applied sciences

MDPI

Article

Exploring Multi-Stage GAN with Self-Attention for Speech Enhancement

Bismark Kweku Asiedu Asante [1,*], Clifford Broni-Bediako [2] and Hiroki Imamura [1,*]

[1] Graduate School of Science and Engineering, Soka University, Hachioji City 192-8577, Japan
[2] RIKEN Center for Advanced Intelligence Project, Nihonbashi, Chuo City 103-0027, Japan
* Correspondence: e18d5201@soka-u.jp (B.K.A.A.); imamura@soka-u.jp (H.I.)

Abstract: Multi-stage or multi-generator generative adversarial networks (GANs) have recently been demonstrated to be effective for speech enhancement. The existing multi-generator GANs for speech enhancement only use convolutional layers for synthesising clean speech signals. This reliance on convolution operation may result in masking the temporal dependencies within the signal sequence. This study explores self-attention to address the temporal dependency issue in multi-generator speech enhancement GANs to improve their enhancement performance. We empirically study the effect of integrating a self-attention mechanism into the convolutional layers of the multiple generators in multi-stage or multi-generator speech enhancement GANs, specifically, the ISEGAN and the DSEGAN networks. The experimental results show that introducing a self-attention mechanism into ISEGAN and DSEGAN leads to improvements in their speech enhancement quality and intelligibility across the objective evaluation metrics. Furthermore, we observe that adding self-attention to the ISEGAN's generators does not only improves its enhancement performance but also bridges the performance gap between the ISEGAN and the DSEGAN with a smaller model footprint. Overall, our findings highlight the potential of self-attention in improving the enhancement performance of multi-generator speech enhancement GANs.

Keywords: speech enhancement; generative adversarial network (GAN); multi-stage GAN; multi-generator GAN; self-attention mechanism

Citation: Asiedu Asante, B.K.; Broni-Bediako, C.; Imamura, H. Exploring Multi-Stage GAN with Self-Attention for Speech Enhancement. *Appl. Sci.* **2023**, *13*, 9217. https://doi.org/10.3390/app13169217

Academic Editors: Dong Wang and Andrew Abel

Received: 23 June 2023
Revised: 23 July 2023
Accepted: 25 July 2023
Published: 14 August 2023

1. Introduction

Speech and audio signal processing applications, such as speech recognition [1] and hearing aids [2], require clean speech signals [3]. However, real-world speech signals are inevitably impacted by background noise which can distort speech quality and intelligibility. Speech enhancement algorithms aim at approximating clean speech signals from distorted speech signals by removing the background noise contained in the noisy signal [4]. The remarkable success in deep learning has inspired the speech and audio signal processing research community to shift from their traditional speech enhancement algorithms [5,6] to deep neural networks (DNNs)-based algorithms [7]. These include convolutional neural networks (CNNs) [8,9] and long short-term memory (LSTM) recurrent neural networks (RNNs) [10–12], which are discriminative methods, and generative methods, such as variation auto-encoder (VAE) [13–15] and generative adversarial networks (GANs) [1,16,17]. The GAN-based methods have been demonstrated to be more promising for speech enhancement tasks [18], and they are more robust to different types of noise [19–21] compared to their discriminative counterparts [22,23]. In the wake of the seminal work, speech enhancement GAN (SEGAN), by Pascual et al. [16], there have been several improvements to the GAN-based speech enhancement methods [18]. For example, self-attention SEGAN (SASEGAN) [23] was introduced to learn temporal dependencies across the signal sequence, and MetricGAN [24] directly optimized the generator with respect to evaluation metrics such as PESQ and STOI to improve the performance in SEGAN. Furthermore, several loss

functions, including relativistic loss function [17], metric loss function [25], and Wasserstein loss function [25], have also been proposed to stabilize the SEGAN training process. These methods are based on the conventional GAN [26]. Multi-stage GANs which involve the use of multiple generators [27] or multiple discriminators [28] to generate samples in multiple stages or levels have become increasingly popular in the field of speech enhancement. Multi-stage GANs have been adopted to refine noisy input signals in speech enhancement tasks [22,29,30]. In [22], multiple generators are employed to refine noisy input signals, whereas [29,30] utilized multiple discriminators to address acoustic degradation, such as noise, reverb, and equalization distortion, aiming to enhance speech clarity and intelligibility. Moreover, MelGAN [31] has shown the effectiveness of multi-stage GANs in high-quality mel-spectrogram inversion.

The multi-stage GANs have demonstrated successful performance in speech enhancement. However, the existing multi-stage GANs for speech enhancement rely on convolutional backbones [27,28], which may not be optimal for capturing temporal dependencies of an input signal sequence [32,33]. This work considers the temporal dependency problem of the convolutional backbone of multi-stage GANs. The self-attention mechanism [34] has successfully been used for temporal dependency modelling in acoustic [35] and speech recognition [32,36] tasks. Compared to RNN and LSTM, self-attention is computationally efficient and more flexible in modelling temporal dependencies of long input signal sequences [25,35]. Motivated by the work in SASEGAN [23], which adopted the concept of non-local attention [37,38] to optimise the performance of SEGAN [16], we introduce the self-attention mechanism in multi-stage GANs for speech enhancement tasks. In this work, we consider multi-stage GANs of multiple generators for speech enhancement (i.e., multi-generator speech enhancement GANs). Here, we aim to optimise the enhancement performance of iterated SEGAN (ISEGAN) and deep SEGAN (DSEGAN), the multi-stage GAN speech enhancement algorithms introduced by Phan et al. [22] (see Figure 1). We leverage the sequential modelling capability of self-attention to infuse the multiple generators of ISEGAN (two shared generators) and DESEGAN (two independent generators) with the power of capturing temporal dependency across an input signal sequence. The main contributions of this paper are summarized as follows:

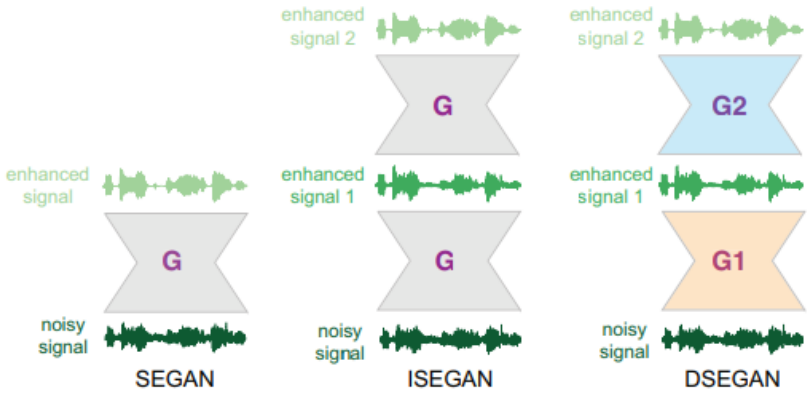


Figure 1. A representation of SEGAN with a single generator *G* (conventional GAN) and multi-stage GANs of multiple generators: ISEGAN with two shared generators *G* and DSEGAN with two independent generators *G1* and *G2* [22] (Figure adapted from [22]).

- To enhance the ISEGAN and DSEGAN networks, we incorporate the self-attention mechanism inspired by the implementation in [23,37]. We refer to these enhanced versions as ISEGAN-Self-Attention and DSEGAN-Self-Attention, respectively.
- Across the commonly used objective evaluation metrics, the proposed ISEGAN-Self-Attention and DSEGAN-Self-Attention demonstrated a better speech enhancement performance in all than the ISEGAN and the DSEGAN baselines, respectively.

- We also demonstrate that with the self-attention mechanism, the ISEGAN can achieve competitive enhancement performance with the DESGAN using only half of the model footprint of the DSEGAN.
- Furthermore, we investigate the effect of the self-attention mechanism applied at different stages of the multiple generators in the ISEGAN and the DSEGAN networks with respect to their enhancement performance.

The rest of the paper is organized as follows: Section 2 presents the background of the study. The proposed ISEGAN-Self-Attention and DSEGAN-Self-Attention are presented in Section 3. Section 4 describes the experimental setup of the study. The results are presented and discussed in Section 5. Finally, Section 6 concludes the paper with some future directions.

