Diffusion-based speech enhancement with a weighted generative-supervised learning loss

Jean-Eudes Ayilo, Mostafa Sadeghi, Romain Serizel

DIFFUSION-BASED SPEECH ENHANCEMENT WITH A WEIGHTED GENERATIVE-SUPERVISED LEARNING LOSS

Jean-Eudes Ayilo, Mostafa Sadeghi, Romain Serizel

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

ABSTRACT

Diffusion-based generative models have recently gained attention in speech enhancement (SE), providing an alternative to conventional supervised methods. These models transform clean speech training samples into Gaussian noise, usually centered on noisy speech, and subsequently learn a parameterized model to reverse this process, conditionally on noisy speech. Unlike supervised methods, generative-based SE approaches often rely solely on an unsupervised loss, which may result in less efficient incorporation of conditioned noisy speech. To address this issue, we propose augmenting the original diffusion training objective with an ℓ_2 loss, measuring the discrepancy between ground-truth clean speech and its estimation at each diffusion time-step. Experimental results demonstrate the effectiveness of our proposed methodology.

***Index Terms*—** Speech enhancement, diffusion models, generative modeling, supervised learning.

1. INTRODUCTION

Diffusion models are a recent class of generative models that have brought significant improvements in image and audio synthesis [1, 2]. Their underlying mechanism is to gradually turn training samples into noise, and then learn a parameterized model to revert this process, thus enabling data generation from pure noise. These models are also gaining increasing interest in the speech enhancement (SE) task, whose goal is to recover a clean speech signal recorded in adverse acoustic environments. In this context, diffusion models aim at learning the distribution of speech data, encoding their temporal-spectral characteristics, in order to infer clean speech from noisy observation [3]. Additionally, this distribution is learned conditionally on the associated noisy speech data to guide the data generation process [3–6].

This generative-based SE approach is systematically different from the prevailing supervised counterpart [7, 8], which learns a deep neural network (DNN) to directly estimate clean speech or a time-frequency mask from noisy input by minimizing a supervised loss, e.g., mean squared error (MSE). In contrast, diffusion-based SE usually follows the standard unsupervised (generative-based) loss used in diffusion models, with the difference that noisy speech is provided as an additional input. While this approach could potentially be advantageous, as it models the intrinsic properties of clean speech contrary to supervised methods, we hypothesize that it might not efficiently incorporate a measure of the goodness of the estimated clean speech.

The current study aims at addressing this issue, and bridging the performance gap between the supervised and diffusion-based approaches by combining the best of the two worlds. To this end, we propose to add an ℓ_2 loss to the original generative-based diffusion loss. This extra supervised loss measures the distance between ground-truth clean speech and its estimation at each diffusion time-step. In doing so, we hope to combine the effectiveness of diffusion models in unseen noise conditions and the strength of supervised methods in seen noise conditions. Experiments are performed to evaluate the effectiveness of the proposed approach against both supervised and standard diffusion-based approaches. The results show promising performance of the proposed training methodology.

In the rest of the paper, we present an overview of diffusion-based SE methods in Section 2, followed by a review of [4]. Our proposed methodology is discussed in Section 3. Next, Section 4 presents the experiments and results, and Section 5 concludes the paper.

2. RELATED WORK

2.1. Diffusion-based speech enhancement

The fundamental concept behind diffusion models involves two main phases. Initially, within a forward process, clean data are progressively distorted by adding (usually Gaussian) noise, eventually resulting in entirely noise data following a tractable distribution like a standard Gaussian. Subsequently, through a reverse process, a DNN is trained to sequentially produce clean data, beginning from random noise sampled from the prior distribution.

Many recent studies have already used the diffusion principle for speech enhancement. [3] used Gaussian Markov

This research 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).

chains to model the forward and reverse processes, where the mean of the forward Gaussian Markov chain is a linear interpolation between the clean speech and the associated noisy one. The training objective is obtained by minimizing the Kullback–Leibler (KL) divergence between the forward and reverse Markov chains, which leads to an objective where the trained network learns to predict both Gaussian noise and non-Gaussian noise. [5] added an auxiliary classification loss to the loss function of [3] to perform noise classification and help the model better use the noise information, which resulted in improved performance. [4] leveraged stochastic differential equation (SDE) [1] to model the forward and reverse processes, and used denoising score matching [9] as the training objective.

All these works condition the diffusion process on the noisy speech to take into account the non-Gaussian nature of environmental noise. More precisely, to estimate clean speech, the reverse process is performed starting from a Gaussian noise centred on the noisy speech, which allows for iteratively recovering an enhanced version. [6] proposed to extend the forward process by incorporating a deterministic, progressive degradation of the clean speech through linear interpolation between the clean speech and the noisy speech. The training objective here consists in minimizing an ℓ_1 loss between the clean speech and the reconstruction of the clean speech at a given step of the forward degradation.

