Received 11 August 2023; revised 16 January 2024; accepted 24 January 2024. Date of publication 19 March 2024; date of current version 18 June 2024. The review of this article was arranged by Associate Editor Romain Serizel.

Digital Object Identifier 10.1109/OJSP.2024.3379070

Causal Diffusion Models for Generalized Speech Enhancement

JULIUS RICHTER [1] (Student Member, IEEE), SIMON WELKER [1,2] (Student Member, IEEE), JEAN-MARIE LEMERCIER [1] (Graduate Student Member, IEEE), BUNLONG LAY [1] (Graduate Student Member, IEEE), TAL PEER [1] (Student Member, IEEE), AND TIMO GERKMANN [1] (Senior Member, IEEE)

[1]Signal Processing (SP), Universität Hamburg, 20148 Hamburg, Germany
[2]Center for Free-Electron Laser Science, DESY, 22527 Hamburg, Germany

CORRESPONDING AUTHOR: JULIUS RICHTER (email: julius.richter@uni-hamburg.de).

This work was supported in part by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169), DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) under Grant HIDSS-0002, and in part by the Federal Ministry for Economic Affairs and Climate Action, project under Grant 01MK20012S, AP380.

ABSTRACT In this work, we present a causal speech enhancement system that is designed to handle different types of corruptions. This paper is an extended version of our contribution to the "ICASSP 2023 Speech Signal Improvement Challenge". The method is based on a generative diffusion model which has been shown to work well in scenarios beyond speech-in-noise, such as missing data and non-additive corruptions. We guarantee causal processing with an algorithmic latency of 20 ms by modifying the network architecture and removing non-causal normalization techniques. To train and test our model, we generate a new corrupted speech dataset which includes additive background noise, reverberation, clipping, packet loss, bandwidth reduction, and codec artifacts. We compare the causal and non-causal versions of our method to investigate the impact of causal processing and we assess the gap between specialized models trained on a particular corruption type and the generalized model trained on all corruptions. Although specialized models and non-causal models have a small advantage, we show that the generalized causal approach does not suffer from a significant performance penalty, while it can be flexibly employed for real-world applications where different types of distortions may occur.

INDEX TERMS Causal processing, diffusion models, generalized speech enhancement.

I. INTRODUCTION

Speech enhancement algorithms typically address the problem of recovering clean speech signals from mixtures that contain speech and additive background noise [1]. Common approaches estimate the speech signal using a filter in the time-frequency domain to reduce noise while avoiding speech distortions [2]. In real-world speech communication, however, there are numerous other factors besides additive background noise that can degrade speech signals. These include reverberation, transmission errors, limited bandwidth, codec artifacts, and non-linearities of recording and playback equipment. To address this broad range of corruptions, several recent attempts have been made to find *generalized* (sometimes called *universal*) speech enhancement models [3], [4], [5]. This can be achieved by training a single model on a dataset comprising multiple corruptions [6], or employing a regeneration strategy, where a two-step approach is used to first enhance and then synthesize speech signals [7].

Classical speech processing tasks include denoising and dereverberation, which can be described as an (approximate) inversion of an additive or convolutive degradation model. For this purpose, deep learning-based methods often employ *predictive models* to either accomplish regression of clean waveform samples [8] and/or spectral components [9], or to estimate a multiplicative mask that is applied to the input spectrogram [10]. In contrast, tasks such as packet loss concealment or bandwidth extension are missing-data problems, where regression and masking-based approaches often reach

their limitations [11]. These missing-data problems can be interpreted as audio inpainting tasks which require the generation of new signal content, i.e., filling the gaps of lost packets or missing frequency bands in a speech signal. Consequently, the use of *generative models* is a natural choice for these tasks. Per definition, generative models aim at learning the prior distribution of the target data. The learned prior distribution is then used to estimate clean speech given a corrupted input that is assumed to lie outside the learned distribution.

Generative and predictive speech enhancement models have been shown to differ in the type of distortions they introduce into the clean speech estimates [12]. While predictive methods are prone to noise leakage and speech distortions, generative models tend to hallucinate when presented with challenging inputs, resulting in speech-like sounds with poor articulation and no semantic meaning [13]. However, since generative models leverage the learned speech prior, the enhanced speech will resemble the characteristics of the clean speech training data which is typically noise-free and of high quality. Moreover, due to their ability to deal with non-additive corruption types and missing data problems while still performing well on denoising and dereverberation tasks [11], generative speech enhancement models are good candidates for generalized speech enhancement.

Recently, diffusion-based generative models [14], or simply *diffusion models*, have been successfully applied to the task of speech enhancement [13], [15], [16]. Diffusion models define data recovery and generation problems as an iterative denoising task, in which a deep neural network (DNN), also called the *score model*, learns to remove the Gaussian noise that was progressively added in a forward diffusion process. To enable the conditional generation of clean speech given a corrupted signal, it was proposed to use a task-adapted diffusion process and condition the score model on the corrupted input [15], [16]. Currently, most diffusion models employ convolutional-based U-Net architectures as score models [13], [15]. However, standard U-Net-like architectures do not have a causal structure, meaning that the network considers information from the future to make predictions about the current time frame. This makes the resulting diffusion model unsuitable for online (and potentially real-time capable) signal processing where causality is a critical requirement.

