# Generative Models as Data-driven Priors for Speech Enhancement

Mostafa Sadeghi

MULTISPEECH team
Inria Nancy - Grand Est

November 2024

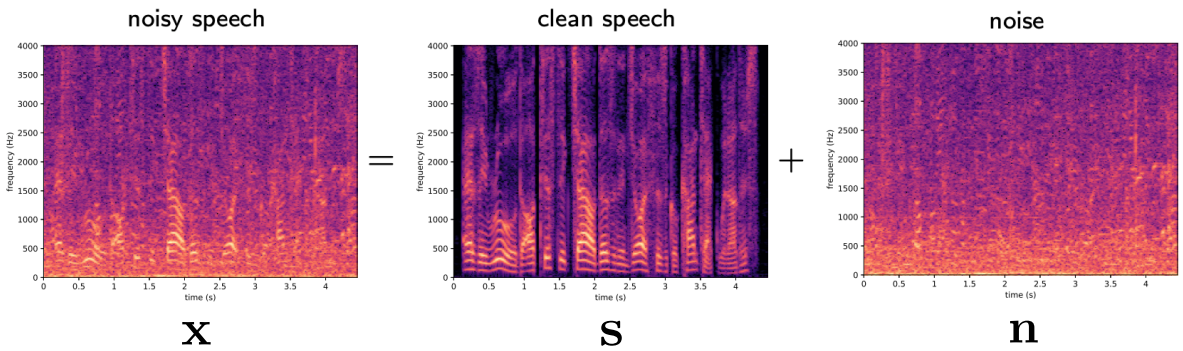
MULTISPEECH

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* $\mathbf{s}$ *from a noisy observation* $\mathbf{x} = \mathbf{s} + \mathbf{n}$.

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

# SE approaches

▷ **Supervised:** Model $p_\Theta(\mathbf{s}|\mathbf{x})$, and learn $\Theta$ 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).

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

[2] Richter, 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 $p_\Theta(\mathbf{s}|\mathbf{x}) \propto \underbrace{p_\psi(\mathbf{x}|\mathbf{s})}_{\text{Inference}} \cdot \underbrace{p_\theta(\mathbf{s})}_{\text{Training}}$, and learn $\Theta = \theta \cup \psi$:
  - Training - Learn speech's prior distribution $p_\theta(\mathbf{s})$
  - Inference - Model $p_\psi(\mathbf{x}|\mathbf{s})$, and infer $\mathbf{s}$ 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)

$$p_\phi(\mathbf{x}|\mathbf{s}) = \mathcal{N}_{\mathbb{C}}\Big(\mathbf{s}, \mathrm{diag}(\mathbf{v}_\phi)\Big), \qquad \mathbf{v}_\phi = \mathrm{vec}(\mathbf{WH})$$

# 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

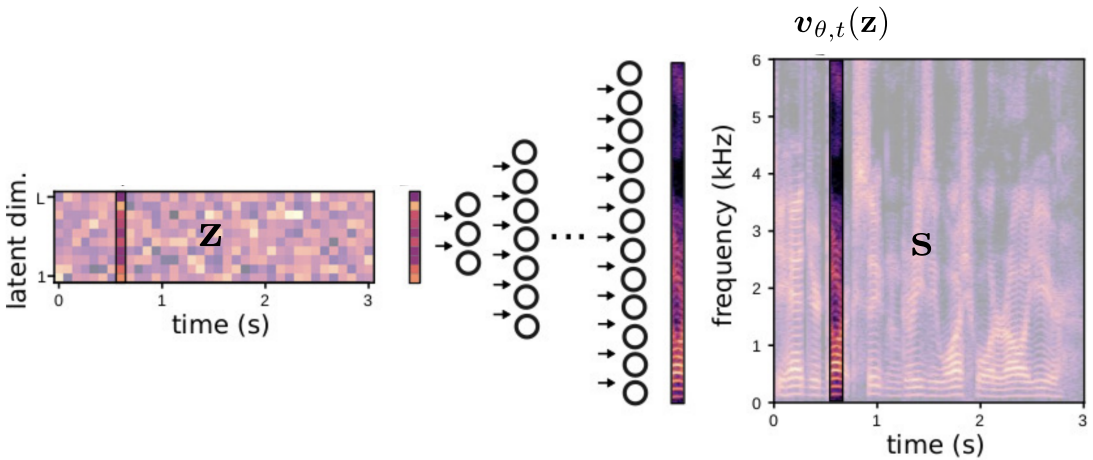
Short-time Fourier transform (STFT) of speech signal [Leglaive *et al.*, 2020]:

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

**Generative speech prior**

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

- $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$: sequence of real-valued latent vectors
- $\boldsymbol{v}_{\theta,t}(.)$ 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:

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

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

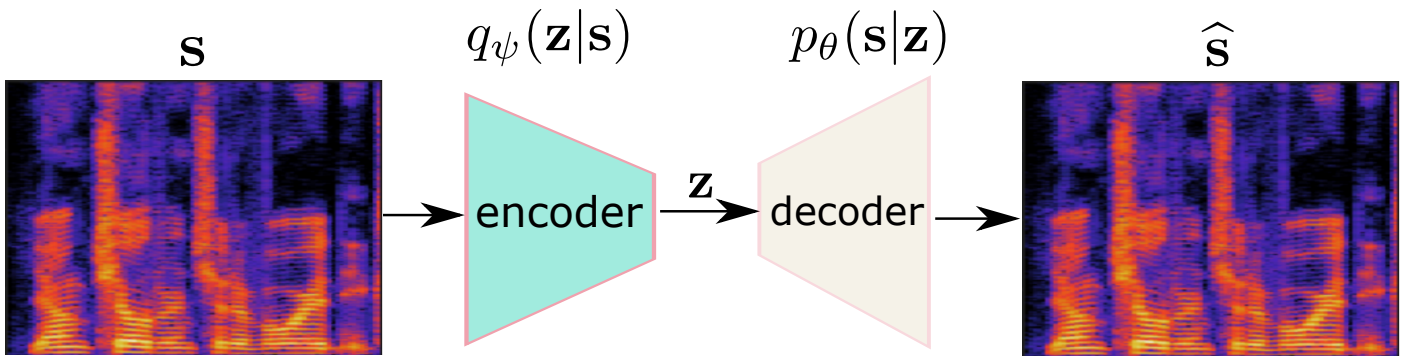
$$\log p_\theta(\mathbf{s}) \geq \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{s})} \left[ \log \frac{p_\theta(\mathbf{s}|\mathbf{z}) p_\theta(\mathbf{z})}{q_\psi(\mathbf{z}|\mathbf{s})} \right]$$

