Diffusion-based speech enhancement with a weighted generative-supervised learning loss

Jean Eudes Ayilo, Mostafa Sadeghi, Romain Serizel

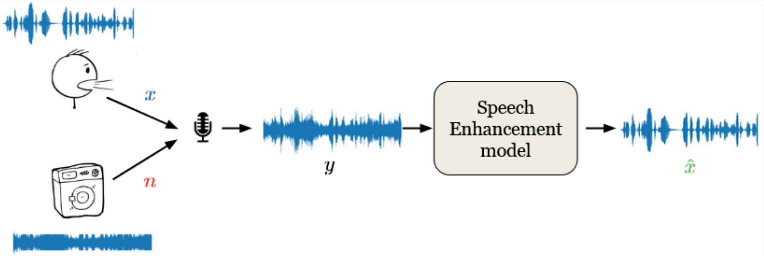
Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)

April 14-19, Seoul, Republic of Korea

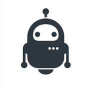
Introduction

Speech Enhancement (SE)



Given noisy speech observation $y = x + n$ in time domain (resp. $\mathbf{y} = \mathbf{x} + \mathbf{n}$ in time-frequency domain), estimate the clean speech signal x (resp $\mathbf{x}$).

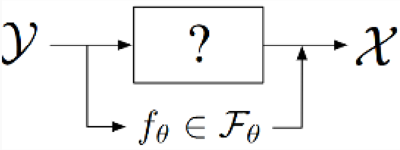
Various applications:



SE approaches

Data-driven approaches based on DNNs:

- **Predictive approach**: learn a mapping function between pairs of noisy ($\mathcal{Y}$) and clean ($\mathcal{X}$) speech signals



  - ▷ good performance on seen noises
  - ▷ need large dataset to achieve better generalization on unseen noises

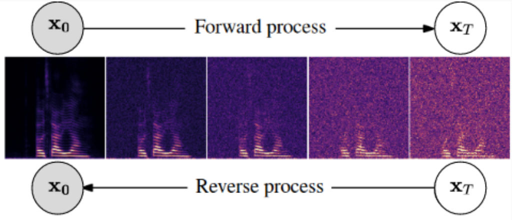
- **Generative approach** (and recently diffusion models) : model (conditional/unconditional) clean speech distribution and at inference, sample from the posterior distribution

Score-based generative model for SE

Observed mixture (in Short Time Fourier Transform):

$\mathbf{y} = \mathbf{x}_0 + \mathbf{n}$ where $\mathbf{x}_0$, $\mathbf{y}$, $\mathbf{n} \in \mathbb{C}^d$

Score-based generative model for SE (SGMSE+) in Richter et al. (2023)[1] i.e.



Richter et al. (2023)

[1] Richter, Julius, et al. "Speech enhancement and dereverberation with diffusion-based generative models." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)

Score-based generative model for SE

- Forward process: $\mathrm{d}\mathbf{x}_t = \gamma\left(\mathbf{y} - \mathbf{x}_t\right)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$

Score-based generative model for SE

- Forward process: $\mathrm{d}\mathbf{x}_t = \gamma\left(\mathbf{y} - \mathbf{x}_t\right)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$

  ▷ Solution to the forward SDE: Gaussian process $\{\mathbf{x}_t\}_{t=1}^{T}$

  Thanks to its transition kernel, sample any $\mathbf{x}_t$ following:

$$\mathbf{x}_t = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + \left(1 - \mathrm{e}^{-\gamma t}\right)\mathbf{y} + \mathbf{e}_t \tag{1}$$

$$\text{where } \mathbf{e}_t \sim \mathcal{N}_{\mathbb{C}}(\mathbf{e}_t; \mathbf{0}, \sigma(t)^2\mathbf{I})$$

Score-based generative model for SE

- Forward process: $\mathrm{d}\mathbf{x}_t = \gamma\left(\mathbf{y} - \mathbf{x}_t\right)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$

  ▷ Solution to the forward SDE: Gaussian process $\{\mathbf{x}_t\}_{t=1}^{T}$

  Thanks to its transition kernel, sample any $\mathbf{x}_t$ following:

$$\mathbf{x}_t = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + \left(1 - \mathrm{e}^{-\gamma t}\right)\mathbf{y} + \mathbf{e}_t \tag{1}$$

where $\mathbf{e}_t \sim \mathcal{N}_{\mathbb{C}}(\mathbf{e}_t; \mathbf{0}, \sigma(t)^2\mathbf{I})$

- Reverse process: $\mathrm{d}\mathbf{x}_t = \left[-\gamma\left(\mathbf{y} - \mathbf{x}_t\right) + g(t)^2 \underbrace{\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{y})}_{\text{score function}}\right]\mathrm{d}t + g(t)\mathrm{d}\overline{\mathbf{w}}_t$

Need to approximate the intractable score function: $\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{y})$

