Clarity or Confusion: The Double-Edged Sword of Diffusion-based Speech Enhancement

Anonymous submission to Interspeech 2024

Abstract

Diffusion-based generative models (DGMs) have recently garnered lots of attention also in the speech processing field following the promising results achieved in computer vision. Recent studies [1–5] suggest that DGMs have better generalization and produces less intrusive artifacts than discriminative methods. However, to our best knowledge, no systematic experimental review exists on this subject with a fairer comparison with state-of-the-art (SotA) discriminative models. In this work we test the generalization capability of DGM for SE, comparing three different models on 4 datasets with mismatch characteristics. We evaluate the performance using signal level, perceptual and semantical characteristics of the enhanced speech 8 metrics. Our results show that DGMs methods do not have yet a clear advantage over SotA discriminative ones for SE and are also more prone to hallucinating the content of the speech especially in highly mismatched conditions such as real-world spontaneous speech.

Index Terms: speech enhancement, speech denoising, generative modeling, diffusion-based generative models

1. Introduction

In the past four years, diffusion generative models (DGMs) [6–8] have become the go-to generative method. Several previous works have applied DGMs successfully for speech enhancement (SE) tasks, e.g. speech denoising [1, 2], joint denoising and dereverberation [3], or, more generally, even speech restoration/regeneration [4, 5] and target speaker extraction/enhancement [9, 10]. Compared to discriminative approaches, in these works, DGMs have been found to have the particularly desirable property of better generalization [1]. They have also been found to be less prone to introduce unpleasant artifacts [3, 4], and, in fact, they generally achieve better perceptual metrics.

However, they have also been found to be prone to some peculiar artifacts by two very recent works [11, 12]. Such work observed that they can introduce phoneme substitution and mumbled speech, and that these artifacts are not usually measured from non-intrusive performance metrics such as DNSMOS [13]. To detect such artifacts more reliably [11] proposes to use the phoneme-level Levenshtein distance, while [12] resorts to word accuracy.

As such, it is not yet clear how current SotA DGMs for SE actually compare with their discriminative counterpart and, if a gap exists, what are the problems that needs to be addressed in the near future. In fact, in most aforementioned previous works there lacks a true comparison with a really SotA discriminative method. Moreover, with the exception of [1] the experimental validation is performed mainly on synthetic datasets only and often limited to scenarios that closely match the training set domain.

Based on the findings in [11, 12], in this work, we compare DGM approaches in a more comprehensive manner with more up-to-date SotA discriminative approaches. For our comparison, we purposely built 4 evaluation sets with different levels of mismatch: different noise domain, different noise domain and signal-to-noise ratio (SNR) and, also, even real-world "in-the-wild" spontaneous speech. Our goal is to investigate the generalizability of both discriminative and DGMs and their behavior under different challenging conditions. In order to evaluate comprehensively the performance, we make use of up to 8 metrics both intrusive (e.g. PESQ [14] and SDR [15]), nonintrusive and ones that can detect semantical distortions in the enhanced speech (e.g. word-error-rate and phoneme error rate).

As another additional contribution, introduce for the SE task a novel mixture-conditioned diffusion-based framework (MC-DiffSE) which follows the paradigm of text-conditioned audio generative models [16]. That is, differently from [1–3] the diffusion process is defined by a conditional stochastic differential equation (SDE). This model allows us to also study the impact of 6 reverse diffusion samplers mainly introduced in the computer vision field; some of which requiring as little as 5 reverse diffusion steps.

Overall our results partially confirm that DGMs overall can obtain better non-intrusive perceptual metrics than discriminative models. However, compared to discriminative approaches, they are also more prone to failure on real-world spontaneous speech data as they are usually trained with high-quality recording speech from e.g. audiobooks. The higher WER scores obtained by DGMs compared to discriminative approaches again also confirm the findings of [11, 12], of DGMs suffering from hallucinations causing babbled speech. Here we observe that this phenomenon is amplified as the test data is more mismatched, and, again, overall the generalization is poorer than what is obtained with SotA discriminative approaches.

