

Generative Models as Data-driven Priors for Speech Enhancement

Mostafa Sadeghi

MULTISPEECH team
Inria Nancy - Grand Est

November 2024



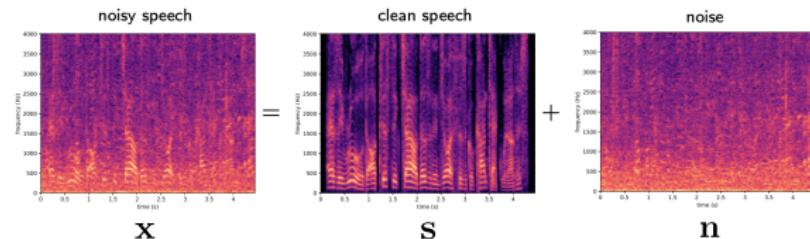
MULTISPEECH
Speech Modeling for Facilitating Oral-Based Communication



- 1 Speech Enhancement
- 2 Variational autoencoder as speech prior
- 3 Diffusion models as speech priors

Speech Enhancement

Speech Enhancement



Estimate the clean speech s from a noisy observation $x = s + n$.

- Close/distant conversations, listening comfort, hearing assistive devices.
- Automatic speech recognition for virtual assistants, social robots.

SE approaches

- ▷ **Supervised:** Model $p_{\Theta}(s|x)$, and learn Θ from pairs of **noisy-clean** data.
 - **Regressive-based:** Map noisy speech to a clean estimate or predict a time-freq. mask.¹
 - **Generative-based:** Incorporate speech generative models (e.g., conditioned on noisy speech,² as a refiner, etc.).
- ✓ Good performance on *seen data* (matched condition).
- ✗ Poor generalization on *unseen data* (mismatched condition).

¹Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation", IEEE ACM/TASLP, 2019.

²Richter, J., et al., "Speech enhancement and dereverberation with diffusion-based generative models," in IEEE/ACM TASLP, 2023.

SE approaches

▷ **Unsupervised:** Speech enhancement without training on paired data.

- **Generative priors:** Model $p_{\Theta}(s|x) \propto \underbrace{p_{\psi}(x|s)}_{\text{Inference}} \cdot \underbrace{p_{\theta}(s)}_{\text{Training}}$, and learn $\Theta = \theta \cup \psi$:
 - **Training** - Learn speech's prior distribution $p_{\theta}(s)$
 - **Inference** - Model $p_{\psi}(x|s)$, and infer s using posterior sampling
- **Unpaired training:** Mixture invariant training,³ domain adaptation,⁴ etc.

✓ Better generalization than supervised methods.

✗ Lagging behind supervised methods on seen data.

³ Wisdom, S., et al., "Unsupervised sound separation using mixture invariant training", NeurIPS, 2020.

⁴ Jiang, W., et al., "Unsupervised Speech Enhancement Using Optimal Transport and Speech Presence Probability", IEEE ACM/TASLP, 2024.

Unsupervised generative-based SE

Estimate clean speech \mathbf{s} by directly sampling from the **intractable** posterior:

$$p_\phi(\mathbf{s}|\mathbf{x}) \propto p_\phi(\mathbf{x}|\mathbf{s}) \cdot p_{\theta^*}(\mathbf{s})$$

▷ Modeling framework:

- $p_{\theta^*}(\mathbf{s})$: Distribution learned by a *generative model* on **clean** speech data
- $p_\phi(\mathbf{x}|\mathbf{s})$: Noise model, e.g., non-negative matrix factorization (NMF)

$$p_\phi(\mathbf{x}|\mathbf{s}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{s}, \text{diag}(\mathbf{v}_\phi)\right), \quad \mathbf{v}_\phi = \text{vec}(\mathbf{WH})$$

Inference framework: Expectation-maximization

▷ Iterative **Expectation Maximization (EM)**:

$$\max_{\phi} \mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{x})} \{\log p_{\phi}(\mathbf{x}|\mathbf{s})\}$$

① **E-step:** *Draw clean estimates*

$$\hat{\mathbf{s}}_k \sim p_{\phi_{k-1}}(\mathbf{s}|\mathbf{x}) \quad \rightarrow \text{posterior sampling}$$

② **M-step:** *Maximize likelihood*

$$\phi_k \leftarrow \operatorname{argmax}_{\phi} \log p_{\phi}(\mathbf{x}|\hat{\mathbf{s}}_k) \quad \rightarrow \text{NMF update}$$

Variational autoencoder as speech prior

VAE-based generative modeling

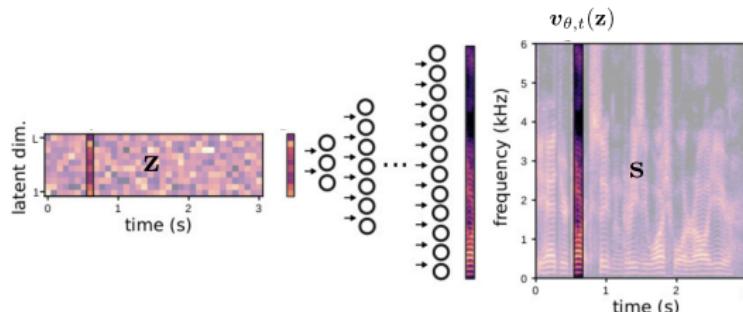
Short-time Fourier transform (STFT) of speech signal [Leglaive *et al.*, 2020]:

$$\mathbf{s} \triangleq \mathbf{s}_{1:T} = \{\mathbf{s}_1, \dots, \mathbf{s}_T\}, \quad \mathbf{s}_t = [s_{ft}]_{f=1}^F \in \mathbb{C}^F$$

Generative speech prior

$$\forall t : \quad p_\theta(\mathbf{s}_t | \mathbf{z}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \text{diag}(\mathbf{v}_{\theta,t}(\mathbf{z}))\right), \quad p(\mathbf{z}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_T\}$: sequence of real-valued **latent vectors**
- $\mathbf{v}_{\theta,t}(\cdot)$ is parameterized by a **decoder (generative) neural network**.



