Supplementary Material for: Unsupervised Speech Enhancement with Diffusion-based Generative Models

Berné Nortier, Mostafa Sadeghi, Romain Serizel

This supplementary provides additional details, discussions and parameter studies that serve to expand our work presenting the Unsupervised Diffusion-Based Speech Enhancement (UDiffSE) framework [5]. In Part A, we detail the Predictor-Corrector (PC) sampler for solving the reverse process and in Part B, we describe the updates of the non-negative matrix factorization (NMF) noise matrices. In Part C, we explain the non-linear amplitude transform mentioned in Section 3.2 of the article and then continue to measure and discuss the impact of different parameters on performance in Part D. Finally, we present some visual results and related discussion in Part E.

A PC Sampling for the reverse SDE

Recall that the unconditional stochastic differential equation (SDE) describing the corruption of a clean speech signal has (under some regularity conditions) the associated reverse SDE

$$\mathrm{d}\mathbf{s}_t = [\mathbf{f}(\mathbf{s}_t)\mathrm{d}t - g(t)^2\nabla_{\mathbf{s}_t}\log p_t(\mathbf{s}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{1}$$

One may solve this reverse SDE numerically using a number of different sampling techniques, such as Euler-Maruyama, Annealed Langevin Dynamics [8], probability flow ODE's [9] or PC [9]. In this paper, we restrict our attention to the PC-sampling scheme which proposes to discretise the reverse SDE sampling and additionally leverage information provided by the score function. In essence, it comprises a two-step scheme that first takes one discrete reverse diffusion step (the *predictor*), followed by a Langevin sampling step (the *corrector*) to correct the time-marginal. To avoid instabilities about $t = 0$, we follow standard practice by only running the reverse process to a minimum time $t_{\min}$. More precisely, for a discretisation of the interval $[t_{\min}, 1]$ into N intervals of length $\Delta\tau$, one step of the unconditional approximated discretised reverse diffusion process at time τ is given by

$$\mathbf{s}_\tau = \mathbf{s}_{\tau+\Delta\tau} - \mathbf{f}_{\tau+\Delta\tau}(\mathbf{s}_{\tau+\Delta\tau}) + g^2_{\tau+\Delta\tau}\mathbf{S}_{\theta^*}(\mathbf{s}_{\tau+\Delta\tau}, t_{\tau+\Delta\tau}) + g_{\tau+\Delta\tau}\mathbf{z}_{\tau+\Delta\tau}. \tag{2}$$

Song et al. [10] then propose to exploit the properties of the score to refine $\mathbf{s}_t$. This is done by adjusting the sample's marginal distribution to ensure that resulting samples have the same time-marginals as do solution trajectories of the reverse-time SDE. In practice, we perform only 1 corrector step for every predictor step. This is done via one iteration of the Langevin Dynamics, a form of score-based MCMC

$$\mathbf{s}_\tau = \mathbf{s}_{\tau+\Delta\tau} + \epsilon_{\tau+\Delta\tau}\mathbf{S}_{\theta^*}(\mathbf{s}_{\tau+\Delta\tau}) + \sqrt{2\epsilon_{\tau+\Delta\tau}}\boldsymbol{\zeta} \tag{3}$$

for step size $\epsilon_\tau := (\sigma_\tau/2)^2$ and with $\boldsymbol{\zeta} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{I})$.

B NMF-updates for the M-step

We assume an NMF-based parameterisation of the noise as

$$\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathrm{diag}(\mathrm{vec}(\mathbf{WH}))), \tag{4}$$

where $\mathbf{n} \in \mathbb{C}^d$ with $d = KF$ and K, F indicating the time and frequency dimensions of the short-time Fourier transform (STFT) transform, respectively. We assume $\mathbf{W} \in \mathbb{R}_+^{K \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times F}$ for some known rank r. Combining this with the noise observation model, the NMF-based likelihood $p_\phi(\mathbf{x}|\mathbf{s})$ is parameterised as $p_\phi(\mathbf{x}|\mathbf{s}) = \mathcal{N}_{\mathbb{C}}(\mathbf{s}, \mathrm{diag}(\boldsymbol{v}_\phi))$ with $\boldsymbol{v}_\phi = \mathrm{vec}(\mathbf{WH})$. Denoting by $\hat{\mathbf{s}}$ the clean speech estimate obtained using the conditional posterior sampling in the E-step, we now learn the noise parameter $\phi = \{\mathbf{W}, \mathbf{H}\}$ by solving the following problem

