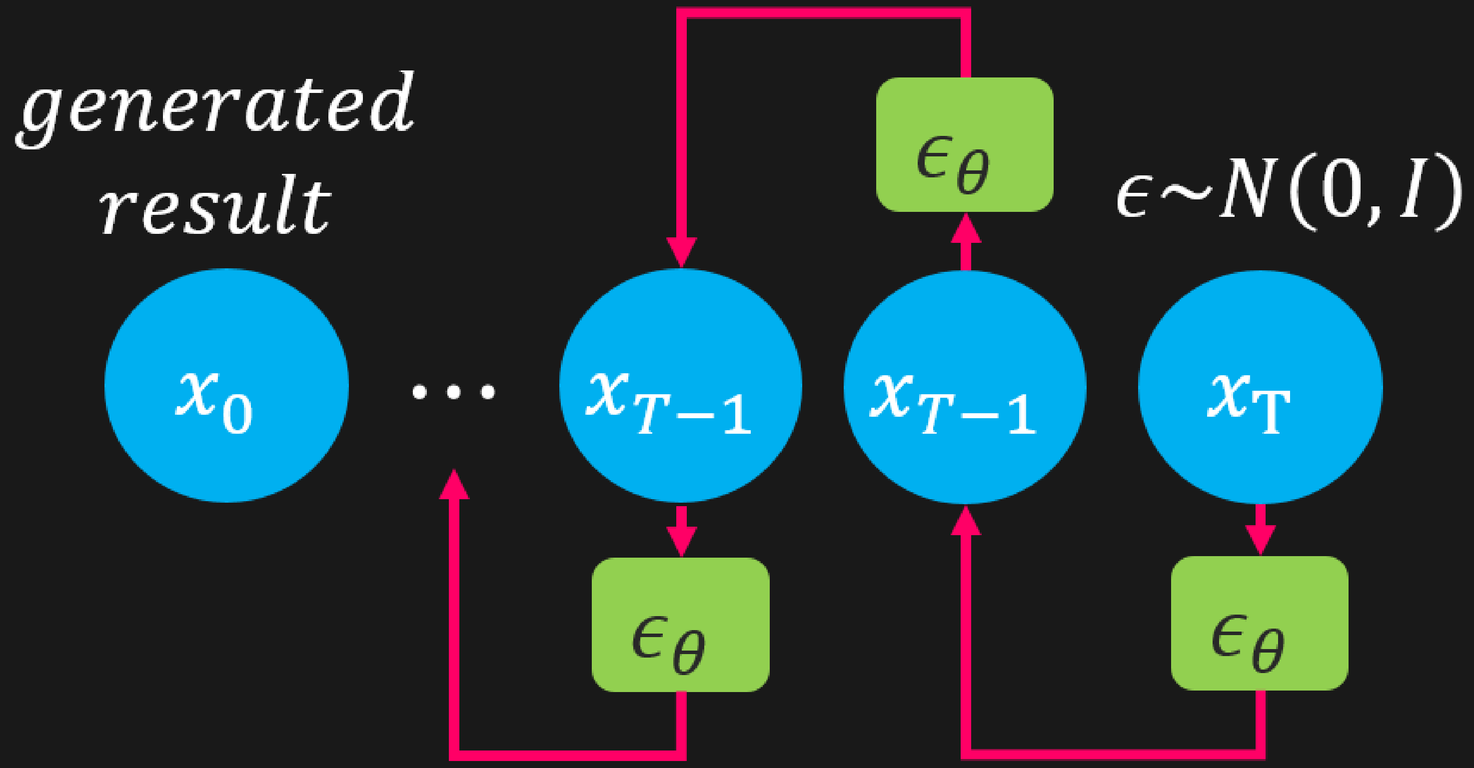
$$x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