2. Background

2.1. Conventional SEGAN

In a speech enhancement task, a given speech or raw audio signal with noise can be represented as $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{n} \in \mathbb{R}^T$, where $\mathbf{x} \in \mathbb{R}^T$ is the clean signal and $\mathbf{n} \in \mathbb{R}^T$ represents a noisy signal that corrupts the speech signal. This noisy signal is often considered additive background noise. The main goal of speech enhancement is to remove the noisy signal in the raw audio signal by finding an enhancement mapping function f such that $f(\tilde{\mathbf{x}}) : \tilde{\mathbf{x}} \rightarrow \mathbf{x}$. One such mapping function that has been adopted with great success is the conditional-based GAN method [39,40].

The conditional-based GAN method was first employed in speech enhancement in the seminal work by Phan et al. [16], which is referred to as speech Enhancement GAN (SEGAN). In SEGAN, the generator G (for emphasis, a single generator) is provided with a corrupted raw audio signal as the enhancement mapping function such that $\hat{\mathbf{x}} = G(\mathbf{z}, \tilde{\mathbf{x}})$, where $\mathbf{z}$ is a latent variable. Then, the generator G is trained to produce the enhanced output signal $\hat{\mathbf{x}}$ simultaneously with the discriminator D, which learns to distinguish between the enhanced output signal $\hat{\mathbf{x}}$ and the real clean signal $\mathbf{x}$ by classifying the pair of signals $(\mathbf{x}, \tilde{\mathbf{x}})$ as real and $(\hat{\mathbf{x}}, \tilde{\mathbf{x}})$ as fake pair. The training procedure is illustrated in Figure 2, and the Equations (1) and (2) are the least-squares objective functions for training the discriminator D and the generator G, respectively.

$$\min_D \mathcal{L}_{LS}(D) = \frac{1}{2}\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{data}(\mathbf{x},\tilde{\mathbf{x}})}(D(\mathbf{x},\tilde{\mathbf{x}}) - 1)^2 + \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}_{data}(\tilde{\mathbf{x}})} D(G(\mathbf{z},\tilde{\mathbf{x}}),\tilde{\mathbf{x}})^2, \quad (1)$$

$$\min_G \mathcal{L}_{LS}(G) = \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}_{data}(\tilde{\mathbf{x}})}(D(G(\mathbf{z},\tilde{\mathbf{x}}),\tilde{\mathbf{x}}) - 1)^2 + \lambda||G(\mathbf{z},\tilde{\mathbf{x}}) - \tilde{\mathbf{x}})||_1. \quad (2)$$

Several improvements have been introduced to the SEGAN algorithm focusing on improving the quality of the enhanced speech signals [17,23,24,41–43]. For instance, Deepak and Verhulst [17] presented an improvement in enhancement by introducing a cost function with a gradient penalty. The self-attention module [37] was incorporated into the SEGAN to prevent the convolutional layers from obscuring the temporal dependency in an input sequence of signals [23]. Additionally, MetricGAN [24] also demonstrated that evaluation metrics can be used to improve the performance of conditional GAN-based method in speech enhancement by using the metrics that are directly related to speech signals (e.g., PESQ and STOI) rather than Euclidean distances to compute the loss in the SEGAN.

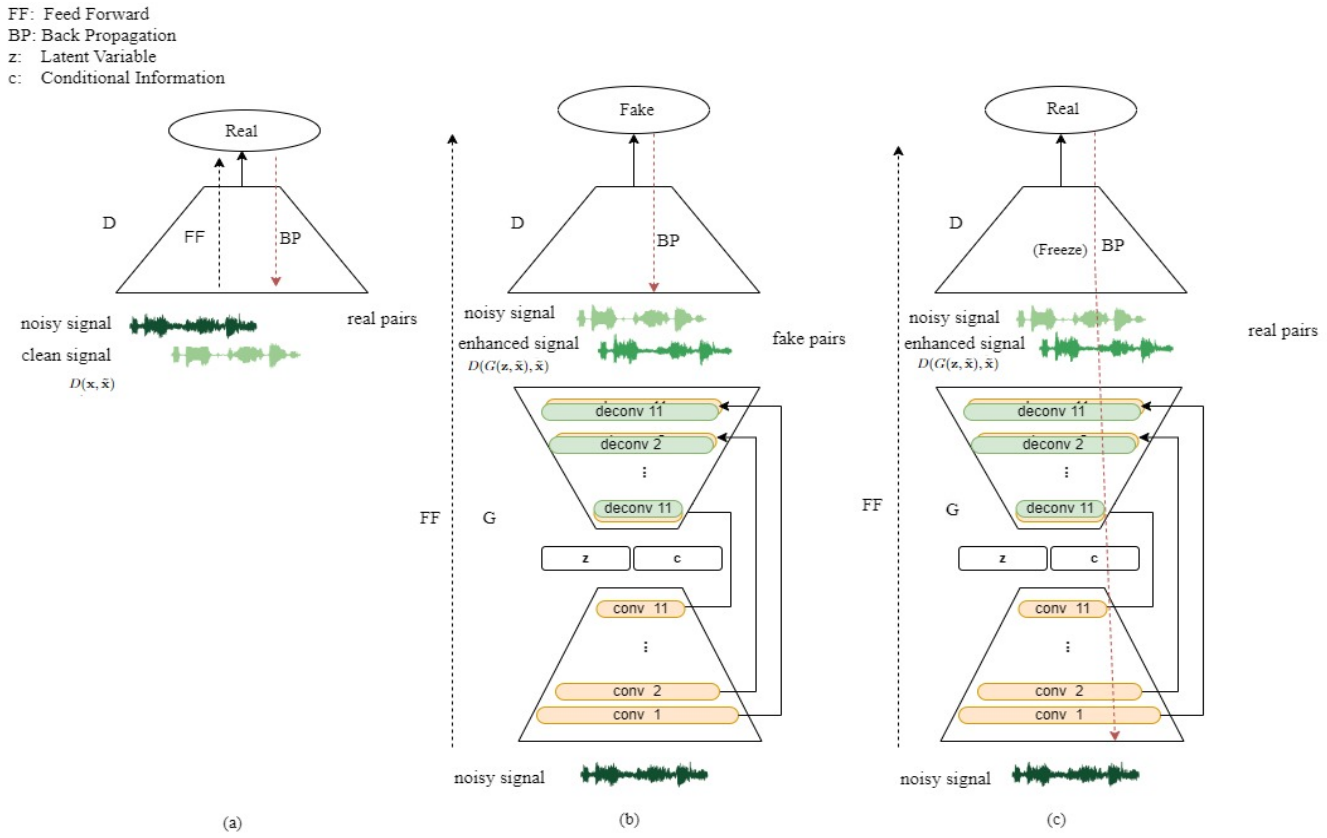


Figure 2. Illustration of the training process of a conventional SEGAN. In (**a**), the training process updates the discriminator D to distinguish real samples, and in (**b**), the training process updates only the discriminator D to distinguish fake samples, while in (**c**), the training process updates only the generator G to generate real samples.

2.2. SEGAN with Multiple Generators

Based on the principle of multiple generators chained to build the generator of a GAN [27], which has shown improved performance in image reconstruction, Phan et al. [22] introduced multi-stage GANs of multiple generators (i.e., ISEGAN and DSEGAN) for speech enhancement. The ISEGAN and DSEGAN learn multiple enhancement mappings with a chained generator $\mathcal{G}$ composes of N generators such that $\mathcal{G} = G_1 \rightarrow G_2 \rightarrow \cdots \rightarrow G_N$, where $N > 1$ to perform a speech enhancement tasks. The use of multiple generators aims to leverage the diversity of their outputs to enhance the overall quality of generated speech. In ISEGAN, the generators share their parameters, resulting in a common mapping function that is iteratively applied at all enhancement stages. Thus, ISEGAN generators can be considered as a single *iterated generator* $\mathcal{G}$ with N iterations. Sharing the generators' parameters makes the memory footprint of the ISEGAN smaller. Unlike the ISEGAN, the parameters of the DSEGAN generators are independent, which allows for flexible learning of different enhancement mappings at different stages of the network. In contrast to ISEGAN generators, the generators of DSEGAN can also be considered as a *deep generator* $\mathcal{G}$ with a depth of N. Figure 3 illustrates ISEGAN and DSEGAN with the number of generators $N = 2$. When the number of generators $N = 1$, both the ISEGAN and the DSEGAN can be viewed as the conventional SEGAN (see Figure 2).