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

# VAE training

**Training objective**

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



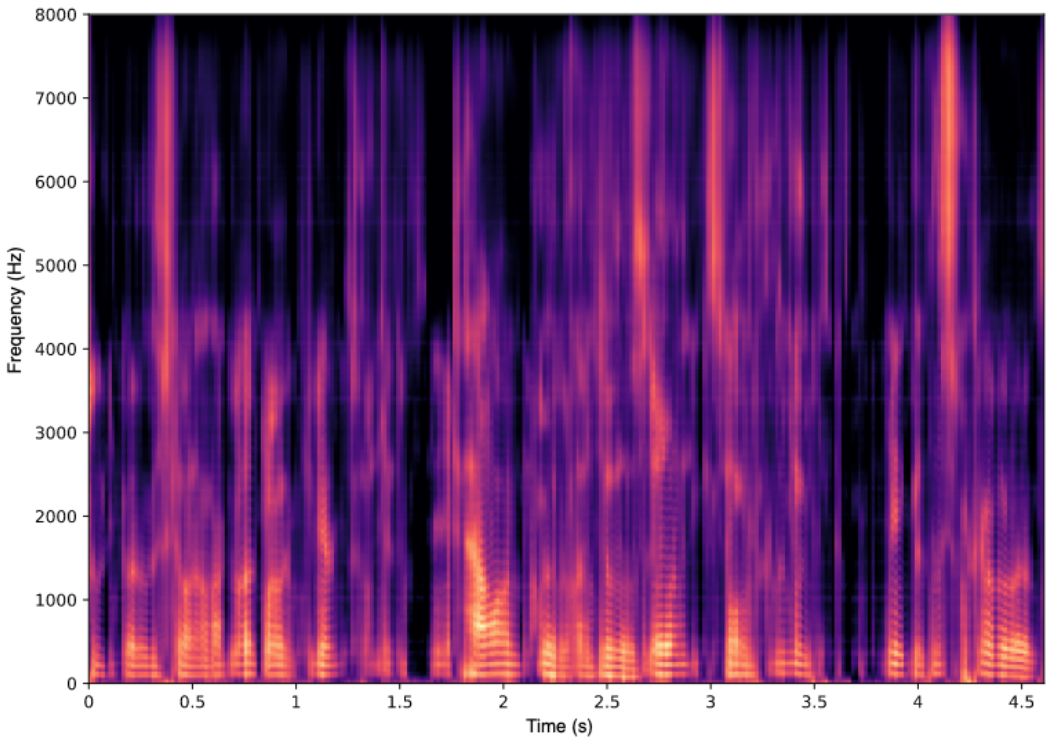
# Sampling from VAE

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




`https://dynamicalvae.github.io/`

# Speech Enhancement

## Clean speech model

Pre-trained decoder network: $p_{\theta^*}(\mathbf{s}_t|\mathbf{z}) = \mathcal{N}_{\mathbb{C}}\Big(\mathbf{0}, \text{diag}(\boldsymbol{v}_{\theta^*,t}(\mathbf{z}))\Big)$

## Noise model

Non-negative Matrix Factorization (NMF):

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

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

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

# Parameter estimation

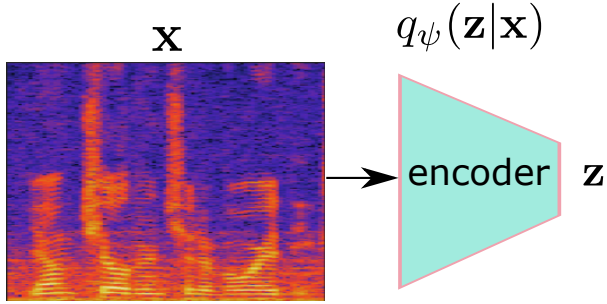
**Variational EM (VEM):** $\max_{\phi} \ \mathbb{E}_{p_\phi(\mathbf{z}|\mathbf{x})}\{\log p_\phi(\mathbf{x}|\mathbf{z})\}$ (a)

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

- **E-step:** Approximate $p_\phi(\mathbf{z}|\mathbf{x})$ using variational inference[5]

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

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**

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

Or, approximately

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

# 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** $j = 1, \cdots, J$ **do**
4: **E-step:** $\mathbf{z} \leftarrow \texttt{Sampler}_\phi(\mathbf{z}, \mathcal{H})$
5: **M-step:** $\phi \leftarrow \text{argmax}_\phi \log p_\phi(\mathbf{x}|\mathbf{z})$
6: **end for**
7: **Clean speech estimation:** $\hat{\mathbf{s}} = \left\{ \frac{\boldsymbol{v}_{\theta,t}(\mathbf{z})}{\boldsymbol{v}_{\theta,t}(\mathbf{z}) + [\mathbf{WH}]_t} \odot \mathbf{x}_t \right\}_{t=1}^T$

---

[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 method[8]
- We develop an MHEM algorithm for VAE-based speech enhancement.

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

1. Initial states $\mathbf{z}^{(0)} = \left\{\mathbf{z}_1^{(0)}, \cdots, \mathbf{z}_T^{(0)}\right\}$
2. Next samples: $\tilde{\mathbf{z}}_t^{(k)}|\mathbf{z}_t^{(k-1)} \sim \mathcal{N}(\mathbf{z}_t^{(k-1)}, \sigma^2\mathbf{I}), \quad \forall t$
3. 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:

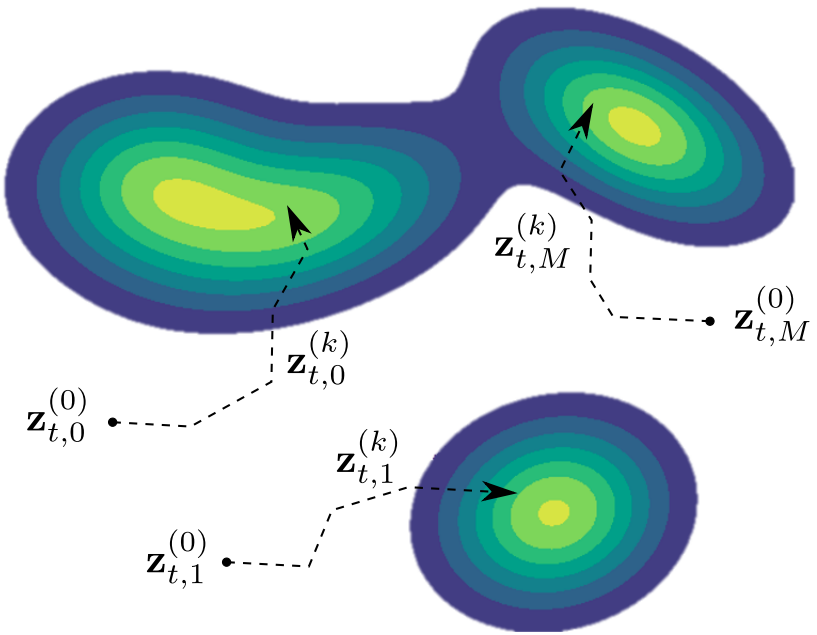
$$\mathbf{z}_{t,i}^{(k)}|\mathbf{z}^{(k-1)} \sim \mathcal{N}\left(\mathbf{z}_{t,i}^{(k-1)} + \frac{\eta}{2} f_\phi(\mathbf{z}^{(k-1)}), \eta\mathbf{I}\right)$$

or

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

[9] R. 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:

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

  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.