In this paper, we present a causal speech enhancement method that is based on generative diffusion models and is designed to handle different types of corruptions. This is an extended version of our previous publication [17], which we contributed to the "ICASSP 2023 Speech Signal Improvement Challenge" ("SIG challenge") [18]. Compared to our original work on diffusion-based speech enhancement [13], we modify the network architecture to meet the causality requirement and now operate on super-wideband speech (32 kHz). By adjusting the short-time Fourier transform (STFT), we achieve an *algorithmic latency* of 20 ms, which we define as the latency introduced by frame-based processing, i.e., the latency that remains when using an infinitely fast processing device. Moreover, we address the more generalized speech enhancement problem by training the model on different corruption types. For this purpose, we generate a new dataset that contains additive background noise, reverberation, clipping, packet loss, bandwidth reduction, codec artifacts, as well as different combinations thereof. In the experiments, we evaluate on each individual corruption set and compare the model trained on all corruptions with specialized models only trained on the specific corruption type. Furthermore, we compare the causal approach against the non-causal counterpart. The results suggest that while specialized models and non-causal models have certain benefits, the causal model trained on all corruptions can keep up despite the causality constraint and provides a single model for generalized speech enhancement.

II. METHOD

In this section, we summarize our contributions to diffusion-based speech enhancement [13], [16] and highlight all modifications that were carried out to meet the causality requirement.

A. DATA REPRESENTATION

We represent all audio signals in the complex-valued STFT domain. Thus, comparable to complex spectral mapping [19], the diffusion model aims at estimating the clean real and imaginary spectrograms from the corrupted ones. Formally, we operate on complex spectrograms that are elements of $\mathbb{C}^{K\times F}$, where K denotes the number of time frames dependent on the audio length, and F is the number of frequency bins. To compensate for the typically heavy-tailed distribution of STFT speech amplitudes [20], we apply an amplitude transformation

$$\widetilde{c} = \beta |c|^{\alpha} \mathrm{e}^{i\angle(c)} \tag{1}$$

to all complex STFT coefficients c, where $\angle(\cdot)$ represents the angle of a complex number, $\alpha \in (0, 1]$ is a compression exponent which brings out frequency components with lower energy (e.g. fricative sounds of unvoiced speech) [21], and $\beta \in \mathbb{R}_+$ is a scaling factor to roughly normalize amplitudes within [0,1].

Hereafter, we denote clean speech as $\mathbf{x}_0 \in \mathbb{C}^d$ and its corrupted version as $\mathbf{y} \in \mathbb{C}^d$, both of which represent flattened and amplitude-transformed spectrograms consisting of $d = KF$ complex coefficients.

B. DIFFUSION MODEL

Following our original approach [13], we use a task-adapted diffusion process for the conditional generation of clean speech $\mathbf{x}_0$ given a corrupted input $\mathbf{y}$. To this end, we define a forward stochastic process whose mean interpolates linearly between $\mathbf{x}_0$ and $\mathbf{y}$. Using the continuous-time formulation for diffusion models [14], the forward process is modeled as the solution to the stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{2}$$

where $\mathbf{x}_t \in \mathbb{C}^d$ denotes the process state at time $t \in [0, T]$, $\gamma \in \mathbb{R}$ controls the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and $g(t) \in \mathbb{R}$ is

the diffusion coefficient that controls the amount of Gaussian noise induced by a standard Wiener process $\mathbf{w}$. Note that t is only used to index the stochastic process and is completely unrelated to the time dimension of the audio signal.

The forward process can be time-inverted [14], resulting in a corresponding reverse process

$$\mathrm{d}\mathbf{x}_t = \left[-\gamma\left(\mathbf{y}-\mathbf{x}_t\right) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t\left(\mathbf{x}_t|\mathbf{y}\right)\right]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{3}$$

where the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is intractable and therefore approximated by a *score model* $\mathbf{s}_\theta$, parameterized by θ.

To train the score model $\mathbf{s}_\theta$, we use the denoising score matching objective [22], which approximates the score function by estimating the Gaussian noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ injected by the Wiener process, resulting in the loss function

$$\arg\min_\theta \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})}\left[\left\|\mathbf{s}_\theta(\mathbf{x}_t,\mathbf{y},t) + \frac{\mathbf{z}}{\sigma_t}\right\|_2^2\right], \tag{4}$$

where the expectation is approximated by sampling all random variables at each training step. For a complete derivation of the loss function, we refer to our previous work [13].

Once the score model is trained, the reverse process (3) can be solved by replacing the score function with its approximation $\mathbf{s}_\theta$ and using a numerical SDE solver. For this purpose, we use the predictor-corrector sampler [14] with the configuration described in Section IV-B and initialize the reverse process with

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_T; \mathbf{y}, \sigma_T^2\mathbf{I}\right)\,, \tag{5}$$

where σ_T^2 corresponds to the noise power at time step $t = T$. The enhancement process is then based on iterating through the reverse process starting at $t = T$ and ending at $t = 0$.

C. NETWORK ARCHITECTURE

As a score model, we use a modified version of the Noise Conditional Score Network (NCSN++) [14]. The network is a U-Net-like encoder-decoder architecture based on 2D convolutions, which takes complex spectrograms $\mathbf{x}_t$ and $\mathbf{y}$, and the process timestep t as input. Real and imaginary parts are considered as separate channels and the convolutions are performed over time and frequency. For the non-causal baseline, we use the same network configuration as in SGMSE+ [13], except that attention is only used in the bottleneck of the network, resulting in 64.8 M parameters.

To design a causal version of NCSN++, we apply the following modifications to the architecture. First, we modify the padding in the 2D convolutions and truncate the output such that the convolution along the time dimension is causal. An example of this procedure is illustrated in Fig. 1 for a 1D convolution of kernel size 3. Next, we replace all group normalization layers with cumulative group normalization [23], aggregating statistics recursively. Downsampling in the time dimension is performed with strided convolutions and corresponding upsampling with transposed strided convolutions.