$$\min_D \mathcal{L}_{LS}(D) = \frac{1}{2}\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{zzdata}(\mathbf{x},\tilde{\mathbf{x}})}(D(\mathbf{x},\tilde{\mathbf{x}})-1)^2 + \sum_{n=1}^{N}\frac{1}{2N}\mathbb{E}_{\mathbf{z}_n\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}_{data}(\tilde{\mathbf{x}})}D(G_n(\mathbf{z}_n,\hat{\mathbf{x}}_{n-1}),\tilde{\mathbf{x}})^2, \tag{3}$$

$$\min_D \mathcal{L}_{LS}(\mathcal{G}) = \sum_{n=1}^{N}\frac{1}{2N}\mathbb{E}_{\mathbf{z}_n\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{data}(\tilde{\mathbf{x}})}(D(G_n(\mathbf{z}_n,\hat{\mathbf{x}}_{n-1}),\tilde{\mathbf{x}})-1)^2 + \sum_{n=1}^{N}\lambda_n||G_n(\mathbf{z}_n,\hat{\mathbf{x}}_{n-1})-\mathbf{x})||_1. \tag{4}$$

Like the conventional SEGAN, the chained generators $\mathcal{G}$ in both ISEGAN and DSEGAN are conditional-based GAN generators. Given an enhancement stage n, the generator $G_n \in \mathcal{G}$ is provided with the output $\hat{\mathbf{x}}_{n-1}$ of generator $G_{n-1} \in \mathcal{G}$ to produce an enhanced signal $\hat{\mathbf{x}}_n$ such that $\hat{\mathbf{x}}_n = G_n(\mathbf{z}_n, \hat{\mathbf{x}}_{n-1})$, where $1 \leq n \leq N$ and $\mathbf{z}_n$ is the latent representation. Hence, the corrupted raw audio signal is considered as $\hat{\mathbf{x}}_0 \equiv \tilde{\mathbf{x}}$ and the final enhanced signal of the last generator $G_N \in \mathcal{G}$ is $\hat{\mathbf{x}} \equiv \hat{\mathbf{x}}_N$. Here, the discriminator D is trained to classify the pair of signals $(\mathbf{x}, \tilde{\mathbf{x}})$ as real and $(\hat{\mathbf{x}}_1, \tilde{\mathbf{x}})$, $(\hat{\mathbf{x}}_2, \tilde{\mathbf{x}})$, ..., $(\hat{\mathbf{x}}_N, \tilde{\mathbf{x}})$ as fake pairs of signals. Equations (3) and (4) are the least-squares objective functions for training the discriminator D and the chained generators $\mathcal{G}$, respectively.

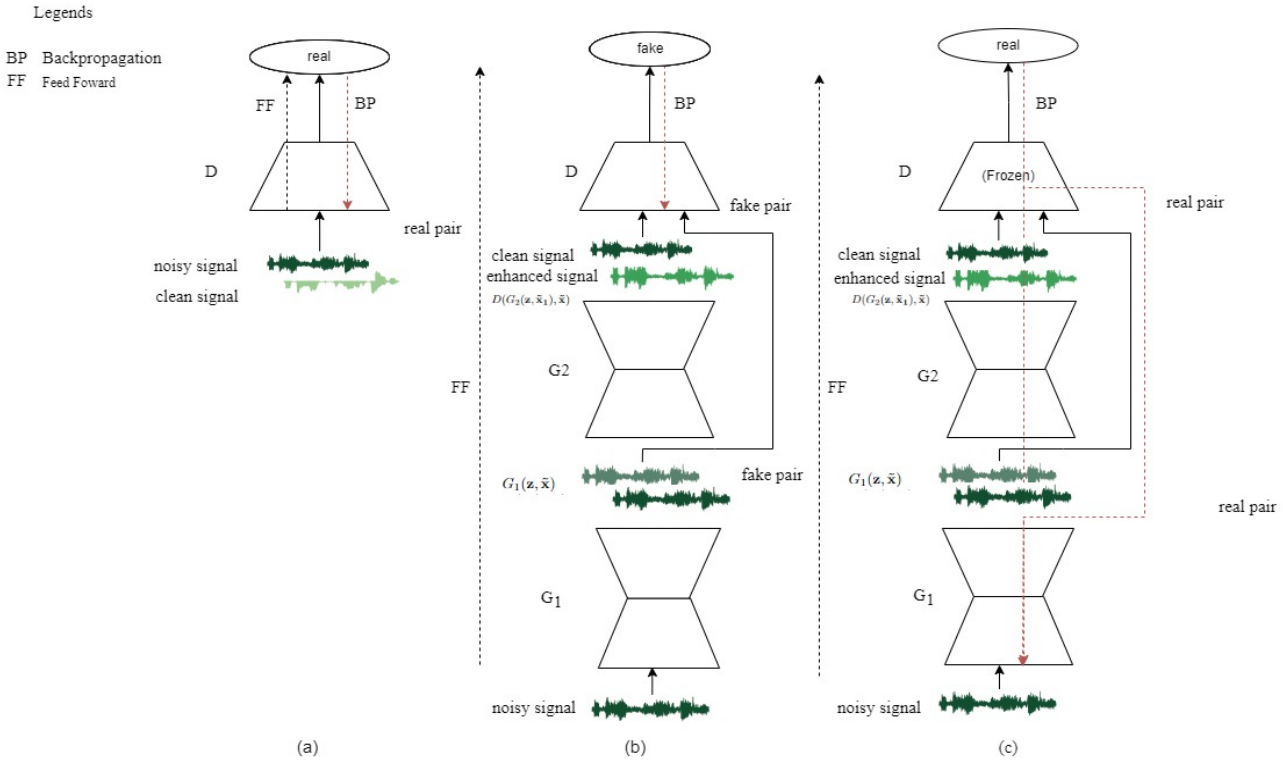


Figure 3. Illustration of the training process of a multi-generator speech enhancement GAN with two generators. In (**a**), the discriminator D is trained to classify $(\mathbf{x}, \tilde{\mathbf{x}})$ as real pairs, and in (**b**), the discriminator D is trained to classify the pairs $(\tilde{\mathbf{x}}_\mathbf{1}, \bar{\mathbf{x}}_\mathbf{1})$ and $(\tilde{\mathbf{x}}_\mathbf{2}, \bar{\mathbf{x}}_\mathbf{2})$ generated by G_1 and G_2, respectively, as fake. Then, in (**c**), the generators G_1 and G_2 are trained to fool the discriminator D with the pairs $(\tilde{\mathbf{x}}_\mathbf{1}, \bar{\mathbf{x}}_\mathbf{1})$ and $(\tilde{\mathbf{x}}_\mathbf{2}, \bar{\mathbf{x}}_\mathbf{2})$ as real.

3. Multi-Generator SEGAN with Self-Attention

3.1. Self-Attention Block

The concept of the attention mechanism was first introduced in Bahdanau et al. [44] to address the alignment and translation problem in sequence-to-sequence modelling. The idea behind attention is to enable the decoder in an encoder–decoder model to access all the encoded input by introducing attention weights, which focus on the positions containing relevant information for generating output tokens. Since its inception, the self-attention [34] variant has been adopted in speech and audio signal processing. Self-attention was first adopted in acoustics models [35], and later in GAN models [23] (i.e., SEGAN) for speech enhancement.

Here, we briefly describe the underlying framework of the self-attention block [23,37] (see Figure 4) that we adopt to investigate the impact of self-attention on the efficiency and quality of enhanced speech produced by SEGAN with multi-generators. Given an output $\mathbf{F} \in \mathbb{R}^{L \times C}$ (i.e., feature map) of a convolutional layer in SEGAN, where L is the time dimension and C is the number of channels, the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ ma-

trices are, respectively, obtained by the transformations presented as follows: $\mathbf{Q} = \mathbf{F}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{F}\mathbf{W}_K, \mathbf{V} = \mathbf{F}\mathbf{W}_V$, where $\mathbf{W}_Q \in \mathbb{R}^{C \times \frac{C}{k}}, \mathbf{W}_K \in \mathbb{R}^{C \times \frac{C}{k}}$, and $\mathbf{W}_V \in \mathbb{R}^{C \times \frac{C}{k}}$ are weights matrices obtained through 1×1 convolutional layer operation of $\frac{C}{k}$ filters, and k is a scalar used to reduce the channel dimension C of the feature space for memory reduction. To further improve memory efficiency, the time dimensions of $\mathbf{K}$ and $\mathbf{V}$ are reduced by a factor of p. Hence, the sizes of the matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ become $\mathbf{Q} \in \mathbb{R}^{L \times \frac{C}{k}}, \mathbf{K} \in \mathbb{R}^{\frac{L}{p} \times \frac{C}{k}}$, and $\mathbf{V} \in \mathbb{R}^{\frac{L}{p} \times \frac{C}{k}}$, respectively.