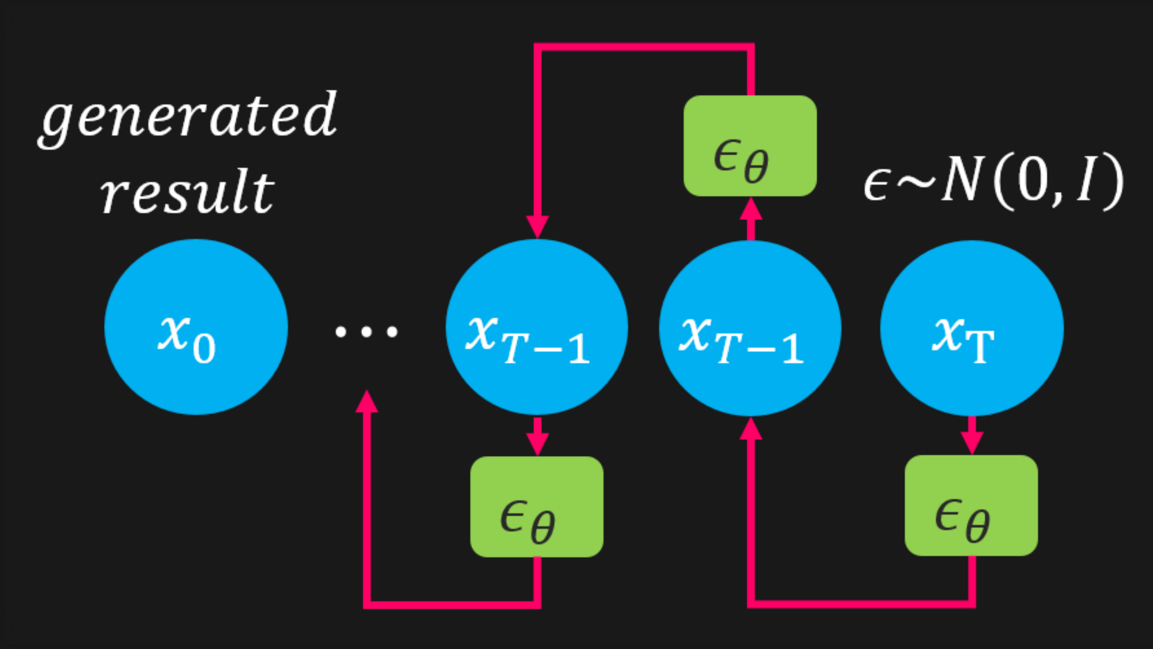
Train



$$\nabla_\theta \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2$$

Reverse Process





$$\begin{aligned} x_{t-1} &= \mu_\theta(x_t, t) + \sigma_t z \\ &= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z \end{aligned}$$

$$
\begin{aligned}
\frac{1}{\sqrt{\alpha_t}} x_t &= \frac{\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon}{\sqrt{\alpha_t}} \\
&= \sqrt{\bar{\alpha}_{t-1}} x_0 + \frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\alpha_t}} \epsilon
\end{aligned}
$$

$$let\ \epsilon_\theta(x_t, t) = \epsilon$$

$$\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}}\epsilon - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon = \frac{\alpha_t - \bar{\alpha}_t}{\sqrt{a_t}\sqrt{1-\bar{a}_t}}\epsilon$$

$$= \tilde{\sigma}\epsilon$$

$$\sigma_t^2 = \frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t} \ for\, t > 1 \ and\, \sigma_1^2 = 1 - \alpha_1$$

$$\tilde{\sigma}\epsilon + \sigma_t z \sim N(0, \tilde{\sigma}^2 + \sigma_t^2)$$

（統計獨立的常態隨機變數相加）

$$\text{且}\ \tilde{\sigma}^2 + \sigma_t^2 = 1 - \bar{\alpha}_{t-1}$$

$$\rightarrow \tilde{\sigma}\epsilon + \sigma_t z \sim N(0, 1 - \bar{\alpha}_{t-1})$$

$$
\begin{aligned}
x_{t-1} &= \mu_\theta(x_t, t) + \sigma_t z \\
&= \sqrt{\bar{\alpha}_{t-1}} x_0 + \tilde{\sigma}\epsilon + \sigma_t z \\
&= \sqrt{\bar{\alpha}_{t-1}} x_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon'
\end{aligned}
$$

$$
\epsilon' \sim N(0, I)
$$

DiffSE Model

$$\epsilon_\theta(x_t, t) \rightarrow \epsilon_\theta(x_t, t, \mathbf{condition})$$