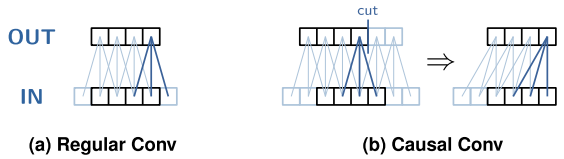

FIGURE 1. To obtain a causal convolution kernel, zero padding (light grey elements) is modified and the output is truncated. (a) Regular Conv. (b) Causal Conv.

Note that the strided and transposed strided convolutions remain causal operations with the padding described above. Up- and downsampling in the frequency dimension are realized with 1D finite impulse response filters [24]. Finally, all attention layers as well as the progressive growing path are removed from the network, resulting in 55.7 M parameters in total. Considering all modifications, we would like to point out that the comparison between the causal and non-causal approach is also a comparison of different network architectures, respectively a comparison of the method in [13] with the proposed causal method.

Contrary to the SIG challenge submission [17], we choose not to apply an automatic gain control. Instead, we train the score model on various input gains. For the reverberation corruption set, however, we noted that the causal diffusion model had difficulties in estimating the correct output gain, resulting in enhanced files with monotonically increasing gain over the whole utterance. We addressed this issue by training the score model on normalized clean speech targets while maintaining the varying input level. Thus, the diffusion model learns to generate normalized clean speech conditioned on variable input levels which results in learning a simpler target distribution. This approach has shown improved robustness for reverberant data.

III. GENERALIZED SPEECH ENHANCEMENT

In this work, we extend the classical speech enhancement task of removing additive background noise and address a more generalized class of speech enhancement. More specifically, we consider the task of reconstructing clean speech that has been degraded by a set of different signal modifications, which we will describe in detail below. Besides training the model on different corruptions, we augment the data during training with variable input gains to make the model more robust against loudness variations.

A. BACKGROUND NOISE

To simulate speech recordings that have been degraded by additive background noise, we assume an additive mixture model

$$y = x + n \tag{6}$$

where the corrupted signal y is the sum of clean speech x and background noise n.

B. REVERBERATION

Reverberation is caused by room acoustics and is characterized by multiple reflections on the room enclosures [25]. Here, we assume a convolutive corruption model

$$y = x * h \tag{7}$$

where the clean speech x is convolved with a room impulse response (RIR) h representing the acoustic path between the source and the listener.

C. CLIPPING

Clipping distortion occurs when the amplitude of a signal exceeds the maximum range that can be faithfully represented by, e.g., a fixed precision digital system or an electrodynamic microphone with limited membrane displacement range. This is modeled by the corruption model

$$y = \min(\max(x, A_{\min}), A_{\max}) \tag{8}$$

where the speech signal x is clipped between a minimum amplitude $A_{\min}$ and a maximum amplitude $A_{\max}$.

D. PACKET LOSS

In digital speech communication, such as voice-over-IP, data packets can get lost because of different transmission breakdown causes such as software corruption, interference, or jammed network nodes. Following the work in [26], we assume packets of length 20 ms and replace the signal with zeros if a packet is lost.

E. BANDWIDTH REDUCTION AND CODECS

Bandwidth reduction occurs when an audio signal at a high sampling rate is converted to a lower rate, e.g., to reduce the data transmission rate (bitrate) in a telecommunication system. The corruption process is linear and typically involves an anti-aliasing low-pass filter followed by a decimation operation:

$$y = \text{Resample}(x * a,\ f_s^{\text{h}},\ f_s^{\text{l}}) \tag{9}$$

where a is the impulse response of the anti-aliasing filter, f_s^{h} is the original high sampling rate and f_s^{l} is the lower target sampling rate.

Modern digital speech communication systems also employ source coding techniques to reduce the required bitrate. By considering specific signal characteristics (e.g. speech) and psychoacoustic models, lossy coding schemes offer a large reduction in bitrate but generally result in audible distortions and artifacts in the decoded signal, which become more severe as the compression ratio increases. In this work, we simulate the effect of coding by applying a speech or audio encoder, $\text{Enc}(\cdot)$, and the corresponding decoder, $\text{Dec}(\cdot)$, subsequently:

$$y = \text{Codec}(x) = \text{Dec}\,(\text{Enc}(x)) \tag{10}$$

TABLE 1. Clean Speech Files Statistics for the MultiCorruption Dataset

		LibriVox [28]	VCTK [29]	Total
# Files	Train	26299	42830	69129
	Valid	414	780	1194
	Test	418	845	1263
# Speakers	Train	220	106	326
	Valid	4	4	8
	Test	4	2	6
Length [h]	Train	75.5	40.2	115.7
	Valid	1.2	0.7	1.9
	Test	1.2	0.6	1.8

F. COMBINATIONS

To simulate the occurrence of multiple corruptions at the same time, we use the aforementioned signal modifications and randomly select corruption chains among plausible candidates, e.g. {Reverb → BackgroundNoise → PacketLoss}.

IV. EXPERIMENTAL SETUP

A. DATA

To train and test our model, we generate a new dataset which we call *MultiCorruption* consisting of pairs of clean and corrupted speech files. We utilize clean speech data from the DNS challenge [27]. More specifically, we use a subset of the clean speech files from the LibriVox recordings [28] ("*read_speech*" directory) and the entire VCTK speech data [29]. To select the subset of the LibriVox recordings, we use speakers who have between 100 and 140 utterances and make the dataset gender-balanced by using an open-source gender recognition method.[1] This results in 228 speakers (114 male / 114 female), of which 220 are used for training and 4 each for validation and testing. The VCTK corpus contains 110 speakers with approximately equal numbers of male and female speakers. Following previous works [13], [15], [30], we are using speakers "*p226*" and "*p287*" for validation, speakers "*p232*" and "*p257"* for the test, and the remaining speakers for training. Table 1 contains further details regarding the clean speech data.