VAE training

Training data

Sequences of STFT time frames of *clean speech signals*: $\mathbf{s} = \{\mathbf{s}^{(i)} \in \mathbb{C}^{F \times T}\}_i$

Training objective

Maximum marginal likelihood:

$$\max_{\theta} \log p_{\theta}(\mathbf{s}) = \max_{\theta} \log \int p_{\theta}(\mathbf{s}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

⚠️ *Intractable!* Maximize a variational lower bound [Kingma and Welling 2014]:

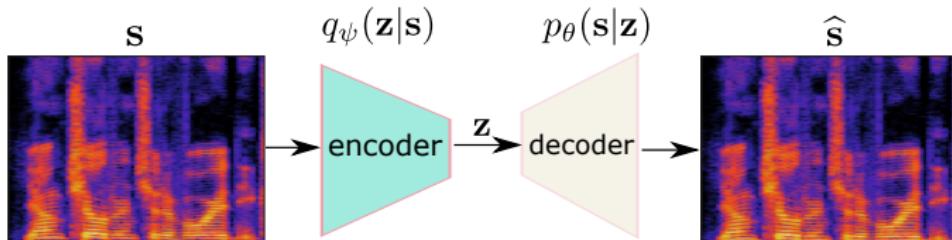
$$\log p_{\theta}(\mathbf{s}) \geq \mathbb{E}_{q_{\psi}(\mathbf{z}|\mathbf{s})} \left[\log \frac{p_{\theta}(\mathbf{s}|\mathbf{z})p_{\theta}(\mathbf{z})}{q_{\psi}(\mathbf{z}|\mathbf{s})} \right]$$

☞ $q_{\psi}(\mathbf{z}|\mathbf{s}) \approx p_{\theta}(\mathbf{z}|\mathbf{s})$: an approximate posterior with parameters ψ .

VAE training

Training objective

$$\theta^*, \psi^* = \arg \max_{\theta, \psi} \underbrace{\mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{s})} \left[\log p_\theta(\mathbf{s}|\mathbf{z}) \right]}_{\text{Reconstruction term}} - \underbrace{D_{\text{KL}} \left(q_\psi(\mathbf{z}|\mathbf{s}) \parallel p(\mathbf{z}) \right)}_{\text{Regularization term}}$$

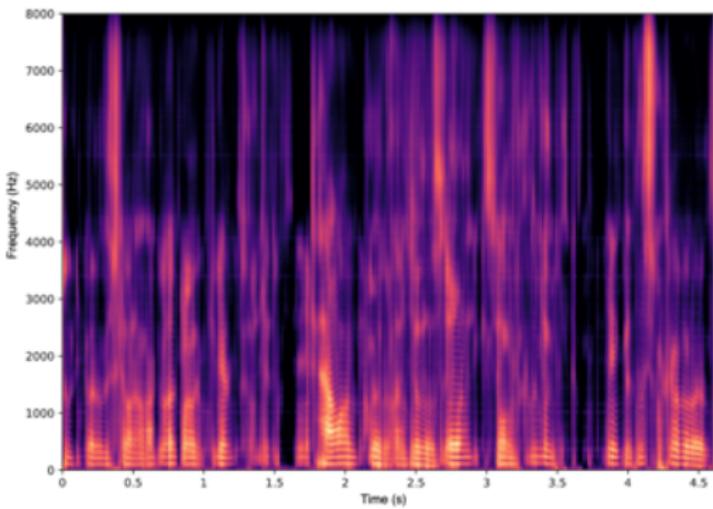


Sampling from VAE

- ① Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- ② Decode a clean speech $\mathbf{s}|\mathbf{z} \sim p_{\theta^*}(\mathbf{s}|\mathbf{z})$

☞ *The STFT phase spectrogram is estimated with the Griffin-Lim algorithm.*

Sampled speech



Speech Enhancement

Clean speech model

Pre-trained decoder network:

$$p_{\theta^*}(\mathbf{s}_t | \mathbf{z}) = \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \text{diag}(\mathbf{v}_{\theta^*, t}(\mathbf{z})))$$

Noise model

Non-negative Matrix Factorization (NMF):

$$\forall t : \mathbf{n}_t \sim \mathcal{N}_c(\mathbf{0}, \text{diag}([\mathbf{WH}]_t)), \quad \mathbf{W}, \mathbf{H} \geq 0$$

 **Likelihood:**

$$p_\phi(\mathbf{x}_t | \mathbf{z}) = \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \text{diag}(\mathbf{v}_{\theta, t}(\mathbf{z}) + [\mathbf{WH}]_t))$$

Parameters to estimate: $\phi = \{\mathbf{W}, \mathbf{H}\}$

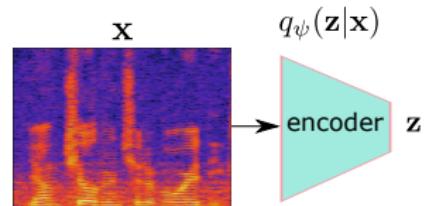
Parameter estimation

Variational EM (VEM): $\max_{\phi} \mathbb{E}_{p_{\phi}(\mathbf{z}|\mathbf{x})} \{\log p_{\phi}(\mathbf{x}|\mathbf{z})\}$ (a)

⚠ Intractable posterior $p_{\phi}(\mathbf{z}|\mathbf{x})$

- **E-step:** Approximate $p_{\phi}(\mathbf{z}|\mathbf{x})$ using variational inference⁵

$$\max_{\psi} \mathbb{E}_{q_{\psi}(\mathbf{z}|\mathbf{x})} [\log p_{\phi}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q_{\psi}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$



- ① Fine-tune the pre-trained encoder on \mathbf{x} .
 - ② Sample from $q_{\psi}(\mathbf{z}|\mathbf{x})$ to approximate (a).
- **M-step:** Update parameters with NMF.

⁵ Leglaive, S. et al. "A recurrent variational autoencoder for speech enhancement," ICASSP, 2020.

Speech estimation

Once the EM iterations converge, an estimate of the clean speech is obtained via:

$$\hat{\mathbf{s}} = \mathbb{E}_{p_{\phi^*}(\mathbf{s}|\mathbf{x})}\{\mathbf{s}\}$$

Wiener-like filtering

$$\forall t : \quad \hat{s}_t = \mathbb{E}_{p_\phi(\mathbf{z}|\mathbf{x})} \left[\frac{\mathbf{v}_{\theta^*,t}(\mathbf{z})}{\mathbf{v}_{\theta^*,t}(\mathbf{z}) + [\mathbf{W}^* \mathbf{H}^*]_t} \odot \mathbf{x}_t \right]$$

Or, approximately

$$\forall t : \quad \hat{s}_t \approx \mathbb{E}_{q_{\psi^*}(\mathbf{z}|\mathbf{x})} \left[\frac{\mathbf{v}_{\theta^*,t}(\mathbf{z})}{\mathbf{v}_{\theta^*,t}(\mathbf{z}) + [\mathbf{W}^* \mathbf{H}^*]_t} \odot \mathbf{x}_t \right]$$

Posterior sampling-based inference algorithms

Motivation

The **VEM** approach is **computationally expensive** for complex encoders.

