Unsupervised speech enhancement with diffusion-based generative models

Berné Nortier, Mostafa Sadeghi, Romain Serizel

UNSUPERVISED SPEECH ENHANCEMENT WITH DIFFUSION-BASED GENERATIVE MODELS

Berné Nortier, Mostafa Sadeghi, Romain Serizel

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

ABSTRACT

Recently, conditional score-based diffusion models have gained significant attention in the field of supervised speech enhancement, yielding state-of-the-art performance. However, these methods may face challenges when generalising to unseen conditions. To address this issue, we introduce an alternative approach that operates in an unsupervised manner, leveraging the generative power of diffusion models. Specifically, in a training phase, a clean speech prior distribution is learnt in the short-time Fourier transform (STFT) domain using score-based diffusion models, allowing it to unconditionally generate clean speech from Gaussian noise. Then, we develop a posterior sampling methodology for speech enhancement by combining the learnt clean speech prior with a noise model for speech signal inference. The noise parameters are simultaneously learnt along with clean speech estimation through an iterative expectation-maximisation (EM) approach. To the best of our knowledge, this is the first work exploring diffusion-based generative models for unsupervised speech enhancement, demonstrating promising results compared to a recent variational auto-encoder (VAE)-based unsupervised approach and a state-of-the-art diffusion-based supervised method. It thus opens a new direction for future research in unsupervised speech enhancement.

***Index Terms*—** Unsupervised speech enhancement, diffusion-based models, expectation-maximisation, posterior sampling.

1. INTRODUCTION

Over the past decade, the speech enhancement (SE) task has been extensively investigated, and numerous novel approaches have been proposed that greatly leverage the advancements and efficacy of deep neural network (DNN) architectures [1]. The majority of these approaches are based on supervised (discriminative) learning of a DNN over training pairs of clean and noisy speech signals, covering different speakers, noise types, and signal-to-noise ratio (SNR) values. Such an approach depends heavily on the number and diversity of training samples and noise conditions, and thus generalisation to unseen (out-of-domain) environments cannot be guaranteed.

Unsupervised SE based on deep generative models presents an alternative approach with improved generalisation performance [2–4]. In contrast to purely supervised methods, the generative-based (unsupervised) framework learns the statistical distribution of clean speech signals and uses it as a prior distribution for inferring the target signal from its noisy observation. In these methods, VAE [5] has been commonly used as a generative clean speech prior, which is combined with a non-negative matrix factorization (NMF)-based observation model to estimate clean speech following a statistical EM framework.

Recently, diffusion-based generative models have emerged as a powerful and state-of-the-art framework to model complex data distributions [6, 7]. These models learn an implicit distribution by estimating the score, i.e., the gradient of the log probability density (with respect to data). This is done by gradually diffusing data samples into noise and then learning a score approximating model that can reverse the noising process for different noise scales. The forward process of corrupting data is modelled as a stochastic differential equation (SDE), which can be reversed and yields a corresponding reverse SDE that depends only on the score of the perturbed data and may easily be solved numerically. Diffusion-based models have been widely applied to the SE task in a supervised way [8–12] by incorporating noisy speech signals in the diffusion process as conditioning information.

In this paper, we develop an *unsupervised* speech enhancement framework leveraging diffusion-based generative models as data-driven priors. Specifically, in a training step, the statistical characteristics of *clean speech signals* are learnt in the complex STFT domain through the use of a score-based diffusion model. At test time, we perform posterior sampling by combining the learnt implicit clean speech prior with a parametric statistical model for noise to infer the clean speech signal. The noise parameters are estimated alongside the clean speech signal by following an iterative EM-based approach. To our knowledge, this is the first work that proposes using diffusion-based generative models for unsupervised SE, and explores their potential. We conduct experiments comparing the proposed framework with a VAE-based unsupervised approach [3] as well as a state-of-the-art diffusion-based supervised method [11]. The results demonstrate the effectiveness and promising performance of the proposed diffusion-based unsupervised approach, paving the path for future research in this direction.