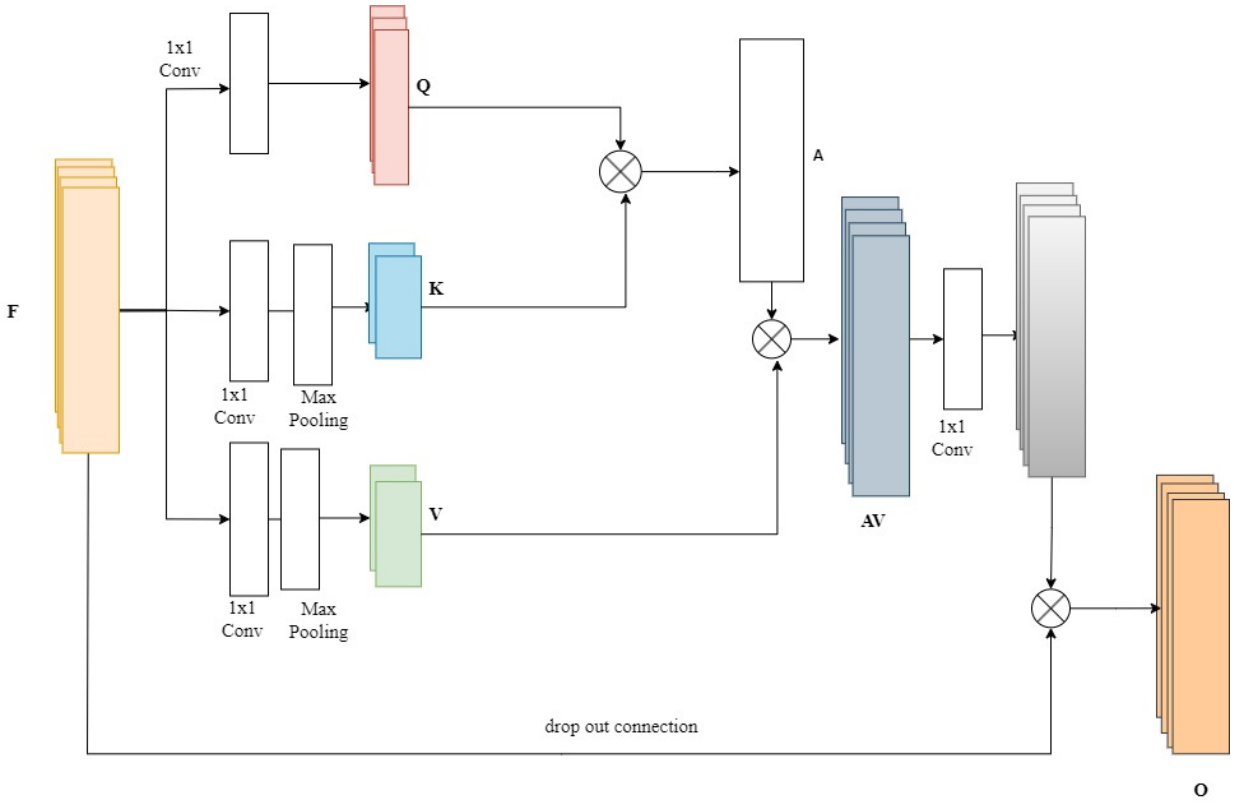


Figure 4. Illustration of the self-attention mechanism we integrated into the generators of the multi-generator speech enhancement GANs (i.e., ISEGAN and DSEGAN). F = input feature, Q = query, K = key, V = value, A = attention map, and O = attentive output.

The attention map $\mathbf{A}$ and attentive output $\mathbf{O}$ are then computed using these matrices. The attention map $\mathbf{A}$ is obtained by computing the softmax of the dot product of the query $\mathbf{Q}$ and the key $\mathbf{K}$ (see Equation (5)), and the attentive output $\mathbf{O}$ is also computed by performing a matrix multiplication of the attention map $\mathbf{A}$ and value matrix $\mathbf{V}$ (see Equation (6)).

$$\mathbf{A} = softmax(\mathbf{Q}\bar{\mathbf{K}}^{\mathsf{T}}), A \in \mathbb{R}^{L \times \frac{L}{p}}, \tag{5}$$

$$\mathbf{O} = (\mathbf{A}\mathbf{V})\mathbf{W}_O, \ \mathbf{W}_O \in \mathbb{R}^{\frac{C}{k} \times C}. \tag{6}$$

where $\mathbf{W}_O$ is a weight matrix of 1×1 convolutional layer of C filters applied to $\mathbf{A}\mathbf{V}$ to restore the size of the attentive output $\mathbf{O}$ to the original size $L \times C$. A shortcut connection is also used to facilitate information propagation.

3.2. ISEGAN-Self-Attention and DSEGAN-Self-Attention Networks

3.2.1. Multi-Generator $\mathcal{G}$

The N generators with self-attention of multi-generator $\mathcal{G} = G_1 \rightarrow G_2 \rightarrow \cdots \rightarrow G_N$, $N > 1$ for both the ISEGAN-Self-Attention and the DSEGAN-Self-Attention net-

works have the same architectural structure, which follows the network architecture in Phan et al. [22]. Figure 5 illustrates an example of ISEGAN-Self-Attention network architecture and DSEGAN-Self-Attention network architecture of a multi-generator $\mathcal{G} = G_1 \rightarrow G_2$.

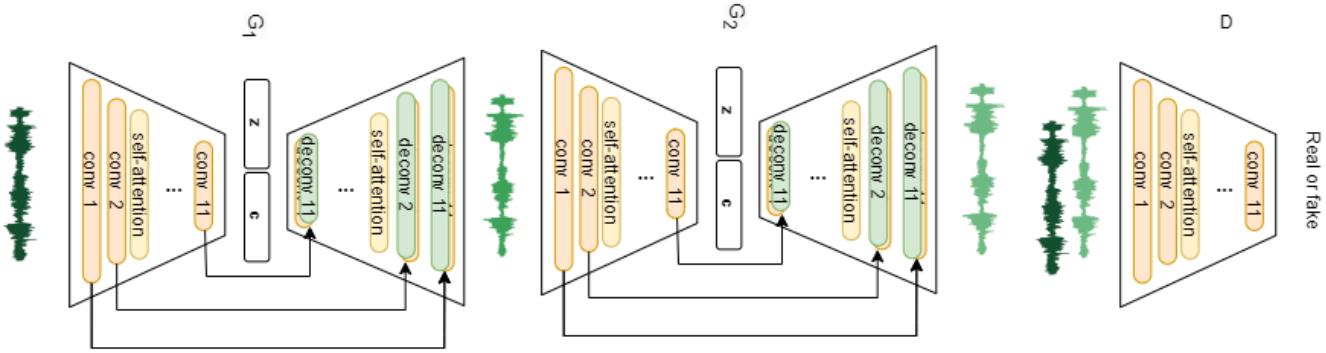


Figure 5. Illustration of the ISEGAN- and DSEGAN-Self-Attention architecture with an encoder of two generators, G_1, G_2, featuring the (de)convolutional layers integrated with self-attention blocks, and a discriminator D.

Each generator G_n, $1 < n \leq N$ in the multi-generator $\mathcal{G}$ network consists of an encoder–decoder architecture with fully convolutional layers, following the implementation in [16,22,23]. The encoder for each generator G_n is composed of 11 one-dimensional stride convolutional layers with a common filter width of 31 and a stride length of 2. This is followed by parametric rectified linear units (PReLUs) [45]. The number of filters used is tailored to suit the requirements of the task at hand as it is increased progressively from 16 to 1024 in the set of {16, 32, 32, 64, 64, 128, 128, 512, 1024}, resulting in feature maps of varying sizes as 8192×16, 4096×32, 2048×32, 1024×64, and so on until 8×1024. The last feature map $\mathbf{c} \in \mathbb{R}^{8\times1024}$ of the decoder is then concatenated with a noise $\mathbf{z} \in \mathbb{R}^{8\times1024}$, which is sampled from the normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, and then provides it to the decoder of generator G_n. The decoder of generator G_n uses deconvolutions to reverse the encoding process by employing the mirror of the encoder architecture as indicated in Figure 5. To facilitate the reconstruction of the waveform from the encoded features, skip connections are used to connect each encoding layer to its corresponding decoding layer. This allows the information from the waveform to bypass the encoding stage and flow directly into the decoding stage [16]. This helps to preserve the details of the waveform and maintain the quality of the reconstructed output. Using skip connections this way, the encoder-decoder architecture can also be more effective in processing waveforms and producing high-quality enhanced speech signals.