$$
\begin{aligned}
\phi &\leftarrow \underset{\{\mathbf{W},\mathbf{H}\}\geq 0}{\mathrm{argmax}} \; \mathbb{E}_{p_\phi(\mathbf{s}|\mathbf{x})}\{\log p_\phi(\mathbf{x}, \mathbf{s})\} \\
&= \underset{\{\mathbf{W},\mathbf{H}\}\geq 0}{\mathrm{argmax}} \; E_{p(\mathbf{s}|\mathbf{x})}\Big[\log p_\phi(\mathbf{x}|\mathbf{s}) + \log p_\theta(\mathbf{s})\Big] \\
&= \underset{\mathbf{W},\mathbf{H}\geq 0}{\mathrm{argmax}} \; \mathbb{E}_{p(\mathbf{s}|\mathbf{x})}\Big[\log p_\phi(\mathbf{x}|\mathbf{s})\Big] \\
&\approx \underset{\{\mathbf{W},\mathbf{H}\}\geq 0}{\mathrm{argmin}} \; \frac{(\mathbf{x} - \hat{\mathbf{s}})^H(\mathbf{x} - \hat{\mathbf{s}})}{\boldsymbol{v}_\phi} + \log(\boldsymbol{v}_\phi)
\end{aligned} \tag{5}
$$

where $(\cdot)^H$ denotes the conjugate transpose operation, and the division is performed element-wise. The expectation is approximated by a Monte-Carlo average and can be solved using different algorithms. We use the multiplicative update rules developed by [1, 7] to obtain the following

$$
\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left(\mathbf{V} \odot (\mathbf{WH})^{\odot -2}\right)\mathbf{H}^\top}{(\mathbf{WH})^{\odot -1}\mathbf{H}^\top}, \tag{6}
$$

$$
\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top\left(\mathbf{V} \odot (\mathbf{WH})^{\odot -2}\right)}{\mathbf{W}^\top(\mathbf{WH})^{\odot -1}}, \tag{7}
$$

where

$$
\mathbf{V} = \Big[|x_{ft} - s_{ft}^*|^2\Big]_{(f,t)} \tag{8}
$$

with x_{ft}, s_{ft} specifiying the elements of the unflattened STFT transforms $\mathbf{x}, \mathbf{s}$ respectively and $\odot$ denotes element-wise operation.

C Amplitude transform for the STFT coefficients

As suggested by [6], we apply an exponential amplitude transformation to balance the heavy-tailed distribution of STFT amplitudes. Specifically, for each complex coefficient c with angle $\angle(c)$, we apply

$$
\tilde{c} = \beta|c|^\alpha \mathrm{e}^{i\angle(c)}. \tag{9}
$$

where the compression constant α brings out lower energy coefficients and β serves to spread the amplitudes. In practice, $\alpha = 0.5$, and $\beta = 0.15$. In Part D, we investigate the effect of the non-linear transform on the speech enhancement performance.

D Parameter studies

We present below a series of parameter studies. All studies are run over a set of samples that consist of equal proportions of noise types (*Café*, *Home*, *Street*, and *Car*) and SNR values (-5 dB, 0 dB, and 5 dB) and a base set of parameters $\lambda = 1.5, n_E = 40, r = 4, b = 2, n_{EM} = 5$. Best-performing settings for each metric are indicated in bold-face in the tables.

D.1 Posterior weighting parameter λ

To explicitly control the contribution of the mixture signal to the generation of an enhanced clean speech sample, we introduce the posterior weighting parameter λ as suggested by [4]. Similar to them, we observe that $\lambda \in [1, 2]$ yields the best results. The UDiffSE framework also seems to be quite sensitive to varying values of λ. Results of this study are shown in Table 1.