👉 We propose efficient posterior sampling methods.

- **Direct sampling** from the intractable posterior $p_\phi(\mathbf{z}|\mathbf{x})$ in the *E-step*
- **Fast and efficient samplers** based on zero/first-order optimization^{6,7}

-
- 1: **Inputs:** $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ (noisy STFT data), \mathcal{H} (hyperparameters).
 - 2: **Initialize:** $\mathbf{z} = \{\mathbf{z}_t\}_{t=1}^T$, $\phi = \{\mathbf{W}, \mathbf{H}\}$.
 - 3: **for** $j = 1, \dots, J$ **do**
 - 4: **E-step:** $\mathbf{z} \leftarrow \text{Sampler}_\phi(\mathbf{z}, \mathcal{H})$
 - 5: **M-step:** $\phi \leftarrow \text{argmax}_\phi \log p_\phi(\mathbf{x}|\mathbf{z})$
 - 6: **end for**
 - 7: **Clean speech estimation:** $\hat{\mathbf{s}} = \left\{ \frac{\mathbf{v}_{\theta,t}(\mathbf{z})}{\mathbf{v}_{\theta,t}(\mathbf{z}) + [\mathbf{WH}]_t} \odot \mathbf{x}_t \right\}_{t=1}^T$
-

⁶ M. Sadeghi and R. Serizel, "Fast and efficient speech enhancement with variational autoencoders," ICASSP, 2023.

⁷ M. Sadeghi and R. Serizel, "Posterior sampling algorithms for unsupervised speech enhancement with recurrent variational autoencoder," ICASSP, 2024.

Metropolis-Hastings-based EM (MHEM)

- An iterative Markov chain Monte Carlo (MCMC) sampling method⁸
- We develop an MHEM algorithm for VAE-based speech enhancement.

Sampling from $p_\phi(\mathbf{z}|\mathbf{x})$ at the E-step:

- ❶ Initial states $\mathbf{z}^{(0)} = \{\mathbf{z}_1^{(0)}, \dots, \mathbf{z}_T^{(0)}\}$
- ❷ Next samples: $\tilde{\mathbf{z}}_t^{(k)} | \mathbf{z}_t^{(k-1)} \sim \mathcal{N}(\mathbf{z}_t^{(k-1)}, \sigma^2 \mathbf{I}), \quad \forall t$
- ❸ Accept the new samples with the probability (relative posteriors):

$$\alpha_t = \min \left(1, \frac{p_\phi(\mathbf{x}_t | \tilde{\mathbf{z}}^{(k)}) p(\tilde{\mathbf{z}}_t^{(k)})}{p_\phi(\mathbf{x}_t | \mathbf{z}^{(k-1)}) p(\mathbf{z}_t^{(k-1)})} \right)$$

☞ The acceptance ratios are computed in parallel using $p_\phi(\mathbf{x} | \mathbf{z}^{(k-1)})$.

⁸ Robert, C. P., et al., Monte Carlo statistical methods, vol. 2, Springer, 1999.

Langevin dynamics-based EM (LDEM)

Needs only $\nabla_{\mathbf{z}} \log p_{\phi}(\mathbf{z}|\mathbf{x})$ (**score function**) for sampling.⁹

$$f_{\phi}(\mathbf{z}) = \nabla_{\mathbf{z}} \log p_{\phi}(\mathbf{z}|\mathbf{x}) = \nabla_{\mathbf{z}} \left(\sum_{t=1}^T \log p_{\phi}(\mathbf{x}_t|\mathbf{z}) + \log p(\mathbf{z}_t) \right)$$

- *Multiple* samples per time-frame:

$$\mathbf{z}_{t,i}^{(0)} | \mathbf{z}_t \sim \mathcal{N}(\mathbf{z}_t, \sigma^2 \mathbf{I}), \quad \forall t, i = 1, \dots M$$

- Next samples via LD:

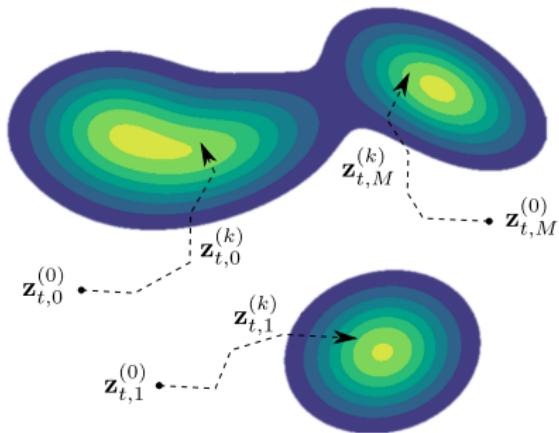
$$\mathbf{z}_{t,i}^{(k)} | \mathbf{z}^{(k-1)} \sim \mathcal{N}\left(\mathbf{z}_{t,i}^{(k-1)} + \frac{\eta}{2} f_{\phi}(\mathbf{z}^{(k-1)}), \eta \mathbf{I}\right)$$

or

$$\mathbf{z}_{t,i}^{(k)} = \mathbf{z}_{t,i}^{(k-1)} + \frac{\eta}{2} f_{\phi}(\mathbf{z}^{(k-1)}) + \sqrt{\eta} \boldsymbol{\zeta}_{t,i}, \quad \boldsymbol{\zeta}_{t,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

⁹R. M. Neal, Bayesian learning for neural networks. Springer Science & Business Media, vol. 118, 2012.

Langevin dynamics-based EM (LDEM)



👉 **Gradient ascent** steps on the score function.

👉 **Noise injection** to better explore the posterior space.

👉 No acceptance/rejection mechanism.

Metropolis-Adjusted Langevin Algorithm (MALA)

Add an **acceptance/rejection mechanism** to LD.¹⁰

- Next samples:

$$\tilde{\mathbf{z}}_t^{(k)} | \mathbf{z}_t^{(k-1)} \sim \mathcal{N}\left(\mathbf{z}_t^{(k-1)} + \frac{\eta}{2} f_\phi(\mathbf{z}_t^{(k-1)}), \eta \mathbf{I}\right)$$

- Accept or reject the new samples:

$$\alpha_t = \min\left(1, \frac{p_\phi(\mathbf{x}_t | \tilde{\mathbf{z}}^{(k)}) p(\tilde{\mathbf{z}}_t^{(k)} | \mathbf{z}^{(k)}) q(\mathbf{z}^{(k)} | \tilde{\mathbf{z}}^{(k)})}{p_\phi(\mathbf{x}_t | \mathbf{z}^{(k-1)}) p(\mathbf{z}_t^{(k-1)} | \mathbf{z}^{(k)}) q(\tilde{\mathbf{z}}^{(k)} | \mathbf{z}^{(k)})}\right)$$

where $q(\mathbf{u}|\mathbf{v})$ is the *transition probability density* from \mathbf{v} to \mathbf{u} :

$$q(\mathbf{u}|\mathbf{v}) \propto \exp\left(-\frac{1}{2\eta} \|\mathbf{u} - \mathbf{v} - \frac{\eta}{2} f(\mathbf{v})\|^2\right)$$

👉 Unlike MH, MALA tends towards higher probability regions.