The self-attention block presented in Section 3.1 is integrated into each generator G_n of both the ISEGAN-Self-Attention network architecture and the DSEGAN-Self-Attention network architecture. The self-attention block is coupled with the convolutional and deconvolutional layers in the generators $G_1, G_2, \ldots, G_N$ of ISEGAN and DSEGAN to construct an ISEGAN-Self-Attention and DSEGAN-Self-Attention, respectively. SASEGAN [23] indicated that it does not matter the number of self-attention blocks placed in the network since it does improve the performance with a little additional memory. Based on this, we added the self-attention block in each generator G_n at the following (de)convolution layers $l = 4, 6, 10$ positions, which were the most effective positions in the SASEGAN. As in the ISEGAN and the DSEGAN setups, the chained generators in the ISEGAN-Self-Attention share the same parameters, and in the DSEGAN-Self-Attention, the generators' parameters are independent. Thus, the ISEGAN-Self-Attention has a smaller memory footprint. However, the parameters of each generator learn independently in the DSEGAN-Self-Attention, allowing each generator to contribute to a specific aspect of the speech enhancement process.

3.2.2. Discriminator D

Following Phan et al. [22], we use the same discriminator architecture for the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks. Remember that both the ISEGAN-Self-Attention network architecture and the DSEGAN-Self-Attention network architecture have multiple generators $G_1, G_2, \ldots, G_N$ but with a single discriminator D (see Figure 5). The architecture of the discriminator D is akin to the encoder part of the generators, except that D accepts a two-channel input. As adopted in previous works [22,23], we also incorporate a virtual batch-normalization [46] prior to the leaky RELU activation [47] with the hyperparameter $\alpha = 0.3$. The self-attention blocks are placed in the same position as in the encoder of the generators. We add a 1D convolutional layer with a filter size of one to the discriminator D to reduce the size of the output of the last convolutional layer from 8×1024 to 8 before a classification is performed with a softmax layer.

4. Experiments

4.1. Baseline and Objective Evaluation

To have a comparative baseline reference, the ISEGAN and the DSEGAN networks [22] are used as the main baseline to compare the performance of the proposed ISEGAN-Self-Attention and DSEGAN-Self-Attention networks. Additionally, the traditional speech enhancement network, SEGAN [16], and the state-of-the-art method, the SASEGAN [23], are selected as baseline networks as well. With the implementations of the self-attention blocks in the proposed ISEGAN-Self-Attention and DSEGAN-Self-Attention networks, we consider the SASEGAN as one of the best gauges to observe how well the self-attention performs in multi-generator settings as compared to a single generator. Also, as adopted in SEGAN, we compare the performance to the noisy signals and to signals filtered via the Wiener method based on a priori signal-to-noise estimation [48]. The objective evaluation metrics used in the selected baselines are adopted to evaluate the quality of the generated enhanced speech signals. These metrics include the five objective signal-quality metrics (PESQ, CSIG, CBAK, COVL, and SSNR) suggested by Loizou [4], as well as the speech intelligibility metric (STOI) introduced by Taal et al. [49]. The five objective signal-quality metrics are computed using the implementation in SEGAN [16]. This set of metrics focuses on assessing various aspects of speech signal quality. Specifically, PESQ and STOI measure perceived speech quality and speech intelligibility, respectively. SSNR and CBAK evaluate the distortion caused by background noise, while COVL provides an overall quality assessment of the speech signal. Additionally, SSNR compares the enhanced speech signal with the original clean speech on a segment-by-segment basis, offering insights into the preservation of speech in various segments [4,49].

4.2. Dataset

The dataset used in the experiments is the noisy dataset for the speech enhancement task, which was put together by Valentini-Botinhao et al. [50] to facilitate the comparison of speech enhancement approaches. It is made of 30 speakers, recorded at 48 kHz, from the Voice Bank corpus [51]. We sub-sampled all the speeches from 48 kHz to 16 kHz. This dataset is used because it offers a variety of learning features for the optimal enhancement of speech signals combined with different noise conditions, and it was adopted to evaluate the baseline methods mentioned in Section 4.1. Following the baseline methods [16,22,23], the noisy training set comprised 40 different conditions which were made by merging 10 types of noise, that is, 2 artificial and 8 obtained from the Demand database [52], at signal-to-noise-ratios (SNRs) of 15, 10, 5, and 0 dB to obtain the noisy conditions in the data. Similarly, for the test set, 20 different conditions were made, merging 5 types of noise obtained from the Demand database with 4 SNRs each at 17.5, 12.5, 7.5, and 2.5 dB.

4.3. Experimental Settings

This work aims at improving the performance of the existing multi-generator SEGAN networks (i.e., ISEGAN and DSEGAN) with a self-attention mechanism. We conduct several

experiments on the ISEGAN and DSEGAN networks with and without the self-attention block illustrated in Figure 4 using the training setup presented in Figure 3. We set three experimental objectives:

- To quantitatively show the effects of a self-attention mechanism in the generators of a multi-generator SEGAN.
- To qualitatively show the effects of a self-attention mechanism in the generation of a multi-generator SEGAN.
- To investigate the overall performance of the model in terms of training the parameters and the generation of the enhanced speech.

We set up the experiments with the number of generators $N = \{2, 3, 4\}$ and with only one discriminator in the ISEGAN network with and without self-attention, and in the DSEGAN network with and without self-attention. The TensorFlow deep learning framework [53] was used to implement the networks, and all the experiments were performed on an 8 GB GPU GeForce RTX 2080 machine. Following previous works [22,23], each network was trained for 100 epochs using the RMSprop optimization technique [54] with a mini-batch size of 64, and a learning rate set to 2×10^{-4}. We sampled raw speech segments sampled at 16 kHz after every epoch to monitor performance during training of the different network setups for $N = \{2, 3, 4\}$ generators to investigate the effects of self-attention in the multi-generator SEGANs for speech denoising and for generating synthetic speech signals. To further understand the impact of the self-attention mechanism in a specific generator of the multi-generator SEGAN setup, we performed an ablation study by placing the self-attention block in one generator at a time and at different convolutional layers of the generator.