2. Diffusion-base Generative Models for SE

The score-based generative modeling (SGM) through stochastic differential equations (SDEs) framework [8] offers a unified approach to express most aforementioned DGM-based SE works. Following this framework, the diffusion process is formalized as a solution to a continuous SDE associated with a continuous forward process $\{\mathbf{x}(t)\}_{t=0}^{T}$ which starts at $t = 0$ and runs up to an arbitrary chosen time T:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \tag{1}$$

where $f(\mathbf{x}, t) : (\mathbb{R}^{\mathrm{d}_{\mathbf{x}}}, \mathbb{R}) \rightarrow \mathbb{R}^{\mathrm{d}_{\mathbf{x}}}$ is the *drift coefficient*, $\mathbf{x} \in \mathbb{R}^{\mathrm{d}_{\mathbf{x}}}$ is the random vector (e.g. speech or image) of interest with dimensionality $\mathrm{d}_{\mathbf{x}}$. $\mathbf{w} \in \mathbb{R}^{\mathrm{d}_{\mathbf{x}}}$ represents a standard Wiener/Brownian process while the scalar $g(t) : \mathbb{R} \to \mathbb{R}$ is the noise schedule that dictates the amount of noise added at each time step t. i.e. by choosing $g(t)$ such that the amount of noise injected increases from $t = 0$ to $t = T$ the probability density function associated with the stochastic process $\{\mathbf{x}(t)\}$ converges to a prior standard distribution. Thus, with the associated reverse SDE is possible to sample from the prior distribution to obtain data points in our target distribution:

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] dt + g(t) d\mathbf{w}, \qquad (2)$$

here dt denotes a negative time-step in the reverse process and $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is in practice approximated with a trainable DNN-based *score function* $s_\theta(\mathbf{x}, t)$ parametrized by θ.

The aforementioned DGM-based SE works differ mainly in how the forward SDE is defined in terms of the drift coefficient and the noise schedule. For example [1] formulates the diffusion process as DDPM [6] and thus their method, following [8] corresponds to the following forward SDE:

$$\mathrm{d}\mathbf{x}_t = -\frac{1}{2}\{m(t)\beta(t)\mathbf{x} + (1 - m(t))\beta(t)\mathbf{y}\}\mathrm{d}t + \beta(t)\mathrm{d}\mathbf{w}, \qquad (3)$$

where $m(t)$ defines an interpolation schedule([8] used a linear schedule) between the mixture noisy signal $\mathbf{y}$ and the ground-truth clean speech $\mathbf{x}$ while $\beta(t)$ defines the noise schedule.

The SGMSE works [2, 3] instead define the drift term as $\mathbf{f}(\mathbf{x}, \mathbf{y}, t) := \gamma(\mathbf{y} - \mathbf{x}_t)$ where γ is a *stiffness* parameter which controls how $\mathbf{x}$ at the current step t drifts towards the noisy mixture $\mathbf{y}$:

$$\mathbf{x}(t) = e^{-\gamma t}\mathbf{x}_{t=0} + (1 - e^{-\gamma t})\mathbf{y} + \sigma(t)\mathbf{z}, \qquad (4)$$

where $\mathbf{z} \sim \mathcal{N}_{\mathbb{R}}(\mathbf{z}; 0, \mathbf{I})$, $\mathbf{I}$ is identity matrix and $\sigma(t)$ is quantity directly related to the noise schedule. Another alternative is instead to define the SE task as a conditional generation task following [7]. The forward and reverse SDEs for the diffusion process remain effectively the same as Eq. 1-2 but the trainable score function $s_\theta(\mathbf{x}, t)$ in the reverse process is now conditioned on the mixture signal $\mathbf{y}$:

$$d\mathbf{x} \approx \left[f(\mathbf{x}, t) - g(t)^2 s_\theta(\mathbf{x} \mid \mathbf{y}, t)\right] dt + g(t) d\mathbf{w}. \qquad (5)$$

We call this approach mixture-conditioned diffusion SE (MC-DiffSE) in the following.

Note that SGMSE [2, 3] and MC-DiffSE are trained in the same manner, using the *score matching* method [8].

3. Experimental Setup

3.1. Evaluation Metrics

As said we adopt a wide variety of metrics to assess SE performance from various aspects; e.g. for signal-level, we use signal-to-distortion ratio (SDR) [15]. DNSMOS [13] overall (D-OVRL), speech intelligibility (D-SIG), and noise suppression (D-BAK) are instead used as a proxy for human perceptual evaluation. Since DNSMOS is non-intrusive it can be also used on real-world data. We also use STOI [17] and PESQ [14] which are intrusive and thus, compared to DNSMOS, more susceptible to estimate vs. reference misalignments that can arise from generative models. We also use word error rate (WER) as computed with Whisper [18] `medium.en` to assess if the SE algorithm does not distort the speech significantly[1] and preserves its semantic content as well as phoneme error rate (PER), following the findings in [11]. This latter is computed using Wav2Vec 2.0 [19] fine-tuned on Commonvoice [20] for phoneme recognition.