Parameter	SI-SDR(dB)	PESQ	STOI	SIG-MOS	BAK-MOS	OVR-MOS
$\lambda = 1$	-3.593 ± 0.857	1.682 ± 0.07	0.358 ± 0.019	3.782 ± 0.059	2.964 ± 0.097	3.128 ± 0.063
$\lambda = 1.5$	**1.345 ± 0.832**	**2.121 ± 0.066**	0.549 ± 0.027	4.266 ± 0.036	**3.427 ± 0.09**	**3.522 ± 0.063**
$\lambda = 2$	0.113 ± 0.8	2.106 ± 0.067	**0.572 ± 0.027**	**4.278 ± 0.042**	3.354 ± 0.091	3.473 ± 0.068
$\lambda = 3$	-1.719 ± 0.773	1.955 ± 0.071	0.529 ± 0.026	4.081 ± 0.054	3.08 ± 0.072	3.307 ± 0.064

Table 1: Effect of the posterior term weighting parameter λ.

We leave for future work a more in-depth exploration of a possible adaptive scheduler for λ.

D.2 Scheduling the addition of the pseudo-likelihood term k

As in [3], we observe that adding the pseudo-likelihood term (i.e. $k = 1$) at every step negatively impacts performance. Indeed, we achieve optimal performance by alternating between one unconditional and one conditional reverse diffusion step. Intuitively, this allows to generate clean speech whilst also retaining fidelity to the mixture signal. Results of this study are presented in Table 2.

Parameter	SI-SDR(dB)	PESQ	STOI	SIG-MOS	BAK-MOS	OVR-MOS
$k = 1$	-1.615 ± 0.744	1.87 ± 0.068	0.522 ± 0.025	4.081 ± 0.05	2.961 ± 0.069	3.252 ± 0.063
$k = 2$	**1.345 ± 0.832**	**2.121 ± 0.066**	**0.549 ± 0.027**	**4.266 ± 0.036**	**3.427 ± 0.09**	**3.522 ± 0.063**
$k = 3$	-0.153 ± 0.75	1.891 ± 0.077	0.521 ± 0.026	4.128 ± 0.043	2.975 ± 0.082	3.398 ± 0.052
$k = 4$	-9.502 ± 1.261	1.229 ± 0.095	0.143 ± 0.014	3.563 ± 0.059	2.316 ± 0.152	2.805 ± 0.078
$k = 5$	-13.985 ± 1.566	1.054 ± 0.079	0.075 ± 0.012	3.656 ± 0.066	2.271 ± 0.157	2.824 ± 0.079

Table 2: Effect of the pseudo-likelihood term inclusion every k steps.

The NMF-parameterisation of the noise term implies that the score in the posterior step explicitly involves on $\boldsymbol{v}_\phi = \text{vec}(\mathbf{WH})$. Since the initialisation of the noise matrices is completely random, it makes little sense to consider the information provided by these matrices on the first expectation-maximisation (EM)-iteration. We investigate this intuition by skipping the posterior step at the first iteration and then alternating as before; indeed, performing merely an unconditional PC-step on the first iteration improves all metrics. Specifically, the scale-invariant signal-to-distortion ratio (SI-SDR) and extended short-time objective intelligibility (ESTOI) metrics improve drastically. Results of this study are shown in Table 3.

Metric	SI-SDR(dB)	PESQ	STOI	SIG-MOS	BAK-MOS	OVR-MOS
$i = 1$	1.345 ± 0.832	2.121 ± 0.066	0.549 ± 0.027	4.266 ± 0.036	3.427 ± 0.09	3.522 ± 0.063
$i = 2$	**3.383 ± 0.72**	**2.246 ± 0.078**	**0.594 ± 0.026**	**4.292 ± 0.038**	**3.991 ± 0.08**	**3.931 ± 0.051**

Table 3: Effect of the iteration at which the posterior step is first performed.