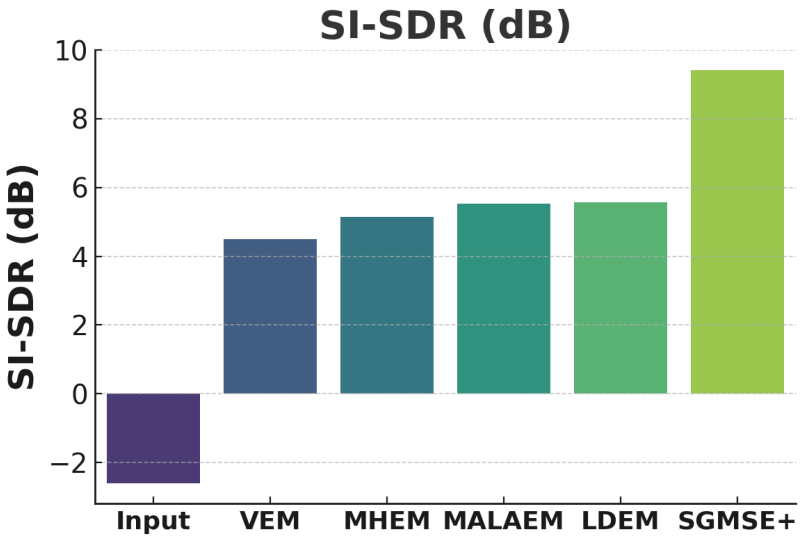
# Experiments

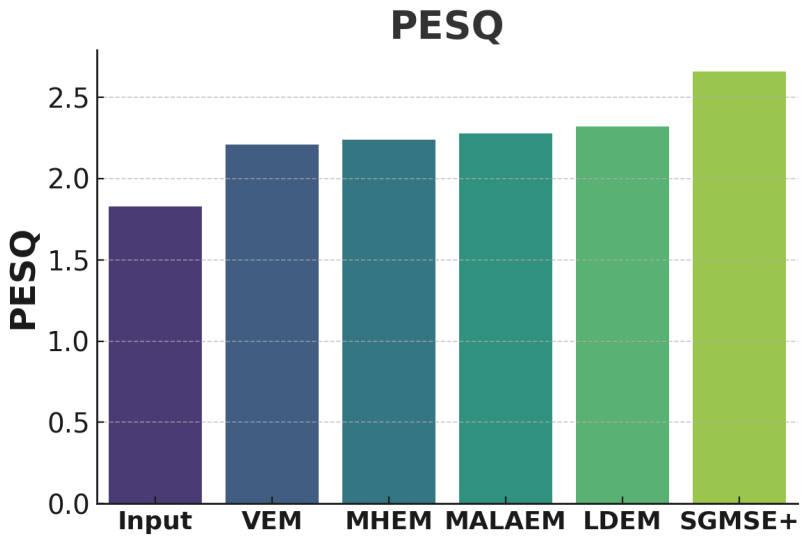
- **Datasets**:
  - WSJ0-QUT *(training & evaluation)*: Matched condition
  - TCD-TIMIT *(evaluation)* : Mismatched condition
- **Parameters**: $K = 1$ (sampling iterations) for LDEM (with $M = 10$ parallel samples), while $K = 10$ for MHEM and MALAEM
- **Baselines**: Pre-trained RVAE[11] (unsupervised) and SGMSE+[12] (supervised).
- **Performance metrics**:
  - SI-SDR (dB) ↑
  - PESQ (perceptual quality) [-0.5, 4.5] ↑
  - STOI (intelligibility) [0, 1] ↑
  - RTF (real-time factor) ↓

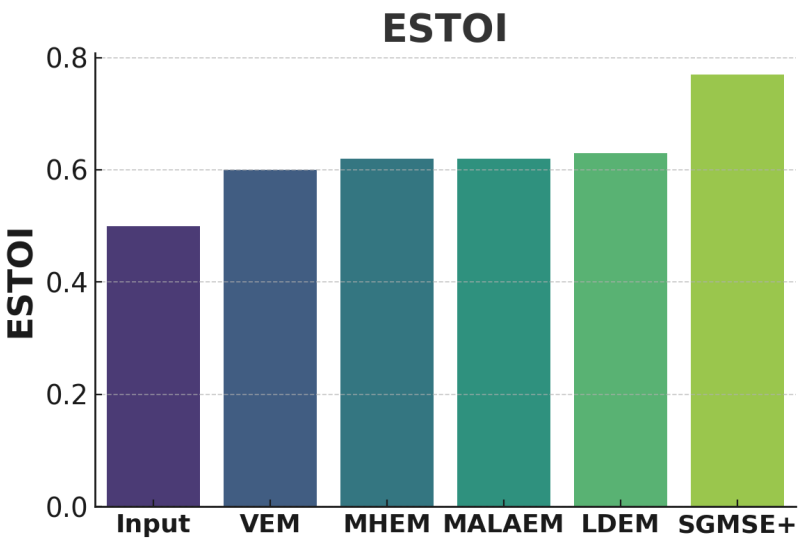
[11] Bie, X., et al., "Unsupervised speech enhancement using dynamical variational autoencoders," IEEE/ACM TASLP, vol. 30, 2022.