3.2. Datasets

To assess the generalization capabilities of the models in exam we created 3 different datasets and, for one of these, an additional version with lower mean SDR. These datasets are described in detail in the following; in Table 1 we report metrics as computed with the input noisy signal.

Dataset	SDR	STOI	PESQ	D-OVRL	WER
LibriFUSS	3.55	0.83	2.13	2.09	10.89
LibriAudioset	4.51	0.81	2.09	1.95	14.5
LibriAudioset-LSNR	2.19	0.72	1.80	1.73	31
CHiME6-CT	N.A.	N.A.	N.A.	2.02	19.87

Table 1: *Datasets input noisy signal metrics. For CHiME6-CT only non-intrusive metrics can be computed.*

3.2.1. LibriFUSS and LibriAudioSet

We created two datasets using LibriSpeech [21] `train-clean-100`, `dev-clean` and `test-clean` as the clean speech source. For the first one, LibriFUSS, we used the FUSS [22] dataset recipe[2] to create background noise soundscapes through Scaper [23] from Freesound [24]. As in [25] we removed from Freesound sound classes that can contain speech and, from LibriSpeech, all speech utterances which have poor recording quality (using DNSMOS). These two are then mixed randomly to create 10 s noisy speech utterances with speech signal-to-noise ratio (SNR) sampled from $\mathcal{U}(-10, 20)$ dB. We created 40k training, 2k validation and 2k test samples.

For LibriAudioSet we followed the same approach but used AudioSet as the noise source. We used Silero VAD [26] to exclude audio-clips that contain speech. For this latter we created another version (LibriAudioSet-LSNR) in which the SNR is sampled instead from $\mathcal{U}(-20, 20)$ dB. We thus created 2 distinct tests sets with $2k$ samples each since we use this only for evaluation.

3.2.2. CHiME-6 Close Talk

To evaluate denoising capability in a real-world scenario with spontaneous speech we use the CHiME-6 corpus eval set [27]. In detail, since we concern here only with speech denoising, as done in the UDASE challenge [28], we used the close talk microphones from each speaker (channel averaged since they are binaural). The result is a dataset with spontaneous speech but not very noisy overall. Contrary to UDASE, we only selected segments which only contain a single speaker as we want to be able to compute the WER in an accurate manner. A total of 509 samples were thus obtained in this way.

3.3. Models and Techniques

We train all models with Adam optimizer [29], on the standard LibriFuss dataset. We describe now each one in detail thereafter.

[1] Whisper is robust enough to handle minor distortions due to its large and diverse training data.

[2] https://github.com/google-research/sound-separation/tree/master/datasets/fuss

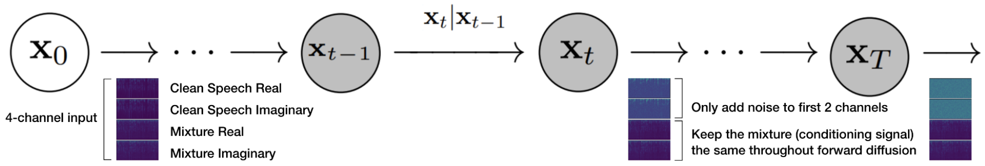

Figure 1: *Illustration of the forward diffusion process of our MC-DiffSE Model*

3.3.1. TF-GridNet

As a more up-to-date, true SotA discriminative baseline we use here TF-GridNet. In detail, following [30] we use the model with the best compromise between performance and computational requirements (Row 7 in [30]), with the exception that here we use a short-time Fourier transform (STFT) window size of 32 ms and 8 ms hop size since we are dealing with 16 kHz data. The model is trained with learning rate (lr) $1e^{-3}$, l2 norm gradient clipping of 1 and batch size 8. The lr is halved if no improvement is observed for 5 epochs.

3.3.2. SGMSE

To ensure our reproducing fidelity, we used the official repository[3] and trained an SGMSE model with the NCSNpp U-Net [3] backbone network ($\sim$ 66 millions parameters) on our LibriFUSS dataset until convergence. The lr is set to $1e^{-4}$ and the batch size is 8.

