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/ICASSP49357.2023.10095126

SPEECH SIGNAL IMPROVEMENT USING CAUSAL GENERATIVE DIFFUSION MODELS

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, Tal Peer, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

ABSTRACT

In this paper, we present a causal speech signal improvement system that is designed to handle different types of distortions. The method is based on a generative diffusion model which has been shown to work well in scenarios with missing data and non-linear corruptions. To guarantee causal processing, we modify the network architecture of our previous work and replace global normalization with causal adaptive gain control. We generate diverse training data containing a broad range of distortions. This work was performed in the context of an "ICASSP Signal Processing Grand Challenge" and submitted to the non-real-time track of the "Speech Signal Improvement Challenge 2023", where it was ranked fifth.

***Index Terms*—** Speech signal improvement, universal speech enhancement, diffusion models, causal processing

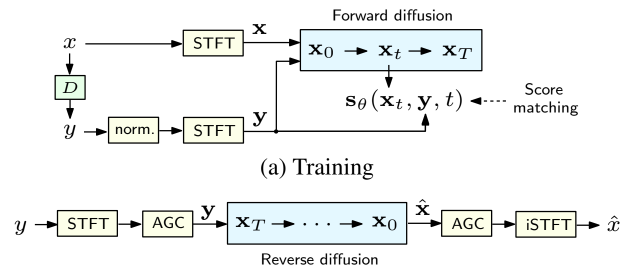
1. INTRODUCTION

High-quality voice communication requires clear audio and natural-sounding speech. However, there are numerous factors that can degrade speech signals including background noise, room acoustics, transmission errors, limited bandwidth, and codec artifacts. Prior works on improving speech signals have typically studied each type of distortion separately. However, there has been some recent interest in developing universal approaches that address a broader range of distortions. These approaches typically use generative modeling, which works particularly well in scenarios with missing data and non-linear corruptions [1].

In this paper, we present our diffusion-based speech enhancement method submitted to the non-real-time track of the "Speech Signal Improvement Challenge 2023" [2] as part of the "ICASSP Signal Processing Grand Challenges". The proposed model extends our previous work [3], incorporating significant modifications in the network architecture to meet the causality requirement and to output super wideband speech. Furthermore, we devise a data corruption approach to generate diverse training data resembling distortions observed in the blind data. Strong variations in loudness are compensated by using causal adaptive gain control.

In the challenge's subjective test, our proposed method yields a final score of 0.445 using ITU-T P.863.2 and P.804 with mean opinion scores (MOSs) of 2.570 (Overall), 2.998 (Signal), 3.765 (Noise), 3.330 (Coloration), 3.674 (Discontinuity), 3.435 (Loudness), and 4.241 (Reverberation). In the subjective test based on ITU-T P.835 our method yields a final score of 0.495 with MOSs of 2.842 (Overall), 3.119 (Signal), and 3.682 (Background). It is interesting to note that, as our approach is generative in nature, it is capable of achieving excellent performance for moderately distorted inputs, while it may generate phonetic confusions and insertions if the input distortions are too strong.

This work has been funded 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) with the Grant-No. HIDSS-0002, and the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380.



(a) Training

(b) Inference

Fig. 1. Proposed system: (a) At training, a corruption model D generates y from clean speech x. The forward diffusion moves from the clean spectrogram $\mathbf{x}_0$ to the corrupted $\mathbf{y}$, while Gaussian noise is gradually added. The score model $\mathbf{s}_\theta$ is learned using score matching. (b) At inference, the data is normalized with causal adaptive gain control (AGC) and the reverse diffusion uses the trained score model.

2. PROPOSED SYSTEM

Fig. 1 shows an overview of our proposed system for speech signal improvement at training and inference time. The diffusion process (Sec. 2.1) is the core of the method and is accompanied by other processing blocks including the short-time Fourier transform (STFT), causal automatic gain control (AGC) (Sec. 2.3), and a data corruption model D used to simulate various distortion types (Sec. 3.1).

2.1. Diffusion process

The diffusion process $\{\mathbf{x}_t\}_{t=0}^T$ for speech enhancement is essentially the same as in our previous works [3], [4], where $\mathbf{x}_t$ is the state of the process at time step t. During training, the forward process moves from clean speech $\mathbf{x}_0 := \mathbf{x}$ to corrupted speech $\mathbf{y}$, while increasing amounts of Gaussian noise are gradually added. At inference, the corresponding reverse process [5] is used to progressively remove the corruption and therefore generate an estimate of $\mathbf{x}$ starting from $\mathbf{x}_T \sim \mathcal{N}_\mathbb{C}(\mathbf{x}_T; \mathbf{y}, \sigma_t^2)$. This reverse process involves the *score function* of $\mathbf{x}_t$, i.e., the gradient of its log-probability. Functioning as a prior for clean speech, it is unavailable at inference time and is thus approximated by a trained deep neural network (DNN) $\mathbf{s}_\theta$ called the *score model*. We make no further adaptations to this process as it is defined for every STFT bin independently, and is thus inherently causal as long as $\mathbf{s}_\theta$ is implemented with a causal network architecture.