The rest of the paper is organised as follows: Section 2 reviews score-based diffusion modelling and VAE-based SE as two closely related problems to our work. The proposed speech generative modelling and enhancement frameworks are detailed in Section 3. Experimental results are then presented in Section 4, followed by a conclusion and suggestions for future lines of work in Section 5.

2. BACKGROUND

2.1. Score-based diffusion models

Diffusion models are a state-of-the-art class of probabilistic generative models that have recently achieved remarkable performance in generating high-quality samples in different applications [7]. These models transform an unknown data distribution p_0 to a tractable prior distribution, usually $\mathcal{N}(\mathbf{0}, \mathbf{I})$, by gradually adding noise to training

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).

data in a forward process. Then, in a reverse process, a parameterised model is learnt to iteratively generate samples starting from noise and transform these into samples from the unknown data distribution. This action of smoothly injecting noise into training samples may be described by a SDE. Specifically, consider a diffusion process $\{\mathbf{s}_t\}_{t\in[0,1]}$, indexed by a continuous time-step variable t, which solves the following general linear SDE

$$\mathrm{d}\mathbf{s}_t = \mathbf{f}(\mathbf{s}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{w}$ denotes a standard Wiener process, the vector-valued $\mathbf{f}$ is the *drift* coefficient term, and the scalar function g is the *diffusion* coefficient. Here, the forward process transforms a clean training sample $\mathbf{s}_0 = \mathbf{s}$ to a noise sample $\mathbf{s}_1$, whose distribution converges to $p_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Under some light regularity conditions [13], the SDE in (1) also has an associated *reverse*-time SDE:

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

where $\bar{\mathbf{w}}$ is a standard Wiener process running backwards in time, $\mathrm{d}t$ is an infinitesimal negative time-step, and $\nabla_{\mathbf{s}_t}\log p_t(\mathbf{s}_t)$ is called the *score* function. In practice, the score is approximated by a time-dependent neural network (NN) $\mathbf{S}_{\theta^*}(\mathbf{s}_t, t) \approx \nabla_{\mathbf{s}_t}\log p_t(\mathbf{s}_t)$, called the *score model*, where θ^* denotes the learnt weights of the NN. By plugging the score model in (2), we can solve the resulting SDE using a variety of solvers to sample from the unknown data distribution [7]. In this paper, we make use of the Predictor-Corrector (PC) sampler [7].

2.2. VAE-based unsupervised speech enhancement

Previous work on unsupervised SE use VAE to learn the prior distribution of clean speech signals, which is then combined with an observation model to estimate clean speech in a statistical framework. Specifically, in the STFT domain, a latent variable-based generative model is assumed as $p_\theta(\mathbf{s}, \mathbf{z}) = p_\theta(\mathbf{s}|\mathbf{z})p_\theta(\mathbf{z})$, where $\mathbf{s}$ denotes the STFT representation of clean speech and $\mathbf{z}$ represents the associated (latent) low-dimensional embedding. Some parameterised Gaussian forms for the generative distributions are usually assumed, whose parameters are learnt over clean speech data, following the evidence lower-bound optimisation principle [5].

For SE, it is assumed that $\mathbf{x} = \mathbf{s} + \mathbf{n}$, where $\mathbf{x}$, $\mathbf{s}$, and $\mathbf{n}$ denote STFT representations of noisy (mixture) speech, clean speech, and background noise, respectively. The likelihood $p_\phi(\mathbf{x}|\mathbf{s})$ is usually a proper complex Gaussian distribution $\mathcal{N}_\mathbb{C}$ with mean $\mathbf{s}$, whose variance is parameterised with a low-rank NMF factorisation. SE then amounts to inferring the latent variable $\mathbf{z}$ associated with $\mathbf{s}$ from $\mathbf{x}$, which necessitates learning the NMF parameters, denoted ϕ, via an EM process formulated below

$$\max_{\phi}\ \mathbb{E}_{p_\phi(\mathbf{z}|\mathbf{x})}\left\{\log p_\phi(\mathbf{x}|\mathbf{z})\right\}. \tag{3}$$

This could be solved using, e.g., the variational EM procedure developed in [3, 14], which approximates $p_\phi(\mathbf{z}|\mathbf{x})$.