¹⁰ G. O. Roberts and O. Stramer, "Langevin diffusions and metropolis-hastings algorithms," Methodol. Comput. Appl. Probab., vol. 4, 2002.

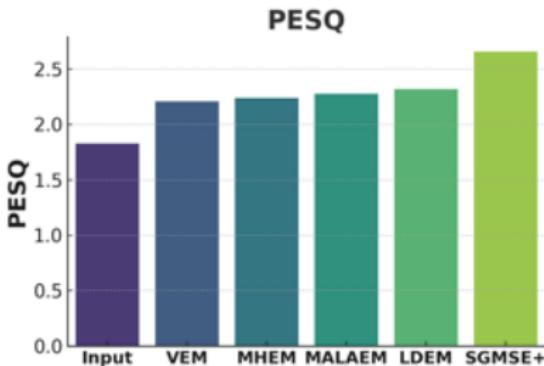
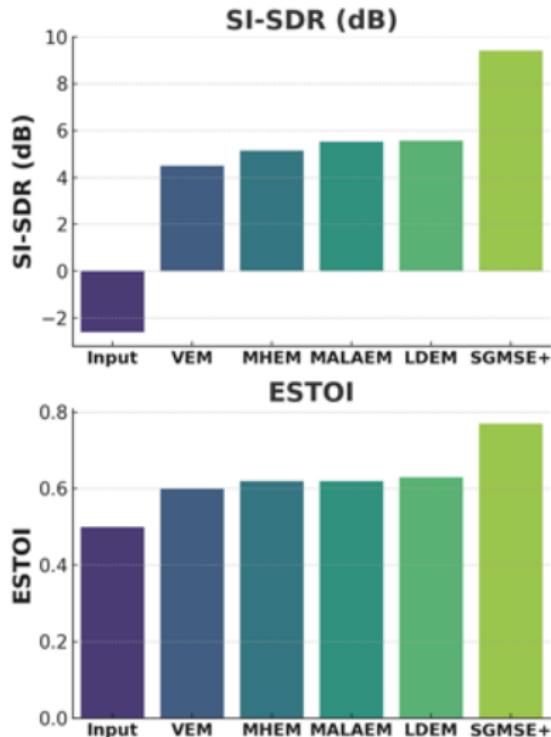
Experiments

- **Datasets:**
 - WSJ0-QUT (*training & evaluation*): Matched condition
 - TCD-TIMIT (*evaluation*) : Mismatched condition
- **Parameters:** $K = 1$ (sampling iterations) for LDEM (with $M = 10$ parallel samples), while $K = 10$ for MHEM and MALAEM
- **Baselines:** Pre-trained RVAE¹¹ (unsupervised) and SGMSE+¹² (supervised).
- **Performance metrics:**
 - SI-SDR (dB) \uparrow
 - PESQ (perceptual quality) [-0.5, 4.5] \uparrow
 - STOI (intelligibility) [0, 1] \uparrow
 - RTF (real-time factor) \downarrow

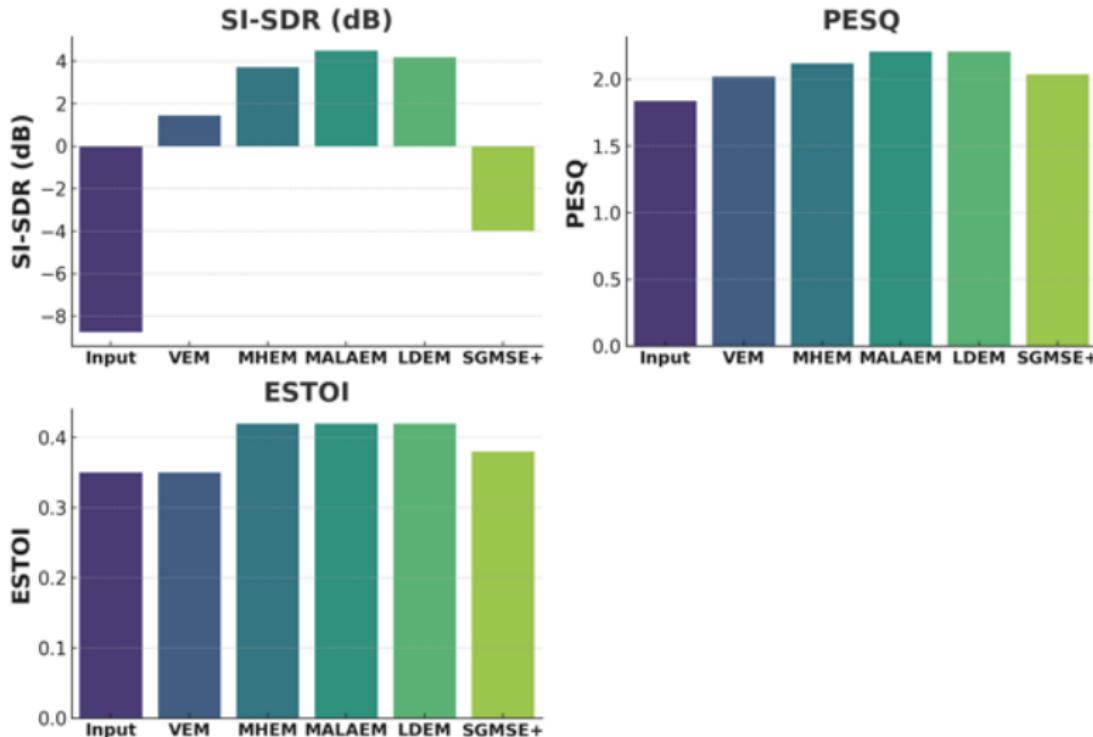
¹¹Bie, X., et al., "Unsupervised speech enhancement using dynamical variational autoencoders," IEEE/ACM TASLP, vol. 30, 2022.