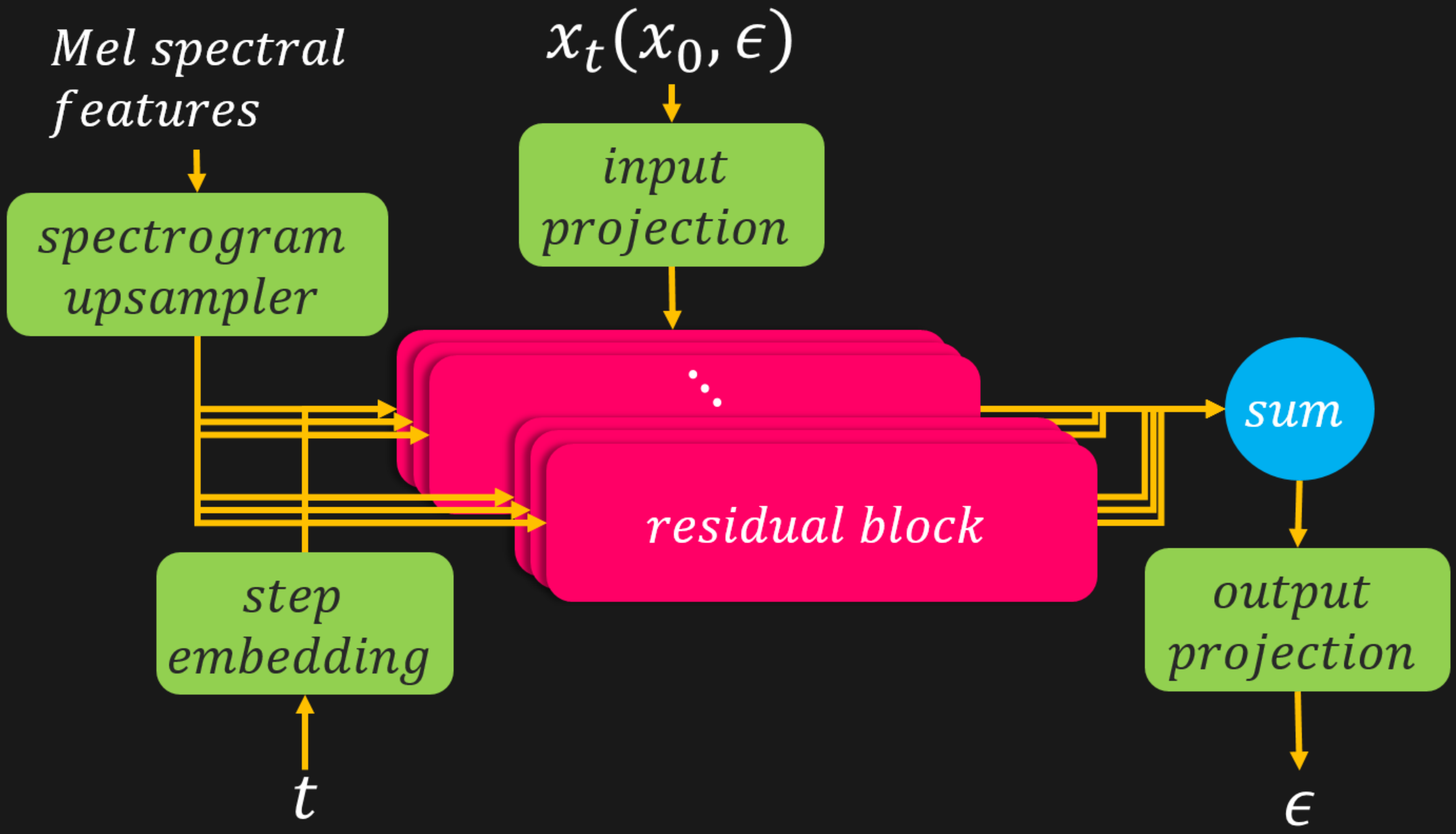
$$condition =$$

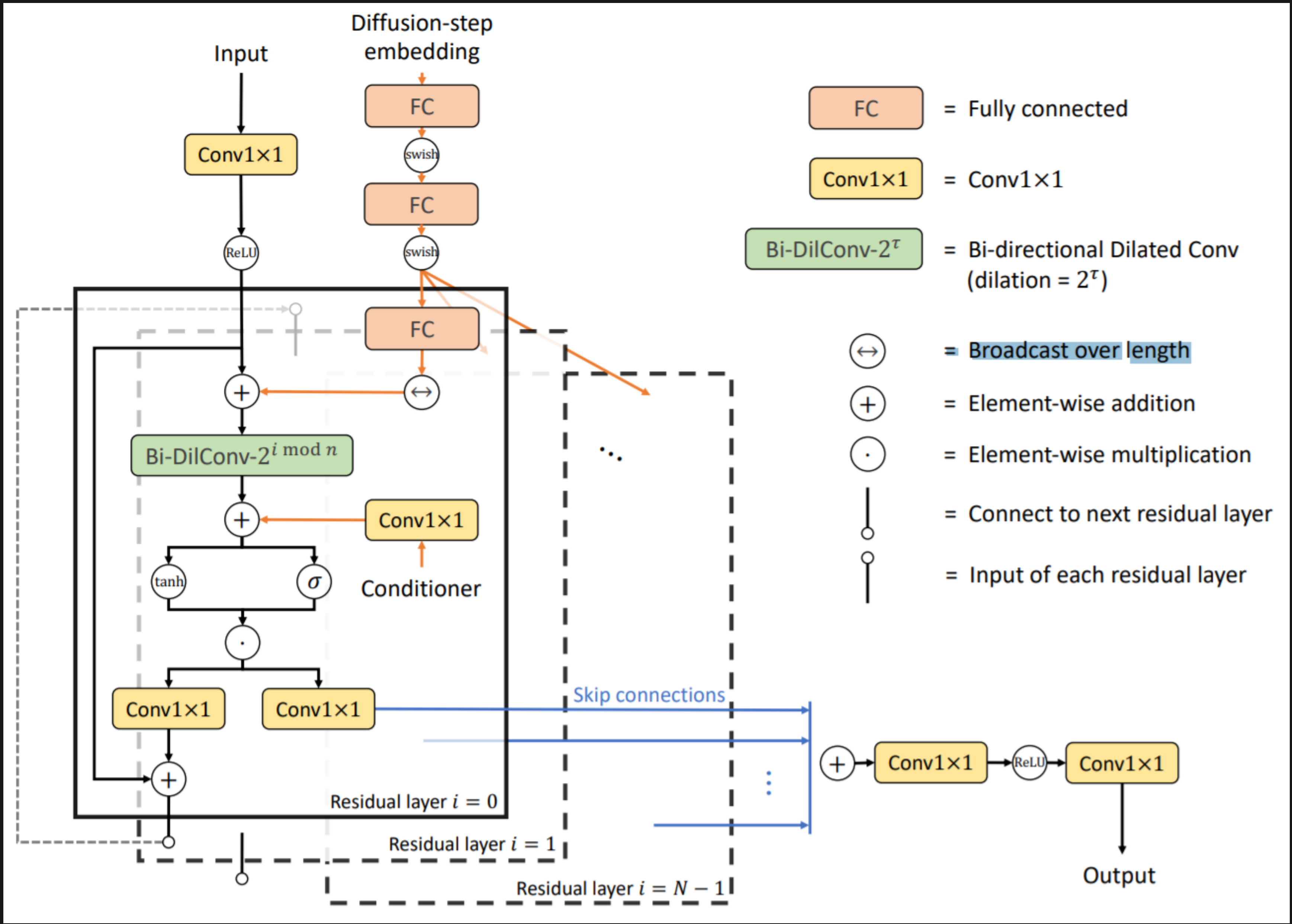
$$if\ pretrain:$$