2.2. Network architecture

We use a modified version of NCSN++ [5] for score estimation. The network is an encoder-decoder architecture based on 2D convolutions, taking complex spectrograms $\mathbf{x}_t$ and $\mathbf{y}$ and the process time t as input. Real and imaginary parts are considered as separate channels and the convolutions are performed over time and frequency.

We apply the following modifications to the architecture to meet the causality constraint: **(1)** Padding in the 2D convolutions is modified so that the convolution along the time-dimension is causal; **(2)** Batch normalization is replaced with cumulative group normalization, aggregating statistics recursively; **(3)** Downsampling in the time dimension is performed with strided convolutions and corresponding upsampling with transposed strided convolutions. Up- and down-sampling in the frequency dimension are realized with finite impulse response filters, as in [3]. **(4)** All attention layers as well as the progressive growing path are removed.

2.3. Automatic gain control

To match the unit-scale training condition of the score model, i.e., normalization of corrupted speech y (see Fig. 1a), we use a causal AGC system before feeding the mixture to the diffusion process, and again after enhancement to maximize loudness, as part of the signal improvement task (see Fig. 1b). To this end, we recursively track the maximum value per magnitude frame averaged over the frequency bins to normalize the spectrogram in a causal manner. We start tracking when speech activity is first detected, using a voice activity detection method based on a causal speech presence probability estimator [6]. The speech probability is fed through an ideal low-pass filter to avoid having high-frequency noise bursts produce false positives. Voice activity is then assumed if the speech presence probability is higher than a threshold $\tau = 0.8$, for a duration of 100ms. When discovering a larger maximum than the previous one, we smooth the normalization with an exponential ramp going from the old value to the new one. Finally, to avoid clipping, we use a causal compressor from the `pedalboard` library[1].

3. EXPERIMENTAL SETUP

3.1. Dataset

We use the VCTK corpus [7] as the clean speech dataset and resample all utterances from 48kHz to our processing sampling frequency 32kHz. Using the `audiomentations` library[2], we simulate several corruptions observed in the blind data, namely stationary and non-stationary noise, reverberation, clipping, gain reduction, packet loss and lossy speech coding (the last two implemented by us). For each clean utterance, a random corruption chain is chosen among plausible candidates, e.g. $\{\text{Reverb} \rightarrow \text{Noise} \rightarrow \text{PacketLoss}\}$, with the parameters of each corruption chosen randomly as well. We use the QUT corpus [8] as the environmental noise dataset and take room impulse responses from the DNS challenge [9] for reverberation.

3.2. Hyperparameters and training configuration

All processing is performed at $f_s = 32\text{kHz}$ and we upsample the processed files to the original 48kHz frequency. We use an STFT with a 638-point Hann window and 160-point hop, which guarantees the global latency to be below 20ms, as we use purely causal processing. We set the lowest two frequency bins of the spectrogram to zero, to remove the DC offset and low frequency noise. The diffusion process hyperparameters are identical to those in [3]. We train the score model $\mathbf{s}_\theta$ with the denoising score matching objective [5] using the Adam optimizer with mini-batch size of 8 and exponential moving average over the parameters with a factor of 0.999. Training takes around two days on two NVIDIA RTX A6000 GPUs.

[1] `https://github.com/spotify/pedalboard`
[2] `https://github.com/iver56/audiomentations`

4. RESULTS

In addition to the challenge's subjective evaluation [2], we use DNSMOS P.835 [10] to evaluate our method on the 500 files in the blind set. Fig. 2 depicts the histograms of DNSMOS scores: (a), (b) speech quality (SIG) and overall quality (OVRL) of the improved files compared to the corrupted files; (c) SIG with and without AGC.

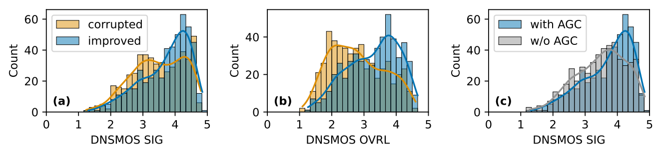


Fig. 2. Speech signal improvement results obtained with DNSMOS.

Computational complexity: The network has about 55.7M parameters and the time to infer a frame on a CPU (Intel Core i7-7800X @ 3.50GHz) takes 0.89s. Please note that the model is designed to run on a GPU, on which the inference time is orders of magnitude faster (0.02 s / frame with an NVIDIA GeForce RTX 2080 Ti).

5. CONCLUSION

In this work, we have built upon our previous work on diffusion-based speech enhancement, with the novel contribution in making the system causal and training the model on different distortion types. The proposed method was ranked fifth in the non-real-time track of the "Speech Signal Improvement Challenge 2023".

6. REFERENCES

[1] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023.

[2] R. Cutler, A. Saabas, B. Naderi, N.-C. Ristea, S. Braun, and S. Branets, "ICASSP 2023 speech signal improvement challenge," 2023. [Online]. Available: `https://arxiv.org/abs/2303.06566`.

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

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

[5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021.

[6] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," *WASPAA*, 2011.

[7] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2019.

[8] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms," in *Interspeech*, 2010.

[9] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 deep noise suppression challenge," in *ICASSP*, 2022.

[10] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2022.