3.3.3. MC-DiffSE

As an additional DGM for comparison, we used the HuggingFace Diffusers library[4] to implement the mixture-conditioned diffusion approach described in Section 2. This model architecture is also U-Net based but has more than one order of magnitude fewer parameters than NCSNpp, bringing it closer to the TF-GridNet number, reported in Table 3. As in SGMSE, the diffusion process is performed in the complex STFT domain, and we adopt the same STFT scaling as SGMSE but, as explained in Section 2 here the score model is fed as an additional signal also the noisy mixture $\mathbf{y}$.

The network is derived from Stable Diffusion [31]. The model takes in input the complex STFT mixture $\mathbf{y} \in \mathbb{C}^{f \times t}$ and the complex STFT speech signal $\mathbf{x} \in \mathbb{C}^{f \times t}$ at various diffusion steps. f denote here the frequency index here and t the frame index. These are fed together into the model as a $\mathbb{R}^{4 \times f \times t}$ tensor by concatenating the real and imaginary part of each complex STFT input, as is illustrated in Figure 1. MC-DiffSE consists of an input convolutional layer converting the four-channel input to 32 channels. Next, are four convolutional downsampling blocks, each outputting a 64-channel latent vector using a 3×3 kernel. We use a self-attention middle block to leverage the mixture conditioning in reconstruction, followed by four upsampling blocks and a final convolutional layer to provide a 2 channel output. This model was trained for $100,000$ steps, a batch size of 16, and a cosine lr decay scheduler starting from $5.12e^{-4}$. Code for our model is publicly available on GitHub [5].

4. Results and Discussion

4.1. Reverse Diffusion Sampler Evaluation

Works such as [32] have shown that DGMs can generate images with higher quality and conditioning control than previous SotA models. However, the iterative nature of their generation leads to extended inference times. Whereas Generative Adversarial Models (GANs) or Variational Autoencoders (VAEs) based models can generate in a single pass, most DGMs require tens to hundreds of forward passes to generate a satisfying image. On the other hand, several reverse diffusion samplers have been proposed recently (e.g. DPM single and multi step [33] and DEI multi step [34]) that could allow for a reduced number of steps without sacrificing performance. We performed sampler comparison for sampling to address this drawback and more fully explore this space.

We selected the following samplers for this study: DDPM [6], DDIM [35] PNDM, [36] DPM Single-Step, DPM Multi-Step [33], and DEI Multi-Step [34] samplers. We split these samplers into two main categories. Traditional sampling with longer sample times: DDPM, DDIM, and PNDM. Contrast to modern sampling, as few as five sampling steps are required: DPM single step sampling, DPM multi step sampling, and DEI multi step sampling. In Table 2 we compare these six different reverse diffusion samplers.

Since this study is computationally demanding, we used a smaller 200 samples subset of LibriFUSS test set. We can see that DDPM consistently performs best across all quality metrics. However it is also the sampler which is more computational demanding. Some samplers allow for a reduced numbers of steps leading to a significant inference speed advantage, which becomes comparable to TF-GridNet inference speed. This however comes at the cost of noticeable performance loss. Our choice of DDPM for final sampling does not invalidate these other samplers, to the contrary this study shows they are able to produce similar quality with a 60× speed up. We include some examples for the different samplers generation online.[6]

Model	Steps	Speed	SDR	STOI	PESQ	D-OVRL	WER
PNDM	200	3.298 s	9.113	0.887	2.364	2.364	0.212
DDPM	200	3.005 s	**10.238**	**0.929**	**2.522**	**2.545**	**0.198**
DDIM	200	3.082 s	9.355	0.894	2.332	2.490	0.233
DPM Multi	40	0.527 s	10.051	0.877	2.324	2.341	0.230
DEIS Multi	40	0.533 s	9.288	0.886	2.350	2.384	0.226
DPM Single	**5**	**0.056 s**	9.196	0.877	2.284	2.333	0.222

Table 2: *Performance of MC-DiffSE for different reverse diffusion samplers on a* 200 *samples subset of LibriFUSS test set.*

4.2. Matched-Condition Evaluation

In Table 3 we report results on LibriFUSS test set, including the number of parameters and inference time for one 10 s input as computed on a A100 40GB NVIDIA GPU. Since all models were trained on LibriFUSS, this set is fully-matched with the training. As such it is an ideal situation that seldom happens in real-world applications.

We can observe that, as expected, TF-GridNet obtains overall the best performance and is also much faster in inference as

[3] https://github.com/sp-uhh/sgmse

[4] https://huggingface.co/docs/diffusers

[5] https://github.com/mcdiffuse/mcdiffuse_code

[6] our GitHub repository.