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

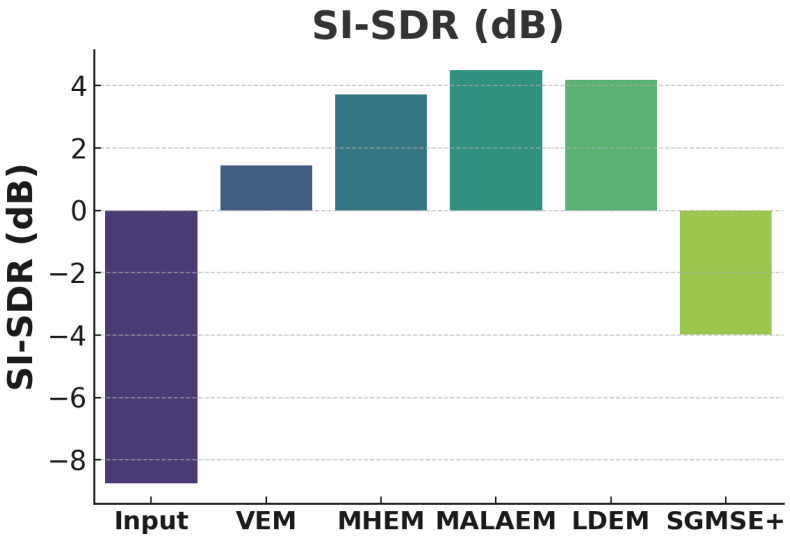
# Results: Matched conditions

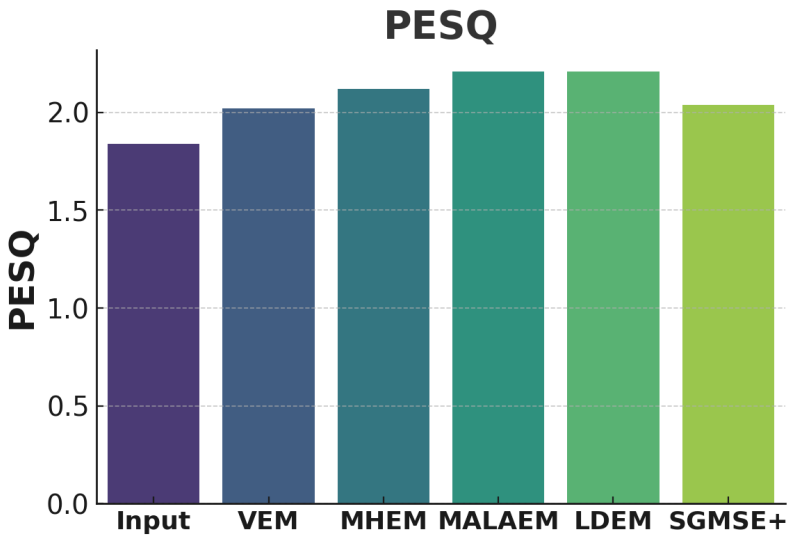


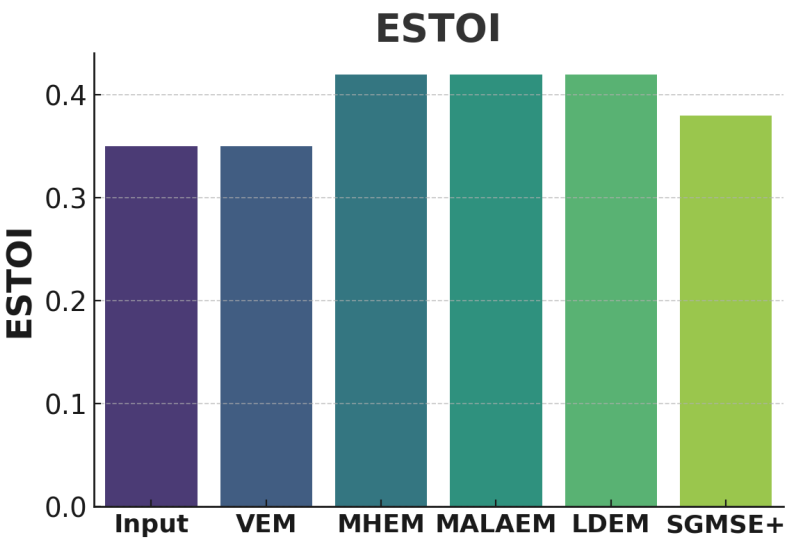



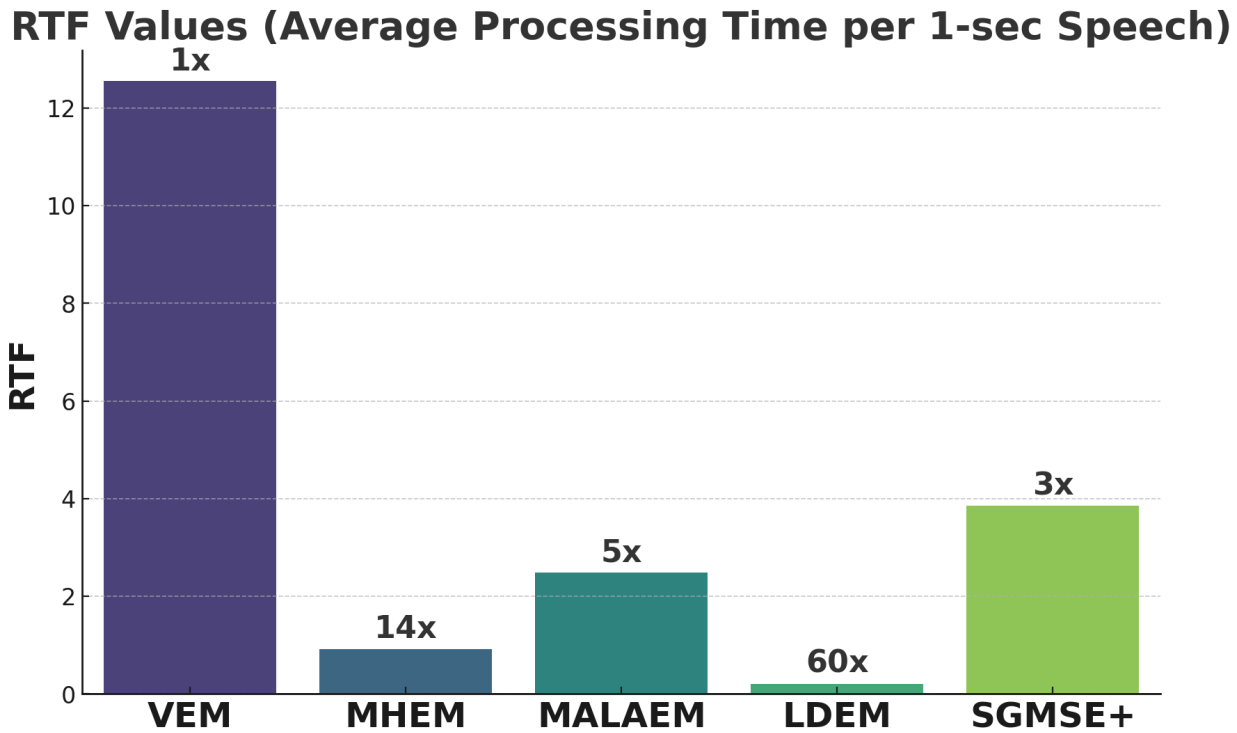
# Results: Mismatched condition



# Results: Real-time Factor (RTF)



# Diffusion models as speech priors

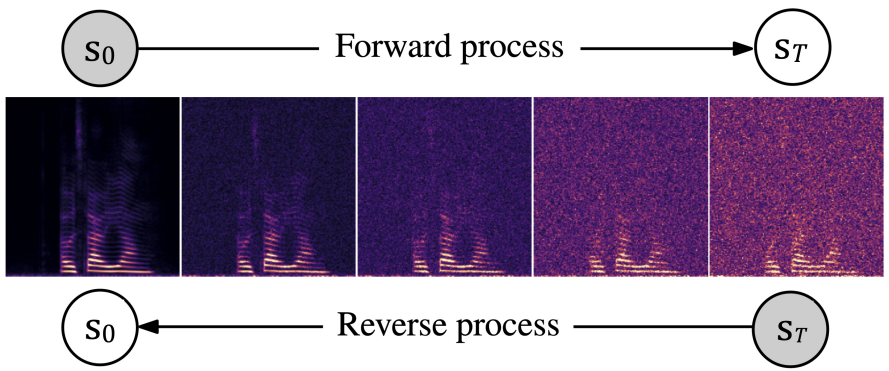
# Diffusion-based speech generative model

▷ Diffusion model for complex-valued clean speech STFT:

- **Forward process:**[13]

$$\mathrm{d}\mathbf{s}_t = \mathbf{f}(\mathbf{s}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \quad \mathbf{f}(\mathbf{s}_t) = -\gamma\mathbf{s}_t$$

$$\boxed{\mathbf{s}_t = \mathrm{e}^{-\gamma t}\mathbf{s} + \sigma(t)\boldsymbol{\epsilon}} \quad \boldsymbol{\epsilon}: \text{ Gaussian noise}$$




- **Reverse process:**

$$\mathrm{d}\mathbf{s}_t = [\mathbf{f}(\mathbf{s}_t) - g(t)^2 \nabla_{\mathbf{s}_t} \log p_t(\mathbf{s}_t)]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

[13] Song, 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:

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

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

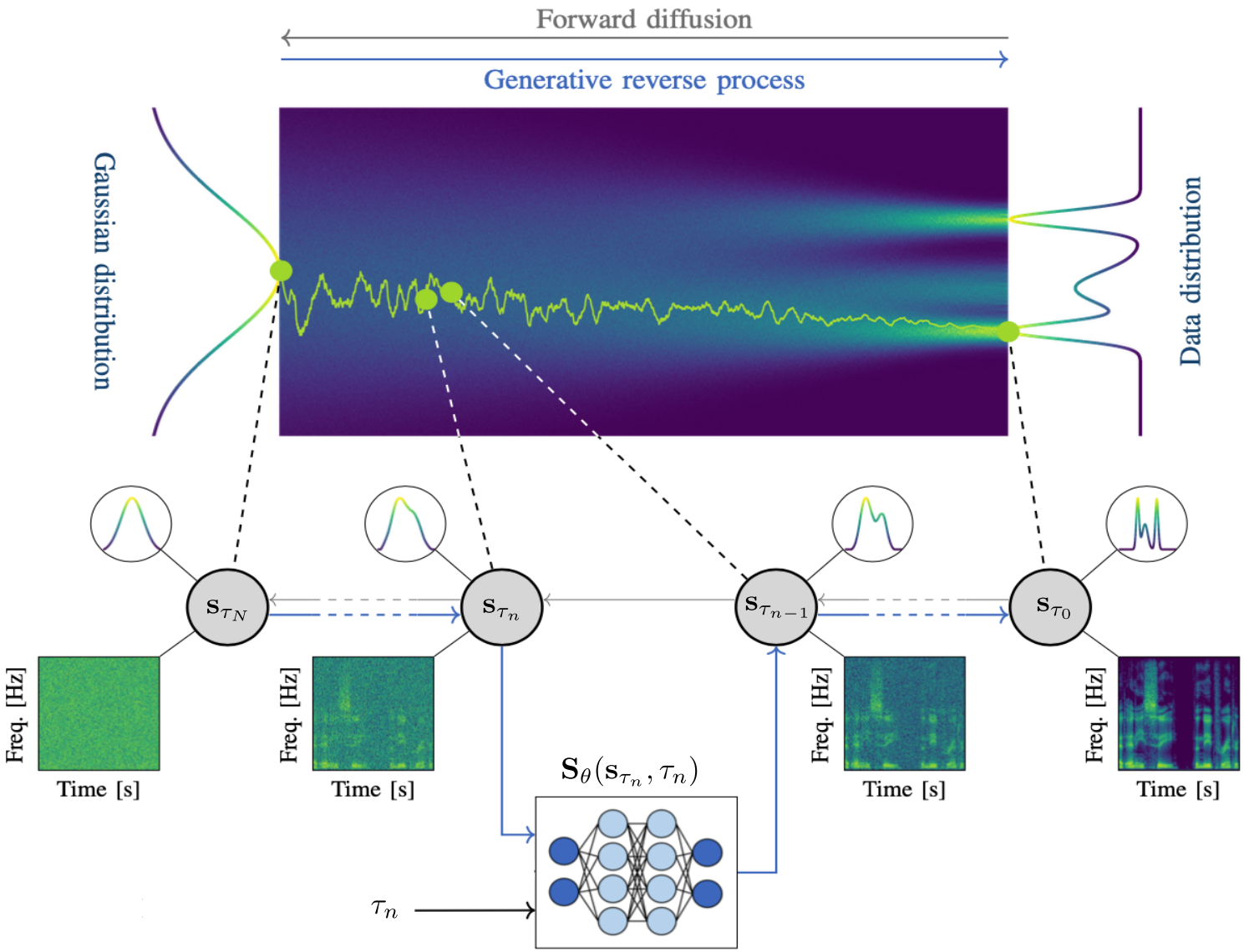
$$\theta^* = \underset{\theta}{\operatorname{argmin}} \, \mathbb{E}_{t,\mathbf{s},\boldsymbol{\zeta},\mathbf{s}_t|\mathbf{s}} \left[ \| \mathbf{S}_\theta(\mathbf{s}_t, t) + \frac{\boldsymbol{\zeta}}{\sigma(t)} \|_2^2 \right], \quad \boldsymbol{\zeta} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{I})$$