¹²Richter, J., et al., "Speech enhancement and dereverberation with diffusion-based generative models," in IEEE/ACM TASLP, vol. 31, June 2023.

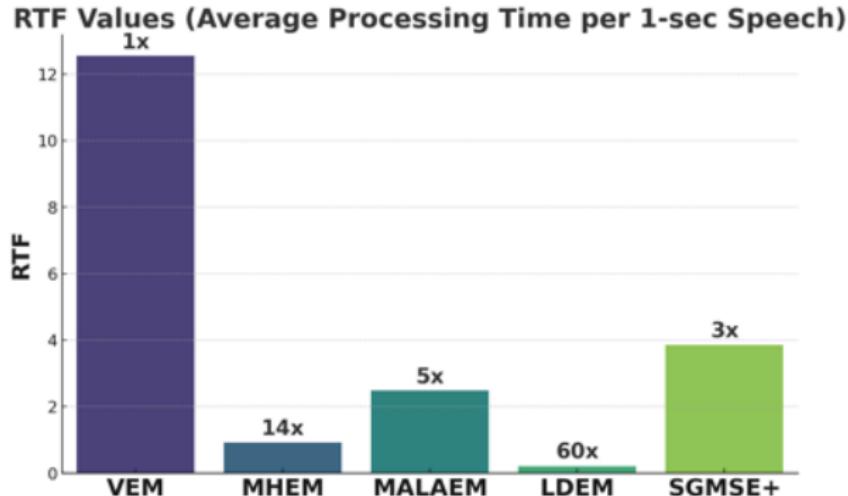
Results: Matched conditions



Results: Mismatched condition



Results: Real-time Factor (RTF)



Diffusion models as speech priors

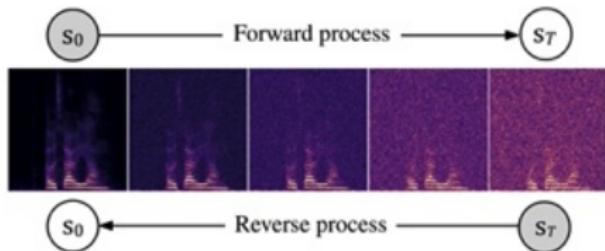
Diffusion-based speech generative model

- Diffusion model for complex-valued **clean speech** STFT:

- **Forward process:**¹³

$$ds_t = f(s_t)dt + g(t)d\omega, \quad f(s_t) = -\gamma s_t$$

$$s_t = e^{-\gamma t} s_0 + \sigma(t)\epsilon \quad \epsilon : \text{Gaussian noise}$$



- **Reverse process:**

$$ds_t = [f(s_t) - g(t)^2 \nabla_{s_t} \log p_t(s_t)] dt + g(t)d\omega$$

¹³Song, Y., et al., "Score-based generative modelling through stochastic differential equations", ICLR, 2021.

Approximating the score

Approximate the score function to enable sampling from the prior:

$$\begin{aligned} d\mathbf{s}_t &= [\mathbf{f}(\mathbf{s}_t) - g(t)^2 \nabla_{\mathbf{s}_t} \log p_t(\mathbf{s}_t)] dt + g(t) d\mathbf{w} \\ &\approx [\mathbf{f}(\mathbf{s}_t) - g(t)^2 \mathbf{S}_{\theta^*}(\mathbf{s}_t, t)] dt + g(t) d\mathbf{w} \end{aligned}$$

- ➊ Learn $\mathbf{S}_\theta(\mathbf{s}_t, t)$:

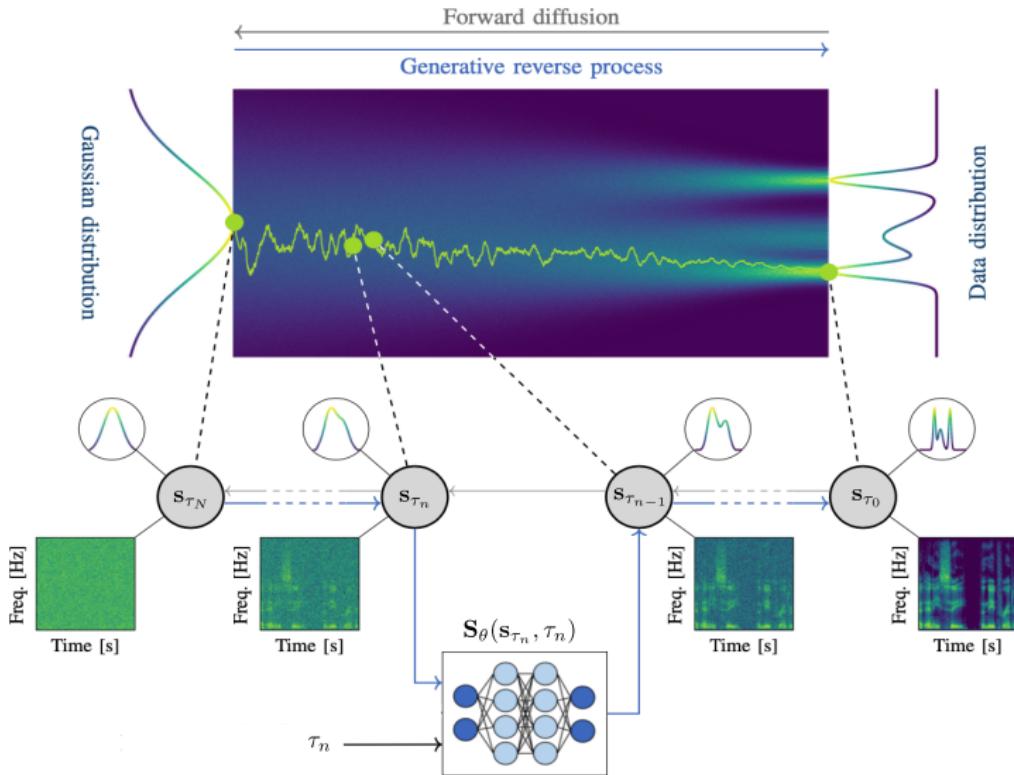
$$\theta^* = \operatorname{argmin}_\theta \mathbb{E}_{t, \mathbf{s}, \zeta, \mathbf{s}_t | \mathbf{s}} \left[\|\mathbf{S}_\theta(\mathbf{s}_t, t) + \frac{\zeta}{\sigma(t)}\|_2^2 \right], \quad \zeta \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{I})$$