To generate the corrupted speech files, we define six corruption datasets (*Noise*, *Reverb*, *Clipping*, *PacketLoss*, *BWR-Codecs*, and *Combinations*), corresponding to the signal modifications listed in Section III. Note that we group bandwidth reduction and codecs into one corruption set as codecs also often include limited bandwidth. Each corruption set is generated using all clean speech files. Thus the total length of training data for the model trained on all corruptions results in 694.2 h. For additive background noise, we use noise files from the DNS challenge [27] and mix clean speech and noise at a segmental signal-to-noise ratio

[1] [Online]. Available: https://github.com/x4nth055/gender-recognition-by-voice

(SNR)[2] uniformly sampled in $[-10, 15]$ dB. Reverberant data is generated using RIRs also taken from the DNS challenge [27], but cropped such that the direct path occurs at the very beginning. This helps guide the training process because no delay is introduced between the reverberant and clean speech file. For clipping, we define $A_{\min}$ and $A_{\max}$ as percentile thresholds, such that $A_{\min}$ correspond to the threshold of the k-th percentile, where k is uniformly sampled between $[0, 30]$, and $A_{\max}$ corresponds to the threshold of the $(100 - k)$-th percentile. For packet loss, we randomly sample the amounts of consecutive packets lost per drop between $[1, 5]$, and sample the number of drops per second between $[3, 6]$. For bandwidth reduction, 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 32 kHz with polyphase filtering. For the codec corruption, we use the AMR-NB, AMR-WB, and EVS speech communication codecs, as well as MP3 coding with bitrates varying between 6.6 kb/s and 64 kb/s. Note that the AMR-NB and AMR-WB codecs implicitly include bandwidth reduction to 8 kHz and 16 kHz, respectively. In addition to the single corruptions, we also simulate various plausible combinations, using the same random corruption parameters as mentioned above.

Besides using the *MultiCorruption* dataset, we train and test our model on a 32 kHz version of the publically available *Voicebank-Demand* dataset [30]. This dataset is often used as a benchmark for single-channel speech enhancement. The utterances are artificially contaminated with eight real-recorded noise samples from the DEMAND database [31] and two artificially generated noise samples (babble and speech shaped) at SNRs of 0, 5, 10, and 15 dB. The test utterances are mixed with different noise samples at SNR levels of 2.5, 7.5, 12.5, and 17.5 dB.

B. HYPERPARAMETER SETTINGS

All processing is performed at a sampling frequency of 32 kHz. We use an STFT with a 638-point Hann window and 160-point hop, which results in an algorithmic latency of 20 ms, as we use purely causal processing. The diffusion process hyperparameters are identical to those in [13]. We use 30 diffusion steps for the reverse diffusion and adopt the predictor-corrector-scheme [14] with one step of annealed Langevin dynamics correction and a step size of 0.5. Please note that we have not optimized the sampler configuration for the individual corruption types, which could potentially improve the performance, as it was the case for dereverberation in [13]. We train the score model $\mathbf{s}_\theta$ with the denoising score matching objective [14] using the Adam optimizer with a learning rate of 10^{-4}. We train each model on two NVIDIA RTX A6000 GPUs using an effective batch size of $2 \times 8 = 16$ and an exponential moving average over the parameters with a factor of 0.999. To allow a fair comparison between the specialized and generalized models, we train all models for 300 k training steps which takes around three days. Different from all other conducted experiments, the training for the Voicebank-Demand dataset is trained on four NVIDIA RTX A5000 GPUs with an effective batch size of $4 \times 4 = 16$ lasting for about two days.

C. EVALUATION METRICS

We describe here the standard speech enhancement metrics we used to assess the performance of the proposed method. A distinction is made between *intrusive* metrics computed by algorithms rating the processed signal in relation to the clean reference signal, and *non-intrusive* metrics which can be employed to evaluate real recordings when no clean reference is available.

Intrusive metrics include POLQA [32] for predicting speech quality which takes values from 1 (poor) to 5 (excellent) as usual for mean opinion scores (MOS). We also report PESQ [33], which is the predecessor of POLQA and is still widely used in the research community. The PESQ score lies between 1 (poor) and 4.5 (excellent). We further use ESTOI [34] as an intrusive measure of speech intelligibility. This metric yields values between 0 and 1, with higher values indicating better intelligibility. We report the speech mode of ViSQOL [35], a full-reference metric for perceived audio quality that is based on a model trained with data from subjective tests. Moreover, we calculate standard energy-based metrics including the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [36], and the log spectral distance (LSD). Both are measured in dB, with higher values for SI-SDR and lower values for LSD indicating better performance.

For non-intrusive speech quality assessment, we use the model DNSMOS P.835 [37] which is based on a listening experiment according to ITU-T P.835 [38] and provides three MOS scores: speech quality (SIG), background noise quality (BAK), and the overall quality (OVRL) of the audio. Furthermore, we use Wav-to-Vec MOS (WVMOS) [39] which is a non-intrusive MOS prediction method for speech quality evaluation using a fine-tuned wav2vec2.0 model [40].

V. RESULTS

In this section, we present the results of the proposed causal speech enhancement method and compare its performance against the non-causal counterpart. First, we present results for background noise removal using the standardized Voicebank-Demand dataset. Then, we investigate the method's generalization capability on different corruption types using the *MultiCorruption* dataset. Finally, we report the results evaluating on the SIG challenge blind set.