$$Clean\ Mel\ Spectrogram$$

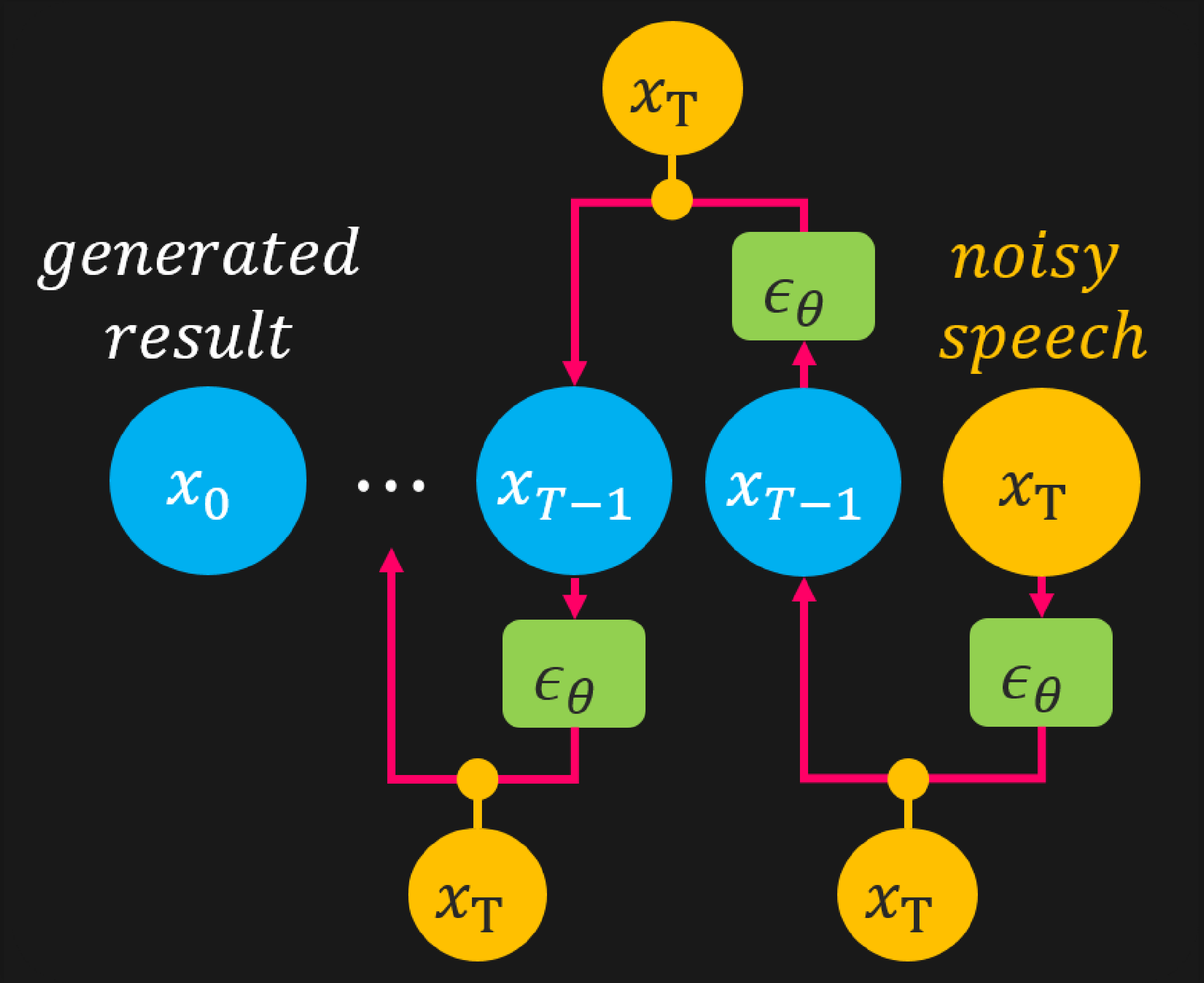
$$else:$$

$$Noisy\ Mel\ Spectrogram$$


Mel spectral features
$x_t(x_0, \epsilon)$
input projection
spectrogram upsampler
sum
residual block
step embedding
output projection
t
ϵ