D.3 Reverse diffusion steps n_E

To solve the reverse diffusion process, the interval $[t_{\min}, 1]$ is discretised into n_E intervals of equal lengths. We saw during early stages of development that making use of more sophisticated discretisation techniques, such as the recent Karras sampler [2], did not improve performance. The process is sensitive to the number of discretisations and we postulate that this is because the reverse process spends too much time in noisy regions and does not allow to sufficiently create clean speech. A higher value for n_E also increases processing time as the reverse step accesses the learnt score

function $\mathbf{S}_{\theta^*}(\mathbf{s}_t, t)$ at each iteration, and we opt to set the final value of $n_E = 30$. Results of this study are presented in Table 4.

Parameter	SI-SDR(dB)	PESQ	STOI	SIG-MOS	BAK-MOS	OVR-MOS
$n_E = 30$	1.235 ± 0.798	2.061 ± 0.065	**0.56 ± 0.027**	**4.312 ± 0.037**	**3.529 ± 0.093**	**3.547 ± 0.071**
$n_E = 40$	1.345 ± 0.832	**2.121 ± 0.066**	0.549 ± 0.027	4.266 ± 0.036	3.427 ± 0.09	3.522 ± 0.063
$n_E = 50$	**1.54 ± 0.85**	2.058 ± 0.074	0.53 ± 0.027	4.184 ± 0.044	3.268 ± 0.105	3.513 ± 0.065
$n_E = 60$	0.894 ± 0.842	1.932 ± 0.072	0.481 ± 0.026	4.089 ± 0.045	2.971 ± 0.132	3.372 ± 0.077
$n_E = 70$	-2.51 ± 0.8	1.624 ± 0.07	0.371 ± 0.021	3.848 ± 0.06	2.158 ± 0.112	2.883 ± 0.065
$n_E = 80$	-6.771 ± 0.932	1.432 ± 0.075	0.257 ± 0.019	3.473 ± 0.067	2.208 ± 0.181	2.652 ± 0.072

Table 4: Effect of the number of reverse diffusion steps.

D.4 Batch size b

We investigate the effect of enhancing parallel versions of a sample and then averaging over the batch. Increasing batch size always increases performance, but it is the biggest contributor to a slower processing time. As such, we do not experiment with batches of sizes bigger than $b = 4$. However, preliminary studies do suggest that additional parameter tuning could allow for a smaller batch size, but we relegate this to future work. Results of this study are shown in Table 5.

Parameter	SI-SDR(dB)	PESQ	STOI	SIG-MOS	BAK-MOS	OVR-MOS
$b = 1$	0.401 ± 0.846	2.046 ± 0.062	0.518 ± 0.026	4.231 ± 0.045	3.334 ± 0.086	3.509 ± 0.063
$b = 2$	1.345 ± 0.832	2.121 ± 0.066	0.549 ± 0.027	4.266 ± 0.036	3.427 ± 0.09	3.522 ± 0.063
$b = 3$	1.605 ± 0.815	2.154 ± 0.066	0.565 ± 0.026	4.272 ± 0.038	3.467 ± 0.088	3.546 ± 0.061
$b = 4$	**2.922 ± 0.752**	**2.229 ± 0.074**	**0.592 ± 0.027**	**4.292 ± 0.035**	**3.741 ± 0.097**	**3.723 ± 0.071**

Table 5: Effect of the batch size (number of parallel estimates for a given speech signal).

D.5 NMF-rank r

On average, we see that a lower NMF-rank also increases the metrics. We opt for a rank of 4 which also allows greater consistency of comparison between UDiffSE and variational auto-encoder (VAE). This is because VAE uses r=8 with T=512, whereas for UDiffSE we work with T=256. Our specific choice for the rank thus balances the expressive power of the NMF model in VAE and UDiffSE.