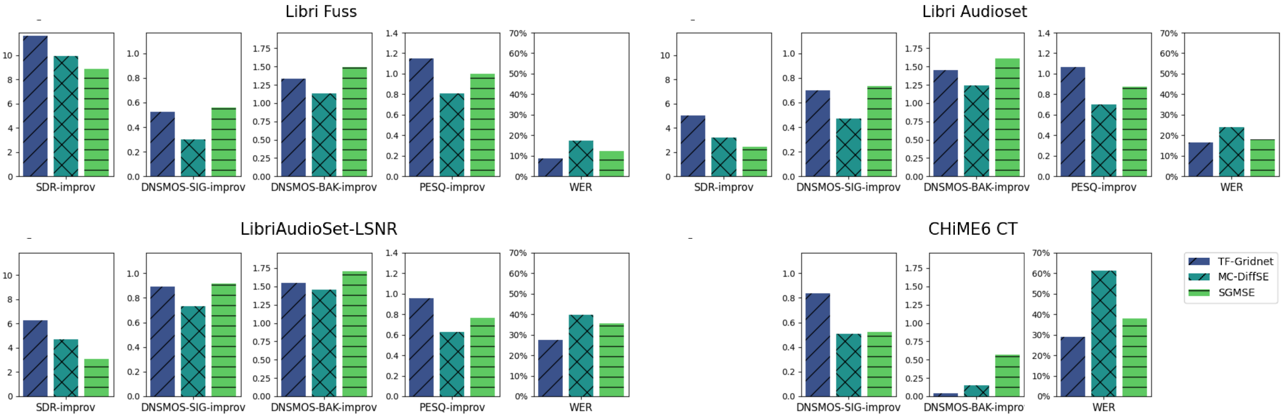

Figure 2: *TF-Gridnet, SGMSE, and MC-DiffSE evaluation on all datasets. We report, for each metric, the average improvement over the mixture. Only non-intrusive metrics can be reported for the CHiME6-CT dataset.*

it does not have to perform any iterative procedure. This confirms what has been observed in [1] where discriminative models were also been observed to be better with in-domain data. Between the two DGMs models, SGMSE performs better. MC-DiffSE has slightly higher SDR but worse metrics overall, in particular, regarding WER.

Model	# Param.	Speed	SDR	STOI	PESQ	D-OVRL	WER
TF-GridNet	3.7 M	**0.14 s**	**15.162**	**0.938**	**3.279**	**2.910**	**0.088**
SGMSE	66 M	3.88 s	12.438	0.922	3.130	2.99	0.123
MC-DiffSE	2.5 M	3.01 s	13.509	0.911	2.935	2.609	0.174

Table 3: *Performance of the different models on the matched-conditions test set. MC-DiffSE uses the DDPM sampler.*

4.3. Robustness Evaluation

In Fig 2 we report a bar plot for all the 4 datasets in exam. Here we report both DNSMOS-BAK and SIG separately as we want to assess separately intelligibility and noise suppression capabilities. We are mainly interested to see how performance changes when 1) the noise-type domain is shifted (LibriAudioSet) 2) both noise-type and SNR is changed (LibriAudioSet-LSNR) and 3) the models are tested on "in-the-wild" spontaneous speech data (CHiME6-CT) with mild noise levels.

We can observe that as the evaluation data is shifted further from the training domain, overall performance degrades for all models in exam. Regarding intrusive metrics (SDR and PESQ), again, as expected the discriminative model comes on top on all scenarios. A less expected result is that the same is also observed for WER, as TF-GridNet obtains significantly the best results across the 4 scenarios. On the other hand, SGMSE seems to exhibit overall better DNSMOS-BAK than the other approaches, suggesting better noise suppression. These two trends are surprisingly quite constant across the scenarios in exam. Another evident result is that MC-DiffSE performs rather poorly compared to the other two. This could be due to the reduced parameter count which may be insufficient for competing with SotA generative modeling techniques. On the other hand, it serves as a useful reference to see how discriminative and DGMs methods compare when they have similar number of parameters. SGMSE also obtains better DNSMOS-SIG scores on all datasets except for CHiME-6 where TF-GridNet obtains a much better improvement. This latter result coupled with the rather large WER difference ($\sim 10\%$) compared to TF-GridNet suggests that SGMSE and also MC-DiffSE may struggle more with spontaneous speech and are particularly susceptible to this kind of mismatch compared to discriminative methods.