- ➋ Numerically sample from the prior $p_{\theta^*}(\mathbf{s})$

☞ The reverse process can be solved by the *Predictor-Corrector (PC) sampler*¹⁴

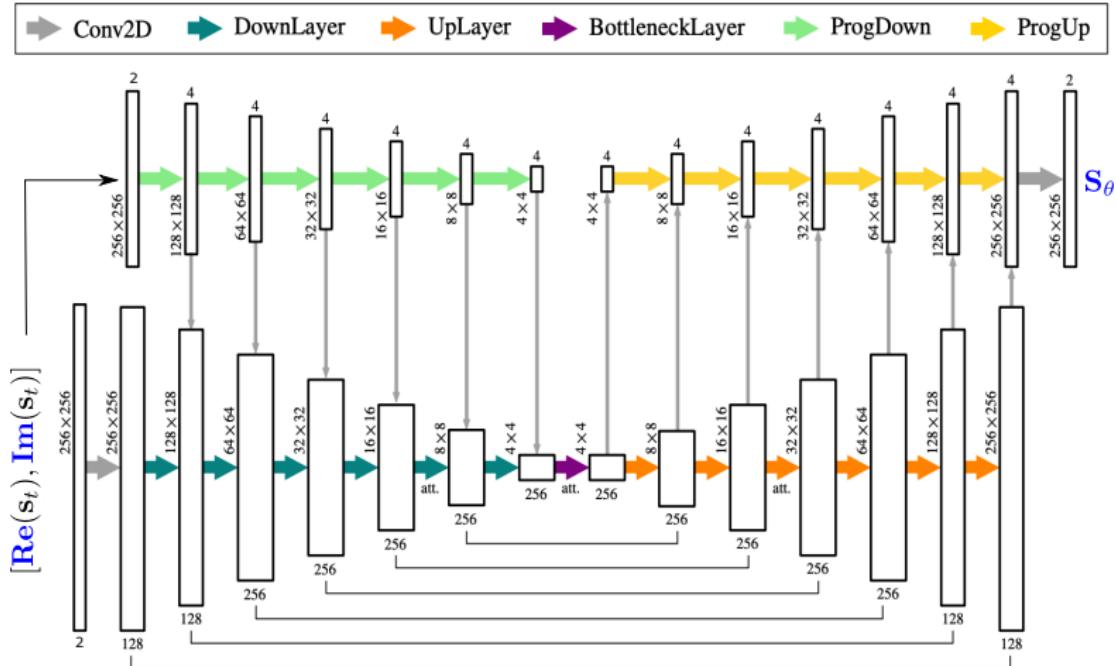
¹⁴ Song, Y., et al., "Score-based generative modelling through stochastic differential equations", ICLR, 2021.

Diffusion modeling - summary



Adapted from: Lemercier, J. M., et al. "Diffusion Models for Audio Restoration." arXiv:2402.09821, 2024.

Score model - architecture

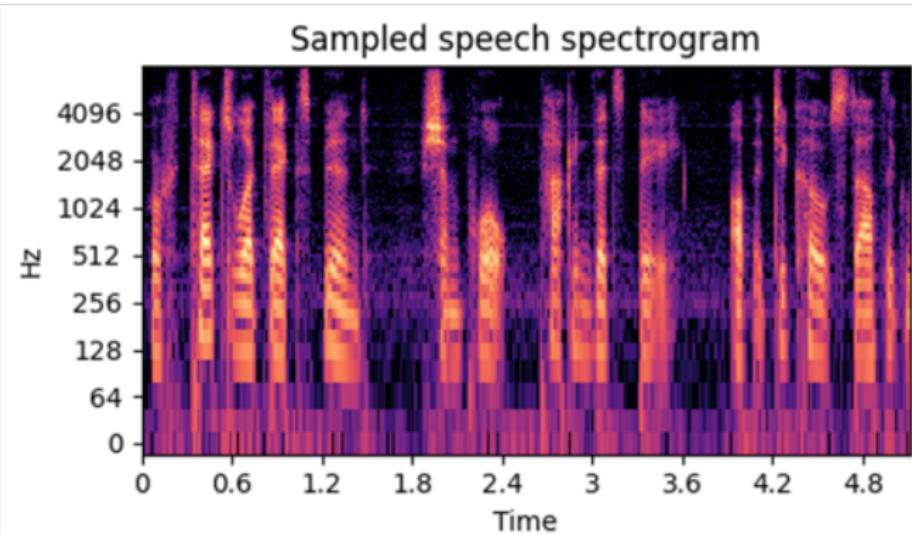


Adapted from: Richter J., et al., "Speech enhancement and dereverberation with diffusion-based generative models," in IEEE/ACM TASLP, 2023.

Sampling from the prior

- ➊ Sample a Gaussian noise $s \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{I})$
 - ➋ Run the reverse (denoising) process starting with s
-

Sampled speech



SE phase as EM approach

Once the *prior score model* is trained, SE is performed via EM.¹⁵

E-step: Replace the prior with the posterior in the reverse process:

$$\begin{aligned}\mathrm{d}\mathbf{s}_t &= \left[\mathbf{f}(\mathbf{s}_t) - g(t)^2 \nabla_{\mathbf{s}_t} \log p_t(\mathbf{s}_t | \mathbf{x}) \right] dt + g(t) d\mathbf{w} \\ &= \left[\mathbf{f}(\mathbf{s}_t) - g(t)^2 \left(\nabla_{\mathbf{s}_t} \log p_\phi(\mathbf{x} | \mathbf{s}_t) + \nabla_{\mathbf{s}_t} \log p_t(\mathbf{s}_t) \right) \right] dt + g(t) d\mathbf{w}\end{aligned}$$

⚠️ *Intractable*, time-dependent likelihood!

Approximation by the “*noise-perturbed pseudo-likelihood*”¹⁶

$$\tilde{p}_\phi(\mathbf{x} | \mathbf{s}_t) \sim \mathcal{N}_{\mathbb{C}} \left(\frac{\mathbf{s}_t}{\delta_t}, \frac{\sigma(t)^2}{\delta_t^2} \mathbf{I} + \text{diag}(\mathbf{v}_\phi) \right), \quad \delta_t = e^{-\gamma t}$$

¹⁵ B. Nortier, M. Sadeghi, and R. Serizel, “Unsupervised speech enhancement with diffusion-based generative models,” ICASSP 2024.

¹⁶ X. Meng and Y. Kabashima, “Diffusion model based posterior sampling for noisy linear inverse problems,” 2022.