Parameter	SI-SDR(dB)	PESQ	STOI	SIG-MOS	BAK-MOS	OVR-MOS
$r = 4$	**2.026 ± 0.81**	2.157 ± 0.072	**0.564 ± 0.028**	4.262 ± 0.04	**3.574 ± 0.093**	**3.619 ± 0.064**
$r = 8$	1.345 ± 0.832	2.121 ± 0.066	0.549 ± 0.027	**4.266 ± 0.036**	3.427 ± 0.09	3.522 ± 0.063
$r = 16$	1.81 ± 0.817	**2.16 ± 0.069**	0.561 ± 0.027	4.265 ± 0.04	3.554 ± 0.095	3.609 ± 0.068
$r = 32$	1.213 ± 0.833	2.118 ± 0.065	0.548 ± 0.027	4.248 ± 0.039	3.391 ± 0.088	3.509 ± 0.064

Table 6: Effect of the NMF-rank.

D.6 Investigating the effect of the non-linear transform

We train a model without the non-linear amplitude transform (9), and although it is able to generate clean speech (a subjective heuristic for measuring the prior model's quality), it does not enable the enhancement of speech to a level of quality as good as that of the same model trained *with* the transform. That is, the non-linear model performs much better at drawing samples from the clean speech distribution (that is, voice-like babbling with no background interference). To showcase this visually, we provide spectograms sampled from the prior in Fig. 1. These spectograms sampled from the prior provide qualitative evidence that the linear model (no transform) succeeds at sampling speech-like signals, but with more background noise and with less sharpness than its non-linear counterpart (with non-linear transform). More informative audio samples will be available in our code repository.

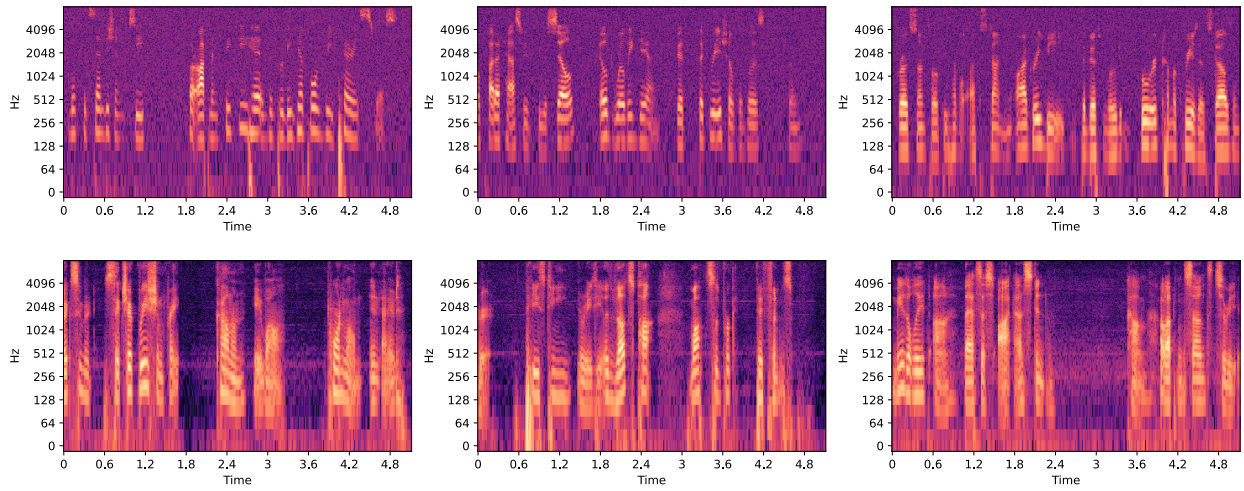

Figure 1: Spectograms of 3 generated clean speech samples each for model without non-linear transform (top) and with the transform (bottom)

This serves as motivation to investigate the speech enhancement performance of our UDiffSE framework for different priors. We compare a model without non-linear transform with another model trained with this transform. Interestingly, the non-linear model performs markedly better. This can be seen in Table 7.
We do also attempt to perform SE by only activating-deactivating the transform whenever input is fed to the score model $\mathbf{S}_{\theta^*}$. This fails completely, indicating that the entire diffusion system is closely linked with the non-linear transform, as all the involved variables have been transformed during the model training. We hope to reach a greater theoretical understanding of the impact of the non-linear transform in future work.

