# Posterior sampling algorithms for unsupervised speech enhancement with recurrent variational autoencoder

Mostafa Sadeghi, Romain Serizel

# POSTERIOR SAMPLING ALGORITHMS FOR UNSUPERVISED SPEECH ENHANCEMENT WITH RECURRENT VARIATIONAL AUTOENCODER

*Mostafa Sadeghi, Romain Serizel*

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

## ABSTRACT

In this paper, we address the unsupervised speech enhancement problem based on recurrent variational autoencoder (RVAE). This approach offers promising generalization performance over the supervised counterpart. Nevertheless, the involved iterative variational expectation-maximization (VEM) process at test time, which relies on a variational inference method, results in high computational complexity. To tackle this issue, we present efficient sampling techniques based on Langevin dynamics and Metropolis-Hasting algorithms, adapted to the EM-based speech enhancement with RVAE. By directly sampling from the intractable posterior distribution within the EM process, we circumvent the intricacies of variational inference. We conduct a series of experiments, comparing the proposed methods with VEM and a state-of-the-art supervised speech enhancement approach based on diffusion models. The results reveal that our sampling-based algorithms significantly outperform VEM, not only in terms of computational efficiency but also in overall performance. Furthermore, when compared to the supervised baseline, our methods showcase robust generalization performance in mismatched test conditions.

***Index Terms*—** Unsupervised speech enhancement, deep generative model, variational autoencoder, posterior sampling.

## 1. INTRODUCTION

Speech enhancement is a fundamental signal processing technique, aiming to improve the quality and intelligibility of a noisy speech signal corrupted by acoustic noise [1]. Over the past few years, and with the unprecedented success of deep learning, speech enhancement approaches have shifted from traditional statistical methods to data-driven approaches based on deep neural networks (DNNs) [2–6]. Predominantly, current DNN-based speech enhancement techniques adopt a supervised (discriminative) paradigm, wherein a DNN is trained to map noisy speech inputs to their corresponding clean counterparts, leading to state-of-the-art performance. However, a notable challenge pervasive in these methods concerns generalization to test conditions not encountered during training, such as distinct noise types and noise levels that deviate from training conditions.

In contrast, unsupervised speech enhancement methods based on deep generative models do not learn noise characteristics during the training process [7–10]. Specifically, a deep generative model, most commonly based on the variational auto-encoder (VAE) [11], is trained solely on clean speech signals. This trained model then serves as a prior distribution for estimating clean speech from noisy input using an expectation-maximization (EM) approach. This gives them a generalization advantage over discriminative approaches. However, unsupervised methods remain significantly less explored than their supervised counterparts and suffer from some challenges, including their notably high computational complexity. This complexity originates from the iterative EM process during inference, which requires sampling from an intractable posterior distribution. For instance, the current state-of-the-art method for unsupervised speech enhancement relies on recurrent VAE (RVAE) [8, 12], as a dynamical and more efficient version of the standard VAE. This approach adopts a variational EM (VEM) strategy, involving the fine-tuning of the trained encoder at each EM iteration on the input noisy speech. Its computational complexity thus grows with the complexity (number of parameters) of the encoder.

To address this issue, we propose alternative, more efficient posterior sampling-based methods for speech enhancement with RVAE. The first approach extends the Langevin dynamics EM (LDEM) method for standard, non-dynamical VAE presented in [13] to RVAE. This technique involves sampling from the intractable posterior using gradient descent steps combined with Gaussian noise injection. Additionally, we develop a Metropolis-Hastings (MH) sampling technique [14], relying on a proposal and acceptance/rejection mechanism, to generate a sequence of samples. Lastly, a Metropolis-adjusted Langevin algorithm (MALA) [15] is proposed, combining the strengths of both LDEM and MH methods. We assess the effectiveness of these algorithms for RVAE-based speech enhancement by comparing them to the VEM method and a state-of-the-art supervised speech enhancement approach based on diffusion models [4], in both matched and mismatched test conditions. The results demonstrate that our proposed speech enhancement algorithms outperform VEM significantly in terms of performance and computational efficiency. Furthermore, they exhibit more robust generalization performance when compared to the supervised baseline method.

The paper is organized as follows: In Section 2, we present an overview of unsupervised speech enhancement based on RVAE. Section 3 introduces the proposed posterior sampling methods. Our experimental results are detailed in Section 4. Lastly, Section 5 provides some conclusions.

