Posterior Sampling Algorithms for Unsupervised Speech Enhancement with Recurrent Variational Autoencoder

Mostafa SADEGHI and Romain SERIZEL

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

Overview

- We address **unsupervised speech enhancement** (SE) using a recurrent variational autoencoder (VAE) generative model.
- Inference's bottleneck: **high complexity** of the iterative variational expectation-maximization (VEM) process.
- We propose **efficient sampling-based inference methods** leveraging Langevin dynamics and Metropolis-Hasting algorithms.
- The proposed sampling techniques are shown to improve over the VEM in **speed and performance** significantly.

Unsupervised speech enhancement

Separate the speech and noise signals without training on noise.

Short-time Fourier transform (STFT) domain: $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{b}$

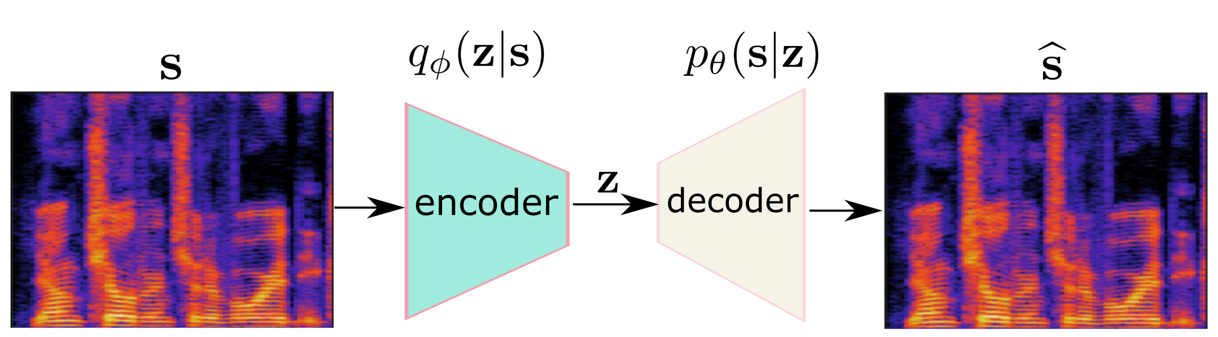
- $\boldsymbol{s} \rightarrow$ clean speech signal with prior $p_\theta(\boldsymbol{s})$
- $\boldsymbol{b} \rightarrow$ noise signal with prior $p_\psi(\boldsymbol{b})$

Training: Learn a parametric prior $p_\theta(\boldsymbol{s})$
Testing: Estimate $\boldsymbol{s}$ using $p_\psi(\boldsymbol{s}|\boldsymbol{x}) \propto p_\psi(\boldsymbol{x}|\boldsymbol{s}) \times p_\theta(\boldsymbol{s})$

Training: learning speech prior

Recurrent VAE (**RVAE**)-based speech generative model [1]:

$$p_\theta(\boldsymbol{s}) = \int p_\theta(\boldsymbol{s}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}, \quad p(\boldsymbol{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$



▷ Learn encoder-decoder parameters over *clean* speech data.

Testing: speech enhancement

Non-negative matrix factorization (**NMF**)-based noise model:

$$p_\psi(\boldsymbol{b}) \sim \mathcal{N}_c\left(\mathbf{0}, \text{diag}(\text{vec}(\mathbf{WH}))\right), \quad \psi = \{\mathbf{W}, \mathbf{H}\}$$

Parameter inference: Variational expectation-maximization (VEM)

- **E-step:** compute posterior $p_\psi(\mathbf{z}|\mathbf{x})$ (Intractable!)
- **M-step:** update parameters:

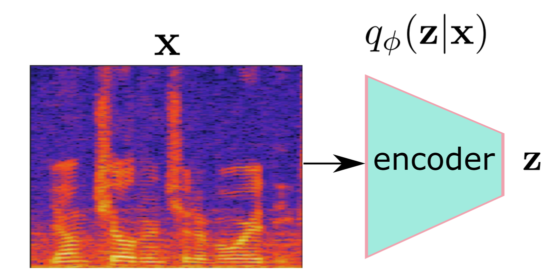
$$\max_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}\{\log p_\psi(\mathbf{x}|\mathbf{z})\}$$

☞ *multiplicative update rules*

VEM-based inference

Computational bottleneck due to the **intractable posterior** during the E-step.

▷ **VEM approach:** fine-tune the pre-trained encoder on $\mathbf{x}$ [1]



Sample from the fine-tuned encoder and estimate the expectation with a **Monte-Carlo average**.

✗ **Computationally expensive**, especially when the encoder has high number of parameters.

Proposed solutions: efficient sampling methods

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

Assume $\mathbf{s} = (\mathbf{s}_1, \cdots, \mathbf{s}_T)$ (STFT time-frames) and associated $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_T)$.

Metropolis-Hastings (MH): Iterative Markov chain Monte Carlo (MCMC) sampling.

- Candidate next samples:

$$\tilde{\mathbf{z}}_t^{(k)}|\mathbf{z}_t^{(k-1)} \sim \mathcal{N}(\mathbf{z}_t^{(k-1)}, \sigma^2\mathbf{I}), \quad \forall t$$

- Accept the new samples with the following probability (*relative posteriors*):

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

Langevin dynamics (LD): Needs only $\nabla_\mathbf{z} \log p_\psi(\mathbf{z}|\mathbf{x})$ (**score function**) for sampling.

