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

Jean Eudes Ayilo, Mostafa Sadeghi, Romain Serizel

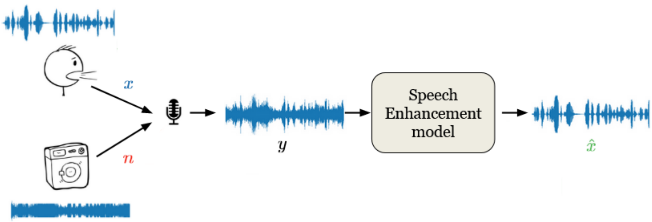
Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

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

April 14-19, Seoul, Republic of Korea

Introduction

Speech Enhancement (SE)



Adapted from info.uni-hamburg.de

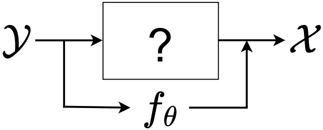
Given noisy speech observation $y = x + n$, estimate the clean speech signal x.

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

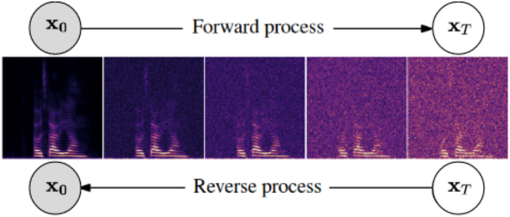
- **Generative approach**: learn (conditional/unconditional) clean speech distribution (using e.g., **diffusion models**) and at inference sample from the posterior distribution

Score-based generative model for SE

Observed mixture (in short time Fourier transform):

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

Score-based generative model for SE (**SGMSE+**)[1]



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:

$$\mathbf{x}_t = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + \left(1 - \mathrm{e}^{-\gamma t}\right)\mathbf{y} + \mathbf{e}_t \ , \qquad \mathbf{e}_t \sim \mathcal{N}_{\mathbb{C}}(\mathbf{e}_t; \mathbf{0}, \sigma(t)^2\mathbf{I})$$

Score-based generative model for SE

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

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

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

$$\mathbf{x}_t = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\,\mathbf{y} + \mathbf{e}_t\ , \qquad \mathbf{e}_t \sim \mathcal{N}_{\mathbb{C}}(\mathbf{e}_t; \mathbf{0}, \sigma(t)^2\mathbf{I})$$

❑ **Reverse process**:

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

Need to approximate the intractable score function

Score-based generative model for SE

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

$$\min_{\theta} \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z}\sim\mathcal{N}_{\mathbb{C}}(\mathbf{z};\mathbf{0},\mathbf{I}),\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}\Bigg[\underbrace{\left\|\sigma(t)\mathbf{s}_\theta\left(\mathbf{x}_t,\mathbf{y},t\right)+\mathbf{z}\right\|^2}_{:=L_\theta(\mathbf{x}_t,\mathbf{y},t,\mathbf{z})}\Bigg] \tag{1}$$

- 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: add an ℓ_2-loss between the ground-truth and an estimate denoted $\hat{\mathbf{x}}_{0,t}$.

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

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

- The new training objective is set to:

$$\min_\theta \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z}\sim\mathcal{N}_\mathbb{C}(\mathbf{0},\mathbf{I}),\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}\left[(1-\alpha_t)L_\theta\left(\mathbf{x}_t, \mathbf{y}, t, \mathbf{z}\right) + \alpha_t \left\|\hat{\mathbf{x}}_{0,t} - \mathbf{x}_0\right\|^2\right] \tag{3}$$

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

[3] C. Hyungjin, et al. "Diffusion Posterior Sampling for General Noisy Inverse Problems." The Eleventh International Conference on Learning Representations. 2022.

