electronics

MDPI

Article

Light-Weight Self-Attention Augmented Generative Adversarial Networks for Speech Enhancement

Lujun Li *, Zhenxing Lu, Tobias Watzel, Ludwig Kürzinger and Gerhard Rigoll

Department of Electrical and Computer Engineering, Technical University of Munich, 80333 Munich, Bavaria, Germany; zhenxing.lu@tum.de (Z.L.); tobias.watzel@tum.de (T.W.); ludwig.kuerzinger@tum.de (L.K.); rigoll@tum.de (G.R.)

* Correspondence: lujun.li@tum.de

Abstract: Generative adversarial networks (GANs) have shown their superiority for speech enhancement. Nevertheless, most previous attempts had convolutional layers as the backbone, which may obscure long-range dependencies across an input sequence due to the convolution operator's local receptive field. One popular solution is substituting recurrent neural networks (RNNs) for convolutional neural networks, but RNNs are computationally inefficient, caused by the unparallelization of their temporal iterations. To circumvent this limitation, we propose an end-to-end system for speech enhancement by applying the self-attention mechanism to GANs. We aim to achieve a system that is flexible in modeling both long-range and local interactions and can be computationally efficient at the same time. Our work is implemented in three phases: firstly, we apply the stand-alone self-attention layer in speech enhancement GANs. Secondly, we employ locality modeling on the stand-alone self-attention layer. Lastly, we investigate the functionality of the self-attention augmented convolutional speech enhancement GANs. Systematic experiment results indicate that equipped with the stand-alone self-attention layer, the system outperforms baseline systems across classic evaluation criteria with up to 95% fewer parameters. Moreover, locality modeling can be a parameter-free approach for further performance improvement, and self-attention augmentation also overtakes all baseline systems with acceptably increased parameters.

Keywords: generative adversarial networks; self-attention mechanism; speech enhancement

check for updates

Citation: Li, L.; Lu, Z.; Watzel, T.; Kürzinger, L.; Rigoll, G. Light-Weight Self-Attention Augmented Generative Adversarial Networks for Speech Enhancement. *Electronics* **2021**, *10*, 1586. https://doi.org/10.3390/electronics10131586

Academic Editor: Valentina E. Balas

Received: 15 June 2021
Accepted: 29 June 2021
Published: 30 June 2021

1. Introduction

Speech enhancement aims to improve speech intelligibility and quality in adverse environments by transforming the interfered speech to its original clean version [1]. Speech enhancement can serve as a front end for downstream speech-related tasks, e.g., speech recognition [2], speaker identification [3], speech emotion recognition [4], etc. In addition, it is also applied successfully in communication systems, e.g., hearing aids [5] and cochlear implants [6]. Classic speech enhancement methods are the Wiener filter [7], time-frequency masking [8], signal approximation [9], spectral mapping [10], etc. Recently, significant improvements in speech enhancement performance have been reported for discriminative deep learning algorithms, e.g., deep neural networks (DNNs) [11], convolutional neural networks (CNNs) [12], and recurrent neural networks (RNNs) [2].

Additionally, generative neural networks (GANs) [13] have been demonstrated to be efficient for speech enhancement [14–19], where the generative training results in fewer artifacts than discriminative models. Conforming to a GAN's principle, the generator G is designated for learning an enhancement mapping that can imitate the clean data distribution to generate enhanced samples. In contrast, the discriminator D plays the role of a classifier that distinguishes the real sample, coming from the dataset that G is imitating, from the fake sample made up by G. Simultaneously, D guides the parameter updating of G towards the distribution of clean speech signals. Nevertheless, most previous attempts had convolutional layers as the backbone, limiting the network's ability in capturing long-range dependencies due to the convolution operator's local receptive field. To remedy this issue, one popular solution is substituting RNNs for CNNs, but RNNs are computationally inefficient, caused by the unparallelization of their temporal iterations.

In 2017, Vaswani et al. [20] proposed the self-attention mechanism, dispensing with RNNs and CNNs entirely. Compared to discriminative deep learning models, self-attention is computationally efficient. Compared to DNNs, it possesses much fewer parameters. Compared to CNNs, it is flexible in modeling both long-range and local dependencies. Compared to RNNs, it is based on matrix multiplication, which is highly parallelizable and easily accelerated. The self-attention mechanism has been successfully used for different human–machine communication tasks [21–26], including the speech enhancement tasks [27,28]. Nevertheless, there are still two problems in the previous works. Firstly, some of them did not adopt adversarial training [28,29], which suffers from unseen distortion derived from handcrafted loss functions. Secondly, some works used discriminative models as the architecture backbone (e.g., DNN [16], CNN [30] or LSTM [19]). However, DNNs are computationally inefficient due to the huge parameter scale. CNNs command the extraordinary ability to model local information, but they experience difficulties in capturing long-range dependencies. RNNs are computationally inefficient, caused by the unparallelization of its temporal iterations.

