ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095258

ANALYSING DIFFUSION-BASED GENERATIVE APPROACHES VERSUS DISCRIMINATIVE APPROACHES FOR SPEECH RESTORATION

Jean-Marie Lemercier[⋆]*, Julius Richter*[⋆]*, Simon Welker*[⋆×]*, Timo Gerkmann*[⋆]

[⋆]Signal Processing (SP), Universität Hamburg, Germany
[×] Center for Free-Electron Laser Science, DESY, Germany
{firstname.lastname}@uni-hamburg.de

ABSTRACT

Diffusion-based generative models have had a high impact on the computer vision and speech processing communities these past years. Besides data generation tasks, they have also been employed for data restoration tasks like speech enhancement and dereverberation. While discriminative models have traditionally been argued to be more powerful e.g. for speech enhancement, generative diffusion approaches have recently been shown to narrow this performance gap considerably. In this paper, we systematically compare the performance of generative diffusion models and discriminative approaches on different speech restoration tasks. For this, we extend our prior contributions on diffusion-based speech enhancement in the complex time-frequency domain to the task of bandwith extension. We then compare it to a discriminatively trained neural network with the same network architecture on three restoration tasks, namely speech denoising, dereverberation and bandwidth extension. We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Code and audio examples can be found online[1].

Index Terms— generative modelling, diffusion models, speech enhancement, dereverberation, bandwidth extension

1. INTRODUCTION

Speech corruptions arise in real-life scenarios and modern communication devices, when clean speech sources are impacted by background noise, interfering speakers, room acoustics and channel degradation. Speech restoration therefore aims at recovering clean speech from the corrupted signal. Traditional speech restoration methods leverage the different statistical properties of the target and interference signals [1]. Data-driven approaches based on machine learning predominately employ discriminative models that learn a single best deterministic mapping between corrupted speech and the corresponding clean speech target [2].

In contrast, generative models implicitly or explicitly learn the target distribution and allow to generate multiple valid estimates instead of a single best estimate as in discriminative approaches [3]. For example, diffusion-based generative models, or simply *diffusion models*, have shown great success in learning the data distribution of natural images [4, 5, 6]. This class of models uses a *forward process* to slowly turn data into a tractable prior, such as a standard normal distribution, and train a neural network to solve the *reverse process* to generate clean data from this prior. These diffusion models can also be used for conditional generation in restoration tasks, which has recently been proposed for speech enhancement and dereverberation [7, 8, 9, 10]. They can in that regard be functionally seen as a mean of generating clean speech based on noisy speech, and can be thus compared to discriminative approaches. However, to make a fair comparison of these two conceptually different approaches, similar network architectures and same training data should be used.

In this work, we present an analysis of a generative diffusion model as compared to its discriminative counterpart sharing the same deep neural network (DNN) architecture, for various speech restoration tasks. We use our previous method which defines the diffusion process in the complex spectrogram domain [7, 8]. We show that the performance gap between the generative and discriminative models varies with respect to the corruption at hand. We evaluate our proposed approaches on the WSJ0 corpus, using various simulated corruptions and recorded background noise. Finally we compare our bandwidth extension model with state-of-the-art bandwidth extension methods on the VCTK corpus.

The remainder of the paper is organized as follows. We first present the three speech restoration problems benchmarked, along with popular solutions for solving them. Then, we introduce diffusion-based generative models using the stochastic differential equation (SDE) formalism. We continue by explaining our experimental setup including data generation and training methods. Finally, we present and discuss our results.

2. SPEECH RESTORATION TASKS AND RELATED WORK

2.1. Speech enhancement

Speech enhancement consists in removing an additive interference n (e.g. background noise or interfering speakers) from the corrupted mixture y to extract the clean speech target s:

$$y = s + n \tag{1}$$

Popular enhancement methods include Wiener-inspired spectral filtering [1], discriminative machine learning methods [2] or generative approaches like denoising variational auto-encoders (VAEs) [11]. Recently, diffusion models were proposed to tackle speech enhancement either in the time domain [12] or in the complex time-frequency (T-F) domain [7, 10, 8].

2.2. Speech dereverberation

Reverberation is caused by room acoustics, and is characterized by multiple reflections on the room enclosures. Late reflections particularly degrade the speech signal and may result in reduced intelligibility [13].

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380, DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).

