



Non-intrusive Speech Quality Assessment with Diffusion Models Trained on Clean Speech

Danilo de Oliveira, Julius Richter, Jean-Marie Lemerrier, Simon Welker, Timo Gerkmann

Signal Processing Group, University of Hamburg, Germany

{danilo.oliveira, julius.richter, jean-marie.lemerrier, simon.welker,
timo.gerkmann}@uni-hamburg.de

Abstract

Diffusion models have found great success in generating high quality, natural samples of speech, but their potential for density estimation for speech has so far remained largely unexplored. In this work, we leverage an unconditional diffusion model trained only on clean speech for the assessment of speech quality. We show that the quality of a speech utterance can be assessed by estimating the likelihood of a corresponding sample in the terminating Gaussian distribution, obtained via a deterministic noising process. The resulting method is purely unsupervised, trained only on clean speech, and therefore does not rely on annotations. Our diffusion-based approach leverages clean speech priors to assess quality based on how the input relates to the learned distribution of clean data. Our proposed log-likelihoods show promising results, correlating well with intrusive speech quality metrics and showing the best correlation with human scores in a listening experiment.

Index Terms: speech quality assessment, diffusion models

1. Introduction

Speech quality estimation is paramount for evaluating algorithms that tackle speech processing tasks, such as speech enhancement, coding, and synthesis. The golden standard for speech quality estimation is widely considered as the mean opinion scores (MOS) obtained during listening experiments, where participants are asked to rate audio samples. However, listening experiments are expensive, time-consuming and can suffer from listener bias if the instructions are not adequately designed. For these reasons, many instrumental metrics have been proposed to attempt to mimic the result of such listening experiments.

Instrumental metrics can be handcrafted, which include signal-based metrics such as scale invariant signal-to-distortion ratio (SI-SDR) [1] or signal-to-noise ratio (SNR), as well as perceptual metrics integrating some modeling of the human auditory model, like Perceptual Evaluation of Speech Quality (PESQ) [2], its successor Perceptual Objective Listening Quality Analysis (POLQA) [3] or Virtual Speech Quality Objective Listener (ViSQOL) [4]. Typically, these metrics are intrusive, i.e., they require a reference clean speech signal matching the test utterance. Recording such a clean reference is, however, impractical in real-life scenarios.

In order to avoid relying on reference speech at inference, learning-based metrics based on deep neural networks (DNNs) have been proposed, requiring reference speech only at training. These methods typically try to predict the MOS provided in large labeled speech datasets. These include for example DNSMOS [5], NISQA [6] and NORESQA-MOS [7]. However, these supervised methods might struggle to predict the quality

of speech for samples that contain characteristics not encountered during training. Furthermore, supervised methods require large labeled datasets, which can be either inaccessible to the speech research community or susceptible to low-quality annotations.

In this paper, we focus on predicting speech quality in an unsupervised fashion, i.e., training our method only on clean speech data. Similar works include SpeechLMscore [8] and VQScore [9]. SpeechLMscore leverages a language model trained on clean speech tokens and computes the likelihood of the test speech sequence according to the language model vocabulary [8]. Lower likelihood will then indicate that the test speech deviates from the clean speech representation of the language model, thereby suggesting low speech quality. In VQScore [9], the authors suggest to train a vector-quantized variational autoencoder (VAE) on clean speech, and inspect the quantization error at the bottleneck of the model. Since the quantized units define a coarse representation of clean speech, observing a large quantization error suggests that the input speech is not well represented by the discrete codebook and therefore it should be considered of low quality.

In this work, we follow similar ideas as SpeechLMscore and VQScore, but instead choose to use a diffusion model [10, 11] for providing us with a representation of clean speech. We compute the likelihood of a test speech by integrating a specific ordinary differential equation (ODE) which, as Song et al. showed [11], provides an exact computation of likelihood using a trained diffusion model. Given that the diffusion model was only trained on clean speech, a test speech utterance of low quality will map to a low-likelihood sample in the terminating Gaussian distribution. In a recent work published during the preparation of this manuscript, Emura [12] showed that the variance of multiple clean speech estimates produced by a diffusion model can be used successfully to estimate the output SI-SDR. However, this work only tests the method on clean speech estimates produced by diffusion-based approaches with the same backbone architecture as in the diffusion models that produce the scores. In comparison to SpeechLMscore [8], our method has no dependence on a choice of speech tokenizer. Rather than using a compressed VAE latent as in VQScore [7], we use diffusion models in a less compressed domain, which are more expressive generative models than VAEs.