A. RESULTS ON VOICEBANK-DEMAND

In Table 2, we report speech enhancement results on the standardized Voicebank-Demand dataset. First, we discuss

[2]For the computation of the segmental SNR, we use the script from the DNS challenge [27].

TABLE 2. Results Obtained for the Voicebank-Demand. Values Indicate Mean and Standard Deviation

Method	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL
Unprocessed	3.00 ± 0.88	2.01 ± 0.73	0.68 ± 0.17	8.4 ± 5.6	3.59 ± 0.68	2.66 ± 0.82	2.66 ± 0.88	2.16 ± 0.66
Causal	3.89 ± 0.65	2.74 ± 0.71	0.75 ± 0.14	16.8 ± 3.2	4.12 ± 0.47	3.13 ± 0.50	3.67 ± 0.45	2.76 ± 0.50
Non-causal	3.98 ± 0.63	2.63 ± 0.74	0.75 ± 0.14	18.3 ± 3.5	4.12 ± 0.48	3.16 ± 0.49	3.81 ± 0.35	2.83 ± 0.49
Non-causal (norm. noisy)	3.94 ± 0.60	2.62 ± 0.66	0.75 ± 0.13	17.8 ± 3.5	4.11 ± 0.46	3.16 ± 0.50	3.74 ± 0.41	2.81 ± 0.50
Non-causal (strided convs)	3.88 ± 0.60	2.69 ± 0.68	0.74 ± 0.14	16.3 ± 3.4	4.08 ± 0.48	3.13 ± 0.51	3.62 ± 0.47	2.74 ± 0.50

TABLE 3. Results Obtained for the Noise Test Set. Values Indicate Mean and Standard Deviation

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	1.76 ± 0.83	1.38 ± 0.50	0.57 ± 0.21	3.1 ± 10.0	2.7 ± 0.9	2.8 ± 1.0	2.2 ± 0.8	2.1 ± 0.7	1.1 ± 2.1
Causal	Noise	2.76 ± 1.10	2.30 ± 0.85	0.74 ± 0.19	14.0 ± 9.6	3.7 ± 0.8	3.5 ± 0.2	3.9 ± 0.3	3.2 ± 0.3	3.6 ± 0.9
Causal	All	2.60 ± 1.14	2.14 ± 0.84	0.71 ± 0.21	11.8 ± 11.4	3.6 ± 0.8	3.5 ± 0.3	3.9 ± 0.4	3.1 ± 0.3	3.4 ± 1.0
Non-causal	Noise	2.99 ± 1.08	2.46 ± 0.88	0.76 ± 0.18	15.0 ± 9.3	3.9 ± 0.7	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.8 ± 0.6
Non-causal	All	2.68 ± 1.08	2.18 ± 0.86	0.71 ± 0.21	12.4 ± 11.1	3.7 ± 0.8	3.4 ± 0.3	4.0 ± 0.2	3.1 ± 0.3	3.6 ± 0.9

ablations using different normalization techniques. The non-causal model does not use any normalization before processing, whereas the method *Non-causal (norm. noisy)* is normalized based on the noisy mixture. We observe that these two methods differ only in a small amount in all reported metrics with a slight advantage for the model without normalization. Second, we show that changing the up- and downsampling in the NCSN++ architecture to non-causal strided and transposed strided convolutions results in a performance reduction mostly audible in the amount of background noise reduction. This is also reflected in the metrics as *Non-causal (strided convs)* shows a 0.19 lower performance in BAK compared to the non-causal method using the original FIR filter as in [13]. Last, we see that enforcing the strided convolution to be causal as shown in Fig. 1 has no negative impact on performance (see last row in Table 2). In summary, when accumulating the ablations we observe that the proposed causal configuration performs on par with the original non-causal NCSN++.

B. RESULTS ON MULTIPLE CORRUPTIONS

To systematically analyze how well the proposed method performs on different corruption types, we test the model on the individual corruption sets described in Section III. For both causal and non-causal settings, we compare the generalized model trained on the entire dataset (denoted as *All*) with specialized models only trained on the respective corruption set. Listening examples can be found online.[3]

We start by looking at the evaluation on the *Noise* corruption set which contains additive background noise (see Table 3). It can be seen that the non-causal method performs generally better than the causal method. This behavior is expected:

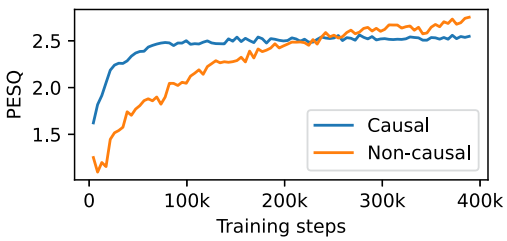

FIGURE 2. Training convergence on the Noise corruption set for the causal and the non-causal model.

due to speech being more temporally correlated than noise, the non-causal model can effectively use future information to disentangle speech and noise. However, the performance difference between the causal and non-causal variants is not significant for the generalized models trained on all data, where the non-causal model leads by a margin of only 0.08 in POLQA, 0.04 in PESQ, and 0.6 dB in SI-SDR. Listening to the enhanced files confirms that both methods sound comparable. For the specialized models, however, this difference is somewhat more pronounced, which can also be confirmed by informal listening, with fewer speech distortions for the non-causal model for challenging input.

Interestingly, we have noted different convergence behavior during training for the causal and the non-causal models. Fig. 2 shows the speech enhancement performance in PESQ evaluated on 20 randomly selected utterances from the validation set after each epoch. While the causal model converges faster, the non-causal model eventually surpasses the causal model. This different behavior is potentially due to the modified network architecture and may result from different up- and downsampling with strided and transposed strided convolutions as opposed to regular convolutions with FIR filters (see Section II-C).