[1] https://uhh.de/inf-sp-sgmsemultitask

The corruption model is then convolutive, as the clean speech s is convolved with a room impulse response (RIR) h representing the acoustic path between the source and the listener:

$$y = s * h \tag{2}$$

Single-channel dereverberation methods range from spectral enhancement [14], inverse filtering [15], and cepstral processing [16] to machine learning algorithms using DNNs in the complex T-F domain [17] or in the time-domain [18].

2.3. Bandwidth extension

Audio super-resolution, or bandwidth extension, aims at converting a low-sampling rate signal back to a version sampled at a higher rate, regenerating time resolution, high-frequency content and audio quality. The corruption process is linear and involves an anti-aliasing low-pass filter followed by a decimation operation:

$$y = \mathrm{Resample}(s * a,\ f_s^{\mathrm{up}},\ f_s^{\mathrm{low}}) \tag{3}$$

where a is the anti-aliasing filter impulse response, f_s^{up} the original high sampling rate and f_s^{low} the low sampling rate.

Several discriminative methods were proposed to tackle bandwidth extension for speech signals [19, 20]. Generative approaches based on neural vocoders using generative adversarial networks (GANs) were also proposed [21, 22, 23]. A continuous-time diffusion model in the time-domain was proposed in [24].

3. SCORE-BASED DIFFUSION MODELS FOR SPEECH RESTORATION

Score-based diffusion models are defined by three components: a forward diffusion process, a score estimator and a sampling method for inference.

3.1. Forward and reverse processes

The stochastic forward process $\{\mathbf{x}_t\}_{t=0}^{T}$ is modeled as the solution to a SDE, in the Itô sense [25, 26]:

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w} \tag{4}$$

where $\mathbf{x}_t$ is the current state of the process indexed by a continuous time variable $t \in [0,T]$ with the initial condition $\mathbf{x}_0$ representing clean speech. As our process is defined in the T-F domain, the variables in bold are assumed to be one-dimensional vectors in $\mathbb{C}^d$ containing the coefficients of a flattened complex spectrogram, whereas variables in regular font represent real scalar values. The stochastic process $\mathbf{w}$ is a standard d-dimensional Brownian motion, which implies that $\mathrm{d}\mathbf{w}$ is a zero-mean Gaussian random variable with standard deviation $\sqrt{\mathrm{d}t}$ for each T-F bin.

The *drift* function $\mathbf{f}$ and *diffusion* coefficient g as well as the initial condition $\mathbf{x_0}$ and the final diffusion time T define uniquely the Itô process $\{\mathbf{x}_t\}_{t=0}^{T}$. Under some regularity conditions on $\mathbf{f}$,g allowing a unique and smooth solution to the Kolmogorov equations associated to (4), the reverse process $\{\mathbf{x}_t\}_{t=T}^{0}$ is another diffusion process defined as the solution of a SDE, with the following form [27, 26]:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\,\mathrm{d}\bar{\mathbf{w}}, \tag{5}$$

where $\mathrm{d}\bar{\mathbf{w}}$ is a d-dimensional Brownian motion for the time flowing in reverse and $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the *score function*, i.e. the gradient of the logarithm data distribution for the current process state $\mathbf{x}_t$.

Speech restoration tasks can be considered as conditional generation tasks, i.e. generation of clean speech $\mathbf{x_0}$ conditioned by the corrupted speech $\mathbf{y}$. In [7, 8] we proposed to incorporate the conditioning directly into the diffusion process by defining the forward process as the solution to the following Ornstein-Uhlenbeck SDE [25]:

$$\mathrm{d}\mathbf{x}_t = \underbrace{\gamma\,(\mathbf{y} - \mathbf{x}_t)}_{:=\,\mathbf{f}(\mathbf{x}_t,\mathbf{y})}\,\mathrm{d}t + \underbrace{\left[\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{t}\sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}\right]}_{:=\,g(t)}\mathrm{d}\mathbf{w}, \tag{6}$$

with γ a *stiffness* hyperparameter, and $\sigma_{\min}$ and $\sigma_{\max}$ two hyperparameters controlling the *noise scheduling*, that is, the amount of Gaussian white noise injected at each timestep of the process.