We demonstrate that the proposed speech quality estimation correlates well with traditional intrusive metrics such as POLQA, SI-SDR and SNR. In particular, our method has a higher correlation to these metrics compared to SpeechLMscore on the traditional VoiceBank-DEMAND noisy speech benchmark [13]. Furthermore, we show that, in contrast to SpeechLMscore and VQScore, our method rates utterances processed by speech enhancement baselines in a similar fash-

ion as intrusive and supervised DNN-based non-intrusive metrics. Code is available online ¹.

2. Score-based likelihood estimation

Score-based generative models [11] are continuous-time diffusion models relying on stochastic differential equations. Such models can learn complex, high-dimensional data distributions such as human speech [14], natural images [10, 11] or music [15]. Score-based models can be considered as iterative Gaussian denoisers. At training time, a so-called *forward diffusion process* maps the target data distribution $p(\mathbf{x})$ to a tractable Gaussian distribution by gradually adding Gaussian-distributed noise to the data $\mathbf{x} \in \mathbb{R}^n$. New data is then generated following the *reverse diffusion process*, which iteratively denoises an initial Gaussian sample until a sample belonging to the target distribution emerges.

Song et al. [11, App. D] show that every stochastic diffusion process has a corresponding deterministic process described by an ODE whose trajectories share the same marginals $p_t(\mathbf{x})$ as the original process. This specific ODE is named the *probability flow ODE*, and it continuously increases (forward in time) or decreases (backward) the level of noise in the data. Karras et al. [16] formulate the probability flow ODE as

$$d\mathbf{x} = -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})dt, \quad (1)$$

where $\sigma(t)$ is a noise schedule defining the level of noise at time t and $\dot{\sigma}(t)$ is its derivative with respect to t . $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the *score function*, a vector field pointing in the direction of higher density of data:

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = \frac{D_{\theta}(\mathbf{x}; t) - \mathbf{x}}{\sigma(t)^2} \quad (2)$$

Here, $D_{\theta}(\mathbf{x}; t)$ is a denoiser function implemented as a neural network $F_{\theta}(\mathbf{x}; t)$. In order to stabilize and facilitate the training of the model in spite of varying levels of noise, a series of σ -dependent scaling operations c_{in} , c_{noise} , c_{out} and c_{skip} are used, preconditioning inputs and outputs of the network to have unit variance, as well as a skip connection to avoid amplifying errors in F_{θ} . When $\sigma(t) = t$ as suggested in Karras et al. [16], Equation (1) can be written as

$$d\mathbf{x} = \underbrace{\frac{\mathbf{x} - D_{\theta}(\mathbf{x}; t)}{t}}_{\mathbf{f}_{\theta}(\mathbf{x}; t)} dt, \quad (3)$$

with

$$D_{\theta}(\mathbf{x}; t) = c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t)F_{\theta}(c_{\text{in}}(t)\mathbf{x}; c_{\text{noise}}(t)). \quad (4)$$

In Equation (3) we defined the *drift* $\mathbf{f}_{\theta}(\mathbf{x}; t)$ that is central to the calculations in Equations (5), (6) and (8). Song et al. [11] leverage the probability-flow ODE associated with their stochastic differential equation (SDE) to compute the log-likelihood of the input data. This is similar to the density estimation procedure from neural ODEs, leveraging the *instantaneous change of variables* formula [17]:

$$\frac{\partial \log p_t(\mathbf{x})}{\partial t} = -\text{Tr} \left(\frac{\partial \mathbf{f}_{\theta}(\mathbf{x}; t)}{\partial \mathbf{x}} \right) \quad (5)$$