The recent work [10] identified the condition collapse problem, and proposed an auxiliary conditional generation network for generating reliable condition representations as well as a dual-path parallel network architecture to provide fine-grained condition guidance for the diffusion model. Additionally, a refinement network is trained in a supervised way that takes the enhanced speech returned by the reverse sampling and outputs a refined version. We adopt a strategy different from these approaches by including an ℓ_2 supervision loss in the training objective function.

2.2. Score-based generative model for SE (SGMSE)

In this section, we review the score-based diffusion model proposed in [4], as a closely related approach to our work. Let us consider the flattened short-time Fourier transform (STFT) representations of clean speech, noisy speech, and noise: $\mathbf{x}_0$, $\mathbf{y}$, $\mathbf{n} \in \mathbb{C}^d$, where d denotes the total number of complex-valued time-frequency (TF) bins. We assume that the noisy speech STFT is formed by the following mixture model: $\mathbf{y} = \mathbf{x}_0 + \mathbf{n}$. The objective of SE is then to recover $\mathbf{x}_0$ given $\mathbf{y}$.

As previously mentioned, the forward process of diffusion involves gradually introducing Gaussian noise to the clean speech. This process is modeled by an SDE, and its solution is represented as the stochastic process $\{\mathbf{x}_t\}_t$ [4]:

$$\mathrm{d}\mathbf{x}_t = \underbrace{\gamma\left(\mathbf{y}-\mathbf{x}_t\right)}_{:=\mathbf{f}(\mathbf{x}_t,\mathbf{y})}\mathrm{d}t + \underbrace{\left[\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t\sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}\right]}_{:=g(t)}\mathrm{d}\mathbf{w} \tag{1}$$

where $\mathbf{x}_t$ denotes the process state at time $t \in (0, T]$, $\gamma \in \mathbb{R}$ controls the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and $g(t) \in \mathbb{R}$, with fixed parameters $\sigma_{\min}$ and $\sigma_{\max}$, is the diffusion coefficient that controls the amount of noise induced by a standard Wiener process $\mathbf{w}$. Moreover, $\mathbf{f}(\mathbf{x}_t, \mathbf{y})$ is the drift term, which makes the forward process conditioned on the noisy speech. For numerical stability, the forward process starts at $t_\varepsilon \neq 0$.

The final SE objective is to reverse the above forward process in order to estimate the clean speech. To this end, one needs to find the solution to the following associated reverse process SDE [11]:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t,\mathbf{y}) + g(t)^2\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{y})\right]\mathrm{d}t + g(t)\mathrm{d}\overline{\mathbf{w}} \tag{2}$$

where $\overline{\mathbf{w}}$ denotes a standard Wiener process running backwards in time. In (2), the term $\nabla_{\mathbf{x}_t}\log p_t\left(\mathbf{x}_t|\mathbf{y}\right)$ refers to the conditional score function, which is approximated by a so-called (conditional) score model $\mathbf{s}_\theta\left(\mathbf{x}_t, \mathbf{y}, t\right)$ with parameters denoted θ. The score model can be trained by minimizing a noise-prediction loss [12] as follows

$$\min_\theta \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}\Big[L_\theta\left(\mathbf{x}_t,\mathbf{y},t,\mathbf{z}\right)\Big], \tag{3}$$

where[1]

$$L_\theta\left(\mathbf{x}_t,\mathbf{y},t,\mathbf{z}\right) := \left\|\sigma(t)\mathbf{s}_\theta\left(\mathbf{x}_t,\mathbf{y},t\right)+\mathbf{z}\right\|^2 \tag{4}$$

and $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; \mathbf{0}, \mathbf{I})$, with $\mathcal{N}_{\mathbb{C}}$ denoting the circularly-symmetric complex normal distribution. As the drift term is linear w.r.t $\mathbf{x}_t$, the transition kernel $p_{0t}\left(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y}\right)$ admits a closed-form expression [1]:

$$p_{0t}(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y}) = \mathcal{N}_{\mathbb{C}}\Big(\mathbf{x}_t; \boldsymbol{\mu}\left(\mathbf{x}_0,\mathbf{y},t\right), \sigma(t)^2\mathbf{I}\Big), \tag{5}$$

where

$$\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t) = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + \left(1-\mathrm{e}^{-\gamma t}\right)\mathbf{y} \tag{6}$$

and

$$\sigma(t)^2 = \frac{\sigma_{\min}^2\Big(\left(\sigma_{\max}/\sigma_{\min}\right)^{2t} - \mathrm{e}^{-2\gamma t}\Big)\log(\sigma_{\max}/\sigma_{\min})}{\gamma + \log(\sigma_{\max}/\sigma_{\min})}. \tag{7}$$

Once the score model is trained, it is plugged in (2), replacing the score function, and the resulting reverse SDE is solved by a Predictor-Corrector sampling procedure [1] to iteratively generate clean speech's estimates.