5. Results and Discussion

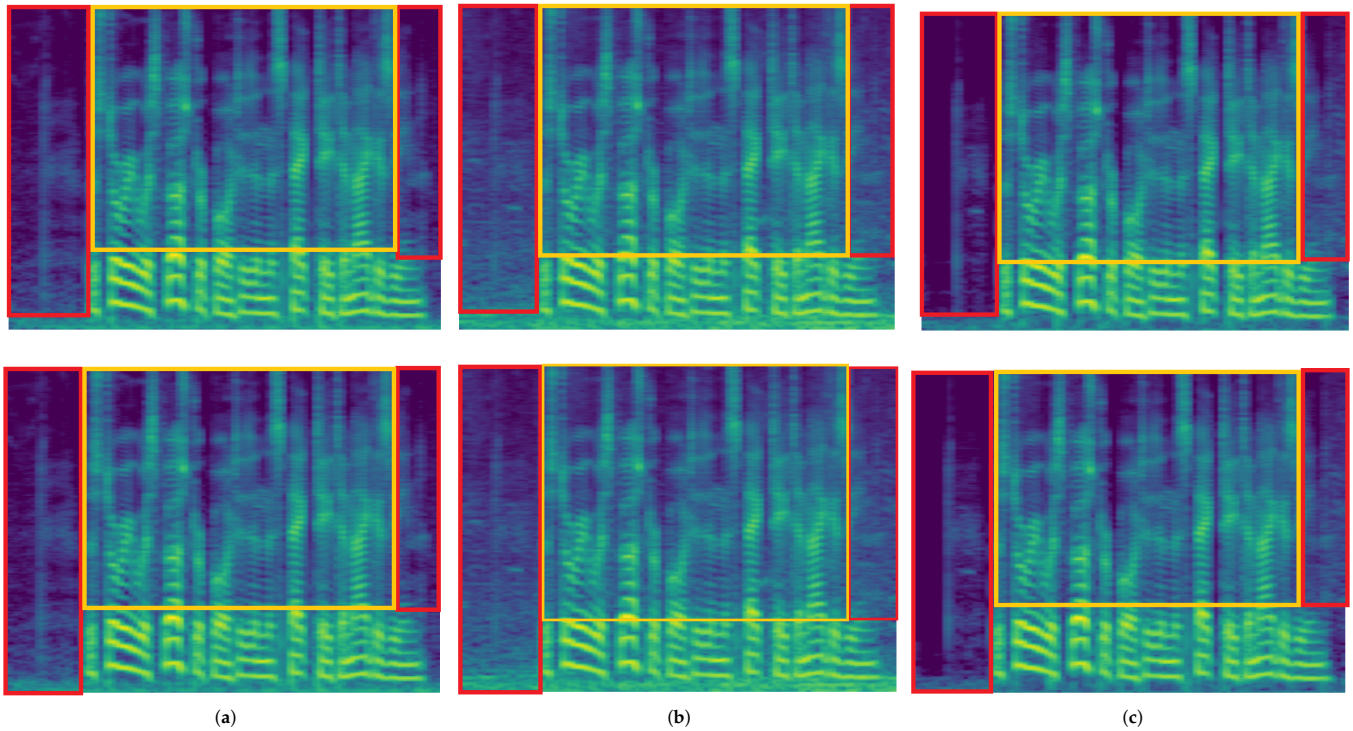
This section presents the experimental results and ablation study on the effectiveness of self-attention in the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks. The results are reported quantitatively (see Tables 1 and 2) and through spectrograms (see Figure 6) to show the quality of the enhanced speech signals produced by the networks. The quantitative results presented in Tables 1 and 2 are based on the averaged objective evaluations of the ISEGAN-Self-Attention and the DSEGAN-Self-Attention, respectively, on the 824 wave files in the test set of the noisy dataset introduced by Valentini-Botinhao et al. [50]. In Table 1, we compare the results of the ISEGAN-Self-Attention with the attention blocks at the fourth and fifth convolutional layers of the encoder and the decoder to the baseline results reported in Phan et al. [22,23] and, likewise, the same for the results of the DSEGAN-Self-Attenation network presented in Table 2.

Table 1. Performance comparison between ISEGAN-Self-Attention and the baseline methods based on the objective evaluation metrics. The highest score per metric is highlighted in **bold**, and the second best is underlined.

Metric	Noisy	Weiner	SEGAN	SASEGAN	ISEGAN			ISEGAN-Self-Attention		
					$N = 2$	$N = 3$	$N = 4$	$N = 2$	$N = 3$	$N = 4$
PESQ	1.97	2.22	2.19	2.34	2.24	2.19	2.21	**2.66**	2.58	2.59
CSIG	3.35	3.23	3.39	**3.52**	3.23	2.96	3.00	3.40	3.35	3.38
CBAK	2.44	2.68	2.90	3.04	2.95	2.88	2.92	3.15	3.09	**3.18**
COVL	2.63	2.67	2.76	2.91	2.69	2.52	2.55	**3.04**	2.97	3.01
SSNR	1.68	5.07	7.36	8.05	8.17	8.11	8.86	**9.04**	8.90	**9.04**
STOI	92.10	-	93.12	93.32	93.29	**93.35**	93.29	93.30	**93.35**	93.32

Table 2. Performance comparison between DSEGAN-Self-Attention and the baseline methods based on the objective evaluation metrics. The highest score per metric is highlighted in **bold** and the second best is underlined.

Metric	Noisy	Weiner	SEGAN	SASEGAN	DSEGAN			DSEGAN-Self-Attention		
					$N = 2$	$N = 3$	$N = 4$	$N = 2$	$N = 3$	$N = 4$
PESQ	1.97	2.22	2.19	2.34	2.35	2.39	2.37	**2.71**	2.64	2.67
CSIG	3.35	3.23	3.39	3.52	3.55	3.46	3.50	**3.58**	3.37	3.54
CBAK	2.44	2.68	2.90	3.04	3.10	3.11	3.10	**3.15**	2.98	3.06
COVL	2.63	2.67	2.76	2.91	2.93	2.90	2.92	**3.11**	2.94	3.07
SSNR	1.68	5.07	7.36	8.05	8.70	8.72	8.59	**9.19**	9.01	9.08
STOI	92.10	-	93.12	93.32	93.25	93.28	**93.49**	93.48	93.30	93.36



(**a**) (**b**) (**c**)

Figure 6. Spectrograms of a speech signal generated using Librosa [55]: (**a**) the target clean speech signals, (**b**) the noisy speech signals, and (**c**) the enhanced speech signals with the different networks. The enhanced speech signal on the top row is produced by the ISEGAN-Self-Attention and the one on the bottom row is produced by DSEGAN-Self-Attention. The regions marked red are regions with high noise levels with little or no intelligible speech signal, and the regions marked with yellow rectangles contain intelligible speech signals with little additive noise. It can be seen in (**c**) that the networks are able to remove additive noise in the red rectangles as well as in the regions marked with yellow rectangles. The spectrograms in (**c**) show that the speech signals are preserved while the networks remove the noise.

As expected, the objective evaluations of both the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks showed an improvement over the ISEGAN and the DSEGAN without self-attention in all the speech quality metrics. For example, when $N = 2$, the ISEGAN-Self-Attention network achieved average improvements of 18.75%, 6.78%, and 10.64% in PESQ, COVL, and SSNR, respectively, over ISEGAN without self-attention; with $N = 4$, average improvements of 8.90% and 0.03% in CBAK and SSNR, respectively, were also achieved (see Table 1). The DSEGAN-Self-Attention also gained average improvements of 15.31%, 0.84%, 1.61%, 6.14%, and 5.63% in PESQ, CSIG, CBAK,

COVL, and SSNR, respectively, over DSEGAN without self-attention (see Table 2). However, in terms of speech intelligibility (STOI), on average, no significant improvement was achieved in both SEGAN-Self-Attention and DSEGAN-Self-Attention over ISEGAN and DSEGAN, respectively.

Furthermore, the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks improved over other baseline methods, including the SEGAN [23] and the SASEGAN [22]. For the ISEGAN-Self-Attention, with $N = 2$, we observed average improvements of 17.57%, 6.12%, and 17.56% in PESQ, COVL, and SSNR, respectively, over SEGAN and SASEGAN. And when $N = 4$, average improvements of 7.13% and 17.56% in CBAK and SSNR, respectively, were achieved over SEGAN and SASEGAN as well. Similarly, in Table 2, with $N = 2$, the DSEGAN-Self-Attention improved over the SASEGAN averagely by 19.77%, 5.79%, 6.12%, 9.78%, and 19.51% in PESQ, CSIG, CBAK, COVL, and SSNR, respectively. In terms of speech intelligibility (STOI), both ISEGAN-Self-Attention and DSEGAN-Self-Attention networks outperformed SEGAN and SASEGAN. For the various multi-generator $\mathcal{G}$ setups, when $N = 2$, the ISEGAN-Self-Attention and the DSEGAN-Self-Attention on average achieved better results in all the five objective signal-quality metrics (PESQ, CSIG, CBAK, COVL, and SSNR) compared to when $N = 3$ and $N = 4$. In the case of STOI, the ISEGAN-Self-Attention improved the enhancement performance when $N = 3$, whereas DSEGAN-Self-Attention competed with the baseline methods. Similarly, when $N = 4$, the ISEGAN-Self-Attention performed better than the DSEGAN-Self-Attention in CBAK and SSNR signal-quality metrics. Overall, the results suggest that $N = 2$ is the best setup for both ISEGAN-Self-Attention and DSEGAN-Self-Attention for enhancing noisy speech signals. Figure 6 presents the spectrograms to visualise the quality and intelligibility of the enhanced speech signals in the $N = 2$ setup of ISEGAN-Self-Attention and DSEGAN-Self-Attention networks. From the spectrograms, we can see that the networks have the capability to remove the additive noise from the speech signal. This suggests that by leveraging the self-attention mechanism, the networks are able to learn temporal dependency in an input sequence of signals.