As a further analysis, as suggested in [11], we also compute the phoneme-level error rate (PER) for our best discriminative model (TF-Gridnet) and our best generative model (SGMSE). As shown in Table 4, where we also report the average error components, the phoneme-level errors are much closer to the word level, as here the effect of the implicit language model of Whisper is absent. We can observe that, for both models, most of the errors come from phone substitutions, and in general SGMSE has negligible more insertions and deletions than TF-GridNet. These results are not decisive and suggest the need for better speech quality and semantical metrics that can also be computed on spontaneous real-world speech. Phoneme recognition might not be suitable for this data, as its difficulty can obscure the sources of errors in SE algorithm outputs.[7]

Model	PER	Substitutions	Insertions	Deletions
TF-Gridnet	**0.552**	**9.498**	**5.126**	**0.702**
SGMSE	0.561	9.546	5.450	0.807

Table 4: *Phoneme level error metrics for highest performing discriminative model vs. highest performing generative model. Scores computed on CHiME6-CT real-world test set.*

5. Conclusion

In this work, we probed the generalization of DGMs and discriminative models for speech enhancement using 4 different datasets with different degrees of mismatch, including real-world data. We also developed a novel SE DGM which uses the input noisy mixture as a conditioning signal to aid in the diffusion process. Additionally, this study has provided valuable insights into the performance of cutting-edge reverse diffusion sampling algorithms, highlighting the various trade-offs involved. Our findings suggest that while DGMs show promises as suggested by their superior denoising capabilities, in order to be competitive with SotA discriminative methods for SE task, they could benefit from new methods such as incorporating language modeling to aid its preservation of speech semantics and better generalization to real-world spontaneous speech.

[7] More PER and WER qualitative examples are on our Github

6. References

[1] 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*, pp. 7402–7406.

[2] 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.

[3] 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.

[4] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A Diffusion-Based Stochastic Regeneration Model for Speech Enhancement and Dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2724–2737.

[5] 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.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[7] Y. Song and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," in *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[9] T. Nguyen, G. Sun, X. Zheng, C. Zhang, and P. C. Woodland, "Conditional diffusion model for target speaker extraction," *arXiv preprint arXiv:2310.04791*, 2023.

[10] N. Kamo, M. Delcroix, and T. Nakatan, "Target speech extraction with conditional diffusion model," *Proc. of Interspeech*, 2023.

[11] J. Pirklbauer, M. Sach, K. Fluyt, W. Tirry, W. Wardah, S. Moeller, and T. Fingscheidt, "Evaluation metrics for generative speech enhancement methods: Issues and perspectives," in *Speech Communication; 15th ITG Conference*, 2023, pp. 265–269.

[12] D. de Oliveira, J. Richter, J.-M. Lemercier, T. Peer, and T. Gerkmann, "On the behavior of intrusive and non-intrusive speech enhancement metrics in predictive and generative settings," in *Speech Communication; 15th ITG Conference*. VDE, 2023, pp. 260–264.

[13] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022*. IEEE, 2022, pp. 886–890.

[14] 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 *2001 ICASSP (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[15] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469.

[16] J. B. Li, J. S. Michaels, L. Yao, L. Yu, Z. Wood-Doughty, and F. Metze, "Audio-journey: Efficient visual+LLM-aided audio encodec diffusion," in *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023. [Online]. Available: https://openreview.net/forum?id=vzMXsTCdFB

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023, pp. 28 492–28 518.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 ICASSP*. IEEE, 2015, pp. 5206–5210.

[22] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *ICASSP 2021*. IEEE, 2021, pp. 186–190.

[23] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[24] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *ISMIR 2017. p. 486-93.*, 2017.

[25] E. Tzinis, J. Casebeer, Z. Wang, and P. Smaragdis, "Separate but together: Unsupervised federated learning for speech enhancement from non-iid data," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 46–50.

[26] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[27] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[28] S. Leglaive, L. Borne, E. Tzinis, M. Sadeghi, M. Fraticelli, S. Wisdom, M. Pariente, D. Pressnitzer, and J. R. Hershey, "The chime-7 udase task: Unsupervised domain adaptation for conversational speech enhancement," *arXiv preprint arXiv:2307.03533*, 2023.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[32] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[33] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," 2023.

[34] Q. Zhang and Y. Chen, "Fast sampling of diffusion models with exponential integrator," 2023.

[35] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2022.

[36] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=PlKWVd2yBkY