Parameter	SI-SDR(dB)	PESQ	STOI	SIG-MOS	BAK-MOS	OVR-MOS
No transform	1.834 ± 0.881	2.228 ± 0.071	**0.596 ± 0.027**	4.251 ± 0.042	3.522 ± 0.094	3.543 ± 0.068
Non-linear transform	**2.922 ± 0.752**	**2.229 ± 0.074**	0.592 ± 0.027	**4.292 ± 0.035**	**3.741 ± 0.097**	**3.723 ± 0.071**

Table 7: Comparing models trained with and without the non-linear transform (9).

E Futher discussion and figures

E.1 Examples of enhanced speech as spectograms

In Fig. 2, we present a triad of spectograms as visual illustrations of the success of our denoising method.

E.2 Convergence of metrics over EM iterations

As opposed to VAE, which needs in the order of 100 EM-iterations to learn the noise variance matrices, our method convergences to a learnt representation of the $\mathbf{W}, \mathbf{H}$ matrices within as few as 5 iterations. This indicates the power of the generative prior to create clean speech. We illustrate this by plotting the evolution of the normed distance between the true and approximated noise variance matrices $\|W^* - \hat{W}\|^2$ and $\|H^* - \hat{H}\|^2$ averaged over the entire Wall Street Journal (WSJ) test set in Figs. 3 and 4 below.
We also plot the evolution of three instrumental metrics PESQ, SI-SDR and STOI across 10 EM-iterations and average across the entire WSJ and TCD-TIMIT test sets respectively and note that the variance over the later iterations is very small. See Figures 5 and 6. We note that there is even a slight decrease in performance after iterations 4 or 5.
Since it is expensive to increase the number of EM-iterations, we set the parameter $n_{EM} = 5$.