3. PROPOSED FRAMEWORK

3.1. Diffusion-based speech generative modeling

Following [11], we work with the complex-valued STFT representations of speech signals and apply an exponential amplitude transformation to balance the heavy-tailed distribution of STFT amplitudes.

Like VAE, the diffusion-based generative model is independently defined for each time-frequency (TF) bin. Therefore, as done in [11], all the vector-valued variables $\mathbf{s}_t$ in boldface contain flattened TF representations of speech signals. For concrete instantiations of the forward and reverse SDE ((1) and (2) respectively), we use an alternative form of the well-known variance-preserving stochastic differential equation (VESDE) [15] inspired by [11], and adapt it to obtain the drift and diffusion coefficients as follows

$$\mathbf{f}(\mathbf{s}_t) = -\gamma\mathbf{s}_t, \qquad g(t) = \sigma_{\text{max}}\Big(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\Big)^t\sqrt{2\log\Big(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\Big)}, \tag{4}$$

where γ is a constant parameter, and σ_{min} and σ_{max} are parameters defining the noise schedule of the Wiener process. The SDE in (1) then has the *perturbation kernel* defined below, which allows one to sample $\mathbf{s}_t$ directly given $\mathbf{s}_0$

$$p_{0t}(\mathbf{s}_t|\mathbf{s}_0) = \mathcal{N}_\mathbb{C}(\delta_t\mathbf{s}_0, \sigma(t)^2\mathbf{I}), \tag{5}$$

where $\delta_t = \mathrm{e}^{-\gamma t}$ and the variance term $\sigma(t)^2$ is given by

$$\sigma(t)^2 = \frac{\sigma_{\text{min}}^2\Big((\sigma_{\text{max}}/\sigma_{\text{min}})^{2t} - {\delta_t}^2\Big)\log(\sigma_{\text{max}}/\sigma_{\text{min}})}{\gamma + \log(\sigma_{\text{max}}/\sigma_{\text{min}})}. \tag{6}$$

To learn the NN parameters θ, a weighted Fisher divergence [15] between the true and approximated score is solved, which, after some mathematical manipulation, leads to the following training objective [11]

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\,\mathbb{E}_{t,\mathbf{s},\boldsymbol{\zeta},\mathbf{s}_t|\mathbf{s}}\Big[\|\mathbf{S}_\theta(\mathbf{s}_t, t) + \frac{\boldsymbol{\zeta}}{\sigma(t)}\|_2^2\Big], \tag{7}$$

where $\boldsymbol{\zeta} \sim \mathcal{N}_\mathbb{C}(\mathbf{0}, \mathbf{I})$, i.e., complex-valued Gaussian noise.

3.2. Diffusion-based unsupervised speech enhancement

We now describe the unsupervised SE framework based on diffusion-based generative models. The prior clean speech distribution $p = p(\mathbf{s})$ is unknown, but can be obtained by training a diffusion-based generative model as described in Section 3.1, yielding an implicit prior, as opposed to the explicit VAE-based speech prior modelling framework. This implicit diffusion-based speech prior only allows for iterative sampling, without an explicit density form. As such, the SE procedure adopted in VAE-based modelling cannot directly be used for diffusion-based learnt speech priors. Assuming the same observation model as before, i.e., $\mathbf{x} = \mathbf{s} + \mathbf{n}$, and NMF-based likelihood parameterisation, we here propose to sample from the following intractable posterior distribution to estimate the clean speech $\mathbf{s}$ directly