The interpretation of our forward process in Eq. (6), visualized on Fig. 1, is as follows: at each time step and for each T-F bin independently, an infinitesimal amount of corruption is added to the current process state $\mathbf{x}_t$, along with Gaussian noise with standard deviation $g(t)\sqrt{\mathrm{d}t}$. Given an initial state $\mathbf{x}_0$ and $\mathbf{y}$, the Itô forward process corresponding to the solution of (6) admits a Gaussian distribution for the process state $\mathbf{x}_t$ called *perturbation kernel*:

$$p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}\big(\mathbf{x}_t; \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t), \sigma(t)^2\mathbf{I}\big), \tag{7}$$

where $\mathcal{N}_{\mathbb{C}}$ denotes the circularly-symmetric complex normal distribution and $\mathbf{I}$ the identity matrix. Given the simple Gaussian kernel, closed-form solutions for the mean $\boldsymbol{\mu}$ and variance $\sigma(t)^2$ can be determined [25]:

$$\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\mathbf{y}, \tag{8}$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2\big((\sigma_{\max}/\sigma_{\min})^{2t} - \mathrm{e}^{-2\gamma t}\big)\log(\sigma_{\max}/\sigma_{\min})}{\gamma + \log(\sigma_{\max}/\sigma_{\min})}. \tag{9}$$

3.2. Score function estimator

When performing inference by sampling through the reverse SDE in Eq. (5), the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is not readily available. Thus, it is approximated by a DNN $\mathbf{s}_\theta$, called the *score model*. In particular, given the Gaussian form of the perturbation kernel $p_{0t}(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y})$ and the regularity conditions exhibited by the mean and variance, a *denoising score matching* objective can be used to train the score model $\mathbf{s}_\theta$ [28].

The score function of the perturbation kernel is:

$$\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y}) = -\frac{\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)}{\sigma(t)^2}. \tag{10}$$

Therefore we can reparameterize the denoising score matching objective as follows [26]:

$$\begin{aligned}\mathcal{J}(\theta) &= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{y},\{\mathbf{x}_t|\mathbf{x}_0,\mathbf{y}\}}\left[\|\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) - \nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y})\|_2^2\right] \\ &= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{y},\mathbf{z}}\left[\left\|\mathbf{s}_\theta\big(\big[\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) + \sigma(t)\mathbf{z}\big], \mathbf{y}, t\big) + \frac{\mathbf{z}}{\sigma(t)}\right\|_2^2\right],\end{aligned} \tag{11}$$

using $\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t) + \sigma(t)\mathbf{z}$, with $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. t is sampled uniformly in $[t_\epsilon, T]$ where t_ϵ is a minimal diffusion time used to avoid numerical instabilities.

3.3. Inference through reverse sampling

At inference time, we first sample an initial condition of the reverse process, corresponding to $\mathbf{x}_T$, with:

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma^2(T)\mathbf{I}), \tag{12}$$

This sample corresponds to corrupted speech $\mathbf{y}$ to which we add Gaussian noise with variance $\sigma(t)^2$, which approximates the training condition.

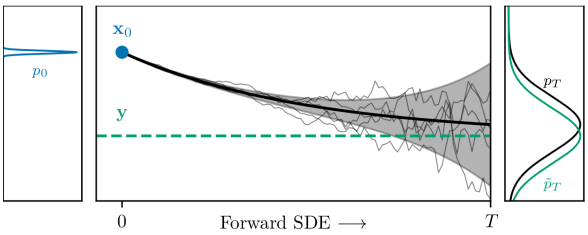

Fig. 1. *Visualization of the forward process* (6)*. Mean curve is in solid black and variance is represented by the greyed area. Several realizations of the diffusion process are represented by thin black lines.*

Conditional generation is then performed by solving the following *plug-in reverse SDE* from $t = T$ to $t = 0$, where the score function is replaced by its estimator $\mathbf{s}_\theta$, assuming the latter was trained e.g. according to Section 3.2:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2\, \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t + g(t)\, \mathrm{d}\bar{\mathbf{w}}, \qquad (13)$$

where $\mathbf{f}$ and g are the drift and diffusion terms defined in (6).

We use classical numerical solvers based on discretization of (13) according to a N points grid of the interval $[0,T]$. Since each reverse diffusion step calls the score network, the inference time of diffusion models is higher than their discriminative counterparts, by two orders of magnitude in our case. Fast inference schemes are discussed in the literature and are outside of the scope of this paper.

4. EXPERIMENTAL SETUP

4.1. Data

We use the WSJ0 corpus [29] for most experiments to ensure easier comparison between tasks. For comparison to bandwidth extension baselines, we use the VCTK corpus [30]. All data generation methods are accessible via our web page[2].

Speech Enhancement: The WSJ0+Chime dataset is generated using clean speech extracts from the Wall Street Journal corpus and noise signals from the CHiME3 dataset [31]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance with a signal-to-noise ratio (SNR) sampled uniformly between 0 and 20dB.