SE phase as EM approach: UDiffSE

E-step: Solve the reverse process to get $\hat{\mathbf{s}}$

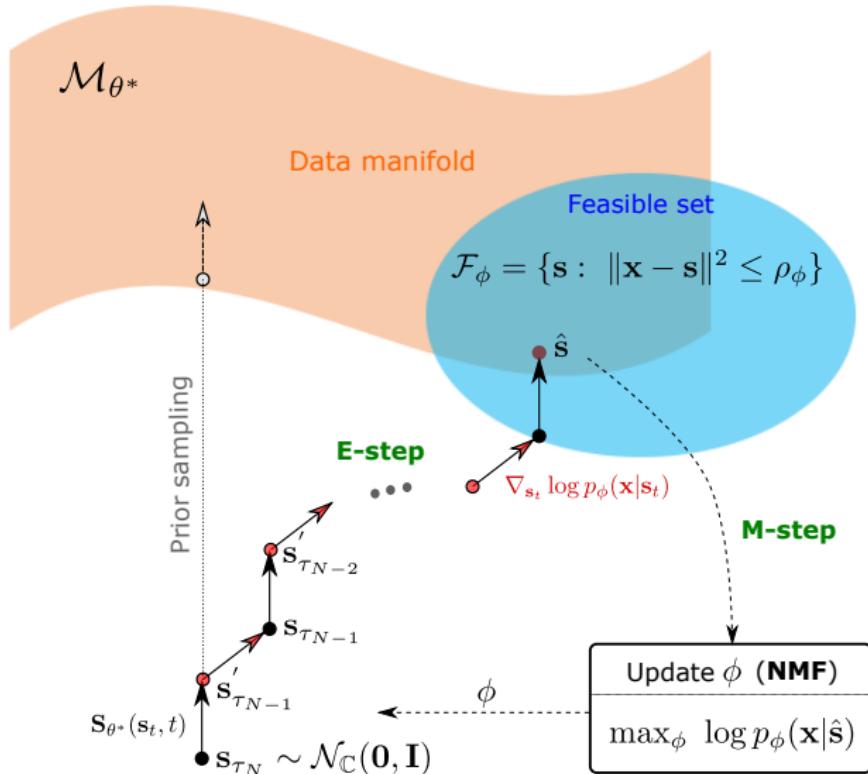
$$d\mathbf{s}_t = \left[\mathbf{f}(\mathbf{s}_t) - g(t)^2 \left(\lambda \nabla_{\mathbf{s}_t} \log \tilde{p}_\phi(\mathbf{x}|\mathbf{s}_t) + \mathbf{S}_{\theta^*}(\mathbf{s}_t, t) \right) \right] dt + g(t)d\mathbf{w}$$

☞ λ : weighting parameter to balance prior and likelihood terms.

M-step: Update noise parameters via NMF

$$\begin{aligned}\phi^* &\leftarrow \operatorname{argmax}_{\mathbf{v}_\phi(i) \geq 0} \log p_\phi(\mathbf{x}|\hat{\mathbf{s}}) \\ &= \operatorname{argmin}_{\mathbf{v}_\phi(i) \geq 0} \sum_i \frac{(\mathbf{x} - \hat{\mathbf{s}})_i^H (\mathbf{x} - \hat{\mathbf{s}})_i}{\mathbf{v}_\phi(i)} + \log(\mathbf{v}_\phi(i))\end{aligned}$$

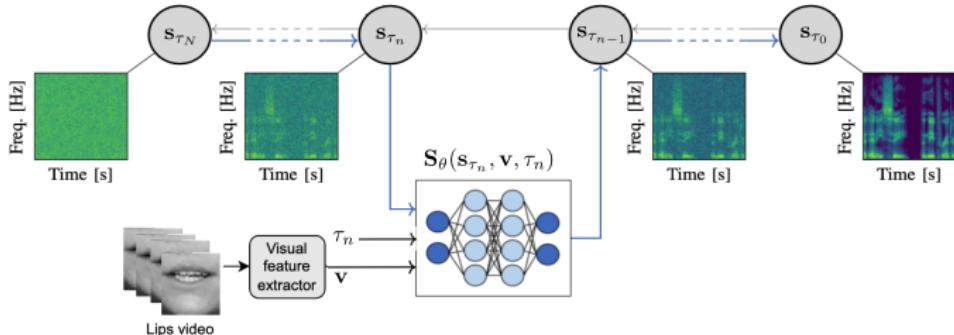
UDiffSE - A schematic view



Audio-visual speech enhancement



☞ Integrate the speaker's **video** into the **score model**.¹⁷



¹⁷ J.-E. Ayilo, M. Sadeghi, R. Serizel, and X. Alameda-Pineda, "Diffusion-based Unsupervised Audio-visual Speech Enhancement", hal-04718254, 2024.

Experiments

- **Datasets**

- Training: WSJ0 (~ 25 hrs)
- Testing: WSJ0-QUT (1.5hrs), TCD-TIMIT (45mins)

- **Evaluation Metrics**

- Objective measures: SI-SDR, ESTOI, PESQ
- (Pseudo)-subjective measures: DNS-MOS (SIG, BAK, OVRL)

- **Baselines:** RVAE¹⁸ and SGMSE+¹⁹, FlowAVSE²⁰

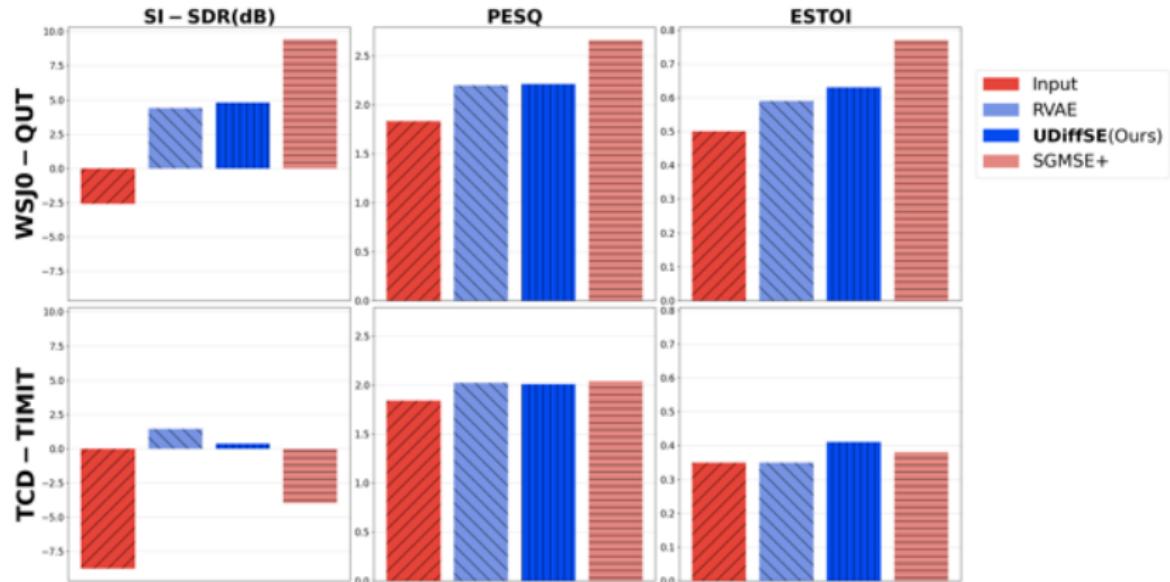
- **Models architecture.** Multi-resolution U-Net as in SGMSE+.