$$p_\phi(\mathbf{s}|\mathbf{x}) \propto p_\phi(\mathbf{x}|\mathbf{s})p_{\theta^*}(\mathbf{s}), \tag{8}$$

where θ^* denotes the diffusion model's pretrained, and thus fixed, parameters. We model the noise by $\mathbf{n} \sim \mathcal{N}_\mathbb{C}(\mathbf{0}, \mathrm{diag}(\mathrm{vec}(\mathbf{WH})))$ where $\mathbf{W}$, $\mathbf{H}$ are low-rank matrices with non-negative entries and rank r and $\mathrm{vec}(\mathbf{WH})$ denotes the vectorised form of $\mathbf{WH}$. The likelihood $p_\phi(\mathbf{x}|\mathbf{s})$ then writes as $p_\phi(\mathbf{x}|\mathbf{s}) = \mathcal{N}_\mathbb{C}(\mathbf{s}, \mathrm{diag}(\mathbf{v}_\phi))$, where $\mathbf{v}_\phi = \mathrm{vec}(\mathbf{WH})$. Learning the NMF parameters, i.e., $\phi = \{\mathbf{W}, \mathbf{H}\}$, is done by solving

$$\max_{\phi}\ \mathbb{E}_{p_\phi(\mathbf{s}|\mathbf{x})}\left\{\log p_\phi(\mathbf{x}|\mathbf{s})\right\}. \tag{9}$$

An overview of the proposed Unsupervised Diffusion-Based Speech Enhancement (UDiffSE) approach is provided in Algorithm 1. The following sections detail the E-step and M-step.

Algorithm 1 UDiffSE

$\phi_0 = \{\mathbf{W}_0, \mathbf{H}_0\}$
for $k = 1, \ldots, K$ **do**
  $\hat{\mathbf{s}} \sim p_{\phi_{k-1}}(\mathbf{s}|\mathbf{x})$ ▷ *(E-Step)*
  $\phi_k \leftarrow \text{argmax}_\phi \log p_\phi(\mathbf{x}|\hat{\mathbf{s}})$ ▷ *(M-Step)*
end for
return $\hat{\mathbf{s}}$

3.2.1. E-Step

Given a current estimate of ϕ, the E-step (posterior sampling) entails the generation of speech samples from the posterior distribution $p_\phi(\mathbf{s}|\mathbf{x})$ to approximate the expectation in (9). This is done via the construction of a stochastic process $\{\mathbf{s}_t|\mathbf{x}\}_{t\in[0,1]}$ by conditioning the original process $\{\mathbf{s}_t\}_{t\in[0,1]}$ on the observation $\mathbf{x}$ to obtain an estimate $\hat{\mathbf{s}} \sim p_\phi(\mathbf{s}|\mathbf{x})$. To this end, we modify the reverse SDE (2) as follows

$$\begin{aligned}\mathrm{d}\mathbf{s}_t &= \Big[\mathbf{f}(\mathbf{s}_t)\mathrm{d}t - g(t)^2\nabla_{\mathbf{s}_t}\log p_t(\mathbf{s}_t|\mathbf{x})\Big]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \\ &= \Big[\mathbf{f}(\mathbf{s}_t)\mathrm{d}t - g(t)^2(\nabla_{\mathbf{s}_t}\log p_t(\mathbf{x}|\mathbf{s}_t) + \nabla_{\mathbf{s}_t}\log p_t(\mathbf{s}_t))\Big]\mathrm{d}t \\ &\quad + g(t)\mathrm{d}\bar{\mathbf{w}} \end{aligned} \tag{10}$$

where again the score function can be approximated by $\mathbf{S}_{\theta^*}(\mathbf{s}_t, t)$. However, the conditional score function $\nabla_{\mathbf{s}_t}\log p_\phi(\mathbf{x}|\mathbf{s}_t)$ is, in fact, intractable to compute in closed form due to its dependence on time. That is,