Input
Diffusion-step embedding
FC
swish
FC
swish
Conv1×1
ReLU
FC
Bi-DilConv-$2^{i \bmod n}$
Conv1×1
Conditioner
tanh
σ
Conv1×1
Conv1×1
Residual layer $i = 0$
Residual layer $i = 1$
Residual layer $i = N - 1$
Skip connections
Conv1×1
ReLU
Conv1×1
Output
FC = Fully connected
Conv1×1 = Conv1×1
Bi-DilConv-2^{τ} = Bi-directional Dilated Conv (dilation = 2^{τ})
= Broadcast over length
= Element-wise addition
= Element-wise multiplication
= Connect to next residual layer
= Input of each residual layer

Supportive Reverse Process



$$x_T = y = noisy\ speech$$

$$\hat{\mu}_\theta(x_t, t) = (1 - \gamma_t)\mu_\theta(x_t, t) + \gamma_t\sqrt{\bar{\alpha}_{t-1}}y$$

$$\hat{\sigma}_t = max(\sigma_t - \gamma_t\sqrt{\bar{\alpha}_{t-1}}, 0)$$

$$x_{t-1} = \hat{\mu}_\theta(x_t, t) + \hat{\sigma}_t z$$

Experiments

VoiceBank DEMAND Dataset

	Train	Test
Speaker	28	2
SNR	0、5、10、15 dB	2.5、7.5、12.5、17.5 dB
Sampling Rate	16k Hz	

Hyper Params

Param	**Value**
T_{Base}	50
T_{Large}	200
$\beta_t(base)$	$1 \times 10^{-4}\ to\ 0.05$
$\beta_t(large)$	$1 \times 10^{-4}\ to\ 0.02$
$\beta_t(fast\&base)$	$[1e-4, 1e-3, 1e-2, 0.05, 0.2, 0.5]$
$\beta_t(fast\&large)$	$[1e-4, 1e-3, 1e-2, 0.05, 0.2, 0.7]$

Results

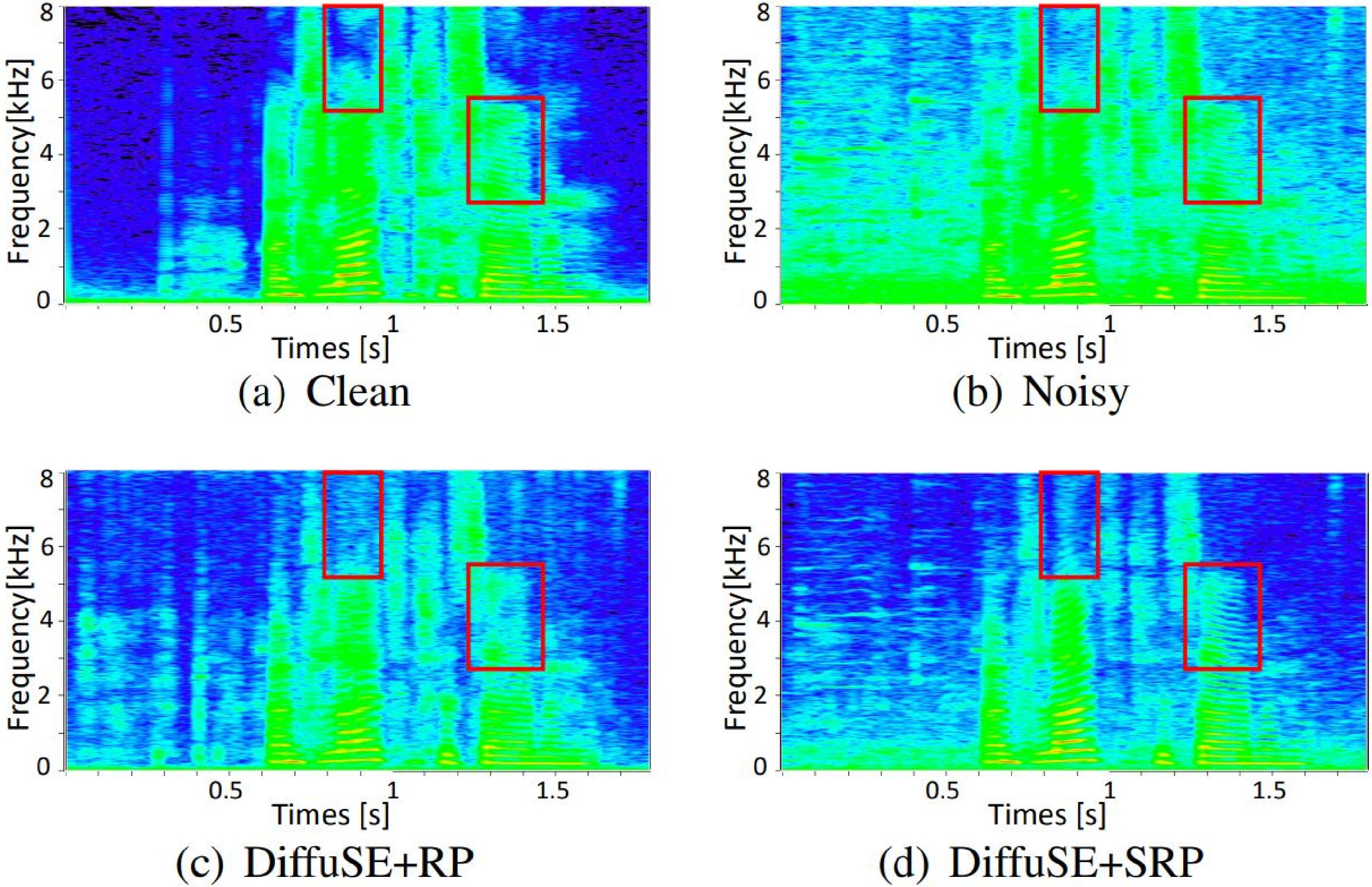
- RP：Reverse Process
- $RP\text{-}N_{in}$：使用 Noisy Speech 而非 Gaussian Noise 作為輸入
- $RP\text{-}N_{out}$：將 RP 生成的輸出與 Noisy Speech 以 4:1 的比例混和
- $RP\text{-}N_{in+out}$：$RP\text{-}N_{in}$ 與 $RP\text{-}N_{out}$ 一同使用
- SRP：Supportive Reverse Process