Speech Dereverberation: The WSJ0+Reverb dataset is generated using clean speech data from the WSJ0 dataset and convolving each utterance with a simulated RIR. We use the PyRoomAcoustics engine [32] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a target T_{60} between 0.4 and 1.0 seconds and room length, width and height in [5,15]×[5,15]×[2,6] m. A dry version of the room is created with the same geometric parameters with a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target.

Bandwidth Extension: The WSJ0+BWR dataset is built with clean speech extracted from the WSJ0 corpus and a similar bandwidth reduction recipe as in [21, 23]. We pick an anti-aliasing filter type among Chebyshev, Butterworth, Elliptic and Bessel and a filter order among$\{2,4,8\}$. Decimating is then realized with a down-scaling factor sampled in$\{2,4,8\}$. The utterance is then resampled at the original 16 kHz with polyphase filtering. To compare against other baselines, we generate VCTK+BWR by replacing WSJ0 with VCTK as the base speech corpus, which we first resample to 16kHz, and use the same process as explained above.

4.2. Hyperparameters and training configuration

Data representation: Utterances are transformed using a short-time Fourier transform (STFT) with a window size of 510, a hop length of 128 and a Hann window. Square-root magnitude compression is carried on the spectrogram. For training, sequences of 256 STFT frames (i.e. 2s) are randomly extracted from the full-length utterances and normalized with respect to the corrupted mixture before being fed to the network.

Forward diffusion: Defined in (6), the stiffness parameter is fixed to γ=1.5, the extremal noise levels to $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 0.5$. The minimal diffusion time defined in (11) is set to $t_\epsilon = 0.03$ as in [8].

Network architecture: The original architecture used for score estimation in [8] is the NCSN++ network proposed in [26]. NCSN++ is a multiresolution U-Net structure which includes in each layer a series of ResNet blocks using 2D convolutions, group normalization and fixed down/upsampling. Attention mechanism is used in the bottleneck, and the network leverages a parallel progressive growing path in addition to the skip connections. The noisy speech spectrogram $\mathbf{y}$ and the current diffusion process estimate $\mathbf{x}_t$ real and imaginary channels are stacked and fed to the network as input. The model is made noise-conditional by feeding each ResNet block with an encoded version of the current noise level $\sigma(t)$. More details about the architecture can be found in [8, 26]. For the generative model proposed in this paper, denoted as *SGMSE+M*, we use a lighter configuration of the NCSN++ architecture called NCSN++M . Ablation studies were designed to halve the number of parameters with almost no degradation, resulting in a network capacity of roughly 27.8M parameters. For the discriminative approach, denoted simply as *NCSN++M* in the following, the noise-conditioning layers are removed This ablation removes only 1.8% of the original number of parameters, which hardly modifies the network capacity.

Training configuration: We train the DNN for a maximum of 300 epochs using early stopping with a patience of 10 epochs. The generative approach SGMSE+M is trained with the denoising score matching criterion (11), and discriminative NCSN++M uses a simple mean-square error loss on the complex spectrogram. We use the Adam optimizer with a learning rate of 10^{-4} and an effective batch size of 16. We track an exponential moving average of the DNN weights with a decay of 0.999.

Inference: 50 time steps are used for reverse inference, adopting the predictor-corrector scheme [26] with one step of annealed Langevin dynamics correction.

4.3. Evaluation metrics

For instrumental evaluation of the speech enhancement and dereverberation performance, we use Perceptual Evaluation of Speech Quality (PESQ) [33] , extended short-term objective intelligibility (ESTOI) [34] and scale-invariant signal to distortion ratio (SI-SDR) [35] . For bandwidth extension we also include log spectral distance (LSD) as a common metric used in the literature. However, it must be stated that the aforementioned instrumental metrics may relate poorly with listening experiments, especially for bandwidth extension. We therefore complement our metrics benchmark with WV-MOS [23][3], which is a DNN-based mean opinion score (MOS) estimation, and was used by the authors for assessment of bandwidth extension performance. For comparability purposes to baselines on the VCTK corpus, we use regular STOI [36] instead of its extended version .

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1. Speech enhancement

In Table 1, we report speech enhancement performance on the WSJ0+Chime dataset. We notice that the generative SGMSE+M produces higher quality samples as measured by WV-MOS and PESQ.

[2]https://uhh.de/inf-sp-sgmsemultitask

[3]https://github.com/AndreevP/wvmos

Table 1. Results for denoising on WSJ0+Chime data.