## 2. BACKGROUND

### 2.1. RVAE as a deep speech prior

We denote by $\mathbf{s} \triangleq \mathbf{s}_{1:T} = \{\mathbf{s}_1, \ldots, \mathbf{s}_T\}$ a sequence of clean speech time-frequency representations computed using short-time Fourier transform (STFT), where $\mathbf{s}_t = [s_{ft}]_{f=1}^{F} \in \mathbb{C}^F$. RVAE [12], as a latent variable-based deep generative model, considers the following

This work was supported by the French National Research Agency (ANR) under the project REAVISE (ANR-22-CE23-0026-01). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria, and including CNRS, RENATER, and several universities as well as other organizations (see https://www.grid5000.fr).

generative model for the speech time frames $\mathbf{s}$:

$$p_\theta(\mathbf{s}, \mathbf{z}) = \prod_{t=1}^{T} p_\theta(\mathbf{s}_t|\mathbf{z})p(\mathbf{z}_t) \tag{1}$$

where $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$, $\mathbf{z}_t \in \mathbb{R}^L$ ($L \ll F$), are low-dimensional latent variables associated with $\mathbf{s}$. Moreover,

$$p_\theta(\mathbf{s}_t|\mathbf{z}) = \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\boldsymbol{v}_{\theta,t}(\mathbf{z}))) \tag{2}$$

is a circularly-symmetric complex Gaussian distribution, with a diagonal covariance matrix whose entries, given by $\boldsymbol{v}_{\theta,t}(\mathbf{z})$, are modeled by a recurrent neural network (RNN), called decoder. Here, $\boldsymbol{v}_{\theta,t}(\mathbf{z})$ refers to the output at time frame $t$ of the RNN with $\mathbf{z}$ as the input. This dynamical modeling makes RVAE more efficient than the standard VAE. Similar to VAE, the prior $p(\mathbf{z}_t)$ is set to a standard Gaussian distribution.

Training the generative model (1) involves learning the RNN parameters $\theta$ following an EM procedure. The intractable posterior $p_\theta(\mathbf{z}|\mathbf{s})$ is approximated with a parametric Gaussian distribution as follows

$$q_\phi(\mathbf{z}|\mathbf{s}) = \prod_{t=1}^{T} q_\phi(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{s}_{t:T}), \tag{3}$$

where similarly as in (2), the mean and variance are modeled via an RNN, called encoder, with parameters denoted $\phi$. The encoder and decoder parameters, i.e., $\{\theta, \phi\}$, are then jointly learned by optimizing an evidence lower-bound [11].

### 2.2. Variational EM for speech enhancement

The observation model for speech enhancement is assumed to be $\mathbf{x}_t = \mathbf{s}_t + \mathbf{b}_t$, with $\mathbf{b}_t$ corresponding to noise. As a statistical model for clean speech $\mathbf{s}_t$, the pretrained RVAE model, i.e., $p_\theta(\mathbf{s}, \mathbf{z})$ is used. Moreover, noise is modeled based on a non-negative matrix factorization (NMF) model [7], where a circularly symmetric Gaussian form $p_\psi(\mathbf{b}_t) \sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}([\mathbf{WH}]_t))$ is considered. The non-negative matrices $\mathbf{W}, \mathbf{H}$ form the noise parameters $\psi$ to be learned from $\mathbf{x}$. This is done following an EM approach, that is

$$\psi^* = \underset{\psi}{\mathrm{argmax}}\ \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}\left\{\log p_\psi(\mathbf{x}, \mathbf{z})\right\}, \tag{4}$$

where $p_\psi(\mathbf{x}, \mathbf{z}) = \prod_t p_\psi(\mathbf{x}_t|\mathbf{z})p(\mathbf{z}_t)$, with likelihood computed as $p_\psi(\mathbf{x}_t|\mathbf{z}) = \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\boldsymbol{v}_{\theta,t}(\mathbf{z}) + [\mathbf{WH}]_t))$ [12]. Here, the posterior $p_\psi(\mathbf{z}|\mathbf{x})$ in (4), needed for the E-step, is intractable to compute. The variational EM (VEM) approach proposed in [12] fine-tunes the pretrained encoder $q_\phi(\mathbf{z}|\mathbf{s})$ on $\mathbf{x}$ at each E-step, to serve as an approximation of $p_\psi(\mathbf{z}|\mathbf{x})$. This approach aligns with the principles of standard variational inference methods. Then, at the M-step, using latent variables sampled from the approximate posterior, the NMF parameters are updated by optimizing (4). Once the EM steps converge, the speech signal is estimated as $\hat{\mathbf{s}} = \mathbb{E}_{p_{\psi^*}(\mathbf{s}|\mathbf{x})}\{\mathbf{s}\}$.