Base DiffuSE	Schedule	PESQ	CSIG	CBAK	COVL
Noisy	-	1.97	3.35	2.44	2.63
PR	Fast	1.96	3.13	2.22	2.52
	Full	1.97	3.21	2.22	2.57
PR-N_{in}	Fast	2.07	3.21	2.57	2.62
	Full	2.05	3.27	2.48	2.64
PR-N_{out}	Fast	2.05	3.31	2.21	2.64
	Full	2.12	3.38	2.25	2.72
PR-N_{in+out}	Fast	2.29	3.47	2.67	2.85
	Full	2.31	3.51	2.61	2.88
SRP	Fast	**2.41**	**3.61**	**2.82**	**2.99**
	Full	2.39	3.60	2.79	2.97

Large DiffuSE	Schedule	PESQ	CSIG	CBAK	COVL
Noisy	-	1.97	3.35	2.44	2.63
PR	Fast	2.09	3.29	2.31	2.67
	Full	2.16	3.39	2.31	2.75
PR-N_{in}	Fast	2.18	3.35	2.60	2.74
	Full	2.20	3.42	2.48	2.78
PR-N_{out}	Fast	2.16	3.42	2.30	2.76
	Full	2.17	3.45	2.29	2.78
PR-N_{in+out}	Fast	2.37	3.56	2.69	2.94
	Full	2.33	3.55	2.56	2.91
SRP	Fast	**2.43**	**3.63**	**2.81**	**3.00**
	Full	2.39	3.63	2.75	2.99