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

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

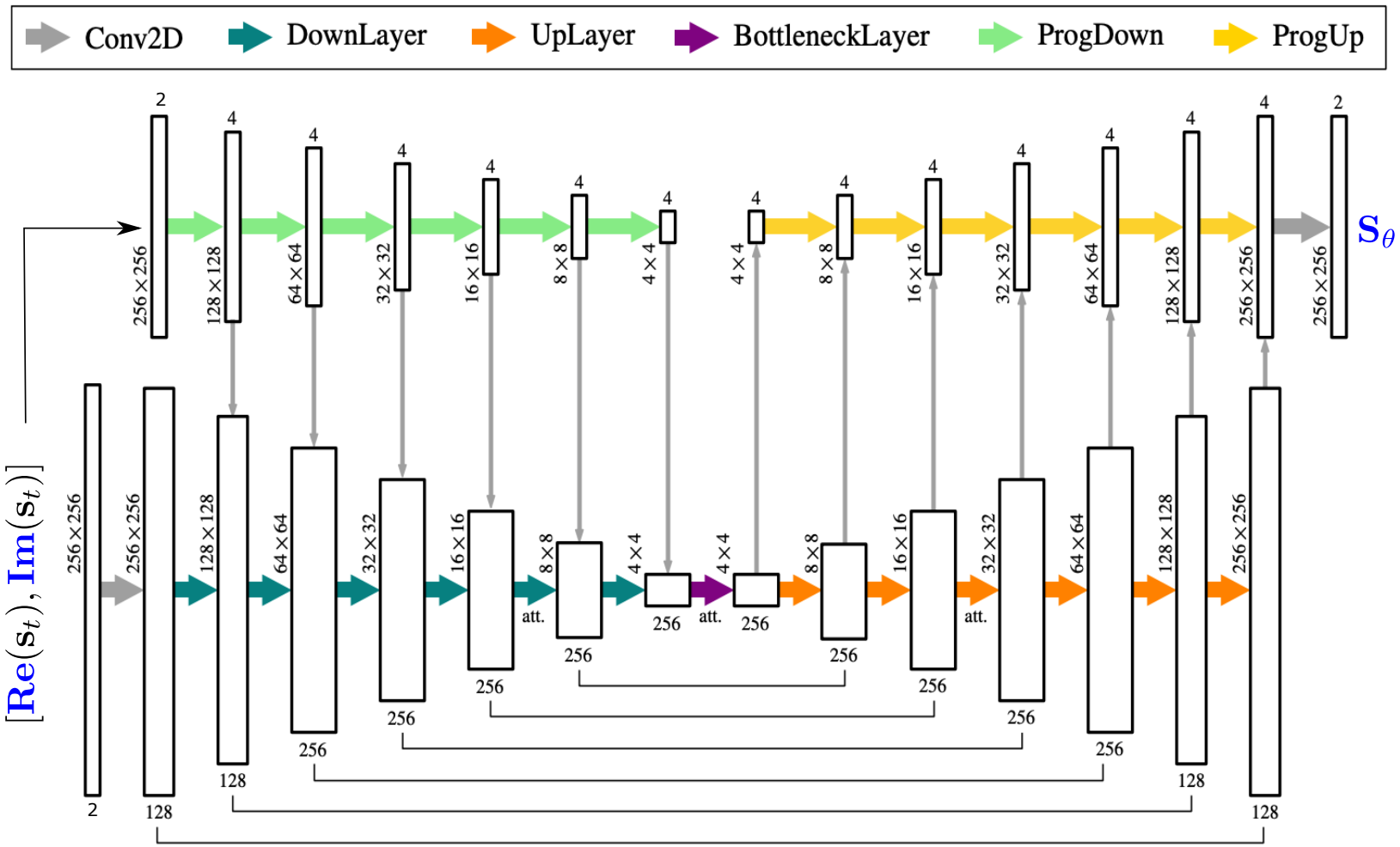
[14] Song, Y., *et al.*, "Score-based generative modelling through stochastic differential equations", ICLR, 2021.