To demonstrate that the self-attention mechanism is capable to learn temporal dependencies in a speech signal, we compare the clean speech signal's amplitude at time t with the pattern of speech features in the enhanced speech's spectrograms at the same time t. We investigate whether the speech features are preserved within the time intervals to determine that the self-attention mechanism captured the temporal dependencies in the speech signal. We expect that whenever there are high frequencies, there should be a high pitch in the audio signals and vice versa. In Figure 7, we can observe the patterns of speech features, which are marked with red-, black-, and yellow-coloured rectangles, demonstrating the preservation of the temporal dependencies in the enhanced speech. The analysis in Figure 7 shows that the integration of self-attention into ISEGAN and DSEGAN helps improve the enhancement of speech signal in terms of signal distortion and speech distortion. The perceived speech quality is preserved, which indicates that the speech components from the inputs are preserved over time. Whereas speech intelligibility is difficult to improve by most speech enhancement algorithms [56], the self-attention mechanism maintains a decent score. The results indicate that the self-attention mechanism significantly improves the temporal dependencies, more so than the general improvement in the generation of the enhanced speech in both ISEGAN-Self-Attention and DSEGAN-Self-Attention as the speech signal progresses with time.

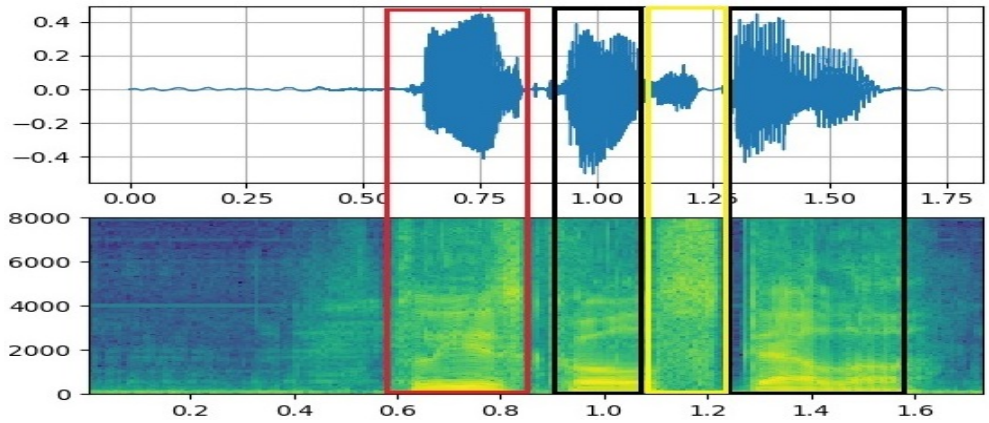


Figure 7. A comparison of the input speech signal, clean speech, and the spectrogram of the enhanced speech. The speech features in the spectrograms are marked with red-, black-, and yellow-coloured rectangles for the corresponding wave signal to demonstrate the preservation of the temporal dependence in the enhanced speech signal.

Ablation Study

To understand the effect of the self-attention blocks in the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks, we ablate the generators in each network to observe the contribution of the self-attention in multi-generator speech enhancement GAN in generating enhanced speech signals. In this study, our focus was to find out which generator with the self-attention blocks in the multi-generator of the networks has the most contribution to the networks' performance. For instance, if the self-attention blocks are only placed in the first generator, how does it contribute to enhancing the speech signals? Here, we only considered the setup for $N = 2$ for both the ISEGAN-Self-Attention and the DSEGAN-Self-Attention. This choice was made on the basis that the setup for $N = 2$ produced the best results, in both quality and intelligibility of the generated signals, among all the experiments we performed. With $N = 2$, we had 4 different cases to investigate (i.e., 2 cases for each network) by removing all the self-attention blocks from one of the two generators in the network to observe the contributions of the remaining self-attention blocks. For example, in the setup for case 1 in the ISEGAN-Self-Attention network, we removed all the self-attention blocks in the second generator and trained the network with the same training settings mentioned in Section 4.

In Table 3, we present the results of the study and compared them with ISEGAN and DSEGAN results. We observed that, in both ISEGAN-Self-Attention and DSEGAN-Self-Attention, having the self-attention blocks in only the first generator (i.e., G_1) produces better-enhanced speech signals than when we place the self-attention blocks only in the second generator (i.e., G_2). This result indicates that the second generator G_2 is performing refining on the predecessor generator G_1; therefore, the first generator G_1 is responsible for capturing the core features and generating almost-enhanced samples which the second generator G_2 just needs to refine. Also, we observed that placing the self-attention blocks in only one generator of the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks did not show much improvement over ISEGAN and DSEGAN and, in most cases, trailed behind ISEGAN and DSEGAN, respectively. To achieve the best performance, the self-attention blocks might be integrated with more than one generator of the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks.

Table 3. The results of the ablation study on $N = 2$ (i.e., 2 generators) setup of the ISEGAN-Self-Attention and the DSEGAN-Self-Attention networks. The results are compared with ISEGAN and DSEGAN. The **boldface** is the setup with the highest score per an objective evaluation metric.

Metric	DSEGAN	ISEGAN	ISEGAN-Self-Attention		DSEGAN-Self-Attention	
			G_1	G_2	G_1	G_2
PESQ	2.71	2.66	**2.63**	2.57	**2.68**	2.64
CSIG	3.58	3.58	**3.51**	3.49	**3.52**	3.50
CBAK	3.15	3.15	**3.09**	3.08	**3.11**	3.08
COVL	3.11	3.04	**3.07**	3.04	**3.09**	3.05
SSNR	9.19	9.04	**9.11**	9.03	**9.09**	9.02
STOI	93.29	93.25	**93.38**	93.32	**93.42**	93.36

6. Conclusions

This paper introduced a self-attention block into multi-generator speech enhancement GANs (ISEGAN and DSEGAN) to improve their temporal dependency capability for enhancing speech signals. The self-attention block can be integrated at different convolutional layers of the multiple generators of the ISEGAN and the DSEGAN networks or given sufficient memory, it can be integrated at all the convolutional layers of the SEGAN's and the DSEGAN's multiple generators. Our experiments demonstrate that integrating a self-attention mechanism into ISEGAN and DSEGAN networks, which we respectively called ISEGAN-Self-Attention and DSEGAN-Self-Attention networks, can significantly improve the performance of the speech enhancement system. The experimental results show that the ISEGAN-Self-Attention and the DSEGAN-Self-Attention significantly improve the enhancement performance of ISEGAN and DSEGAN. In addition, the ISEGAN-Self-Attention and the DSEGAN-Self-Attention outperformed the SEGAN and the SASEGAN baselines in all the objective evaluation metrics. Furthermore, we observe that the use of self-attention helps bridge the performance gap between the ISEGAN and the DSEGAN, demonstrating the effectiveness of the self-attention mechanism for both the shared-parameter multi-generator speech enhancement GAN (ISEGAN) and the non-shared-parameter multi-generator speech enhancement GAN (DSEGAN). Overall, our findings highlight the potential benefits of self-attention as a valuable technique for enhancing the performance of multi-generator speech enhancement systems.

Author Contributions: Conceptualization, B.K.A.A. and C.B.-B.; methodology, B.K.A.A.; investigation, B.K.A.A.; writing—original draft preparation, B.K.A.A. and C.B.-B.; writing—review and editing, H.I. and C.B.-B.; supervision, H.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The work of the first author was supported by the Makiguchi Foundation under the Makiguchi Scholarship for International Students.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SEGAN	Speech Enhancement Generative Adversarial Networks
SASEGAN	Self-Attention Speech Enhancement Generative Adversarial Network
ISEGAN	Iterated Speech Enhancement Generative Adversarial Network
DSEGAN	Deep Speech Enhancement Generative Adversarial Network
PESQ	Perceptual Evaluation of Speech Quality
STOI	Short-Time Objective Intelligibility
CBAK	Composite MOS Predictor for Background-Noise Intrusiveness
SSNR	Segmental Signal-to-Noise Ratio
CSIG	Composite Measure of Signal-to-Distortion Ratio
COVL	Composite MOS Predictor of Overall Signal Quality