¹<https://github.com/sp-uhh/diffusion-sqa>

Following Grathwohl et al. [18], by integrating Equation (5) for the ODE in Equation (3), we get the following expression for the log density of the data:

$$\log p_0(\mathbf{x}) = \log p_T(\mathbf{x}_T) + \int_0^T \text{Tr} \left(\frac{\partial \mathbf{f}_{\theta}(\mathbf{x}_t; t)}{\partial \mathbf{x}_t} \right) dt \quad (6)$$

with initial value $\mathbf{x}_0 := \mathbf{x}$ and $\mathbf{x}_T = \int_0^T \mathbf{f}_{\theta}(\mathbf{x}_t; t)dt$. Additionally, the authors show that the trace of the Jacobian matrix $\frac{\partial \mathbf{f}_{\theta}}{\partial \mathbf{x}_t}$ can be efficiently computed using the Hutchinson estimator [19]:

$$\text{Tr}(\mathbf{A}) = \mathbb{E}_{p(\boldsymbol{\epsilon})}[\boldsymbol{\epsilon}^{\top} \mathbf{A} \boldsymbol{\epsilon}], \quad (7)$$

where the distribution $p(\boldsymbol{\epsilon})$ must satisfy $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ and $\text{Cov}_{p(\boldsymbol{\epsilon})}[\boldsymbol{\epsilon}] = \mathbf{I}$. This avoids the computation of a separate derivative for each element of the diagonal of the Jacobian.

The following coupled initial value problem is then solved:

$$\begin{bmatrix} \mathbf{x}_T \\ \log p_0(\mathbf{x}) - \log p_T(\mathbf{x}_T) \end{bmatrix} = \int_0^T \begin{bmatrix} \mathbf{f}_{\theta}(\mathbf{x}_t; t) \\ \boldsymbol{\epsilon}^{\top} \frac{\partial \mathbf{f}_{\theta}(\mathbf{x}_t; t)}{\partial \mathbf{x}_t} \boldsymbol{\epsilon} \end{bmatrix} dt \quad (8)$$

with initial value $[\mathbf{x}_0 \quad 0]^{\top}$. Both equations are solved simultaneously using the same solver, as explained in Algorithm 1. The vector-Jacobian product $\boldsymbol{\epsilon}^{\top} \frac{\partial \mathbf{f}_{\theta}}{\partial \mathbf{x}_t}$ can be evaluated at roughly the same cost as a computation of $\mathbf{f}_{\theta}(\mathbf{x}_t; t)$ by performing reverse-mode automatic differentiation, having already computed the forward pass when solving for \mathbf{x}_t . $\log p_T$ is obtained from the probability density function of a multivariate Gaussian.

Algorithm 1 Log-likelihood

```

1:  $\mathbf{x} \leftarrow$  Data
2:  $\boldsymbol{\epsilon} \sim$  Rademacher
3:  $\Delta \log p \leftarrow 0$ 
4:  $t \leftarrow 0$ 
5: while  $t < T$  do
6:    $d\mathbf{x} \leftarrow \mathbf{f}_{\theta}(\mathbf{x}; t)$ 
7:    $d\Delta \log p \leftarrow$  Hutchinson( $d\mathbf{x}, \mathbf{x}, \boldsymbol{\epsilon}$ )
8:    $\mathbf{x} \leftarrow \mathbf{x} + d\mathbf{x} \cdot dt$ 
9:    $\Delta \log p \leftarrow \Delta \log p + d\Delta \log p \cdot dt$ 
10:   $t \leftarrow t + \Delta t$ 
11: end while
12: return  $\log p \leftarrow \log p_T + \Delta \log p$ 

```

3. Implementation details

The framework used in this work to train the diffusion model is the denoising score matching formulation proposed by Karras et al. [16]. Our neural network follows the ADM architecture [21], with the architectural and training improvements subsequently proposed in [22], in particular the *magnitude preserving layers*, which keep the magnitudes of activations in the model controlled during training. We reduce the model size by using only 3 resolutions and employing one residual block per resolution, resulting in a model with 49 M parameters. We set the exponential moving average length to 0.08 using *post-hoc* reconstruction method from [22]. As in [16], we employ a second order solver with 32 steps.