Weighted generative-supervised learning loss

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

- trade-off between the generative loss and the supervised loss
  - ▷ In this new proposed objective, α_t is set to :

$$\alpha_t = \frac{\sigma(T)-\sigma(t)}{\sigma(T)-\sigma(t_\varepsilon)} \tag{4}$$

  - ▷ when $\sigma(t) \searrow \alpha_t \nearrow$ and $\sigma(t) \nearrow \alpha_t \searrow$

Experiments

Model architecture and baselines

- Same architecture as the Noise Conditional Score Network (NCSN++) used in SGMSE+ (U-net like architecture)

- Trained models:
  - NCSN++ trained with the generative loss only (SGMSE+) **(baseline)**
  - Supervised version trained with MSE loss **(baseline)**
  - NCSN++ trained with our proposed loss

Hyperparameters setting and Metrics

- Same hyperparameters as in SGMSE+

- Metrics (the higher, the better):
    - 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)

Datasets

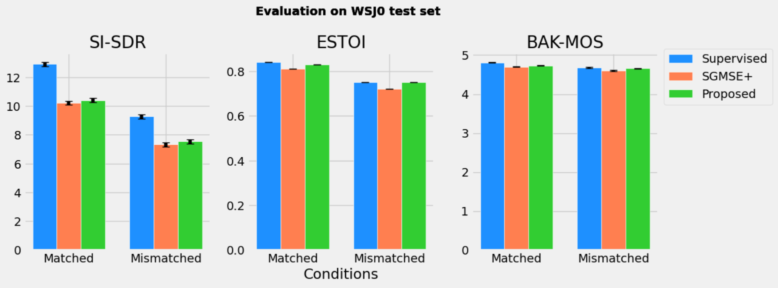
Training and test sets

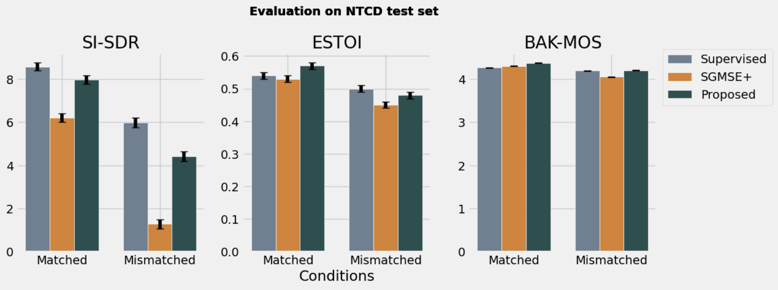
Clean speech dataset	Training noise dataset	Test noise dataset	Total [h] (Train/Test)	SNRs in test [dB]	Noise types in test
NTCD-TIMIT	DEMAND	NTCD-TIMIT	17.15 / 1.18	-5,0,5	(street, living room, cafe, car), white, babble
WSJ0	QUT-Noise	QUT-Noise	29.10 / 1.48	-5,0,5	(street, living room, cafe, car)

Cross data set evaluation

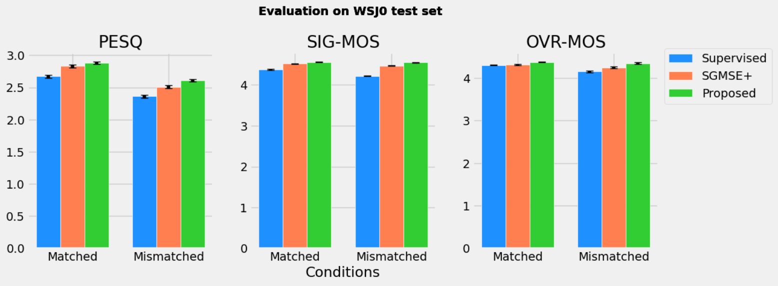
- ❑ **Matched**: Train and Test come from the *same* corpus
- ❑ **Mismatched**: Train and Test come from *different* corpora