[3][Online]. Available: https://uhh.de/inf-sp-causal-diffusion

TABLE 4. Results Obtained for the Reverb Test Set. Values Indicate Mean and Standard Deviation

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unprocessed	-	2.18 ± 0.57	1.70 ± 0.49	0.70 ± 0.13	-2.4 ± 6.6	4.1 ± 0.5	2.8 ± 0.7	3.2 ± 0.9	2.4 ± 0.6	3.0 ± 1.1
Causal	Reverb	3.03 ± 0.77	2.75 ± 0.66	0.86 ± 0.07	7.2 ± 5.4	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.3 ± 0.2	3.9 ± 0.5
Causal	All	2.83 ± 0.74	2.45 ± 0.66	0.82 ± 0.09	3.2 ± 6.5	4.1 ± 0.3	3.5 ± 0.2	4.0 ± 0.1	3.2 ± 0.2	3.7 ± 0.5
Non-causal	Reverb	3.19 ± 0.81	2.83 ± 0.71	0.87 ± 0.08	7.0 ± 6.8	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.3 ± 0.2	3.9 ± 0.5
Non-causal	All	3.05 ± 0.75	2.58 ± 0.66	0.84 ± 0.08	4.4 ± 6.5	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.2	3.9 ± 0.4

TABLE 5. Results Obtained for the Clipping Test Set. Values Indicate Mean and Standard Deviation

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	1.76 ± 0.80	1.63 ± 0.74	0.73 ± 0.13	3.3 ± 4.9	3.4 ± 0.8	3.4 ± 0.4	3.8 ± 0.4	3.0 ± 0.5	2.4 ± 1.4
Causal	Clipping	3.31 ± 0.79	3.15 ± 0.69	0.88 ± 0.07	12.7 ± 5.4	4.4 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.2	4.1 ± 0.4
Causal	All	2.74 ± 0.90	2.58 ± 0.78	0.82 ± 0.09	9.5 ± 5.9	4.1 ± 0.5	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.9 ± 0.5
Non-causal	Clipping	3.32 ± 0.73	2.99 ± 0.65	0.87 ± 0.08	12.6 ± 5.1	4.3 ± 0.4	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.2	4.1 ± 0.4
Non-causal	All	2.83 ± 0.83	2.49 ± 0.69	0.83 ± 0.09	9.9 ± 5.5	3.9 ± 0.5	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.9 ± 0.5

TABLE 6. Results Obtained for the PacketLoss Test Set. Values Indicate Mean and Standard Deviation

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	1.21 ± 0.27	1.31 ± 0.23	0.75 ± 0.07	5.8 ± 3.9	4.0 ± 0.4	2.9 ± 0.5	3.8 ± 0.3	2.6 ± 0.4	3.4 ± 0.6
Causal	PacketLoss	1.79 ± 0.64	2.02 ± 0.53	0.84 ± 0.06	7.4 ± 4.5	4.4 ± 0.2	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	4.1 ± 0.4
Causal	All	1.48 ± 0.55	1.86 ± 0.50	0.82 ± 0.07	6.7 ± 4.3	4.3 ± 0.3	3.4 ± 0.2	4.0 ± 0.1	3.2 ± 0.3	4.0 ± 0.4
Non-causal	PacketLoss	2.03 ± 0.64	1.95 ± 0.51	0.83 ± 0.06	7.7 ± 4.5	4.3 ± 0.3	3.4 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	4.0 ± 0.4
Non-causal	All	1.81 ± 0.62	1.87 ± 0.47	0.83 ± 0.07	7.0 ± 4.3	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	4.0 ± 0.4

Next, in Table 4 we compare the results on the *Reverb* corruption set, which contains noiseless reverberant speech data. We note that the specialized models have an advantage over the generalized models for the intrusive metrics (except for ViSQOL, which seems to be insensitive to reverberation). This reveals that task-specific training on dereverberation helps to improve the performance compared to multi-task training. Furthermore, we observe that non-causal models have a consistent improvement over the causal models, which shows that future information can be effectively leveraged for dereverberation. Nevertheless, the performance of the causal models remains competitive and shows significant improvements with respect to the unprocessed files.

In Table 5, we report the results on the *Clipping* corruption set. For the intrusive metrics, we observe a rather large performance difference between the specialized models and the generalized models trained on all data. In the causal setup, the generalized model sees a drop of 0.57 in POLQA and PESQ, 0.06 in ESTOI, and 3.2 dB in SI-SDR. A possible cause for this is the inclusion of combinations such as clipping followed by bandwidth reduction in the *All* dataset. This leads to utterances where the resulting corruption is relatively different from that of pure clipping since clipping mainly introduces high frequencies that are subsequently removed by bandwidth reduction. Interestingly, we note that the SIG metric for the unprocessed data (i.e. the clipped signals) shows a high value which does not reflect our listening experience, indicating that this metric is not particularly suitable for assessing clipping artifacts.

Table 6 reports the results using the *PacketLoss* corruption set. We can observe a clear improvement in all metrics with respect to the unprocessed files with benefits for the specialized models. This advantage of the specialized model seems to be more pronounced for the causal case, which is also confirmed by informal listening. However, for this typical inpainting task, we actually assumed that the non-causal model would perform significantly better than the causal model since future information should be heavily used by the non-causal model. This is, however, only the case for POLQA and is not reflected in the other metrics. We argue that the causal model still performs comparably well, because the gaps are relatively short (20-100 ms), and the past temporal context given by the receptive field of the causal model still covers enough speech parts to draw inferences to fill the gaps.

In Table 7, we show the results for the *BWR-Codecs* corruption set, which contains bandwidth-reduced audio and codec artifacts. We notice that the unprocessed files already yield good values for speech quality and intelligibility measures.