Method	Type	WV-MOS	PESQ	ESTOI	SI-SDR
Mixture		1.44 ± 1.62	1.70 ± 0.49	0.78 ± 0.14	10.0 ± 5.7
NCSN++M	D	3.65 ± 0.48	2.67 ± 0.69	$\mathbf{0.93 \pm 0.06}$	$\mathbf{19.5 \pm 4.4}$
SGMSE+M	G	$\mathbf{3.77 \pm 0.32}$	$\mathbf{2.94 \pm 0.60}$	0.92 ± 0.06	18.0 ± 5.1

Table 2. Results for dereverberation on WSJ0+Reverb data.

Method	Type	WV-MOS	PESQ	ESTOI	SI-SDR
Mixture		1.78 ± 0.99	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5
NCSN++M	D	2.96 ± 0.38	2.19 ± 0.48	$\mathbf{0.87 \pm 0.05}$	$\mathbf{7.2 \pm 3.7}$
SGMSE+M	G	$\mathbf{3.43 \pm 0.33}$	$\mathbf{2.64 \pm 0.42}$	$\mathbf{0.87 \pm 0.05}$	6.4 ± 4.2

Table 3. Results for bandwidth extension on WSJ0+BWR data.

Method	Type	WV-MOS ↑	ESTOI ↑	LSD ↓
Mixture		2.45 ± 1.01	0.72 ± 0.21	2.31 ± 0.32
AE-NCSN++M	D	2.17 ± 0.93	0.71 ± 0.19	1.81 ± 0.21
NCSN++M	D	2.25 ± 0.87	0.73 ± 0.16	2.21 ± 0.30
SGMSE+M	G	$\mathbf{3.43 \pm 0.48}$	$\mathbf{0.83 \pm 0.13}$	$\mathbf{1.44 \pm 0.17}$

It is however slightly outperformed by discriminative NCSN++M on intelligibility and noise removal. Indeed, in a denoising task, the interference does not share any information with the target speech, making it relatively easy for a discriminative approach to remove the interference without distorting the target. However, we show in the uploaded listening examples that the discriminative approach tends to destroy low-energy speech regions for low SNRs, whereas the generative model does not. A larger benefit of the generative approach is observed when training and testing data have a stronger mismatch [8].

5.2. Speech dereverberation

In Table 2, we report dereverberation results on the WSJ0+Reverb dataset. Here, generative SGMSE+M clearly outperforms discriminative NCSN++M in terms of quality by a large margin on WV-MOS and PESQ, and performs on par on ESTOI and SI-SDR. For dereverberation, in contrast to denoising, the interference model is completely dependent on the target as it is a filtered version of the latter (Eq. (2)). The generative model is able to extract the speech cues and directly reconstructs it with very little reverberation. The discriminative method, however, cannot do so without introducing significant distortions.

5.3. Bandwidth extension

Results on WSJ0+BWR: In Table 3 we report bandwidth extension performance on the WSJ0+BWR dataset. Interestingly, using a STFT representation did not allow the discriminative approach to recreate the lost high-frequency content. The approach simply learnt an identity mapping, and similar results were observed when experimenting with other STFT-based DNN backbones and data. For this discriminative case, we modified the NCSN++M architecture to use a learnt encoder and decoder, as in e.g. [19] . The resulting approach, denoted as AE-NCSN++M in the following, uses a single 1D convolutional layer with 256 filters of length 510 and stride 128, so that the learnt representation is equivalent to the chosen STFT filterbank. As opposed to NCSN++M, AE-NCSN++M is able to generate high-frequency components, however the reconstruction quality is overall poor, which is to be expected given the generative nature of the bandwidth extension task. By learning an approximate identity mapping, NCSN++M performs better than AE-NCSN++M on instrumental metrics although it does not actually perform bandwidth extension.

Table 4. Results for bandwidth extension on VCTK data. * means that the results were taken from [23]. † means that the method was trained on each bandwidth reduction factor separately.