## 3. POSTERIOR SAMPLING ALGORITHMS

In this section, we present our EM-based speech enhancement frameworks, utilizing RVAE as a deep speech prior. These frameworks share a common structure but vary in the E-step, where each employs a distinct strategy to draw samples from the intractable posterior $p_\psi(\mathbf{z}|\mathbf{x})$. We provide a concise summary of the overall speech enhancement process in Algorithm 1. Specifically, the first approach extends the LDEM method, as proposed in [13], to RVAE. The second approach utilizes the Metropolis-Hastings sampling algorithm, while the third algorithm is a Metropolis-adjusted version of LDEM.

**Algorithm 1** EM-based speech enhancement

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

### 3.1. Langevin dynamics (LD)

In the conventional VAE-based speech enhancement method described in [13], the process of sampling from the posterior distribution is carried out independently for each latent variable. To capture temporal dependencies, a total variation (TV) regularization term is introduced. However, in the context of RVAE, latent variables are naturally interconnected through an RNN model, making the TV regularization term redundant.

Langevin dynamics enables the generation of a sequence of samples from the posterior distribution $p_\psi(\mathbf{z}|\mathbf{x})$ solely using its score function, defined as follows:

$$\begin{aligned} f_\psi(\mathbf{z}) &= \nabla_\mathbf{z} \log p_\psi(\mathbf{z}|\mathbf{x}) \\ &= \nabla_\mathbf{z}\Big(\log p_\psi(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z})\Big) \\ &= \nabla_\mathbf{z}\Big(\sum_{t=1}^{T} \log p_\psi(\mathbf{x}_t|\mathbf{z}) + \log p(\mathbf{z}_t)\Big). \end{aligned} \tag{5}$$

In contrast to VAE, this score function cannot be decomposed over individual latent variables, meaning that $f_\psi(\mathbf{z}) \neq \sum_t f_\psi(\mathbf{z}t)$. Consequently, each $\mathbf{z}_t$ must be sampled individually, akin to the sequential Gibbs sampling procedure [14]. This sequential approach would significantly increase complexity. Instead, we adopt a parallel sampling strategy, wherein all latent variables are sampled simultaneously. Furthermore, following the methodology employed in LDEM for VAE, we generate multiple samples for each latent variable to obtain a more robust and efficient approximation of the expectation in (4). Therefore, starting from $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_T)$, we initially draw $M$ distinct samples (states) for each latent variable $\mathbf{z}_t$, denoted as $\bar{\mathbf{z}}^{(0)} = \left\{\mathbf{z}_{t,i}^{(0)}\right\}_{t,i}$, with $t = 1, \ldots T$ and $i = 1, \ldots M$, using a random walk approach by sampling from the following proposal distribution:

$$\mathbf{z}_{t,i}^{(0)}|\mathbf{z}_t \sim \mathcal{N}(\mathbf{z}_t, \sigma^2\mathbf{I}), \quad \forall t, i \tag{6}$$

or $\mathbf{z}_{t,i}^{(0)} = \mathbf{z}_t + \sigma\boldsymbol{\epsilon}_{t,i}$, where $\boldsymbol{\epsilon}_{t,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\sigma^2 > 0$. The next states are produced by sampling from the following distribution:

$$\mathbf{z}_{t,i}^{(k)}|\bar{\mathbf{z}}^{(k-1)} \sim \mathcal{N}(\mathbf{z}_{t,i}^{(k-1)} + \frac{\eta}{2} f_\psi(\bar{\mathbf{z}}^{(k-1)}), \eta\mathbf{I}), \tag{7}$$

or, equivalently

$$\mathbf{z}_{t,i}^{(k)} = \mathbf{z}_{t,i}^{(k-1)} + \frac{\eta}{2} f_\psi(\bar{\mathbf{z}}^{(k-1)}) + \sqrt{\eta}\boldsymbol{\zeta}_{t,i}, \tag{8}$$