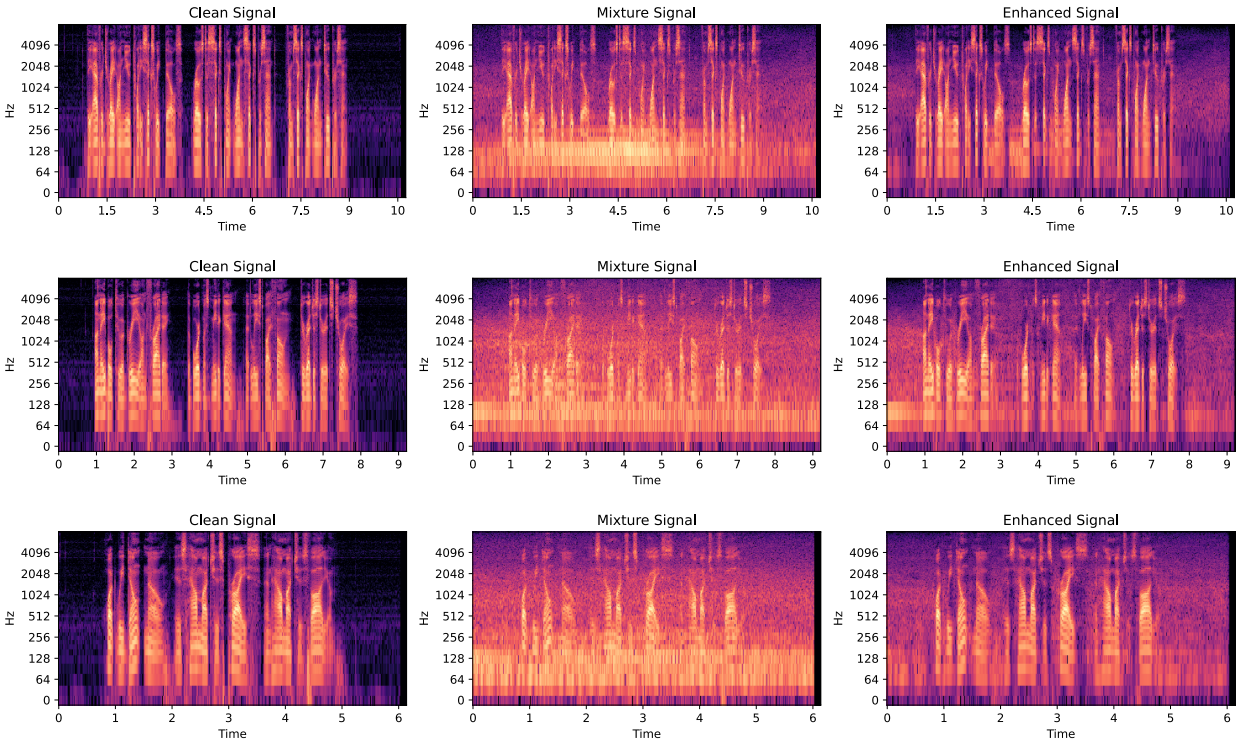

Figure 2: Three spectogram triads to showcase enhanced clean speech as ouput of the UDiffSE framework.

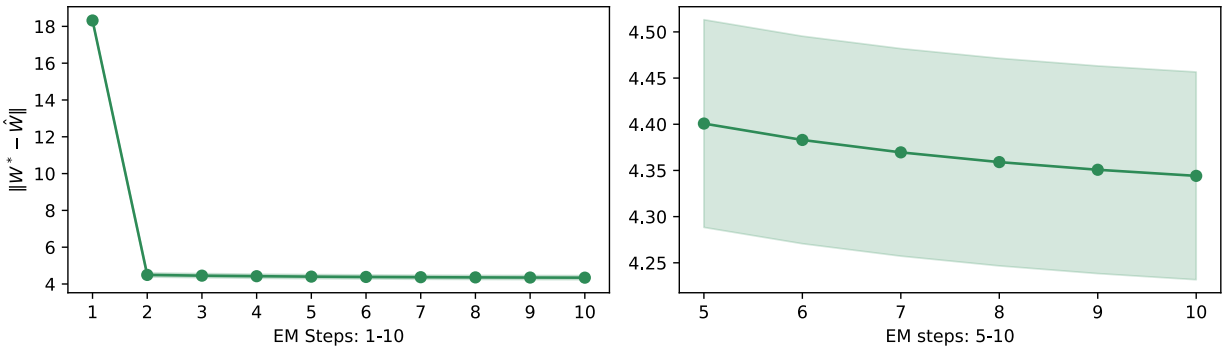

Figure 3: Evolution of normed distance between true W^* and approximated $\hat{W}$, zoomed on the RHS panel, with 1 standard deviation shaded on on both panels.

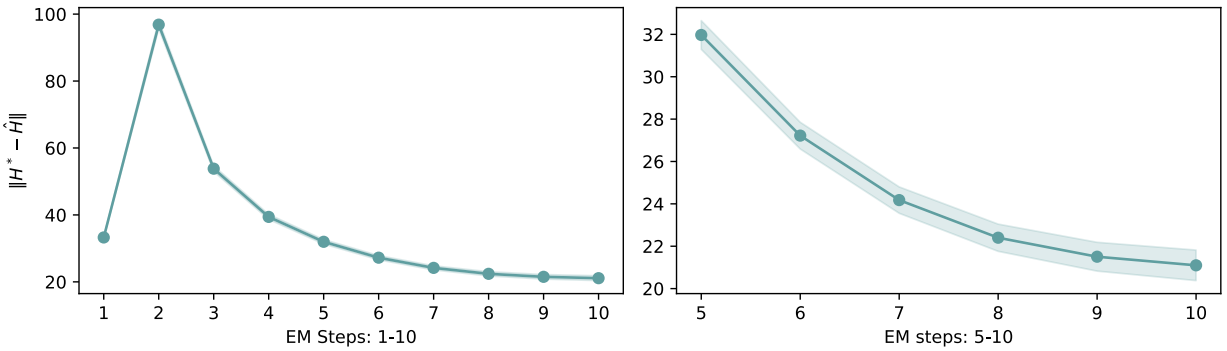

Figure 4: Evolution of normed distance between true H^* and approximated $\hat{H}$, zoomed on the RHS panel, with 1 standard deviation shaded on on both panels.