$$p_\phi(\mathbf{x}|\mathbf{s}_t) = \int p_\phi(\mathbf{x}|\mathbf{s}_0)p_{t0}(\mathbf{s}_0|\mathbf{s}_t)\mathrm{d}\mathbf{s}_0, \tag{11}$$

where $p_{t0}(\mathbf{s}_0|\mathbf{s}_t) \propto p_{0t}(\mathbf{s}_t|\mathbf{s}_0)p(\mathbf{s}_0)$ is intractable. As an approximation, we follow [16] and assume an uninformative prior $p(\mathbf{s}_0)$, which, along with (5), results in

$$\tilde{p}_{t0}(\mathbf{s}_0|\mathbf{s}_t) = \mathcal{N}_\mathbb{C}\Big(\frac{\mathbf{s}_0}{\delta_t}, \frac{\sigma(t)^2}{\delta_t^2}\mathbf{I}\Big). \tag{12}$$

Plugging this approximation in (11) gives us the following *noise-perturbed pseudo-likelihood*

$$\tilde{p}_\phi(\mathbf{x}|\mathbf{s}_t) \sim \mathcal{N}_\mathbb{C}\Big(\frac{\mathbf{s}_t}{\delta_t}, \frac{\sigma(t)^2}{\delta_t^2}\mathbf{I} + \text{diag}(\boldsymbol{v}_\phi)\Big). \tag{13}$$

The conditional reverse process is then approximated as

$$\begin{aligned}\mathrm{d}\mathbf{s}_t = \Big[\mathbf{f}(\mathbf{s}_t)\mathrm{d}t - g(t)^2\mathbf{S}_{\theta^*}(\mathbf{s}_t, t)\Big]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \\ -g(t)^2\nabla_{\mathbf{s}_t}\log\tilde{p}_\phi(\mathbf{x}|\mathbf{s}_t)\mathrm{d}t.\end{aligned} \tag{14}$$

This is exactly the unconditional reverse process (2) for sampling clean speech, plus an additional term which imposes data consistency. We use the change of variables formula and take the gradient to compute $\nabla_{\mathbf{s}_t}\log\tilde{p}(\mathbf{x}|\mathbf{s}_t)$, the *noise-perturbed pseudo-likelihood score* as

$$\nabla_{\mathbf{s}_t}\log\tilde{p}(\mathbf{x}|\mathbf{s}_t) = \frac{1}{\delta_t}\Big[\frac{\sigma(t)^2}{\delta_t^2}\mathbf{I} + \text{diag}(\boldsymbol{v}_\phi)\Big]^{-1}\Big(\frac{\mathbf{s}_t}{\delta_t} - \mathbf{x}\Big). \tag{15}$$

Lastly, we introduce an additional weighting parameter λ to the pseudo-likelihood as in [16] to balance the effect of the mixture signal on the estimated sample. We experimentally observed that performing the full posterior reverse step at each iteration enforces strongly the data consistency condition, causing the sample to converge to the mixture signal. To prevent this, we only perform the posterior step every ℓ iterations. We solve the reverse SDE using a PC sampler [7] - a numeric sampler consisting of a discretisation of (2) - the *predictor* - followed by a Langevin sampling step - the *corrector* - to 'correct' the marginal at time t. The overall E-step is summarised in Algorithm 2. The variable τ denotes discrete time-step in $[0, 1]$. For simplicity, we employ the shorthand σ_τ, $\mathbf{f}_\tau$, g_τ for $\sigma(\tau)$, $\mathbf{f}(\tau)$, $g(\tau)$, respectively.