For estimation of the trace, we sample $\boldsymbol{\epsilon}$ from the Rademacher distribution, which is commonly used with the Hutchinson trace estimator [19]; it is a discrete distribution taking two possible values $\{1, -1\}$, each occurring with equal

Table 1: Pearson (PCC) and Spearman’s rank (SRCC) correlation coefficients between supervised/unsupervised metrics and human-assigned scores on the EARS-WHAM (matched case) and VoiceBank-DEMAND (mismatched) noisy test sets. (–)SpeechLMscore indicates that the correlations have been flipped to agree to the other metrics where higher is better.

Measure		EARS-WHAM				VoiceBank-DEMAND			
		POLQA		SI-SDR		POLQA		SI-SDR	
		PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC
Supervised	NISQA [6]	0.797	0.816	0.689	0.712	0.897	0.896	0.608	0.591
	DNSMOS OVRL [5]	0.667	0.711	0.625	0.650	0.776	0.828	0.542	0.569
Unsupervised	VQScore [9]	0.723	0.755	0.804	0.821	0.837	0.841	0.539	0.537
	(–)SpeechLMscore [8]	0.761	0.779	0.733	0.760	0.702	0.681	0.471	0.428
	Diff. Log-likelihood (ours)	0.640	0.667	0.617	0.633	0.831	0.835	0.489	0.498

Table 2: Pearson (PCC) and Spearman’s rank (SRCC) correlation coefficients between supervised/unsupervised metrics and human-assigned scores on the WSJ0-CHiME3 listening experiment results of Richter et al. [20]. (–)SpeechLMscore indicates that the correlations have been flipped to agree to the other metrics where higher is better.

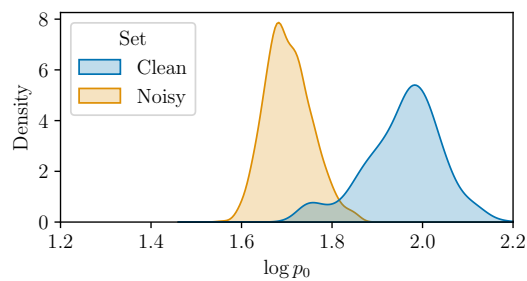
Measure		Listening Scores	
		PCC	SRCC
Supervised	NISQA	0.822	0.840
	DNSMOS OVRL	0.785	0.837
Unsupervised	(–)SpeechLMscore	0.762	0.818
	VQScore	0.798	0.724
	Diff. Log-likelihood (ours)	0.902	0.899

probability, therefore it satisfies the conditions specified in Section 2. The authors of [18] found the variance of the log-likelihood induced by the trace estimator to be lower than 10^{-4} on a validation set, and we confirm this finding on our model by running it with different seeds. We therefore compute the trace using only one noise vector. In order to obtain a consistent and reproducible result, we keep a fixed seed for ϵ . Similarly to Song et al. [11] and other works showing log-likelihood scores, we normalize the likelihoods by the number of elements in the sample; in this case, time-frequency bins.

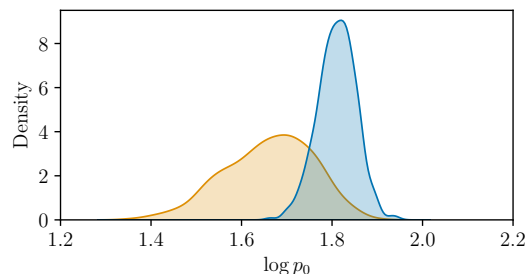
We train the model on the EARS dataset [23], following the train–validation–test split performed for the EARS-WHAM set proposed in the same work [23]. This training set consists of approximately 87 hours of clean, anechoic English speech data with different speaking styles. Here, the audio is downsampled to 16 kHz and transformed into mel spectrograms, with 80 mel bands and Hann windows of length 64 ms with 75% overlap. The dynamic range of the spectrogram is compressed using the logarithm, and the values are then scaled to match mean 0 and standard deviation of 0.5, using the statistics computed on the training set.