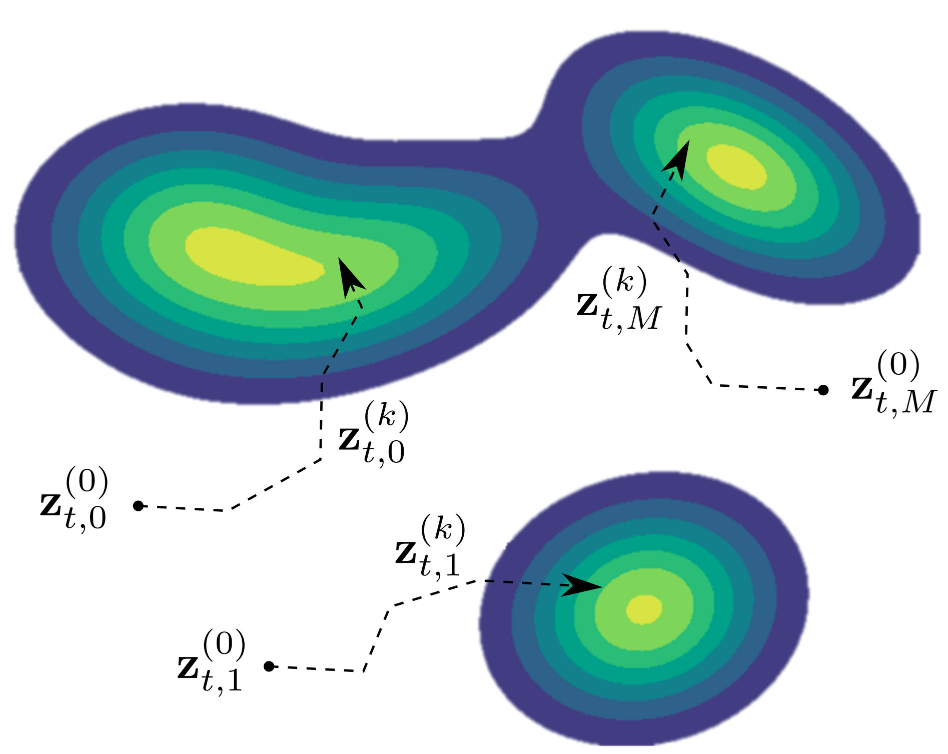
$$f_\psi(\mathbf{z}) = \nabla_\mathbf{z} \log p_\psi(\mathbf{z}|\mathbf{x}) = \nabla_\mathbf{z}\left(\sum_{t=1}^{T} \log p_\psi(\mathbf{x}_t|\mathbf{z}) + \log p(\mathbf{z}_t)\right)$$

- *Multiple* samples per time-frame:

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

- Next samples via LD:

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



☞ **Gradient ascent** steps on score function + **noise injection** to better explore posterior space.

☞ No acceptance/rejection mechanism, unlike MH.

Metropolis-Adjusted Langevin Algorithm (MALA):

Add an acceptance/rejection mechanism to LD.

- Candidate next samples:

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

- Accept or reject the new samples:

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

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.

Experiments

- **Datasets**: WSJ0-QUT (training & evaluation) and TCD-TIMIT (evaluation)
- **Parameters**: $K = 1$ (sampling iterations) for LDEM, while $K = 10$ for MHEM and MALAEM
- **Baseline**: Pre-trained RVAE [1] (unsupervised) and SGMSE+ [2] (supervised).

Table 1: Speech enhancement performance metrics.

Metric		SI-SDR (dB)	PESQ	ESTOI
Input (**WSJ0-QUT**)		-2.60 ± 0.16	1.83 ± 0.02	0.50 ± 0.01
RVAE	VEM [1]	4.50 ± 0.21	2.21 ± 0.02	0.60 ± 0.01
	MHEM	5.15 ± 0.20	2.24 ± 0.02	0.62 ± 0.01
	MALAEM	5.52 ± 0.21	2.28 ± 0.02	0.62 ± 0.01
	LDEM	<u>5.58</u> ± 0.20	<u>2.32</u> ± 0.02	<u>0.63</u> ±0.01
SGMSE+ [2]		**9.41** ± 0.18	**2.66** ± 0.02	**0.77** ± 0.01
Input (**TCD-TIMIT**)		-8.74 ± 0.29	1.84 ± 0.02	0.35 ± 0.01
RVAE	VEM [1]	1.44 ± 0.30	2.02 ± 0.02	0.35 ± 0.01
	MHEM	3.72 ± 0.27	2.12 ± 0.02	**0.42** ± 0.01
	MALAEM	**4.49** ± 0.29	**2.21** ± 0.02	**0.42** ± 0.01
	LDEM	<u>4.18</u> ± 0.29	**2.21** ± 0.02	**0.42** ± 0.01
SGMSE+ [2]		-3.97 ± 0.41	2.04 ± 0.02	0.38 ± 0.01

Table 2: RTF values (average processing time per 1-sec speech).

VEM	MHEM	MALAEM	LDEM	SGMSE+
12.55 ± 0.01	<u>0.92</u> ± 0.01	2.49 ± 0.01	**0.21** ± 0.01	3.85 ± 0.01

▷ Proposed methods surpass VEM in RVAE algorithms, especially in *mismatched* conditions, showing better generalizability.

▷ LDEM consistently scores highest or near-highest in all metrics, underlining its effectiveness.

▷ SGMSE+ excels in *matched* conditions but lags in *mismatched* ones (*generalization issue of supervised methods*).

▷ Proposed methods, especially LDEM, are much faster than VEM.

References

[1] X. Bie, et al., "*Unsupervised speech enhancement using dynamical variational autoencoders*," IEEE/ACM TASLP, vol. 30, pp. 2993–3007, 2022.

[2] J. Richter et al., "*Speech enhancement and dereverberation with diffusion-based generative models*," in IEEE/ACM TASLP, vol. 31, pp. 2351-2364, June 2023.