Score-based generative model for SE

- Learn a score network, by minimizing a noise-prediction loss :

$$\min_{\theta} \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z}\sim\mathcal{N}_{\mathbb{C}}(\mathbf{z};\mathbf{0},\mathbf{I}),\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}\left[\underbrace{\left\|\sigma(t)\mathbf{s}_\theta\left(\mathbf{x}_t,\mathbf{y},t\right)+\mathbf{z}\right\|^2}_{:=L_\theta(\mathbf{x}_t,\mathbf{y},t,\mathbf{z})}\right] \tag{2}$$

- Perform SE, by finding numerical solutions for the plug-in reverse SDE:

$$\mathrm{d}\mathbf{x}_t=\left[-\gamma\left(\mathbf{y}-\mathbf{x}_t\right)+g(t)^2\mathbf{s}_\theta\left(\mathbf{x}_t,\mathbf{y},t\right)\right]\mathrm{d}t+g(t)\mathrm{d}\overline{\mathbf{w}}_t$$

Remark: Contrary to supervision loss, there is no comparison of the generated enhanced speech signals against the ground-truths.

Weighted generative-supervised learning loss

Proposed solution: account for the goodness of fit of the generated speech via an ℓ_2-loss between the ground-truth and an estimate $\hat{\mathbf{x}}_{0,t}$.

- Apply Tweedie's formula[2] [3] to $\mathbf{x}_t$ (eq. 1) and get $\mathbb{E}(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}) = \hat{\mathbf{x}}_{0,t}$

$$\mathrm{e}^{-\gamma t}\hat{\mathbf{x}}_{0,t} + (1 - \mathrm{e}^{-\gamma t})\mathbf{y} \approx \mathbf{x}_t + \frac{\sigma(t)^2}{2}\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) \tag{3}$$

- Taking the ℓ_2 distance between the estimate $\hat{\mathbf{x}}_{0,t}$ and the ground-truth $\mathbf{x}_0$, the new training objective is set to :

$$\min_\theta \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z}\sim\mathcal{N}_{\mathbb{C}}(\mathbf{0},\mathbf{I}),\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}\left[(1-\alpha_t)L_\theta\left(\mathbf{x}_t, \mathbf{y}, t, \mathbf{z}\right) + \alpha_t \left\|\hat{\mathbf{x}}_{0,t} - \mathbf{x}_0\right\|^2\right] \tag{4}$$

---

[2] See B. Efron, "Tweedie's formula and selection bias," Journal of the American Statistical Association, vol. 106, no. 496, pp. 1602–1614, 2011

[3] Chung, Hyungjin, et al. "Diffusion posterior sampling for general noisy inverse problems." arXiv preprint arXiv:2209.14687 (2022)

Weighted generative-supervised learning loss

$$\min_{\theta} \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z}\sim\mathcal{N}_{\mathbb{C}}(\mathbf{z};\mathbf{0},\mathbf{I}),\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}\left[(1-\alpha_t)L_\theta\left(\mathbf{x}_t,\mathbf{y},t,\mathbf{z}\right)+\alpha_t\left\|\hat{\mathbf{x}}_{0,t}-\mathbf{x}_0\right\|^2\right]$$

In this new proposed objective, α_t is set to :

$$\alpha_t = \frac{\sigma(T)-\sigma(t)}{\sigma(T)-\sigma(t_\varepsilon)} \tag{5}$$

- trade-off between the generative loss and the supervised loss
- when $\sigma(t) \nearrow$, $\alpha_t \searrow$ and vice-versa

Experiments

Model architecture and baselines

- Same architecture as the Noise Conditional Score Network (NCSN $++$) used in Richter et al. (2023) i.e. SGMSE $+$)

- Variants of the model in this paper:
  - ▷ NCSN $++$ trained with our proposed loss function
  - ▷ NCSN $++$ trained with the generative loss function only (SGMSE$++$) **(baseline)**
  - ▷ Supervised version trained with MSE loss **(baseline)**

Datasets

Training and test sets

Clean speech dataset	Training noise dataset	Test noise dataset	Total [h] (Train/Test)	SNRs in test [dB]	Noise types in test
NTCD-TIMIT[4]	DEMAND	NTCD-TIMIT	17.15 / 1.18	-5,0,5	(street, living room, cafe, car), white, babble
WSJ0	QUT-Noise	QUT-Noise	29.10 / 1.48	-5,0,5	(street, living room, cafe, car)

Cross data set evaluation

- **Matched**: Train and Test clean speech signals come from the same corpus
- **Mismatched**: Train and Test clean speech signals come from different corpus

[4] A.-H. Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," INTERSPEECH, 2017.