¹⁸Bie, X., et al., "Unsupervised speech enhancement using dynamical variational autoencoders," IEEE/ACM TASLP, vol. 30, 2022.

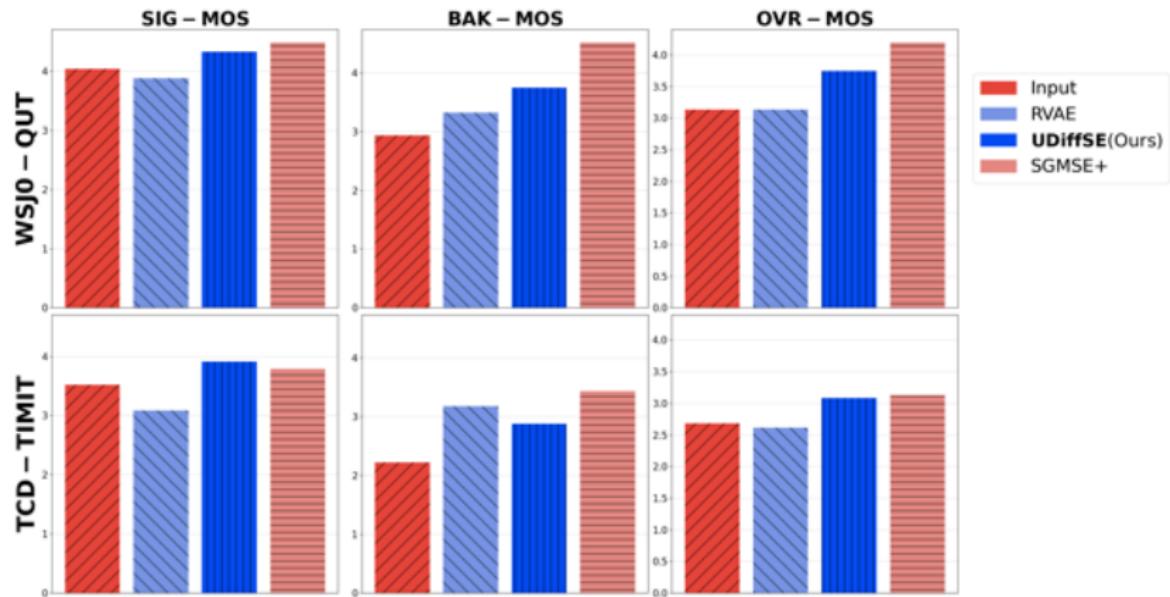
¹⁹Richter, J., et al., "Speech enhancement and dereverberation with diffusion-based generative models," in IEEE/ACM TASLP, vol. 31, June 2023.

²⁰Jung, C., et al., "FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching," INTERSPEECH, 2024.

Results (audio-only SE)

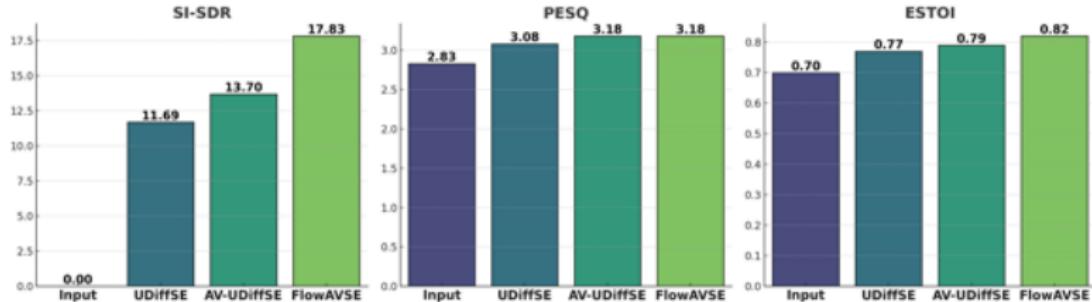


Results (audio-only SE)

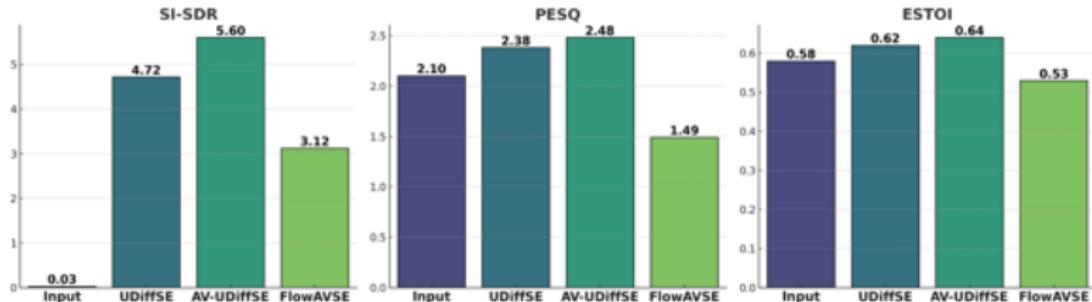


Results (audio-visual SE)

TCD speech + DEMAND noise (matched)



LRS3 speech + NTCD noise (mismatched)



Conclusions

- Data-driven speech priors are versatile tools for many tasks:
 - They exhibit better generalization performance than supervised models.
 - Can be better interpreted than the supervised models.
 - A single pre-trained model can effectively address multiple inverse problems.

However,

- They lag behind the supervised models in matched conditions.
- They are computationally more expensive than the supervised counterparts.
- Sampling methods (e.g., Langevin dynamics) work well for VAE-based SE.
- Diffusion models offer efficient unsupervised SE alternatives to VAEs.

Listening examples



Demo (UDiffSE)

<https://team.inria.fr/multispeech/demos/udiffse/>



Demo (AV-UDiffSE)

https://jeaneudesayilo.github.io/fast_UdiffSE/

Thank you for your attention