TABLE 7. Results Obtained for the BWR-Codecs Test Set. Values Indicate Mean and Standard Deviation

Method	Training	POLQA ↑	PESQ ↑	ESTOI ↑	SI-SDR [dB] ↑	LSD [dB] ↓	ViSQOL ↑	WVMOS ↑
Unproc.	-	3.88 ± 0.74	3.76 ± 0.78	0.91 ± 0.11	11.2 ± 18.6	1.90 ± 0.41	4.4 ± 0.5	3.86 ± 0.60
Causal	BWR+Codecs	4.06 ± 0.65	3.89 ± 0.70	0.94 ± 0.07	17.9 ± 7.7	1.46 ± 0.19	4.6 ± 0.4	4.12 ± 0.39
Causal	All	3.66 ± 0.93	3.59 ± 0.95	0.92 ± 0.10	13.1 ± 14.2	1.45 ± 0.24	4.5 ± 0.5	4.01 ± 0.48
Non-causal	BWR+Codecs	4.13 ± 0.57	3.94 ± 0.60	0.94 ± 0.07	18.4 ± 7.1	1.57 ± 0.26	4.5 ± 0.4	4.13 ± 0.39
Non-causal	All	3.88 ± 0.74	3.68 ± 0.73	0.91 ± 0.10	12.2 ± 17.6	1.46 ± 0.21	4.4 ± 0.4	4.05 ± 0.48

TABLE 8. Results Obtained for the All Test Set. Values Indicate Mean and Standard Deviation

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	2.03 ± 1.09	1.83 ± 1.03	0.68 ± 0.21	2.2 ± 11.9	3.5 ± 1.0	2.9 ± 0.8	3.2 ± 1.1	2.5 ± 0.8	2.3 ± 2.0
Causal	All	2.49 ± 1.13	2.35 ± 0.98	0.77 ± 0.18	7.1 ± 10.7	3.9 ± 0.8	3.5 ± 0.2	4.0 ± 0.3	3.1 ± 0.3	3.6 ± 0.9
Non-causal	All	2.67 ± 1.09	2.39 ± 0.95	0.78 ± 0.19	7.2 ± 11.8	3.9 ± 0.7	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.8 ± 0.6

While the specialized models further improve these metrics, the generalized models slightly reduce them, although they still turn out to be quite high (PESQ and POLQA > 3.5) in absolute terms. It should, however, be noted that PESQ and POLQA have been reported to correlate poorly with listening experiments for bandwidth reduction [11]. We thus also include the log spectral distance (LSD) as a common metric used for bandwidth extension and observe that the causal models perform slightly better in it than the non-causal model. Note that for brevity, we omit SIG, BAK, and OVRL as they have not shown any difference on this corruption set.

Finally, Table 8 shows the average performance using all corruption sets, including different combinations of corruptions. It can be seen, that the non-causal model has a slight advantage over the causal approach but does not suffer from a significant performance penalty except a drop of 0.22 in POLQA.

In summary, we find that task-specific models trained on a specific corruption type typically outperform generalized models using multi-task training. This follows the intuitive explanation that a more complex task leads to worse overall performance for the same network capacity. Similar behavior has also been reported for image restoration tasks [41]. In addition, we observe that the modification to causal processing slightly degrades the performance. This is consistent with the intuition that the causal model can only use information from the past, while the non-causal model can also leverage future information for the estimation.

C. RESULTS ON THE SIG CHALLENGE BLIND SET

In Table 9, we present the results when evaluating the causal and non-causal model on the 500 files from the 2023 SIG challenge's blind test set [18]. Since there is no reference signal available, we can only report non-intrusive metrics. While both models improve all metrics, there is a small advantage for the non-causal model.

TABLE 9. Results Obtained for the Blind Set of the SIG Challenge

Method	SIG	BAK	OVRL	WVMOS
Unprocessed	2.9 ± 0.6	3.4 ± 0.7	2.5 ± 0.6	2.2 ± 1.2
Causal	3.2 ± 0.4	3.8 ± 0.4	2.9 ± 0.4	2.4 ± 0.8
Non-causal	3.3 ± 0.4	4.0 ± 0.2	3.0 ± 0.4	2.7 ± 0.8

D. COMPUTATIONAL COMPLEXITY

With the given reverse sampler configuration, the proposed method requires calling the score model 60 times to process an utterance. While designed to run on a GPU, this currently poses a practical limitation in real-time applications on a CPU. Processing a file of 10 s requires 236 TMACs and takes 30.5 s on a GPU and 30 minutes on a CPU.[4] However, since the method performs strictly causal signal processing, real-time operation is still theoretically possible, assuming that the computational complexity can be decreased in the future. One possible avenue towards decreased complexity is reducing the number of diffusion steps, e.g., as in [42]. This is an active area of research which we plan to focus on in future work.

VI. CONCLUSION

In this work, we have presented a causal speech enhancement method with an algorithmic latency of 20 ms that is based on generative diffusion models. To guarantee causal processing, we have modified the network architecture and removed all non-causal normalization techniques. We addressed a more generalized speech enhancement scenario beyond speech-in-noise by training and testing the model on multiple corruption types. For this purpose, we generated a new corrupted speech dataset which includes additive background noise, reverberation, clipping, packet loss, bandwidth reduction, and codec artifacts. In the experiments, we have conducted evaluations

[4] Average processing time for 10 audio files on an NVIDIA GeForce RTX 2080 Ti GPU, in a machine with an Intel Core i7-7800X CPU @ 3.50 GHz.

on the individual corruption sets and compared the generalized model trained on all data with specialized models trained solely on each specific corruption type. Moreover, we compared the causal diffusion model against the non-causal baseline. Our findings revealed that, although specialized models and non-causal models have a small advantage, the generalized causal approach does not suffer from a significant performance penalty, while being more practicable for real-world applications.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-The-Art*. San Rafael, CA, USA: Morgan & Claypool, 2013.