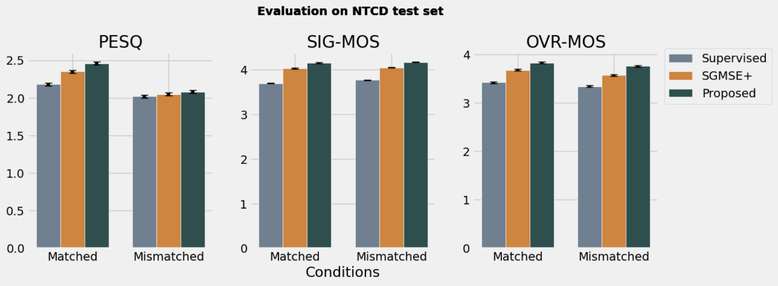
Results


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


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

Results


Evaluation on WSJ0 test set
PESQ
SIG-MOS
OVR-MOS
Matched
Mismatched
Conditions
Supervised
SGMSE+
Proposed


Evaluation on NTCD test set
PESQ
SIG-MOS
OVR-MOS
Matched
Mismatched
Conditions
Supervised
SGMSE+
Proposed

Conclusions

- We addressed training of score-based generative models for speech enhancement
- We integrated a supervised training loss with the generative-based Gaussian noise prediction loss used in a diffusion-based SE.
- Balance appropriately the supervised loss and the generative loss to improve the mapping between clean and noisy speech in a generative approach
- Empirical results showed that this approach combines the strengths of supervised and diffusion-based approaches

Thank you for your attention!

❑ Tweedie's formula

Lemma 1 (Tweedie's formula). *Let $p(\boldsymbol{y}|\boldsymbol{\eta})$ belong to the exponential family distribution*

$$p(\boldsymbol{y}|\boldsymbol{\eta}) = p_0(\boldsymbol{y}) \exp(\boldsymbol{\eta}^\top T(\boldsymbol{y}) - \varphi(\boldsymbol{\eta})), \tag{23}$$

where $\boldsymbol{\eta}$ is the canonical vector of the family, $T(\boldsymbol{y})$ is some function of $\boldsymbol{y}$, and $\varphi(\boldsymbol{\eta})$ is the cumulant generation function which normalizes the density, and $p_0(\boldsymbol{y})$ is the density up to the scale factor when $\boldsymbol{\eta} = \mathbf{0}$. Then, the posterior mean $\hat{\boldsymbol{\eta}} := \mathbb{E}[\boldsymbol{\eta}|\boldsymbol{y}]$ should satisfy

$$(\nabla_{\boldsymbol{y}} T(\boldsymbol{y}))^\top \hat{\boldsymbol{\eta}} = \nabla_{\boldsymbol{y}} \log p(\boldsymbol{y}) - \nabla_{\boldsymbol{y}} \log p_0(\boldsymbol{y}) \tag{24}$$

Taken from C. Hyungjin, et al. "Diffusion Posterior Sampling for General Noisy Inverse Problems." The Eleventh International Conference on Learning Representations. 2022.

Appendices

- How to apply Tweedie's formula to $\mathbf{x}_t$?

  Remind:

$\mathbf{x}_t = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\,\mathbf{y} + \mathbf{e}_t$, $\qquad \mathbf{e}_t \sim \mathcal{N}_{\mathbb{C}}(\mathbf{e}_t; \mathbf{0}, \sigma(t)^2\mathbf{I})$ $\mathbf{e}_t$ is a circularly-symmetric complex normal, so the joint distrbution of its the real and immaginary follows:

$$\begin{bmatrix} \Re(\mathbf{e}_t) \\ \Im(\mathbf{e}_t) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0}_d \\ \mathbf{0}_d \end{bmatrix}, \tfrac{\sigma(t)^2}{2} \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix} \right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{0}_d \\ \mathbf{0}_d \end{bmatrix}, \tfrac{\sigma(t)^2}{2} I_{2d} \right)$$

$$\begin{bmatrix} \Re(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) \\ \Im(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \Im(\mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\,\mathbf{y}) \\ \Re(\mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\,\mathbf{y}) \end{bmatrix}, \tfrac{\sigma(t)^2}{2} I_{2d} \right)$$