During training, we sample segments of 4 seconds in length, with random starting indices. The segments are sampled from the dataset and the model is trained on approximately 37 M sampled segments, with batch size 128. We perform evaluations on the EARS-WHAM test set at 16 kHz. The test set contains 6 speakers and input SNRs randomly sampled in a range of $[-2.5, 17.5]$ dB. We additionally report results on the VoiceBank-DEMAND (VB-DMD) test set [13] also downsampled to 16kHz, as a mismatched condition for the model. The

test set contains two speakers and noise at 2.5, 7.5, 12.5 and 17.5 dB SNR. Lastly, we make use of the samples from the listening experiment of Richter et al. [20], where participants were asked to rate randomly selected samples from the WSJ0-CHiME3 dataset [20], reconstructed by three different speech enhancement models.



(a) EARS-WHAM



(b) VB-DMD

Figure 1: Histogram of diffusion-based log-likelihood values for noisy and clean test data

4. Results and discussion

In Table 1, we show correlations with intrusive metrics on the noisy data from the EARS-WHAM and VB-DMD test sets, compared to the other non-intrusive baselines. We report the Pearson correlation coefficient (PCC) and Spearman’s rank correlation coefficient (SRCC), quantifying linear and monotonic relationships, respectively. While the correlation of the diffusion-based log-likelihood is generally below that of the other metrics on EARS-WHAM, it performs similarly to VQScore and better than SpeechLMscore on VB-DMD.

Figure 1a shows how the distribution of the log-likelihoods

Table 3: Evaluation results on the listening experiment samples from Richter et al. [20], with WSJ0+CHiME3 data. The listening scores are presented in a scale from 0 to 100.

	Intrusive			Non-Intrusive				Listening	
	POLQA \uparrow	PESQ \uparrow	SI-SDR \uparrow	DNSMOS OVRL \uparrow	NISQA \uparrow	SpeechLMscore \downarrow	VQScore \uparrow	Diffusion Log-likelihood \uparrow	Score \uparrow
Clean	—	—	—	3.34 ± 0.12	4.24 ± 0.27	1.35 ± 0.09	0.720 ± 0.007	1.90 ± 0.04	98.9 ± 1.3
SGMSE+	3.32 ± 0.35	2.44 ± 0.33	12.5 ± 1.9	3.26 ± 0.09	3.82 ± 0.29	1.49 ± 0.08	0.715 ± 0.009	1.87 ± 0.03	88.2 ± 3.5
ConvTasNet	3.15 ± 0.35	2.41 ± 0.34	15.4 ± 2.1	3.25 ± 0.15	3.70 ± 0.32	1.53 ± 0.10	0.719 ± 0.012	1.80 ± 0.04	75.9 ± 7.7
MetricGAN+	2.92 ± 0.32	2.61 ± 0.22	6.4 ± 2.2	2.83 ± 0.25	3.34 ± 0.37	1.61 ± 0.11	0.681 ± 0.023	1.65 ± 0.06	49.7 ± 9.9
Noisy	1.97 ± 0.39	1.22 ± 0.10	2.52 ± 1.73	1.89 ± 0.46	1.65 ± 0.59	2.11 ± 0.31	0.610 ± 0.020	1.65 ± 0.04	27.2 ± 4.0

Table 4: Evaluation results on a predictive method (Demucs) and a generative method (SGMSE+). These correspond to the models reported by the EARS-WHAM benchmark [23], with the enhanced outputs downsampled from 48kHz to 16kHz.