# Diffusion modeling - summary




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

# Score model - architecture



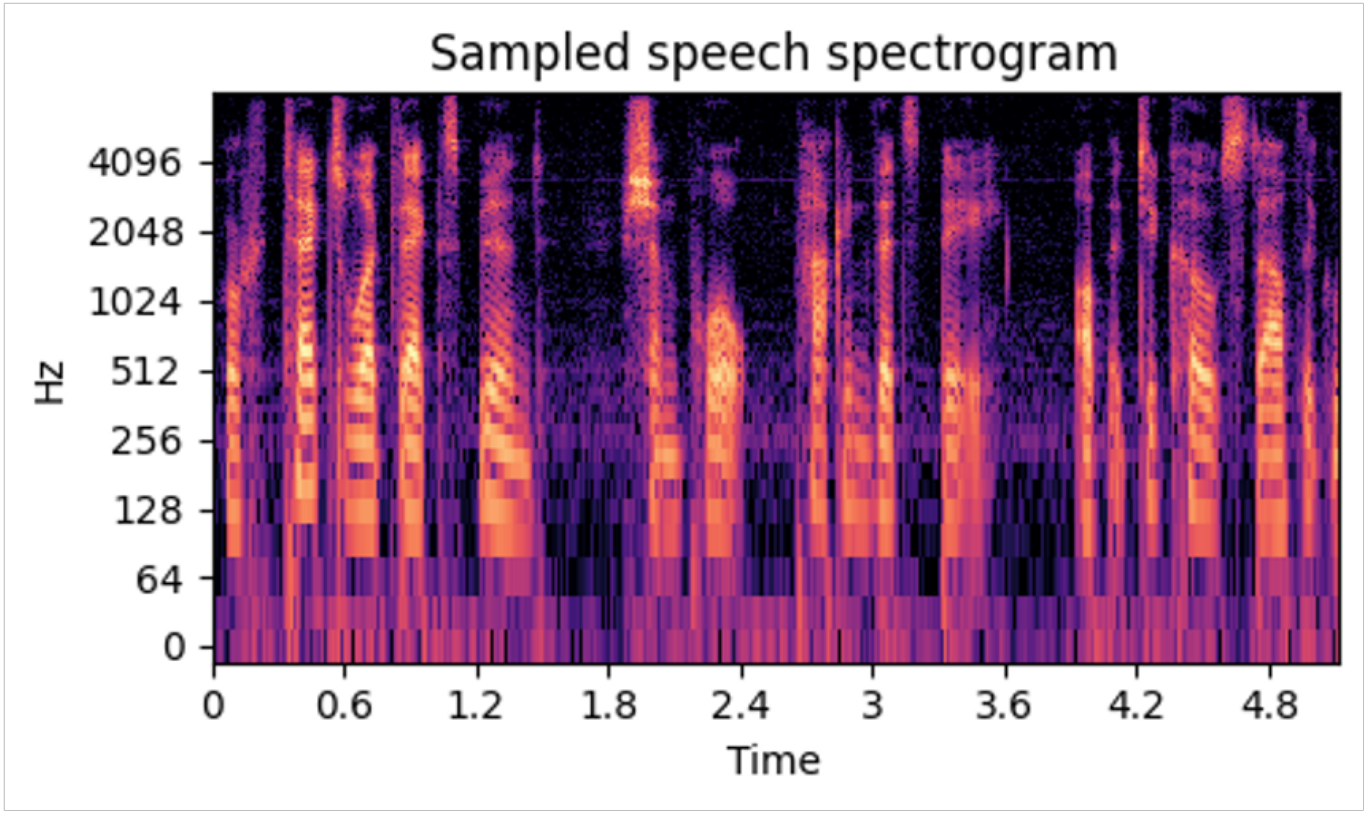

Adapted from: Richter J., et al., "Speech enhancement and dereverberation with diffusion-based generative models," in IEEE/ACM TASLP, 2023.

# Sampling from the prior

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

---

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:

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

⚠ *Intractable*, time-dependent likelihood!

Approximation by the *"noise-perturbed pseudo-likelihood"*[16]

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

[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 $\hat{\mathbf{s}}$

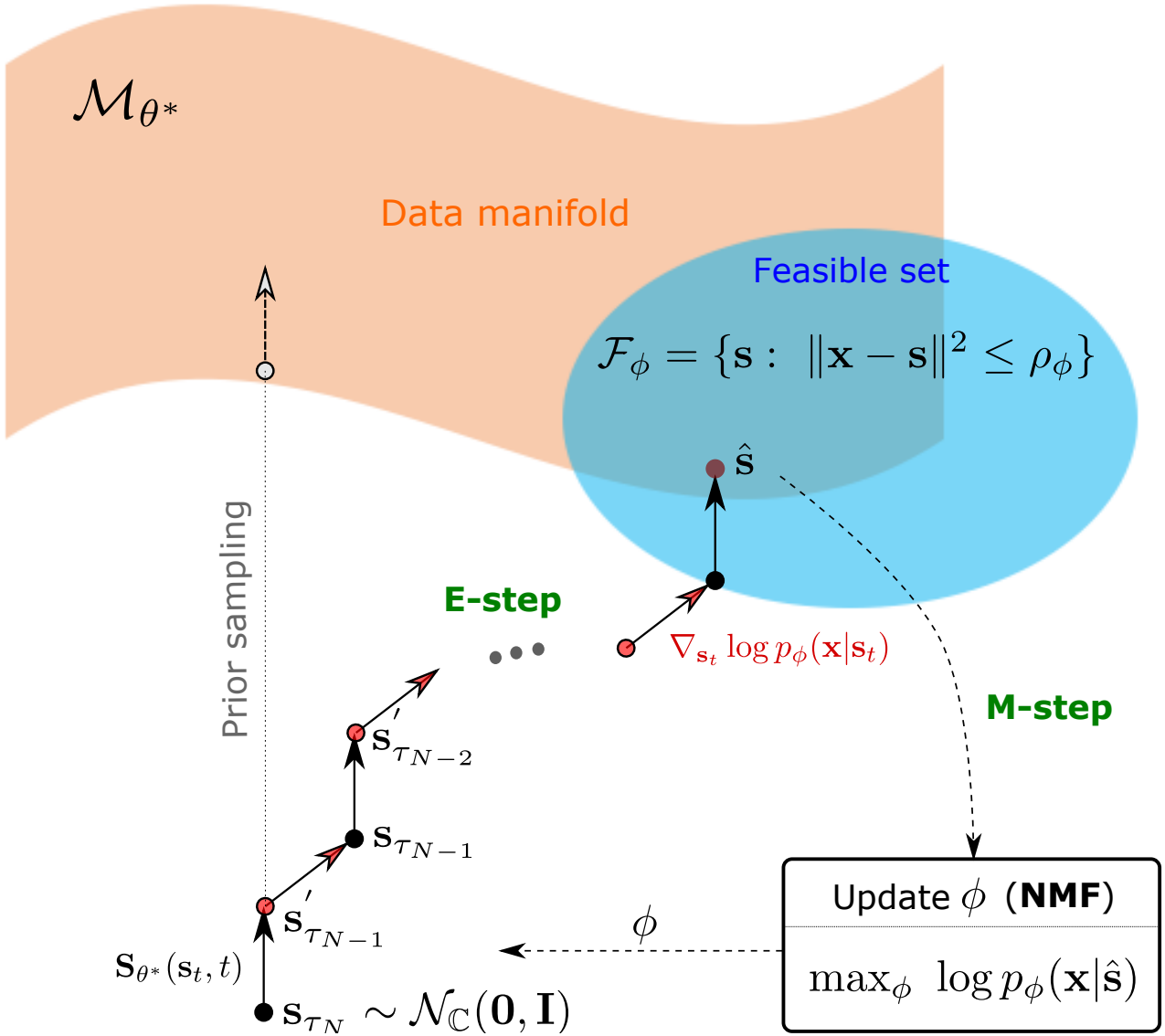
$$\mathrm{d}\mathbf{s}_t = \Big[\mathbf{f}(\mathbf{s}_t) - g(t)^2\Big(\lambda\nabla_{\mathbf{s}_t}\log\tilde{p}_\phi(\mathbf{x}|\mathbf{s}_t) \; + \mathbf{S}_{\theta^*}(\mathbf{s}_t, t)\Big)\Big]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

☞ $\lambda$: weighting parameter to balance prior and likelihood terms.

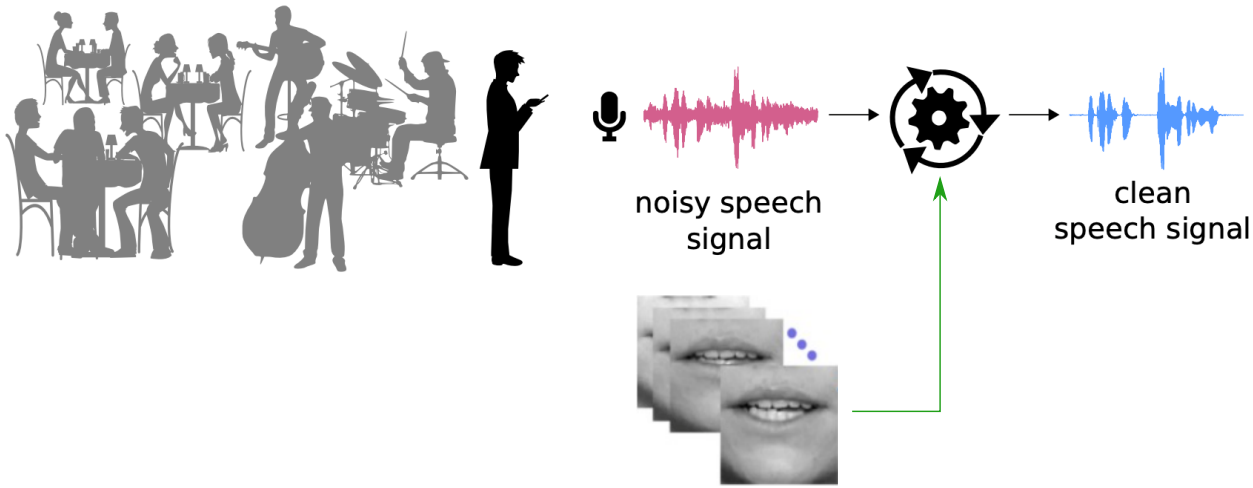
**M-step:** Update noise parameters via NMF