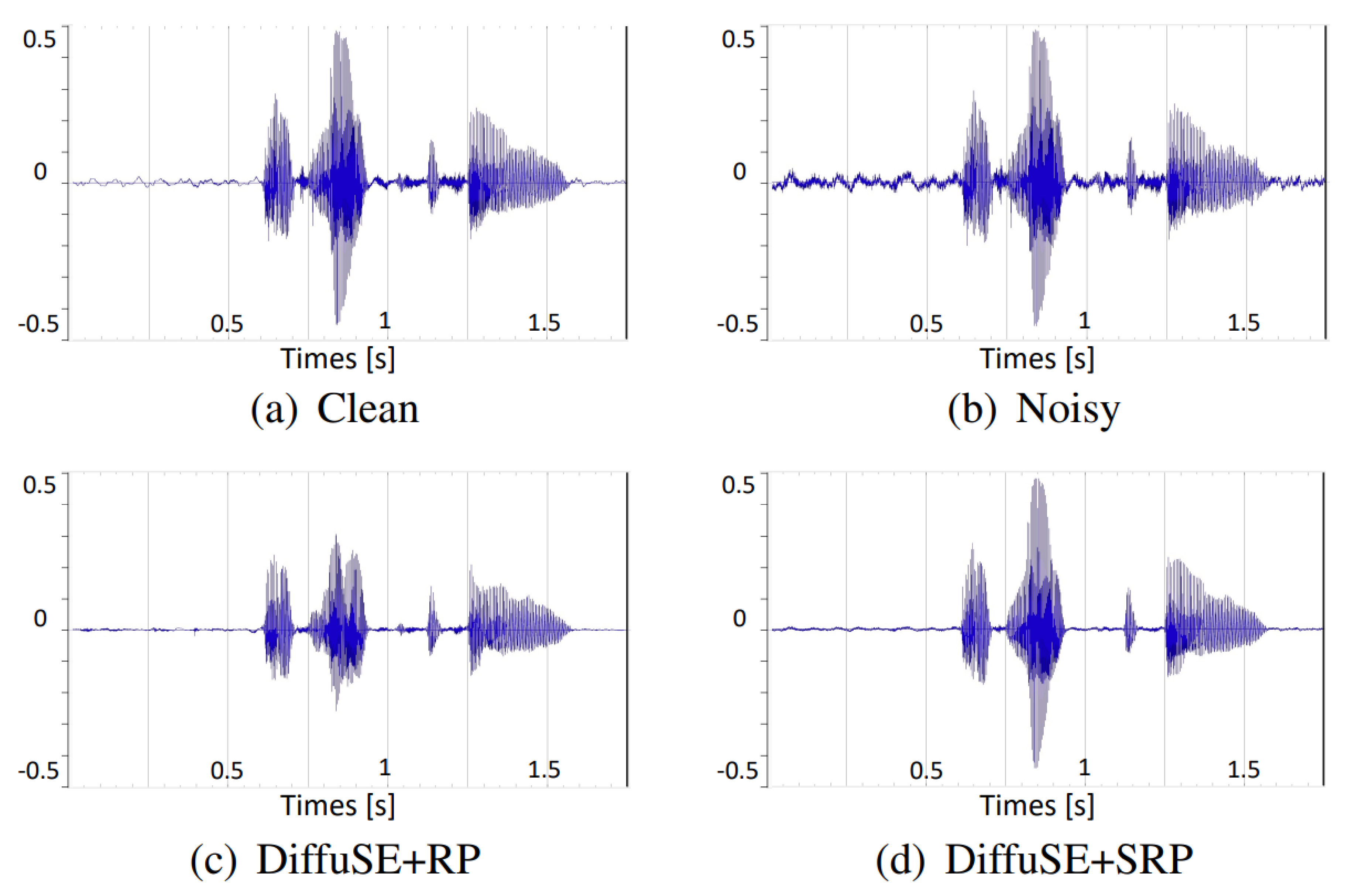
vs Time Domain SOTA

Method	**PESQ**	**CSIG**	**CBAK**	**COVL**
Noisy	1.97	3.35	2.44	2.63
SEGAN	2.16	3.48	2.94	2.80
DSEGAN	2.39	3.46	**3.11**	**3.50**
SE-Flow	2.28	**3.70**	3.03	2.97
DiffuSE(Base)	2.41	3.61	2.82	2.99
DiffuSE(Large)	**2.43**	3.63	2.81	3.00



(a) Clean (b) Noisy

(c) DiffuSE+RP (d) DiffuSE+SRP



(a) Clean (b) Noisy

(c) DiffuSE+RP (d) DiffuSE+SRP

Conclusion

- SRP gets better results than RP by adding (Noisy) Speech information in the reverse process.
- In the reverse process, only a few key steps need to be performed to get good results.
- Can real-world noise be used instead of gaussian noise for training during the diffusion process?