Algorithm 2 Posterior sampling (E-step) of UDiffSE

Require: $N, \mathbf{x}, \ell, \lambda$
$\mathbf{s}_1 \sim \mathcal{N}_\mathbb{C}(\mathbf{x}, \mathbf{I}), \Delta\tau \leftarrow \frac{1}{N}$
for $i = N, \ldots, 1$ **do**
  $\tau \leftarrow \frac{i}{N}$
  $\boldsymbol{\zeta}_c \sim \mathcal{N}_\mathbb{C}(\mathbf{0}, \mathbf{I})$ ▷ *(Corrector)*
  $\mathbf{s}_\tau \leftarrow \mathbf{s}_\tau + \epsilon_\tau\mathbf{S}_{\theta^*}(\mathbf{s}_\tau, \tau) + \sqrt{2\epsilon_\tau}\boldsymbol{\zeta}_c$
  $\boldsymbol{\zeta}_p \sim \mathcal{N}_\mathbb{C}(\mathbf{0}, \mathbf{I})$ ▷ *(Predictor)*
  $\mathbf{s}_\tau \leftarrow \mathbf{s}_\tau - \mathbf{f}_\tau\Delta\tau + g_\tau^2\mathbf{S}_{\theta^*}(\mathbf{s}_\tau, t)\Delta\tau + g_\tau\sqrt{\Delta\tau}\boldsymbol{\zeta}_p$
  if $i \equiv 0 \pmod{\ell}$ **then** ▷ *(Posterior)*
    $\nabla_{\mathbf{s}_\tau}\log\tilde{p}(\mathbf{x}|\mathbf{s}_\tau) \leftarrow \frac{1}{\delta_\tau}\Big[\frac{\sigma_\tau^2}{\delta_\tau^2}\mathbf{I} + \text{diag}(\boldsymbol{v}_\phi)\Big]^{-1}(\frac{\mathbf{s}_\tau}{\delta_\tau} - \mathbf{x})$
    $\mathbf{s}_\tau \leftarrow \mathbf{s}_\tau + \lambda g_\tau^2\nabla_{\mathbf{s}_\tau}\log\tilde{p}(\mathbf{x}|\mathbf{s}_\tau)$
  end if
end for
return $\hat{\mathbf{s}} = \mathbf{s}_0$

3.2.2. M-Step

Having obtained a clean speech estimate $\hat{\mathbf{s}}$ in the E-step, we now consider updating the noise parameters $\phi = \{\mathbf{W}, \mathbf{H}\}$ via (9), approximating the expectation with a Monte-Carlo average using $\mathbf{s} \leftarrow \hat{\mathbf{s}}$:

$$\begin{aligned}\phi &\leftarrow \underset{\{\mathbf{W},\mathbf{H}\}\geq 0}{\text{argmax}}\ \log p_\phi(\mathbf{x}|\hat{\mathbf{s}}) \\ &= \underset{\{\mathbf{W},\mathbf{H}\}\geq 0}{\text{argmin}}\ \frac{(\mathbf{x}-\hat{\mathbf{s}})^H(\mathbf{x}-\hat{\mathbf{s}})}{\boldsymbol{v}_\phi} + \log(\boldsymbol{v}_\phi),\end{aligned} \tag{16}$$

where $(\cdot)^H$ denotes the conjugate transpose operation, and the division is done element-wise. The above problem can be solved using different algorithms, e.g., the multiplicative update rules [17, 18].

4. EXPERIMENTS

In this section, we provide a performance evaluation of our proposed UDiffSE framework as compared against an unsupervised speech enhancement approach based on recurrent VAE (RVAE)[1] [3, 14], as well as a state-of-the-art diffusion-based supervised SE method, called score-based generative model for speech enhancement (SGMSE+)[2] [11].

Evaluation Metrics. To measure the quality of the enhanced speech signals, we use standard instrumental evaluation metrics, including the scale-invariant signal-to-distortion ratio (SI-SDR) in dB [19], the extended short-time objective intelligibility (ESTOI) measure [20] ranging in $[0, 1]$, and the perceptual evaluation of speech quality (PESQ) score [21] ranging in $[-0.5, 4.5]$. We also use the

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

Table 1: Speech enhancement results under both matched and mismatched conditions. 'S': supervised, 'U': unsupervised. Bold and italicised indicate the best and second best performances, respectively.