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

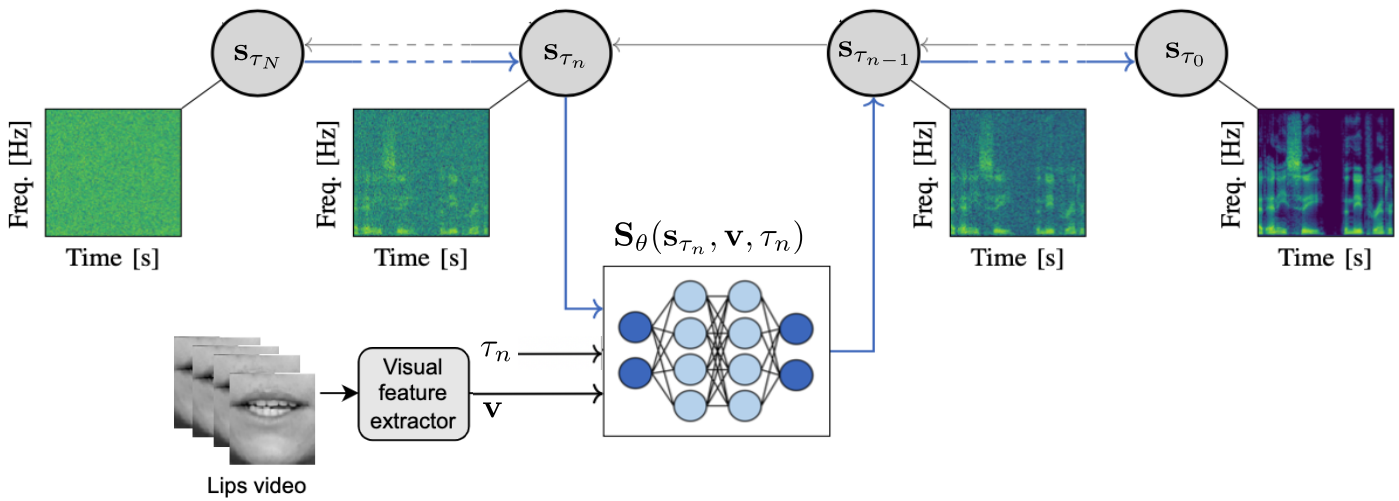
# UDiffSE - A schematic view



# 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

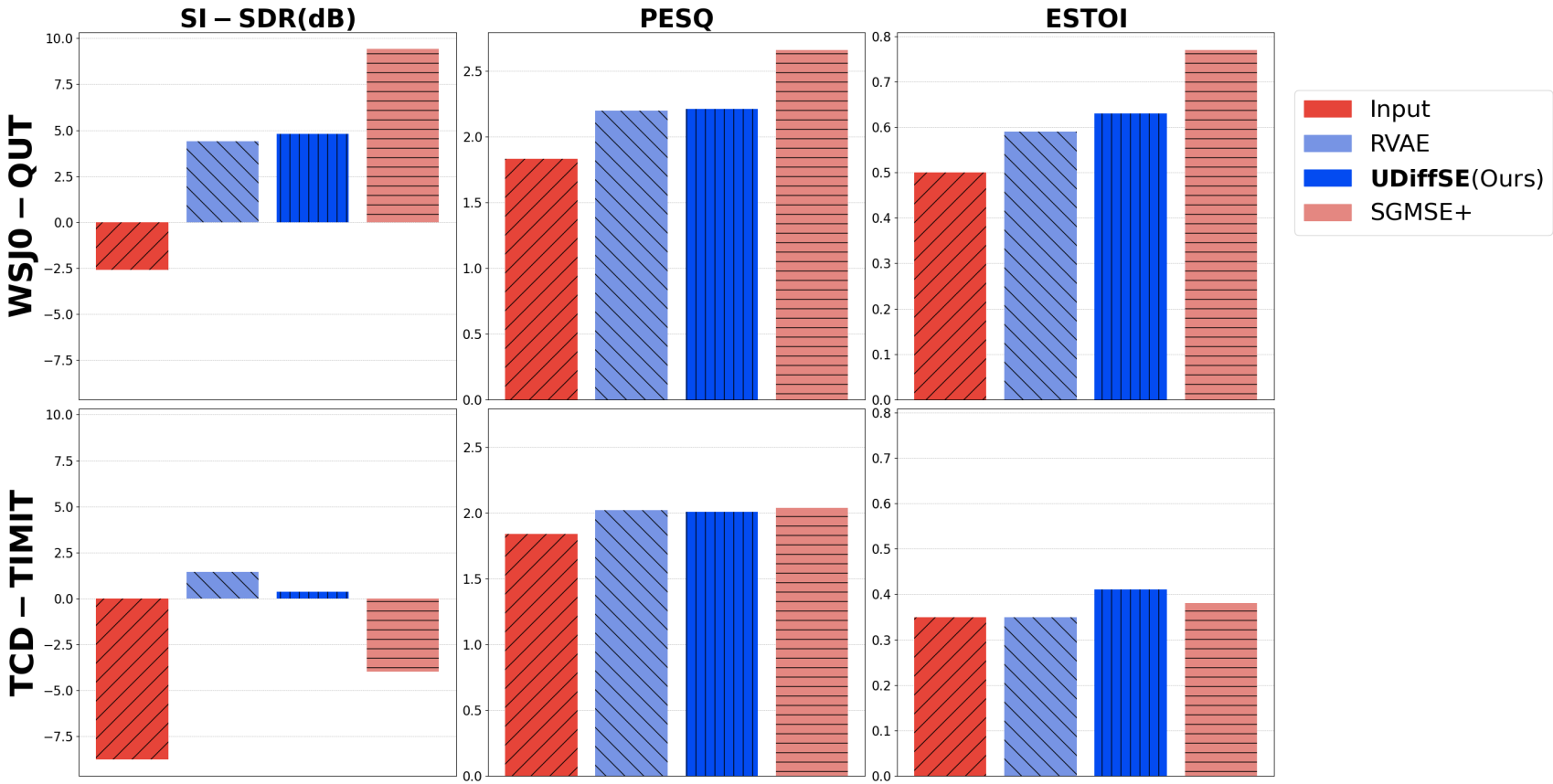
- **Datasets**
  - Training: WSJ0 ($\sim 25$hrs)
  - Testing: WSJ0-QUT (1.5hrs), TCD-TIMIT (45mins)
- **Evaluation Metrics**
  - Objective measures: SI-SDR, ESTOI, PESQ
  - (Pseudo)-subjective measures: DNS-MOS (SIG, BAK, OVRL)
- **Baselines**: RVAE,[18] and SGMSE+,[19] FlowAVSE.[20]
- **Models architecture**. Multi-resolution U-Net as in SGMSE+.