**Algorithm 2** LD sampler

**Inputs:** $\bar{\mathbf{z}}^{(0)} = \left\{\mathbf{z}_{t,i}^{(0)}\right\}_{t,i}$, $\mathcal{H}$ (hyperparameters).
**for** $k = 1, \cdots, K$ **do**
  $\boldsymbol{\zeta} = \left\{\boldsymbol{\zeta}_{t,i}\right\}_{t,i}$, with $\boldsymbol{\zeta}_{t,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\bar{\mathbf{z}}^{(k)} = \bar{\mathbf{z}}^{(k-1)} + \frac{\eta}{2} f_\psi(\bar{\mathbf{z}}^{(k-1)}) + \sqrt{\eta}\boldsymbol{\zeta}$,
**end for**
**Output:** $\bar{\mathbf{z}}^{(K)} = \left\{\mathbf{z}_{t,i}^{(K)}\right\}_{t,i}$

**Algorithm 3** MH sampler

**Require:** $\mathbf{z}^{(0)} = \left\{\mathbf{z}_{t}^{(0)}\right\}_{t}$, $\mathcal{H}$ (hyperparameters).
**for** $k = 1, \cdots, K$ **do**
  $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_t\}_t$, with $\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\tilde{\mathbf{z}}^{(k)} = \mathbf{z}^{(k-1)} + \sqrt{\eta}\boldsymbol{\zeta}$,
  Accept $\tilde{\mathbf{z}}_t^{(k)}$ $(\forall t)$ according to (10)
**end for**
**Output:** $\bar{\mathbf{z}} = \left\{\mathbf{z}^{(k)}\right\}_{k>k_{\text{burn-in}}}$

**Algorithm 4** MALA sampler

**Require:** $\mathbf{z}^{(0)} = \left\{\mathbf{z}_{t}^{(0)}\right\}_{t}$, $\mathcal{H}$ (hyperparameters).
**for** $k = 1, \cdots, K$ **do**
  $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_t\}_t$, with $\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\tilde{\mathbf{z}}^{(k)} = \mathbf{z}^{(k-1)} + \frac{\eta}{2} f_\psi(\mathbf{z}^{(k-1)}) + \sqrt{\eta}\boldsymbol{\zeta}$,
  Accept $\tilde{\mathbf{z}}_t^{(k)}$ $(\forall t)$ according to (12)
**end for**
**Output:** $\bar{\mathbf{z}} = \left\{\mathbf{z}^{(k)}\right\}_{k>k_{\text{burn-in}}}$

where $\boldsymbol{\zeta}_{t,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\eta > 0$ is a step size. The added noise introduces stochasticity, which enhances exploration within the high-density regions of the posterior. The generated samples converge to the true posterior distribution under some regularity conditions [16]. The LD sampler is summarized in Algorithm 2.

### 3.2. Metropolis-Hastings (MH)

Metropolis-Hastings (MH) [17] is a Markov chain Monte Carlo (MCMC) sampling technique for generating a sequence of samples from a probability distribution. It begins with initial states and iteratively proposes new states using a typically Gaussian proposal distribution. Candidate states are accepted or rejected based on defined probabilities.

For RVAE, similarly as done in the LD sampler, the MH algorithm generates samples for all the latent variables in parallel. More precisely, starting from some initial states $\mathbf{z}^{(0)} = \left\{\mathbf{z}_1^{(0)}, \cdots, \mathbf{z}_T^{(0)}\right\}$, at the $k$-th iteration, a sequence of candidate states, denoted $\tilde{\mathbf{z}}^{(k)}$, are sampled from the following proposal distribution

$$\tilde{\mathbf{z}}_t^{(k)} | \mathbf{z}_t^{(k-1)} \sim \mathcal{N}(\mathbf{z}_t^{(k-1)}, \sigma^2 \mathbf{I}), \quad \forall t. \tag{9}$$

Each candidate state $\tilde{\mathbf{z}}_t^{(k)}$ in the sequence $\tilde{\mathbf{z}}^{(k)}$ is then accepted according to the following probability:

$$\alpha_t = \min\left(1, \frac{p_\psi(\mathbf{x}_t|\tilde{\mathbf{z}}^{(k)}) p(\tilde{\mathbf{z}}_t^{(k)})}{p_\psi(\mathbf{x}_t|\mathbf{z}^{(k-1)}) p(\mathbf{z}_t^{(k-1)})}\right) \tag{10}$$