Hyperparameters setting and Metrics

- Same hyperparameters as in SGMSE+:
  - STFT representation : Sampling rate=16KHz, Hann window of size 510, hop length=128
  - SDE

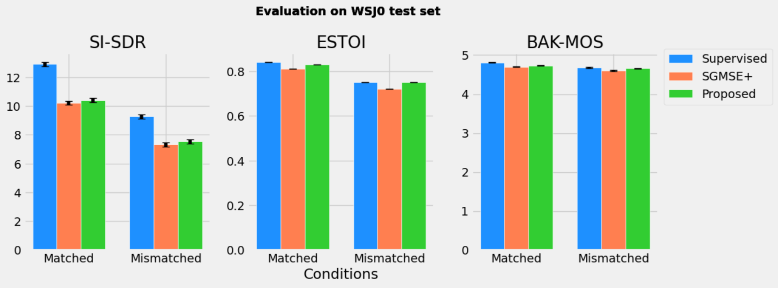
    stiffness parameter: $\gamma = 1.5$

    minimal and maximal noise variance: $\sigma_{\text{min}} = 0.05, \sigma_{\text{max}} = 0.5$

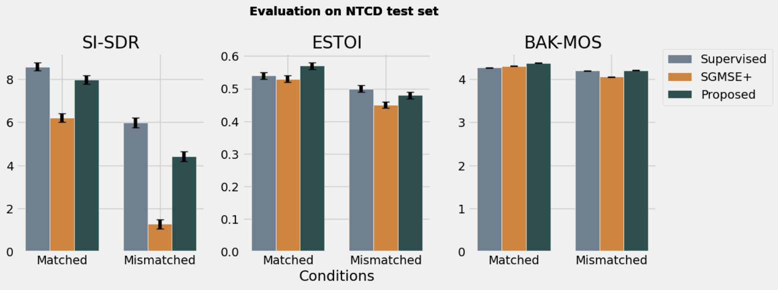
    minimum and maximum process times: $t_\varepsilon = 0.03, T = 1$
  - Number of Predictor-Corrector steps: $N = 30$.

- Metrics:
  - Scale-invariant signal-to-distortion Ratio measured in dB (SI-SDR)
  - Perceptual evaluation of speech quality (PESQ).
  - Extended short-time objective intelligibility (ESTOI).
  - DNSMOS for computing: speech signal quality (SIG), background intrusiveness (BAK), and overall quality (OVR)

Results

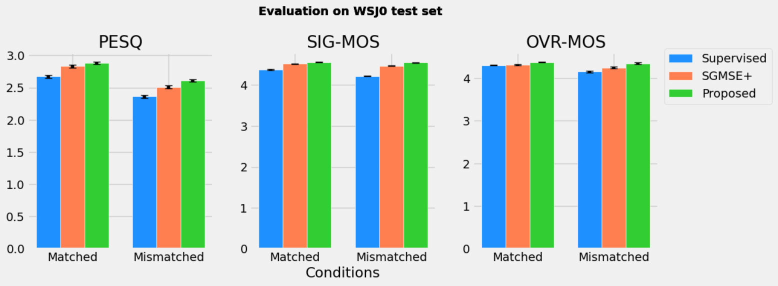


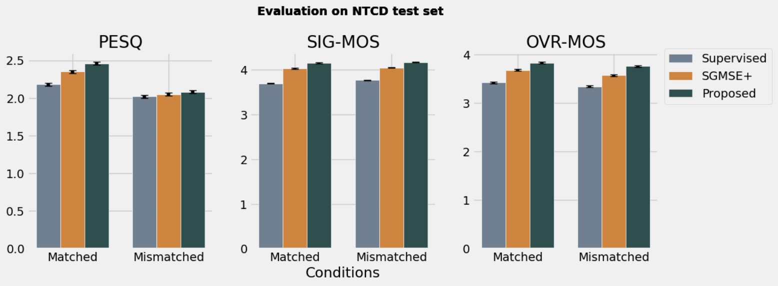
Evaluation on WSJ0 test set
SI-SDR
ESTOI
BAK-MOS
0
2
4
6
8
10
12
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
Matched
Mismatched
Conditions
Supervised
SGMSE+
Proposed



Evaluation on NTCD test set
SI-SDR
ESTOI
BAK-MOS
0
2
4
6
8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
Matched
Mismatched
Conditions
Supervised
SGMSE+
Proposed

Results


Evaluation on WSJ0 test set
PESQ
SIG-MOS
OVR-MOS
3.0
2.5
2.0
1.5
1.0
0.5
0.0
4
3
2
1
0
Matched
Mismatched
Conditions
Supervised
SGMSE+
Proposed


Evaluation on NTCD test set
PESQ
SIG-MOS
OVR-MOS
2.5
2.0
1.5
1.0
0.5
0.0
4
3
2
1
0
Matched
Mismatched
Conditions
Supervised
SGMSE+
Proposed

Conclusions

- **Objective**: account for how good will be the generated speeches, while using a generative diffusion-based loss.
- A weighted loss to compromise the generative loss with a supervised loss between the groundtruth and a clean speech estimates at the current diffusion time-step, is proposed.
- Experimental results showed that this approach combines the strengths of supervised methods and diffusion-based approach and improves performance.

Thank you for your attention!