References

1. Donahue, C.; Li, B.; Prabhavalkar, R. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5024–5028.
2. Fedorov, I.; Stamenovic, M.; Jensen, C.; Yang, L.C.; Mandell, A.; Gan, Y.; Mattina, M.; Whatmough, P.N. TinyLSTMs: Efficient Neural Speech Enhancement for Hearing Aids. *arXiv* **2020**, arXiv:2005.11138.
3. Gold, B.; Morgan, N. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 1st ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1999.
4. Loizou, P.C. *Speech Enhancement: Theory and Practice*, 2nd ed.; CRC Press, Inc.: Boca Raton, FL, USA, 2013.
5. Kolmogorov, A. Interpolation and extrapolation of stationary random sequences. *Izv. Acad. Sci. Ussr* **1941**, *5*, 3–14.
6. Wiener, N. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*; MIT Press: Cambridge, MA, USA, 1964.
7. Bhangale, K.B.; Kothandaraman, M. Survey of Deep Learning Paradigms for Speech Processing. *Wirel. Pers. Commun.* **2022**, *125*, 1913–1949. [CrossRef]
8. Bulut, A.E.; Koishida, K. Low-Latency Single Channel Speech Enhancement Using U-Net Convolutional Neural Networks. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6214–6218.
9. Defossez, A.; Synnaeve, G.; Adi, Y. Real Time Speech Enhancement in the Waveform Domain. *arXiv* **2020**, arXiv:2006.12847.
10. Weninger, F.; Erdogan, H.; Watanabe, S.; Vincent, E.; Roux, J.L.; Hershey, J.R.; Schuller, B. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation, Liberec, Czech Republic, 25–28 August 2015; pp. 91–99.
11. Hu, Y.; Liu, Y.; Lv, S.; Xing, M.; Zhang, S.; Fu, Y.; Wu, J.; Zhang, B.; Xie, L. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. *arXiv* **2020**, arXiv:2008.00264.
12. Saleem, N.; Khattak, M.I.; Al-Hasan, M.; Jan, A. Multi-objective long-short term memory recurrent neural networks for speech enhancement. *J. Ambient Intell. Humaniz. Comput.* **2020**, *12*, 9037–9052. [CrossRef]
13. Leglaive, S.; Girin, L.; Horaud, R. A variance modeling framework based on variational autoencoders for speech enhancement. In Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), Aalborg, Denmark, 17–20 September 2018; pp. 1–6.
14. Sadeghi, M.; Leglaive, S.; Alameda-Pineda, X.; Girin, L.; Horaud, R. Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1788–1800. [CrossRef]
15. Fang, H.; Carbajal, G.; Wermter, S.; Gerkmann, T. Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 676–680.
16. Pascual, S.; Bonafonte, A.; Serrà, J. SEGAN: Speech Enhancement Generative Adversarial Network. *arXiv* **2017**, arXiv:1703.09452.
17. Baby, D.; Verhulst, S. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 106–110.
18. Wali, A.; Alamgir, Z.; Karim, S.; Fawaz, A.; Ali, M.B.; Adan, M.; Mujtaba, M. Generative adversarial networks for speech processing: A review. *Comput. Speech Lang.* **2022**, *72*, 101308. [CrossRef]
19. Wang, P.; Tan, K.; Wang, D.L. Bridging the Gap Between Monaural Speech Enhancement and Recognition With Distortion-Independent Acoustic Modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 39–48. [CrossRef]
20. Li, L.; Kang, Y.; Shi, Y.; Kürzinger, L.; Watzel, T.; Rigoll, G. Adversarial joint training with self-attention mechanism for robust end-to-end speech recognition. *EURASIP J. Audio Speech Music. Process.* **2021**, *2021*, 26. [CrossRef]

21. Feng, T.; Li, Y.; Zhang, P.; Li, S.; Wang, F. Noise Classification Speech Enhancement Generative Adversarial Network. In Proceedings of the 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 4–6 March 2022; pp. 11–16.
22. Phan, H.; McLoughlin, I.V.; Pham, L.; Chén, O.Y.; Koch, P.; De Vos, M.; Mertins, A. Improving GANs for speech enhancement. *IEEE Signal Process. Lett.* **2020**, *27*, 1700–1704. [CrossRef]
23. Phan, H.; Le Nguyen, H.; Chén, O.Y.; Koch, P.; Duong, N.Q.; McLoughlin, I.; Mertins, A. Self-attention generative adversarial network for speech enhancement. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7103–7107.
24. Fu, S.W.; Liao, C.F.; Tsao, Y.; Lin, S.D. MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019.
25. Zhang, Z.; Deng, C.; Shen, Y.; Williamson, D.S.; Sha, Y.; Zhang, Y.; Song, H.; Li, X. On Loss Functions and Recurrency Training for GAN-based Speech Enhancement Systems. *arXiv* **2020**, arXiv:2007.14974.
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
27. Quan, T.M.; Nguyen-Duc, T.; Jeong, W.K. Compressed Sensing MRI Reconstruction Using a Generative Adversarial Network With a Cyclic Loss. *IEEE Trans. Med. Imaging* **2018**, *37*, 1488–1497. [CrossRef] [PubMed]
28. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
29. Su, J.; Jin, Z.; Finkelstein, A. HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. *arXiv* **2020**, arXiv:2006.05694.
30. Su, J.; Jin, Z.; Finkelstein, A. HiFi-GAN-2: Studio-Quality Speech Enhancement via Generative Adversarial Networks Conditioned on Acoustic Features. In Proceedings of the 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 17–20 October 2021; pp. 166–170.
31. Kumar, K.; Kumar, R.; De Boissiere, T.; Gestin, L.; Teoh, W.Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; Courville, A.C. Melgan: Generative adversarial networks for conditional waveform synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
32. Pham, N.Q.; Nguyen, T.S.; Niehues, J.; Müller, M.; Waibel, A.H. Very Deep Self-Attention Networks for End-to-End Speech Recognition. *arXiv* **2019**, arXiv:1904.1337.
33. Yu, G.; Wang, Y.; Zheng, C.; Wang, H.; Zhang, Q. CycleGAN-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 523–529.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
35. Sperber, M.; Niehues, J.; Neubig, G.; Stüker, S.; Waibel, A. Self-Attentional Acoustic Models. *arXiv* **2018**, arXiv:1803.09519.
36. Tian, Z.; Yi, J.; Tao, J.; Bai, Y.; Wen, Z. Self-Attention Transducers for End-to-End Speech Recognition. *arXiv* **2019**, arXiv:1909.13037.
37. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning (PMLR), Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
38. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
39. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
40. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
41. Li, L.; Wudamu; Kürzinger, L.; Watzel, T.; Rigoll, G. Lightweight End-to-End Speech Enhancement Generative Adversarial Network Using Sinc Convolutions. *Appl. Sci.* **2021**, *11*, 7564. [CrossRef]
42. Sarfjoo, S.S.; Wang, X.; Henter, G.E.; Lorenzo-Trueba, J.; Takaki, S.; Yamagishi, J. Transformation of low-quality device-recorded speech to high-quality speech using improved SEGAN model. *arXiv* **2019**, arXiv:1911.03952.
43. Sakuma, M.; Sugiura, Y.; Shimamura, T. Improvement of Noise Suppression Performance of SEGAN by Sparse Latent Vectors. In Proceedings of the 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Taipei, Taiwan, 3–6 December 2019; pp. 1–2.
44. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
46. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
47. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. *Proc. ICML* **2013**, *30*, 3.

48. Scalart, P.; Filho, J. Speech enhancement based on a priori signal to noise estimation. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; Volume 2, pp. 629–632.
49. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
50. Valentini-Botinhao, C.; Wang, X.; Takaki, S.; Yamagishi, J. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. In Proceedings of the 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13–15 September 2016; pp. 146–152.
51. Veaux, C.; Yamagishi, J.; King, S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In Proceedings of the 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Gurgaon, India, 25–27 November 2013; pp. 1–4.
52. Thiemann, J.; Ito, N.; Vincent, E. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In Proceedings of the Meetings on Acoustics ICA2013, Montreal, QC, Canada, 2–7 June 2013; Volume 19, p. 035081.
53. Abadi, M. TensorFlow: Learning functions at scale. In Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming, Nara, Japan, 18–24 September 2016; p. 1.
54. Tieleman, T.; Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* **2012**, *4*, 26–31.
55. McFee, B.; McVicar, M.; Faronbi, D.; Roman, I.; Gover, M.; Balke, S.; Seyfarth, S.; Malek, A.; Raffel, C.; Lostanlen, V.; et al. librosa/librosa: 0.10.0.post2. 2023. Available online: https://zenodo.org/record/7746972 (accessed on 21 May 2023).
56. Loizou, P.C.; Kim, G. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 47–56. [CrossRef] [PubMed]