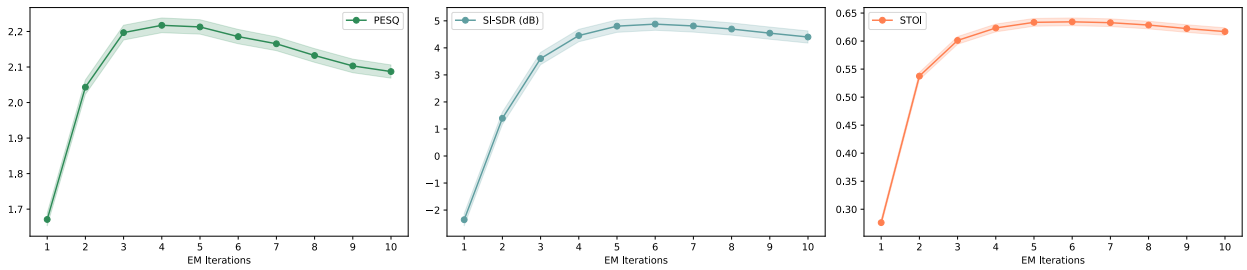

Figure 5: Evolution of 3 objective metrics over EM-iterations for test evaluation in the matched condition (WSJ dataset), with 1 standard deviation shaded on both panels.

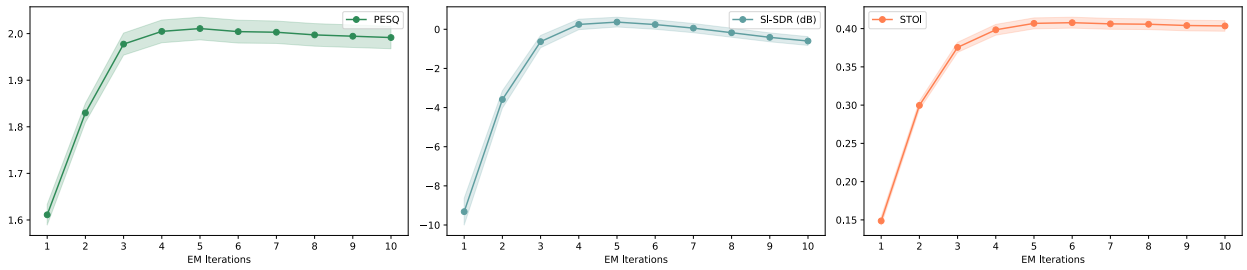

Figure 6: Evolution of 3 objective metrics over EM-iterations for test evaluation in the mismatched condition (TCD-TIMIT dataset), with 1 standard deviation shaded on both panels.

References

[1] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis". In: *Neural computation* 21.3 (2009), pp. 793–830.

[2] Tero Karras et al. "Elucidating the design space of diffusion-based generative models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26565–26577.

[3] Xiangming Meng and Yoshiyuki Kabashima. *Diffusion Model Based Posterior Sampling for Noisy Linear Inverse Problems*. 2023. arXiv: `2211.12343 [cs.LG]`.

[4] Gemma E Moran et al. "Identifiable Variational Autoencoders via Sparse Decoding". In: *arXiv preprint arXiv:2110.10804* (2021).

[5] Berné Nortier, Mostafa Sadeghi, and Romain Serizel. "Unsupervised speech enhancement with diffusion-based generative models". In: *arXiv preprint arXiv:2309.10450* (2023).

[6] Julius Richter et al. "Speech enhancement and dereverberation with diffusion-based generative models". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[7] Mostafa Sadeghi and Xavier 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.

[8] Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. 2020. arXiv: `1907.05600 [cs.LG]`.

[9] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: `2011.13456 [cs.LG]`.

[10] Yang Song et al. *Solving Inverse Problems in Medical Imaging with Score-Based Generative Models*. 2022. arXiv: `2111.08005 [eess.IV]`.