Method	Type	SI-SDR (dB)	PESQ	ESTOI	SIG-MOS	BAK-MOS	OVR-MOS
Input (**WSJ0-QUT**)	-	-2.60 ± 0.17	1.83 ± 0.02	0.50 ± 0.01	4.04 ± 0.01	2.93 ± 0.02	3.13 ± 0.01
RVAE [3, 14]	U	4.39 ± 0.21	2.20 ± 0.02	0.59 ± 0.01	3.88 ± 0.02	3.32 ± 0.02	3.13 ± 0.02
UDiffSE (Ours)	U	*4.80 ± 0.23*	*2.21 ± 0.02*	*0.63 ± 0.01*	*4.33 ± 0.01*	*3.74 ± 0.02*	*3.74 ± 0.02*
SGMSE+ [11]	S	**9.41 ± 0.18**	**2.66 ± 0.02**	**0.77 ± 0.01**	**4.48 ± 0.01**	**4.51 ± 0.01**	**4.19 ± 0.01**
Input (**TCD-TIMIT**)	-	-8.74 ± 0.29	1.84 ± 0.02	0.35 ± 0.01	3.52 ± 0.02	2.22 ± 0.03	2.68 ± 0.01
RVAE [3, 14]	U	**1.44 ± 0.31**	*2.02 ± 0.02*	0.35 ± 0.01	3.08 ± 0.03	*3.18 ± 0.02*	2.61 ± 0.02
UDiffSE (Ours)	U	*0.37 ± 0.25*	2.01 ± 0.02	**0.41 ± 0.01**	**3.91 ± 0.01**	2.88 ± 0.03	*3.08 ± 0.02*
SGMSE+ [11]	S	-3.97 ± 0.41	**2.04 ± 0.03**	*0.38 ± 0.01*	*3.79 ± 0.02*	**3.43 ± 0.02**	**3.13 ± 0.02**

DNS-MOS [22], a non-intrusive speech quality metric, which provides scores for the speech quality (SIG), background noise quality (BAK), and overall quality (OVRL) of speech. For all the metrics, higher values indicate improved performance.

Datasets. To learn the clean speech prior model, we train on the 'si_tr_s' subset of the Wall Street Journal (WSJ) corpus [23], which amounts to roughly 25 hours of data. The STFT is computed using a window size of 510, a hop-length of 128 ($\approx 75\%$ overlap), and a Hann window, which gives $F = 256$ frequency bins. All signals have a sampling rate of 16kHz. To ensure similarity across samples of different length during training, subsamples are randomly selected from a STFT transform so that we get $T = 256$ time frames with start and end positions randomly generated during training.

For performance evaluation, we use the WSJ0-QUT dataset created by [14], comprising 651 synthetic mixtures (about 1.5 hours of noisy speech data) which uses clean speech signals from the 'si_et_05' subset of WSJ dataset and noise signals from the QUT-NOISE corpus [24]. These include *Café*, *Home*, *Street*, and *Car* and have SNR values of -5 dB, 0 dB, and 5 dB. We also evaluate generalisation capability of different methods in mismatched conditions by using pre-computed noisy versions of the TCD-TIMIT data presented in [25]. This set contains noise types *Living Room* (from the second CHiME challenge [26]), *White*, *Car*, and *Babble* (from the RSG-10 corpus [27]) with SNR values of -5 dB, 0 dB, and 5 dB and. This yields 540 test speech signals (or approximately 45 minutes).

Stochastic Differential Equation. The SDE in (4) has parameter values $\gamma = 1.5, \sigma_{\text{min}} = 0.05, \sigma_{\text{max}} = 0.5$. To avoid instabilities around 0, we adopt standard practice and set a minimum process time with $t_{\text{min}} = 0.03$.

Models architecture. We adapt the SGMSE+ architecture developed in [10], which is based on a multi-resolution U-Net structure, by zeroing out their $\mathbf{x}$ term and adapting the channels. RVAE consists of an encoder-decoder architecture composing bidirectional long short-term memory (BLSTM) networks. For both RVAE and SGMSE+, we use the pretrained models that are available in their associated public code repositories.