[1]This is the actual loss used in the implementation of [4]: `https://github.com/sp-uhh/sgmse`.

3. WEIGHTED GENERATIVE-SUPERVISED LEARNING LOSS

The training objective function described in (3) and (4) aims to train the score model by minimizing an ℓ_2 loss between the added Gaussian noise $\mathbf{z}$ and its estimation given by $\hat{\mathbf{z}} = -\sigma(t)\mathbf{s}_\theta\left(\mathbf{x}_t, \mathbf{y}, t\right)$. This suggests that, in its current form, the score model is not informed of the specific SE task it is meant to perform, as it does not explicitly take into account the environmental noise present in $\mathbf{y}$. Instead, the training loss primarily resembles the generative (unsupervised) loss typically utilized in unconditional diffusion models.

To address this issue, we propose the inclusion of a supervised loss as a form of regularization or guidance. This additional loss explicitly reinforces the SE objective during the score model's training. To accomplish this objective, we need to have an estimate of the clean speech at each diffusion time-step t during training, denoted $\hat{\mathbf{x}}_{0,t}$, to be compared against the ground-truth $\mathbf{x}_0$. Such an estimate could be provided using Tweedie's approach [13]. To this end, by combining (5) and (6), we can write

$$\mathbf{x}_t = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + \left(1 - \mathrm{e}^{-\gamma t}\right)\mathbf{y} + \mathbf{e}_t \tag{8}$$

where $\mathbf{e}_t \sim \mathcal{N}_{\mathbb{C}}(\mathbf{e}_t; \mathbf{0}, \sigma(t)^2\mathbf{I})$. If we apply Tweedie's formula independently to the real and imaginary parts of the variables, we obtain the following result:

$$\begin{aligned}\mathrm{e}^{-\gamma t}\hat{\mathbf{x}}_{0,t} + (1 - \mathrm{e}^{-\gamma t})\mathbf{y} &= \mathbf{x}_t + (\sigma(t)^2/2)\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{y}) \\ &\approx \mathbf{x}_t + (\sigma(t)^2/2)\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t). \end{aligned} \tag{9}$$

We then propose to add an ℓ_2 loss between the above expression and the associated ground-truth one, i.e., $\mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\mathbf{y}$, to the original score loss (3). This leads to the following weighted training objective for learning the conditional score model

$$\begin{aligned}&\min_\theta \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}[(1-\alpha_t)\, L_\theta\left(\mathbf{x}_t, \mathbf{y}, t, \mathbf{z}\right) + \\ &\alpha_t \left\| \mathbf{x}_t + \frac{\sigma(t)^2}{2}\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) - (\mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\mathbf{y}) \right\|^2]\end{aligned} \tag{10}$$

which is a weighted loss between the original generative-based training loss in (4) and a supervised ℓ_2 loss. α_t are time-dependent scalar weights taking values in $[0, 1]$. We propose to use the following expression for the weights

$$\alpha_t = \frac{\sigma(T) - \sigma(t)}{\sigma(T) - \sigma(t_\varepsilon)}. \tag{11}$$

In the proposed loss function (10), we aim to make a balance between two essential tasks: conditional score estimation and supervised estimation of clean speech at every time-step t. The parameter α_t is a decreasing function of time, contrary to $\sigma(t)$. Consequently, at earlier stages of the forward diffusion process, i.e., when the noise variance is small, the network is expected to give higher weights to the supervised component of the loss, while in the later stages, the score network is tasked with assigning more importance to the loss associated with Gaussian noise estimation.

4. EXPERIMENTS

Baselines. We evaluate the performance of our proposed weighted generative-supervised learning loss for SE, comparing it to the reference approach referred to as SGMSE+ [4]. Both methods utilize the same network architecture, based on the Noise Conditional Score Network (NCSN++), which consists of a multi-resolution U-Net design. Additionally, we compare these two methods with a purely supervised model that directly predicts the clean speech spectrogram from the noisy speech spectrogram input. For the supervised approach, we also use the NCSN++ network, which is trained by minimizing the MSE loss between the enhanced and clean spectrograms.