Let $u_t$ be drawn from the continuous uniform distribution over $[0, 1]$. Then, if $u_t \leq \alpha_t$, the proposal is accepted and $\mathbf{z}_t^{(k)} \leftarrow \tilde{\mathbf{z}}_t^{(k)}$. Otherwise, the current state is retained $\mathbf{z}_t^{(k)} \leftarrow \mathbf{z}_t^{(k-1)}$. A key observation here is that the acceptance ratios, $\alpha_1, \ldots, \alpha_T$, are computed in parallel, with the same current likelihood $p_\phi(\mathbf{x}|\mathbf{z}^{(k-1)})$ for all the samples. Once a sufficient number of iterations is performed, the initial samples corresponding to the so-called burn-in period are discarded. The overall MH sampler is summarized in Algorithm 3.

### 3.3. Metropolis-Adjusted Langevin Algorithm (MALA)

The Metropolis-Adjusted Langevin Algorithm (MALA) [15] aims at combining the MH and LD mechanisms to achieve a more efficient exploration of the target distribution. MALA follows the same steps as MH with the difference that the candidate states are generated using a proposal distribution guided by LD. More precisely, the proposal distribution takes a similar form as (7), except for the fact that here we do not generate multiple samples per latent variable:

$$\tilde{\mathbf{z}}_t^{(k)} | \mathbf{z}_t^{(k-1)} \sim \mathcal{N}(\mathbf{z}_t^{(k-1)} + \frac{\eta}{2} f_\psi(\mathbf{z}_t^{(k-1)}), \eta \mathbf{I}), \tag{11}$$

Nevertheless, in contrast to LD sampler, which always updates the states according to the update rule (7), MALA considers the updated states as candidates, similar to MH, and accepts them according to the following probability

$$\alpha_t = \min\left(1, \frac{p_\psi(\mathbf{x}_t|\tilde{\mathbf{z}}^{(k)}) p(\tilde{\mathbf{z}}_t^{(k)}) q(\mathbf{z}^{(k)}|\tilde{\mathbf{z}}^{(k)})}{p_\psi(\mathbf{x}_t|\mathbf{z}^{(k-1)}) p(\mathbf{z}_t^{(k-1)}) q(\tilde{\mathbf{z}}^{(k)}|\mathbf{z}^{(k)})}\right) \tag{12}$$

where

$$q(\mathbf{u}|\mathbf{v}) \propto \exp\Big(-\frac{1}{2\eta}\|\mathbf{u} - \mathbf{v} - \frac{\eta}{2} f(\mathbf{v})\|^2\Big) \tag{13}$$

is the transition probability density from $\mathbf{v}$ to $\mathbf{u}$. Unlike the basic MH approach, MALA often suggests moves towards regions of higher probability, thus increasing the probability of their acceptance. The overall MALA sampler is described in Algorithm 4.

## 4. EXPERIMENTS

### 4.1. Baselines

This section presents and discusses the speech enhancement results of the proposed EM-based approaches, i.e., LDEM, MHEM, and MALAEM, with RVAE [12] as the generative model. We compare against the VEM method[1] [8, 12], and SGMSE+[2] [4], as a state-of-the-art diffusion-based speech enhancement method.

### 4.2. Evaluation metrics

To evaluate the speech enhancement performance, we use standard metrics, including the extended short-time objective intelligibility (ESTOI) measure [18], ranging in $[0, 1]$, the perceptual evaluation of speech quality (PESQ) metric [19], ranging in $[-0.5, 4.5]$, and the scale-invariant signal-to-distortion ratio (SI-SDR) metric [20] in dB. For all these metrics, the higher the better. Moreover, as a rough measure of the computational complexity of different methods, we report the average real-time factor (RTF), which is defined as the time (in seconds) required to enhance one second of speech signal. Our experiments were conducted using a machine with an AMD EPYC 7351 CPU and an NVIDIA Tesla T4 GPU.