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

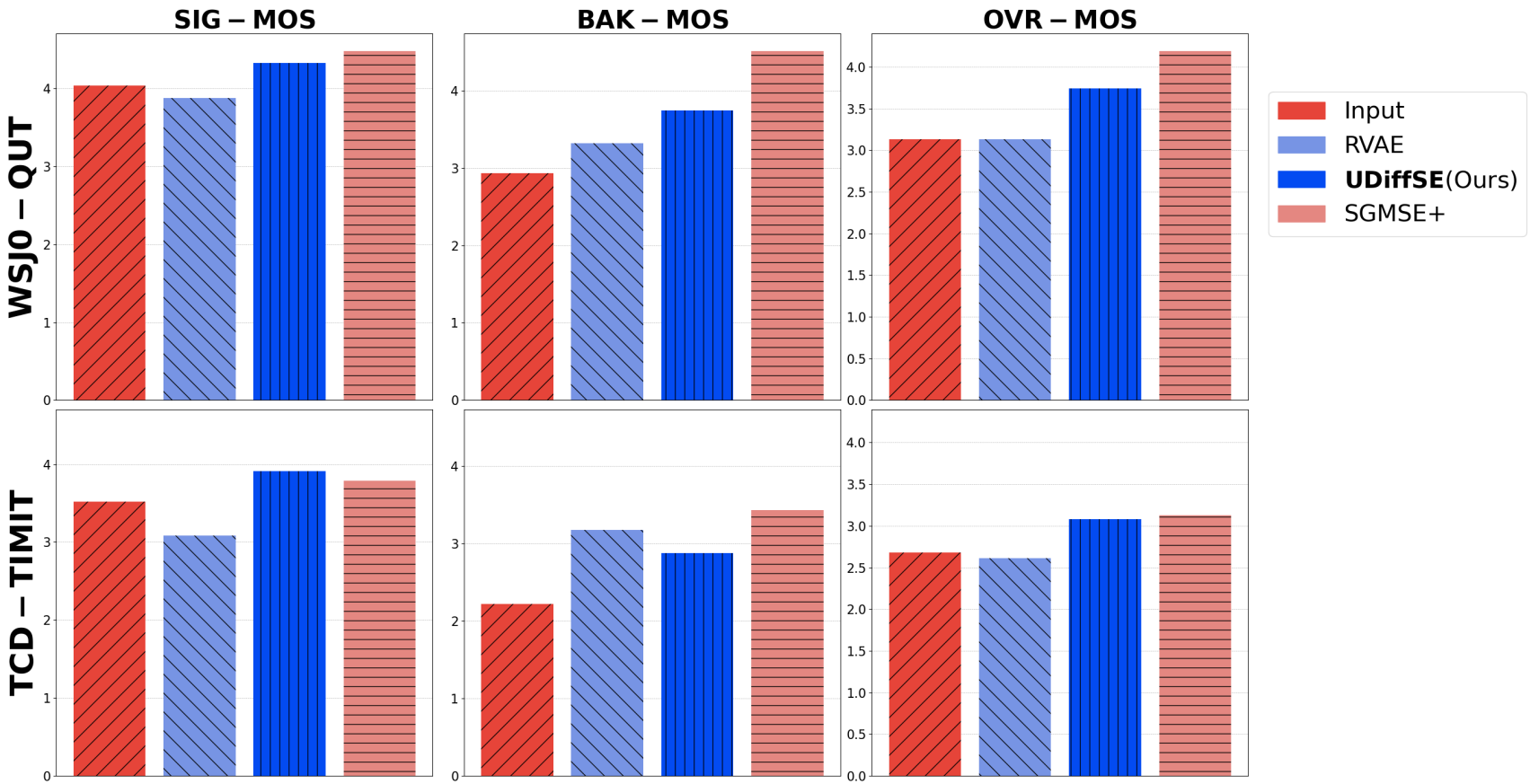
[19] Richter, J., et al., "Speech enhancement and dereverberation with diffusion-based generative models," in IEEE/ACM TASLP, vol. 31, June 2023.

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

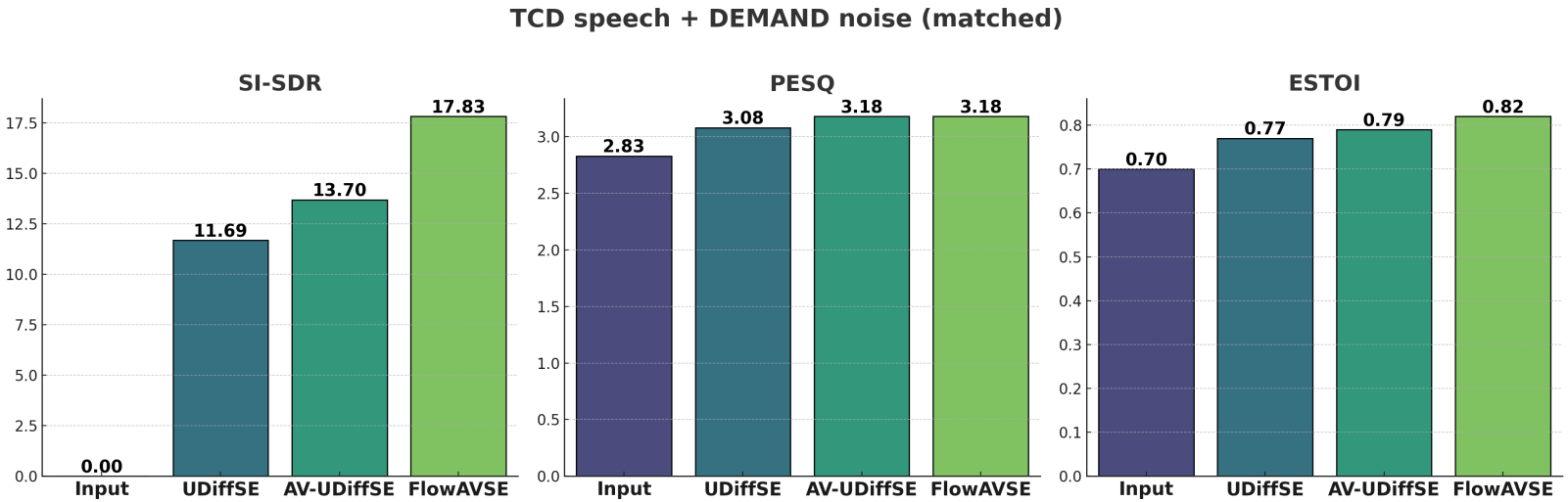
# Results (audio-only SE)

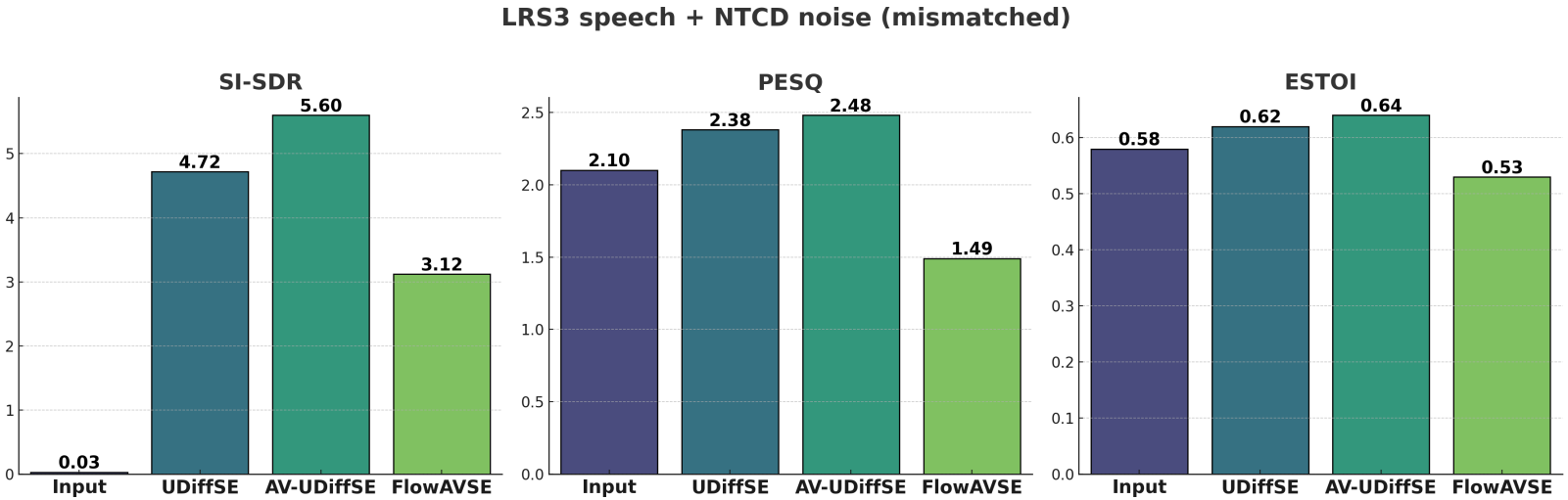


# Results (audio-only SE)



# Results (audio-visual SE)



# 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