Bandwidth	Type	1kHz		2kHz		4kHz	
		WV-MOS	STOI	WV-MOS	STOI	WV-MOS	STOI
Mixture		1.36	0.79	2.34	0.89	3.52	0.99
TUNet† [20]	D	-	-	-	-	3.86	0.98
TFiLM*† [19]	D	1.65	0.81	2.27	0.91	3.49	**1.00**
HiFi++*† [23]	G	**3.71**	**0.86**	**3.95**	**0.94**	4.16	**1.00**
VoiceFixer [21]	G	2.50	0.73	3.35	0.78	3.81	0.83
NuWave2 [24]	G	-	-	-	-	3.76	0.97
SGMSE+M	G	3.25	0.83	3.70	0.93	**4.20**	**1.00**

In contrast, generative SGMSE+M performs much better in all metrics, generating plausible content even when the Nyquist frequency is down to 2kHz. When the Nyquist frequency is down to 1kHz, the approach can struggle with generating the right consonants in some cases. Typically the generation process may mistake $[ch]$ for $[s]$ as the information needed to differentiate those sounds is available only at frequencies way above 1kHz. Integrating a linguistic or visual model could be here envisaged to make the approach robust to this lack of acoustic cues.

Results on VCTK+BWR: In Table 4, we compare the proposed generative SGMSE+M on the VCTK+BWR test set against HiFI++ [23], VoiceFixer[4] [21], TFiLM [19], TUNet[5] [20] and NuWave2[6] [24]. Please note that HiFi++, TFiLM and TUNet are trained on each input bandwidth separately, while our generative model SGMSE+M as well as VoiceFixer and NuWave2 are bandwidth-agnostic. We use the official implementations for all approaches without retraining, except HiFi++ as no code is available, and TFiLM as no multi-speaker model is provided. When a method is not trained to restore speech at 16kHz, we use it at the nominal sampling rate then downsample its output to 16kHz. SGMSE+M achieves on par results with HiFi++ on 4kHz bandwidth and worse on 1kHz and 2kHz bandwidths, which is partially due to the fact that HiFi++ is trained separately for each input bandwidth. Using neural vocoders incorporating speech knowledge, as is the case for HiFi++, also probably helps improve robustness to very low input bandwidths. Against all other approaches than HiFi++, SGMSE+M performs significantly better on almost all metrics and conditions.

6. CONCLUSION

The goal of this work is to analyse the potential benefit of recent diffusion-based generative approaches against discriminative approaches on various speech restoration tasks. For this, we apply our recently proposed diffusion generative model to speech enhancement, dereverberation and bandwidth extension, and compare against a discriminative approach using the same DNN architecture . We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Furthermore, we show that the proposed bandwidth-agnostic method performs slightly worse or on par in comparison with a recent bandwidth-dependent approach, and largely outperforms other discriminative and bandwidth-agnostic generative approaches.

[4]https://github.com/haoheliu/voicefixer

[5]https://github.com/NXTProduct/TUNet

[6]https://github.com/mindslab-ai/nuwave2

7. REFERENCES

[1] T. Gerkmann and E. Vincent, *Spectral Masking and Filtering*. John Wiley & Sons, 2018.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Int. Conf. Machine Learning (ICML)*, Apr. 2015.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Neural Information Proc. Systems (NIPS)*, Dec. 2020.

[6] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Neural Information Proc. Systems (NIPS)*, Dec. 2019.

[7] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Interspeech*, Sept. 2022.

[8] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *arXiv*, 2022.

[9] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2022.

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

[11] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," June 2021.

[12] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Asia-Pacific Signal and Information Processing Association (APSIPA)*, Dec. 2021.

[13] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, vol. 59. Springer, 2011.

[14] E. Habets, *Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*. PhD thesis, 2007.

[15] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2014.

[16] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Sept. 2011.

[17] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[18] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Sept. 2019.

[19] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. Koh, and S. Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations," in *Neural Information Proc. Systems (NIPS)*, Dec. 2019.

[20] V.-A. Nguyen, A. H. T. Nguyen, and A. W. H. Khong, "TUNet: A block-online bandwidth extension model based on transformers and self-supervised pretraining," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2022.

[21] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv*, 2021.

[22] H. Liu, W. Y. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural vocoder is all you need for speech super-resolution," in *Interspeech*, Sept. 2022.

[23] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," *arXiv*, 2022.

[24] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *Interspeech*, Sept. 2022.

[25] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, vol. 82. Journal of the American Statistical Association, 2000.

[26] 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. Learning Repr. (ICLR)*, May 2021.

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

[28] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching.," *Journal of Machine Learning Research*, vol. 6, p. 695–709, 2005.

[29] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," May 2007.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Speech Synthesis Workshop (SSW)*, Sept. 2016.

[31] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015.

[32] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018.

[33] 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. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2001.

[34] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[35] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - Half-baked or well done?," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2019.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.