Training setup. We train the score model $\mathbf{S}_{\theta^*}$ for 220 epochs using an Adam optimiser with a learning rate of 0.0001 and a batch size of 16. Our loss is an exponential moving average of the network's weights, initialised with a decay of 0.999.

EM settings. The reverse process in (14) is solved using a PC sampler with step size $\epsilon_\tau := (\sigma_\tau/2)^2$. The number of reverse sampling steps is set to $N = 30$. The posterior update step is performed every $\ell = 2$ steps, and the NMF variances matrices have rank $r = 4$. For each sample, we perform 5 EM iterations. We observe that performing the same denoising procedure over b samples in parallel and then averaging the result yields much better performance; we thus set the batch size to $b = 4$. The weighting parameter λ is set to 1.5. These parameter choices are motivated by an extensive set of experimental studies provided in the Supplementary Material.

Results. We report our SE results in Table 1. Competing methods are evaluated in the matched and mismatched cases. Inspecting the results, we can make a number of conclusions: As may be expected, the supervised framework outperforms its unsupervised counterpart in the matched case, but at the cost of utilising labelled data. Our UDiffSE framework outperforms the alternative unsupervised RVAE on almost all metrics under both matched and mismatched conditions. In particular, it achieves much higher ESTOI, SIG-MOS, and OVR-MOS scores than RVAE, which is more noticeable in the mismatched condition.

Furthermore, the proposed UDiffSE method outperforms the supervised SGMSE+ framework for both the ESTOI and SIG-MOS metrics in the mismatched condition, with a comparable OVR-MOS score. While all three frameworks have very similar PESQ results in the mismatched case, the unsupervised methods significantly outperform SGMSE+ in terms of SI-SDR (by more than 4 dB). The performance of UDiffSE on the TCD-TIMIT dataset showcases its capacity to generalise to unseen data, which could possibly imply that it has learnt a good representation of general clean speech as the underlying prior. Supplementary material, including audio samples, is available online. [3]

5. CONCLUSION

In this paper, we introduce UDiffSE, an unsupervised generative-based framework to solve the SE task by learning an implicit prior distribution over clean speech data. We do this by defining a continuous diffusion process in the STFT domain in the form of a conditional SDE, and imposing an NMF-based parameterised additive noise model. An EM approach is developed to simultaneously generate clean speech and learn the noise parameters. An approximation of the likelihood term in the E-step then yields a tractable posterior sampling procedure. This method outperforms an unsupervised VAE-based approach to SE for almost all metrics in matched and mismatched test conditions, while showcasing better generalisation performance than a state-of-the-art diffusion-based supervised method. UDiffSE does, however, have the disadvantage of being time-consuming, which originates from the complexity of the reverse diffusion process. Future works include speeding up the reverse process, utilising the recent advancements in diffusion-based image generation, and developing more efficient noise models.

[3] https://github.com/joanne-b-nortier/UDiffSE

6. REFERENCES

[1] 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.

[2] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 716–720.

[3] 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.

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

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, April 2014.

[6] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.

[7] 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*, 2021.

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

[9] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[10] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech 2022*, 2022, pp. 2928–2932.

[11] 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.

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

[13] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[14] 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.

[15] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems*, 2019, pp. 11895–11907.

[16] X. Meng and Y. Kabashima, "Diffusion model based posterior sampling for noisy linear inverse problems," *arXiv preprint arXiv:2211.12343*, 2022.

[17] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[18] M. Sadeghi and X. Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7534–7538.

[19] 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.

[20] 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.

[21] 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 *IEEE international conference on acoustics, speech, and signal processing. Proceedings (ICASSP)*, 2001, vol. 2, pp. 749–752.

[22] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 886–890.

[23] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

[24] 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 Interspeech*, 2015, pp. 3456–3460.

[25] 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.

[26] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 126–130.

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