Appendices

Denoting:

$$\mathbf{X}_t = \begin{bmatrix} \Re(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) \\ \Im(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) \end{bmatrix}, \qquad \mathbf{X}_0 = \begin{bmatrix} \Re(\mathbf{x}_0) \\ \Im(\mathbf{x}_0) \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} \Re(\mathbf{y}) \\ \Im(\mathbf{y}) \end{bmatrix}$$

One can get the exponential family distribution formula (lemma 1) of $\mathbf{X}_t$ which follows a normal distribution:

$$p(\mathbf{X}_t|\mathbf{X}_0, \mathbf{Y}) = C.\exp(\frac{\left\|\mathbf{X}_t - [\mathrm{e}^{-\gamma t}\mathbf{X}_0 + (1-\mathrm{e}^{-\gamma t})\mathbf{Y}]\right\|^2}{\sigma(t)^2})$$

$$p_0(\mathbf{X}_t) = p(\mathbf{X}_t|0, \mathbf{Y}) = C.\exp(\frac{\left\|\mathbf{X}_t - (1-\mathrm{e}^{-\gamma t})\mathbf{Y}\right\|^2}{\sigma(t)^2})$$

We have:

$$\mathbf{T}(\mathbf{X}_t) = \exp(\frac{-2(-\mathrm{e}^{-\gamma t}\mathbf{X}_t + \mathrm{e}^{-\gamma t}(1-\mathrm{e}^{-\gamma t}\mathbf{Y}))}{\sigma(t)^2}), \text{ and } \varphi(\mathbf{X}_0) = \exp(\frac{\|\mathbf{Y}\|^2}{\sigma(t)^2})$$

Appendices

or alternatively:

$$\mathbf{T}(\mathbf{X}_t) = \exp(\frac{-2(-\mathrm{e}^{-\gamma t}\mathbf{X}_t)}{\sigma(t)^2}), \text{ and } \varphi(\mathbf{X}_0) = \exp(\frac{-2\mathrm{e}^{-\gamma t}(1-\mathrm{e}^{-\gamma t})\mathbf{X}_0^T\mathbf{Y} - \|\mathbf{Y}\|^2}{\sigma(t)^2})$$

Then we can apply the Tweedie's formula :

$$(\nabla_{X_t}\mathbf{T}(\mathbf{X}_t))^T\mathbb{E}[\mathbf{X}_0|\mathbf{X}_t, \mathbf{Y}] = \nabla_{X_t}\log p(\mathbf{x}_t) - \nabla_{X_t}\log p_0(\mathbf{X}_t)$$

which leads to equation 2 when summing the real and imaginary parts to get the notation in complex domain.
Note that it is also possible to apply the Tweedie's formula on the real and imaginary parts separately and then sum up.

- Note: If one didn't consider the factor 2 in $\frac{\sigma(t)^2}{2}$ in the Tweedie's formula, the supervised added loss, boils down to :

$$\alpha_t \sigma(t)^2 \|\sigma(t)\mathbf{s}_\theta\left(\mathbf{x}_t, \mathbf{y}, t\right) + \mathbf{z}\|^2$$

and in this case the total loss is:

$$[1+\alpha_t(\sigma(t)^2 - 1)]\|\sigma(t)\mathbf{s}_\theta\left(\mathbf{x}_t, \mathbf{y}, t\right) + \mathbf{z}\|^2$$

Appendices

- Justify performance in terms of PESQ

Denoising Score Matching objective with loss weighting $\lambda(t)$:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t\sim\mathcal{U}(0,T)} \mathbb{E}_{\mathbf{x}_0\sim q_0(\mathbf{x}_0)} \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \frac{\lambda(t)}{\sigma_t^2} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)||_2^2$$

Different loss weightings trade off between model with good perceptual quality vs. high log-likelihood

- *Perceptual quality:* $\lambda(t) = \sigma_t^2$
- *Maximum log-likelihood:* $\lambda(t) = \beta(t)$ *(negative ELBO)*

Taken from Karsten Kreis, Ruiqi Gao, Arash Vahdat. Tutorial : Denoising Diffusion-based Generative Modeling: Foundations and Applications (2022)

(https://cvpr2022-tutorial-diffusion-models.github.io/)

Appendices

❏ Justify performance in terms of SI-SDR

Table 1. Results for denoising on WSJ0+Chime data.