[1] https://github.com/XiaoyuBIE1994/DVAE_SE/
[2] https://github.com/sp-uhh/sgmse

### 4.3. Model architectures

We utilized pretrained models from the respective code repositories for both RVAE and SGMSE+. In the RVAE architecture, the encoder and decoder employ bidirectional long short-term memory (BLSTM) networks with an internal state dimension of 128, and the latent space is of dimension $L = 16$. The input data consists of STFT power spectrograms with a dimension of $F = 513$. This model was trained on the training subset of the Wall Street Journal (WSJ0) corpus. The architecture of SGMSE+ is detailed in [4], and its core network is based on the Noise Conditional Score Network (NCSN++) [21], adapted for processing complex-valued STFT features. The overall model was trained using the same clean training utterances as RVAE, combined with noise signals from the CHiME3 dataset [22]. The input data for SGMSE+ comprises STFT representations with $F = 256$, and training was conducted on sequences with a length of $T = 256$, as opposed to $T = 50$ used for RVAE.

### 4.4. Parameter settings

For the inference parameters of SGMSE+, we adhered to their default values. In the case of RVAE-based methods, we conducted a total of $J = 100$ EM iterations. The learning rate for VEM, with the Adam optimizer, as well as for LDEM and MALAEM (denoted as $\eta$), was consistently set at $0.005$. For LDEM and VEM, we empirically selected $K = 1$, while for MHEM and MALAEM, we opted for $K = 10$, and included a burn-in period of $k_{\texttt{burn-in}} = 5$. Additionally, we set $\sigma^2 = 0.02$ in both (6) and (9).

### 4.5. Evaluation datasets

For performance evaluation, we used the test set from of the WSJ0-QUT corpus [12], created by mixing clean speech signals from WSJ0 (distinct speakers from training) with noise signals from the QUT-NOISE corpus [23]. It includes *Café*, *Home*, *Street*, and *Car* noise types, with signal-to-noise ratio (SNR) values of $-5$ dB, 0 dB, and 5 dB. The test set amounts to 651 noisy speech signals with a total duration of 1.5 hours. Furthermore, we evaluated the generalization performance of various methods under mismatched conditions by incorporating pre-computed noisy versions of the TCD-TIMIT data as introduced in [24]. The noisy test set that we used includes *Living Room (LR)* (from the second CHiME challenge [25]), *White*, *Car*, and *Babble* (from the RSG-10 corpus [26]), with SNR[3] values of $-5$ dB, 0 dB, and 5 dB, yielding 540 test speech signals, with a total duration of about 0.75 hours.

### 4.6. Results

The average speech enhancement metrics, under both matched (WSJ0-QUT) and mismatched (TCD-TIMIT) conditions, are presented in Table 1, with the associated average RTF values reported in Table 2. From these results, we can make several observations. First of all, among the RVAE-based algorithms, the proposed posterior sampling methods outperform VEM with a significant margin. The performance gap is even more remarkable in the mismatched conditions, which demonstrates that VEM is not as generalizable as our proposed methods. This could be due to the fact that VEM relies on fine-tuning the trained encoder on the new data, which might not be efficient. For instance, in the mismatched condition, MALAEM outperforms VEM by around **3 dB** in SI-SDR, **0.19** in PESQ, and **0.07** in ESTOI. Furthermore, the LDEM algorithm consistently stands out with the highest or second-highest scores in all three metrics under both test conditions. The observation that LDEM outperforms MALAEM could be because of their different sampling strategies. That is, LDEM creates multiple parallel sequences of samples at each E-step, whereas MALAEM has only one sequential sequence of samples, where the final retained samples are chosen based on a probabilistic acceptance strategy.

On the other hand, SGMSE+, as the supervised baseline, showcases remarkable performance in the matched condition, achieving much higher speech enhancement metrics than those of the unsupervised RVAE-based methods. Nevertheless, when tested in the mismatched condition, the performance of SGMSE+ drops significantly, under-performing our proposed methods with a large margin. This confirms the generalization dilemma of supervised methods.

In terms of computational complexity at inference time, all the proposed methods achieve much smaller RTF values than VEM, making them more applicable. In particular, the LDEM algorithm demonstrates a competitive RTF of **0.21 sec**, as compared to **12.55 sec** for VEM and **3.85 sec** for SGMSE+, showcasing its high computational efficiency in enhancing speech signals

[3]Here, the protocol used to compute SNR is different than the one used in [12].