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 [14], the extended short-time objective intelligibility (ESTOI) measure [15] ($[0, 1]$), and the perceptual evaluation of speech quality (PESQ) score [16] ($[-0.5, 4.5]$). In addition, we use the DNS-MOS, as a non-intrusive objective evaluation metric [17], which provides three MOS scores: speech signal quality (SIG), background intrusiveness (BAK), and overall quality (OVRL). For all these metrics, the higher, the better.

Datasets. For training and evaluation, we used the WSJ0-QUT [18] and NTCD-TIMIT datasets [19], to allow for a cross-dataset evaluation. The WSJ0-QUT dataset combines the clean speech signals from the WSJ0 dataset [20] with noise signals from the QUT-NOISE dataset [21]. The test subset of WSJ0-QUT comprise 651 synthetic mixtures (roughly 1.5 hours). It is created by taking clean speech signals from the 'si_et_05' subset of WSJ0 (unseen speech samples) and adding noise signals sampled uniformly from the 'verification' set of the QUT-NOISE dataset with signal-to-noise ratio (SNR) values of -5, 0, and 5 dB.

The NTCD-TIMIT dataset comprises 62 English speakers (with/without Irish accent), divided into train, test, and validation subsets, where each speaker utters 98 different sentences. Duration of each utterance is approximately 5 seconds. To create the training and validation datasets, we combined the clean speech signals from the NTCD-TIMIT dataset with various types of noise from the DEMAND dataset [22]. We applied different SNR values, including -10, -5, 0, 5, and 10 dB. Each utterance in the training and validation sets was mixed with three different combinations of DEMAND noises and SNRs, resulting in a total of 12,348 training and 2,352

Table 1: Speech enhancement results (mean ± standard error) for WSJ0-QUT and NTCD-TIMIT under both matched and mismatched conditions. The best average metric value is highlighted in bold, while the second best is italicized.

Training set	Method	SI-SDR (dB)	PESQ	ESTOI	SIG-MOS	BAK-MOS	OVR-MOS
WSJ0-QUT	Input (**WSJ0-QUT**)	-2.60 ± 0.16	1.83 ± 0.02	0.50 ± 0.01	4.04 ± 0.01	2.93 ± 0.02	3.13 ± 0.01
	Supervised	**12.91 ± 0.14**	2.67 ± 0.02	**0.84 ± 0.00**	4.38 ± 0.01	**4.81 ± 0.01**	4.30 ± 0.01
	SGMSE+ [4]	10.21 ± 0.16	*2.83 ± 0.02*	0.81 ± 0.00	*4.52 ± 0.01*	4.70 ± 0.01	*4.31 ± 0.01*
	Proposed	*10.40 ± 0.15*	**2.88 ± 0.02**	*0.83 ± 0.00*	**4.56 ± 0.01**	*4.73 ± 0.00*	**4.37 ± 0.01**
NTCD-TIMIT	Supervised	**9.27 ± 0.15**	2.36 ± 0.02	**0.75 ± 0.00**	4.22 ± 0.01	**4.68 ± 0.01**	4.15 ± 0.02
	SGMSE+ [4]	7.32 ± 0.15	*2.51 ± 0.02*	0.72 ± 0.01	*4.47 ± 0.01*	4.60 ± 0.01	*4.24 ± 0.01*
	Proposed	*7.55 ± 0.14*	**2.61 ± 0.02**	**0.75 ± 0.00**	**4.55 ± 0.01**	*4.66 ± 0.00*	**4.34 ± 0.01**
NTCD-TIMIT	Input (**NTCD-TIMIT**)	-7.81 ± 0.22	1.77 ± 0.02	0.31 ± 0.00	3.51 ± 0.01	2.28 ± 0.02	2.69 ± 0.01
	Supervised	**8.57 ± 0.19**	2.18 ± 0.02	*0.54 ± 0.01*	3.69 ± 0.01	4.26 ± 0.01	3.42 ± 0.02
	SGMSE+ [4]	6.21 ± 0.23	*2.35 ± 0.02*	0.53 ± 0.01	*4.02 ± 0.01*	*4.30 ± 0.01*	*3.68 ± 0.01*
	Proposed	*7.97 ± 0.18*	**2.46 ± 0.02**	**0.57 ± 0.01**	**4.14 ± 0.01**	**4.37 ± 0.01**	**3.83 ± 0.01**
WSJ0-QUT	Supervised	**5.98 ± 0.22**	2.02 ± 0.02	**0.50 ± 0.01**	3.76 ± 0.01	*4.19 ± 0.01*	3.34 ± 0.02
	SGMSE+ [4]	1.28 ± 0.27	*2.05 ± 0.02*	0.45 ± 0.01	*4.04 ± 0.01*	4.05 ± 0.01	*3.57 ± 0.02*
	Proposed	*4.42 ± 0.23*	**2.08 ± 0.02**	*0.48 ± 0.01*	**4.16 ± 0.01**	**4.20 ± 0.01**	**3.76 ± 0.01**