Method	Type	WV-MOS	PESQ	ESTOI	SI-SDR
Mixture		1.44 ± 1.62	1.70 ± 0.49	0.78 ± 0.14	10.0 ± 5.7
NCSN++M	D	3.65 ± 0.48	2.67 ± 0.69	$\mathbf{0.93} \pm \mathbf{0.06}$	$\mathbf{19.5} \pm \mathbf{4.4}$
SGMSE+M	G	$\mathbf{3.77} \pm \mathbf{0.32}$	$\mathbf{2.94} \pm \mathbf{0.60}$	0.92 ± 0.06	18.0 ± 5.1

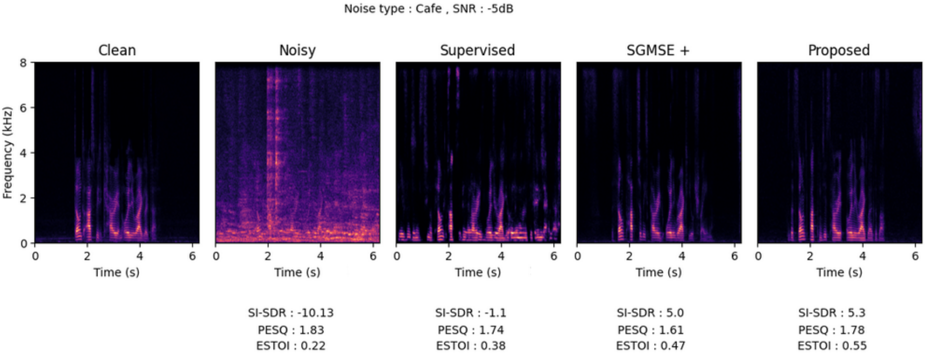
It is however slightly outperformed by discriminative NCSN++M on intelligibility and noise removal. Indeed, in a denoising task, the interference does not share any information with the target speech, making it relatively easy for a discriminative approach to remove the interference without distorting the target. However, we show in the uploaded listening examples that the discriminative approach tends to destroy low-energy speech regions for low SNRs, whereas the generative model does not. A larger benefit of the generative approach is observed when training and testing data have a stronger mismatch [8].

Taken from Lemercier, Jean-Marie, et al. "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration."

ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

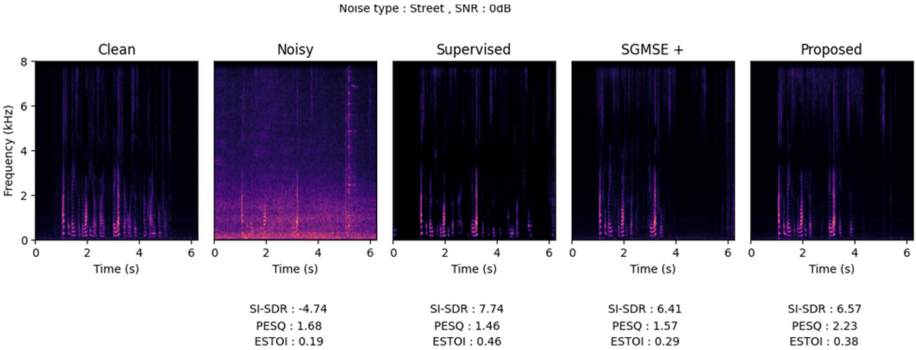
Appendices

❑ Visualisation example



Appendices

❑ Visualisation example



Appendices

- ❑ Visualisation example