**Table 1**: Speech enhancement results under matched (WSJ0-QUT) and mismatched (TCD-TIMIT) test conditions.

| Metric | | SI-SDR (dB) | PESQ | ESTOI |
|---|---|---|---|---|
| Input (**WSJ0-QUT**) | | -2.60 $\pm$ 0.16 | 1.83 $\pm$ 0.02 | 0.50 $\pm$ 0.01 |
| RVAE | VEM [8] | 4.5 $\pm$ 0.21 | 2.21 $\pm$ 0.02 | 0.60 $\pm$ 0.01 |
| | MHEM | 5.15 $\pm$ 0.20 | 2.24 $\pm$ 0.02 | 0.62 $\pm$ 0.01 |
| | MALAEM | 5.52 $\pm$ 0.21 | 2.28 $\pm$ 0.02 | 0.62 $\pm$ 0.01 |
| | LDEM | <u>5.58</u> $\pm$ 0.20 | <u>2.32</u> $\pm$ 0.02 | <u>0.63</u> $\pm$0.01 |
| SGMSE+ [4] | | **9.41** $\pm$ 0.18 | **2.66** $\pm$ 0.02 | **0.77** $\pm$ 0.01 |
| Input (**TCD-TIMIT**) | | -8.74 $\pm$ 0.29 | 1.84 $\pm$ 0.02 | 0.35 $\pm$ 0.01 |
| RVAE | VEM [8] | 1.44 $\pm$ 0.30 | 2.02 $\pm$ 0.02 | 0.35 $\pm$ 0.01 |
| | MHEM | 3.72 $\pm$ 0.27 | 2.12 $\pm$ 0.02 | **0.42** $\pm$ 0.01 |
| | MALAEM | **4.49** $\pm$ 0.29 | **2.21** $\pm$ 0.02 | **0.42** $\pm$ 0.01 |
| | LDEM | <u>4.18</u> $\pm$ 0.29 | **2.21** $\pm$ 0.02 | **0.42** $\pm$ 0.01 |
| SGMSE+ [4] | | -3.97 $\pm$ 0.41 | 2.04 $\pm$ 0.02 | 0.38 $\pm$ 0.01 |

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

| VEM [8] | MHEM | MALAEM | LDEM | SGMSE+ [4] |
|---|---|---|---|---|
| 12.55 $\pm$ 0.01 | <u>0.92</u> $\pm$ 0.01 | 2.49 $\pm$ 0.01 | **0.21** $\pm$ 0.01 | 3.85 $\pm$ 0.01 |

## 5. CONCLUSIONS

In this paper, we present new posterior sampling techniques for EM-based unsupervised speech enhancement using RVAE. These methods serve as viable alternatives to the computationally intensive variational inference-based VEM approach. Our experimental results illustrate the efficiency of the proposed techniques—LDEM, MHEM, and MALAEM—which not only significantly reduce computational complexity but also outperform VEM. Notably, the LDEM algorithm demonstrates high efficiency and competitive enhancement outcomes. In contrast, the supervised baseline, SGMSE+, excels under matched conditions but faces challenges in mismatched scenarios, highlighting generalization limitation of supervised methods. In summary, our proposed methods offer a promising avenue for efficient and effective unsupervised speech enhancement.

## 6. REFERENCES


[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*, John Wiley & Sons, 2018.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[4] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] H. Yen, F. G. Germain, G. Wichern, and J. Le Roux, "Cold diffusion for speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7402–7406.

[7] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[8] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2993–3007, 2022.

[9] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[10] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder.," in *INTERSPEECH*, 2020, pp. 2437–2441.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. International Conference on Learning Representations (ICLR)*, April 2014.

[12] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[13] M. Sadeghi and R. Serizel, "Fast and efficient speech enhancement with variational autoencoders," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.

[15] G. O. Roberts and O. Stramer, "Langevin diffusions and metropolis-hastings algorithms," *Methodology and computing in applied probability*, vol. 4, pp. 337–357, 2002.

[16] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 2011, pp. 681–688.

[17] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*, vol. 2, Springer, 1999.

[18] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2001.

[20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[23] D. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015*, 2015, pp. 3456–3460.

[24] A. H. Abdelaziz et al., "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition.," in *Interspeech*, 2017, pp. 3752–3756.

[25] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 126–130.

[26] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise database," *report IZF*, vol. 3, pp. 1988, 1988.