[3] S. Pascual, J. Serrá, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2019, pp. 1791–1795.

[4] H. Liu et al., "Voicefixer: Toward general speech restoration with neural vocoder," 2021, *arXiv:2109.13731*.

[5] J. Serrá, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," 2022, *arXiv:2206.03065*.

[6] J. Su, A. Finkelstein, and Z. Jin, "Perceptually-motivated environment-specific speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7015–7019.

[7] J. Chen et al., "Gesper: A unified framework for general speech restoration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[8] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2020, pp. 3291–3295.

[9] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

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

[11] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[12] D. d. 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 *Proc. Speech Commun. 15th ITG Conf.*, 2023, pp. 260–264.

[13] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2021.

[15] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7402–7406.

[16] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2022, pp. 2928–2932.

[17] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Speech signal improvement using causal generative diffusion models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[18] R. Cutler, A. Saabas, B. Naderi, N.-C. Ristea, S. Braun, and S. Branets, "ICASSP 2023 speech signal improvement challenge," *IEEE Open J. Signal Process.*, 2024.

[19] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[20] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop Acoust. Echo Noise Control.*, 2010.

[21] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. 44th Int. Conf. Telecom. Signal Process.*, 2021, pp. 72–76.

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

[23] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[24] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7324–7334.

[25] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, vol. 59. Berlin, Germany: Springer, 2011.

[26] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "Interspeech 2022 audio deep packet loss concealment challenge," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2022, pp. 580–584.

[27] H. Dubey et al., "ICASSP 2022 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9271–9275.

[28] "LibriVox: Free public domain audiobooks." Accessed: Jul. 24, 2023. [Online]. Available: https://librivox.org/

[29] J. Yamagishi et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. ISCA Speech Synth. Workshop*, 2016, pp. 146–152.

[31] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *J. Acoustical Soc. Amer.*, vol. 133, no. 5, pp. 3591–3591, 2013.

[32] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," Int. Telecom. Union, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.

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

[35] M. Chinen, F. S. 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. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.

[36] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done ?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 626–630, 2019.

[37] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 886–890.

[38] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Int. Telecom. Union, 2003. [Online]. Available: https://www.itu.int/rec/T-REC-P.835-200311-I/en

[39] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi : A unified framework for bandwidth extension and speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[40] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[41] C. Saharia et al., "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH Conf.*, 2022, pp. 1–10.

[42] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2023, pp. 3809–3813.

JULIUS RICHTER (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the Technical University of Berlin, Berlin, Germany, in 2017 and 2019, respectively. He is currently with the Signal Processing Group, Universität Hamburg, Hamburg, Germany, under the supervision of Prof. Dr.-Ing. Timo Gerkmann. His research interests include deep generative models and multi-modal learning with applications to audio-visual speech processing.

SIMON WELKER (Student Member, IEEE) received the B.Sc. degree in computing in science and M.Sc. degree in bioinformatics degrees from Universität Hamburg, Hamburg, Germany, in 2019 and 2021, respectively. He is currently working toward the Ph.D. degree (second-year) in signal processing with the labs of Prof. Timo Gerkmann, Universität Hamburg and Prof. Henry N. Chapman Center for Free-Electron Laser Science, DESY, Hamburg, researching machine learning techniques for solving inverse problems that arise in speech processing and X-ray imaging.

JEAN-MARIE LEMERCIER (Graduate Student Member, IEEE) received the M.Eng. degree in electrical engineering from Ecole Polytechnique, Paris, France, in 2019, and the M.Sc. degree in communications and signal processing from Imperial College London, London, U.K., in 2020. He is currently with the Signal Processing Group, Universität Hamburg, Hamburg, Germany, under the supervision of Prof. Dr.-Ing. Timo Gerkmann. His research interests include machine learning-based speech enhancement and dereverberation for hearing devices applications. His recent works also include the design and analysis of diffusion-based generative models for various speech restoration tasks.

BUNLONG LAY (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in mathematics from the University of Bonn, Bonn, Germany, in 2015 and 2017, respectively. From 2018 to 2021, he was with Research Institute Fraunhofer FKIE, Wachtberg, where he focused on research in the field of radar signal processing. Since 2021, he has been working toward Ph.D. with the University of Hamburg, Hamburg, Germany. His current research focuses on diffusion-based models for speech enhancement.

TAL PEER (Student Member, IEEE) received the B.Sc. degree in general engineering science and the M.Sc. degree in Electrical engineering from the Hamburg University of Technology, Hamburg, Germany, in 2016 and 2019, respectively. He is currently with the Signal Processing group at Universität Hamburg under the supervision of Prof. Dr.-Ing. Timo Gerkmann. His research interests include phase-aware speech enhancement and phase retrieval for speech and audio applications.

TIMO GERKMANN (Senior Member, IEEE) is currently a Professor for signal processing with the Universität Hamburg, Hamburg, Germany. He has held positions with Technicolor Research and Innovation, University of Oldenburg, Oldenburg, Germany, KTH Royal Institute of Technology, Stockholm, Sweden, Ruhr-Universität Bochum, Bochum, Germany, and Siemens Corporate Research, Princeton, NJ, USA. His research interests include statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. He was the recipient of the VDE ITG award 2022. Mr. Gerkmann was the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and currently a Senior Area Editor of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.