	Intrusive			Non-intrusive				
	POLQA \uparrow	PESQ \uparrow	SI-SDR \uparrow	DNSMOS OVRL \uparrow	NISQA \uparrow	SpeechLMscore \downarrow	VQScore \uparrow	Diffusion Log-likelihood \uparrow
Clean	—	—	—	3.12 ± 0.36	4.11 ± 0.73	1.38 ± 0.18	0.719 ± 0.020	1.96 ± 0.09
SGMSE+ [20]	3.45 ± 0.67	2.53 ± 0.63	16.81 ± 4.50	3.13 ± 0.36	4.22 ± 0.74	1.40 ± 0.19	0.723 ± 0.020	1.92 ± 0.09
Demucs [24]	3.17 ± 0.67	2.40 ± 0.59	16.95 ± 4.37	3.08 ± 0.36	3.72 ± 0.75	1.42 ± 0.20	0.728 ± 0.019	1.84 ± 0.08
Noisy	1.82 ± 0.53	1.25 ± 0.22	6.02 ± 6.12	2.08 ± 0.66	2.02 ± 0.70	2.06 ± 0.26	0.619 ± 0.027	1.70 ± 0.05

for clean EARS-WHAM test data compare to that of the noisy. We note that the likelihoods are densities, meaning that they are not bounded to a maximum value of 1. Consequently, the logarithm is not necessarily negative. The values for clean data have a large standard deviation, which hints that the clean speech in the test set has variations which are not completely modeled by the neural network. Nevertheless, the two distributions are clearly separated, as one would wish for in a quality metric. For the VB-DMD test data, the mean scores of clean speech are lower, which is to be expected, since the data is mismatched from training. Additionally, the distribution is narrower, a finding that can be explained by the less varied expressivity if compared to EARS. Concerning the noisy set, although the data is again mismatched, the larger proportion of higher SNRs mixtures seems to be working in the opposite direction, resulting in a wider distribution.

Table 2 shows the correlations of listening scores from the experiment of Richter et al. [20] with DNN-based metrics. Supervised methods were trained with MOS labels as targets and non-paired methods were trained in an unsupervised way, with only clean speech data. The numbers show the highest correlation between our proposed approach and the scores given by audio experts, confirming the effectiveness of the method.

One important aspect to analyze is how a measure handles corruptions from an enhancement model, which can have different behaviors depending on the training paradigm [25–27]. To paint a complete picture of the models’ performances, we evaluate them with a set of non-intrusive and intrusive metrics [28], presenting the results on a system level. In the non-intrusive group of metrics, we report VQScore and SpeechLMscore (in perplexity, so lower is better), as well as metrics trained in a supervised fashion, with MOS labels. In the intrusive subset, we employ PESQ [2] and POLQA [3], as well as SI-SDR [1]. Table 3 displays the scores of the listening experiment data, indicating that our method’s scores agree with the listeners’ by giving a preference to SGMSE+. Almost all non-intrusive metrics point SGMSE+ as the leading model as well, with the exception of VQScore. Additionally, out of the intrusive metrics, only POLQA also reflects this preference.

For an evaluation with more samples, we compare our method with other known objective metrics on a speech enhancement benchmark. In Table 4 we show such a comparative evaluation on the EARS-WHAM set, comparing the generative method SGMSE+ [20, 23] against the predictive method Demucs [24], both trained on EARS-WHAM data. To make it compatible with the metrics, we downsample the enhanced files to 16kHz. Here we can see that, along with DNSMOS and NISQA, log-likelihood favors SGMSE+ over Demucs, whereas VQScore [9] and SpeechLMscore [8] show a only a minor difference between these two evaluated enhancement methods, and even produce values slightly in favor of Demucs. The log-likelihood is therefore the only non-intrusive method trained without access to paired MOS data that correctly reflects the clear human listener preference for the method SGMSE+ over Demucs reported in [23]. Furthermore, the log-likelihood also shows better alignment with POLQA and PESQ scores.

Even though quality assessment measures generally do not face limited computational constraints, diffusion-based methods such as ours tend to be more computationally demanding than predictive methods. Due to the multiple calls of the score network required to iteratively solve Equation (8), inference is slower than the baselines. At 32 steps, evaluating 100 utterances takes ~ 6 minutes, while VQScore takes ~ 3 seconds. Nevertheless, there is active research in diffusion models that aims to reduce the number of steps without sacrificing performance.