validation mixtures. For the test subset, we retained the same noisy speech signals as provided in the NTCD-TIMIT dataset. These noisy samples were generated by adding six different noise types, including Living Room, White, Cafe, Car, Babble, and Street, with SNRs of -5 dB, 0 dB, and 5 dB. The test set comprises 810 mixtures.

Hyperparameters setting for SDE and STFT. Input data representations, SDE, and STFT representations follow the same settings as in [4]. Specifically, the STFT of the speech data, with a sampling rate of 16 kHz, is computed with a window size of 512, a hop length of 128 (75% overlap) and a Hann window which gives $F = 256$ as the number of frequency bins. For the drift and diffusion coefficients of SDE in (1), the parameters are set as $\gamma = 1.5, \sigma_{\text{min}} = 0.05, \sigma_{\text{max}} = 0.5$. The minimum and maximum process times are set to $t_\varepsilon = 0.03$ and $T = 1$, respectively.

Results. In Table 1, we present the average speech enhancement metrics for all the cross-dataset configuration settings, along with the corresponding standard error of the mean. To clarify, we use the term "matched condition" when the model is trained and tested on the same dataset, while "mismatched condition" refers to cases where the model is evaluated on a test set from a different dataset than the one it was trained on.

From Table 1, we can draw several conclusions. First, in both matched and mismatched conditions, the supervised method consistently outperforms the two diffusion-based methods when considering the SI-SDR, ESTOI, and BAK-MOS metrics. Our proposed method ranks second in performance. Nevertheless, when trained and evaluated on the NTCD-TIMIT dataset, the ESTOI and BAK-MOS metrics show a different trend. In these cases, the supervised method underperforms our proposed method. For the PESQ, SIG-MOS, and OVR-MOS metrics, the proposed method consistently performs the best.

A noteworthy observation from these findings is that in the matched conditions, when the supervised method performs the best, the gap between the supervised and SGMSE+ tends to narrow when the supervision loss is added. This trend also holds for the mismatched conditions.

In summary, our proposed method appears to inherit some capabilities from the supervised method, striving to match its performance in terms of SI-SDR, ESTOI, and BAK-MOS. Simultaneously, it retains and even improves upon the strengths of the baseline SGMSE+ method, resulting in better performance in terms of ESTOI, PESQ, SIG-MOS, and OVR-MOS. This suggests that the supervision loss provides valuable feedback for score estimation. Code is available online.[2]

5. CONCLUSION

In this paper, we addressed the problem of diffusion-based speech enhancement and introduced a supervised training loss component alongside the generative-based Gaussian noise prediction loss. This weighted loss balances the original noise prediction loss with an ℓ_2-based supervision loss, enhancing the mapping between clean and noisy speech by incorporating clean speech estimates at every diffusion time-step. This additional loss aids in training a score model better optimized for speech enhancement. Our experiments, conducted in both matched and mismatched conditions, demonstrate that our approach combines the strengths of supervised methods and diffusion-based approaches, resulting in improved performance. Future research directions involve exploring alternative supervised loss functions to ℓ_2 loss and developing more efficient adaptive weighting mechanisms.

[2] https://github.com/jeaneudesAyilo/weighted_generative_supervised_DiffSE

6. REFERENCES

[1] 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 (ICLR)*, 2021.

[2] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations (ICLR)*, 2020.

[3] 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.

[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*, vol. 31, 2023.

[5] Y. Hu, C. Chen, R. Li, Q. Zhu, and E. S. Chng, "Noise-aware speech enhancement using diffusion probabilistic model," *arXiv preprint arXiv:2307.08029*, 2023.

[6] 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.

[7] 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.

[8] 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.

[9] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[10] W. Tai, F. Zhou, G. Trajcevski, and T. Zhong, "Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 13627–13635.

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

[12] D. P. Kingma and R. Gao, "Understanding diffusion objectives as the elbo with simple data augmentation," in *37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[13] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.

[14] 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, pp. 626–630.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Garofolo, John S., Graff, David, Paul, Doug, and Pallett, David, "CSR-I (WSJ0) Complete," May 2007, Artwork Size: 9542041 KB Pages: 9542041 KB.

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

[22] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*. AIP Publishing, 2013, vol. 19.