5. Conclusion

We proposed a speech quality estimator based on unconditional score-based diffusion models trained on clean speech only. Using the natural likelihood computation abilities of score-based models, the proposed estimator can estimate the quality of speech utterances without having any access to paired data. The resulting measure is non-intrusive and yet correlates well with intrusive metrics on noisy speech benchmarks. When evaluating utterances processed by speech enhancement baselines, our method showed the highest correlation with the scores assigned by human listeners.

6. Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 498394658. We acknowledge the support by the DFG in the transregio project Crossmodal Learning (TRR 169) and DASHH (Data Science in Hamburg – Helmholtz Graduate School for the Structure of Matter) with Grant-No. HIDSS-0002. We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

7. References

- [1] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR - half-baked or well done?” in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2019, pp. 626–630.
- [2] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2001.
- [3] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, “Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I - temporal alignment,” *Journal of the Audio Engineering Society (AES)*, vol. 61, no. 6, pp. 366–384, 2013.
- [4] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O’Gorman, and A. Hines, “ViSQOL v3: An open source production ready objective speech and audio metric,” in *Proc. QoMEX*, 2020.
- [5] C. K. A. Reddy, V. Gopal, and R. Cutler, “DNSMOS P.835: A non-intrusive perceptual speech quality metric to evaluate noise suppressors,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022, pp. 886–890.
- [6] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Interspeech*, 2021, pp. 2127–2131.
- [7] P. Manocha and A. Kumar, “Speech quality assessment through MOS using non-matching references,” in *Interspeech*, 2022.
- [8] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, “Speechlmscore: Evaluating speech generation using speech language model,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.
- [9] S.-W. Fu, K.-H. Hung, Y. Tsao, and Y.-C. F. Wang, “Self-supervised speech quality estimation and enhancement using only clean speech,” in *Int. Conf. on Learning Representations (ICLR)*, 2024.
- [10] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.
- [11] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *Int. Conf. on Learning Representations (ICLR)*, 2021.
- [12] S. Emura, “Estimation of output SI-SDR solely from enhanced speech signals in diffusion-based generative speech enhancement method,” in *EURASIP EUSIPCO*, 2024.
- [13] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust text-to-speech,” in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016.
- [14] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “WaveGrad: Estimating gradients for waveform generation,” *Int. Conf. on Learning Representations (ICLR)*, 2021.
- [15] E. Moliner, J. Lehtinen, and V. Välimäki, “Solving audio inverse problems with a diffusion model,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2023.
- [16] T. Karras, M. Aittala, T. Aila, and S. Laine, “Elucidating the design space of diffusion-based generative models,” in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 35. Curran Associates, Inc., 2022, pp. 26 565–26 577.
- [17] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, “Neural ordinary differential equations,” in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2018, p. 6572–6583.
- [18] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, and D. Duvenaud, “Scalable reversible generative models with free-form continuous dynamics,” in *Int. Conf. on Learning Representations (ICLR)*, 2019.
- [19] M. Hutchinson, “A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines,” *Communications in Statistics - Simulation and Computation*, vol. 19, no. 2, pp. 433–450, 1990.
- [20] J. Richter, S. Welker, J.-M. Lemerrier, B. Lay, and T. Gerkmann, “Speech enhancement and dereverberation with diffusion-based generative models,” *IEEE Trans. on Audio, Speech, and Lang. Process. (TASLP)*, vol. 31, pp. 2351–2364, 2023.
- [21] P. Dhariwal and A. Q. Nichol, “Diffusion models beat GANs on image synthesis,” in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.
- [22] T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine, “Analyzing and improving the training dynamics of diffusion models,” in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 174–24 184.
- [23] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, “EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation,” in *Interspeech*, 2024, pp. 4873–4877.
- [24] S. Rouard, F. Massa, and A. Défossez, “Hybrid transformers for music source separation,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.
- [25] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, “Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.
- [26] D. de Oliveira, J. Richter, J.-M. Lemerrier, 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*, 2023, pp. 260–264.
- [27] 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.
- [28] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, “The PESQetarian: On the relevance of Goodhart’s law for speech enhancement,” in *Interspeech*, 2024, pp. 3854–3858.