To combine the adversarial training and the self-attention mechanism, Zhang et al. [31] proposed the self-attention generative adversarial network for image synthesis, which introduces the self-attention mechanism into convolutional GANs. In their work, the self-attention module is complementary to convolutional layers and helps with modeling long-range and multi-level dependencies across image regions. In the same year, Ramachandran et al. [32] provided the theoretical basis for substituting the self-attention mechanism for discriminative models. They verify that self-attention layers can completely replace convolutional layers and achieve state-of-the-art performance on vision tasks. Afterwards, Cordonnier et al. [33] presented evidence that self-attention layers can perform convolution and attend to pixel-grid patterns similarly to convolutional layers.

Nonetheless, Yang et al. [34] suggested that the self-attention mechanism might fully attend to all elements, dispersing the attention distribution, and thus overlook the relation of neighboring elements and phrasal patterns. Guo et al. [35] indicated that the generalization ability of the self-attention mechanism is weaker than CNNs or RNNs, especially on moderate-sized datasets, and the reason can be attributed to its unsuitable inductive bias of the self-attention structure. To this end, Yang et al. [34] proposed a parameter-free convolutional self-attention model to enhance the feature extraction of neighboring elements and validate its effectiveness and universality. Guo et al. [35] regarded self-attention as a matrix decomposition problem and proposed an improved self-attention module by introducing locality linguistic constraints. Xu et al. [36] proposed a hybrid attention mechanism via a gating scalar for leveraging both the local and global information, and verified that these two types of contexts are complementary to each other.

Inspired by prior works, this paper presents a series of speech enhancement GANs (SEGANs) equipped with a self-attention mechanism in three ways: first, we deploy the stand-alone self-attention layer in a SEGAN. Next, we employ locality modeling on the stand-alone self-attention layer. Finally, we investigate the functionality of the self-attention augmented convolutional SEGAN. We aim to probe the performance of a SEGAN equipped (i) with stand-alone standard self-attention layers, (ii) with stand-alone hybrid (global and local) self-attention layers, and (iii) with self-attention augmented convolutional layers. In addition, we also calculate the parameter scales of these proposed models.

Please note that there are four highlights of our work. Firstly, we deploy the adversarial training to alleviate the distortion introduced by handcrafted loss functions, and hence the enhancement module is supposed to capture more underlying structural characteristics. Secondly, we employ self-attention layers to obtain a more flexible ability to capture both long-range or local interactions. Thirdly, the locality modeling of the self-attention layer is via a parameter-free method. Lastly, we utilize raw speech waveforms as inputs of the system to avoid any distortion introduced by handcrafted features.

We evaluate the proposed systems in terms of various objective evaluation criteria. Systematic experiment results reveal that equipped with the stand-alone self-attention layer, the proposed system outperforms baseline systems in terms of various objective evaluation criteria with up to 95% fewer parameters. In addition, the locality modeling on the stand-alone self-attention layer delivers further performance improvements without increasing any parameter. Moreover, the self-attention augmented SEGAN outperforms all baseline systems and achieves the best results on SSNR and STOI of our work, with acceptably increased parameters.

2. Related works

Pascual et al. [14] open the exploration of generative architectures for speech enhancement, leveraging the ability of deep learning to learn complex functions from large example sets. The enhancement mapping is accomplished by the generator G, whereas the discriminator D, by discriminating between real and fake signals, transmits information to G so that G can learn to produce outputs that resemble the realistic distribution of the clean signals. The proposed system learns from different speakers and noise types, and incorporates them together into the same shared parametrization, which makes the system simple and generalizable in those dimensions.

On the basis of [14], Phan et al. [37] indicate that all existing SEGAN systems execute the enhancement mapping via a single stage by a single generator, which may not be optimal. In this light, they hypothesize that it would be better to carry out multi-stage enhancement mapping rather than a single-stage one. To this end, they divide the enhancement process into multiple stages and each stage contains an enhancement mapping. Each mapping is conducted by a generator, and each generator is tasked to further correct the output produced by its predecessor. All these generators are cascaded to enhance a noisy input signal gradually to yield an refined enhanced signal. They propose two improved SEGAN frameworks, namely iterated SEGAN (ISEGAN) and deep SEGAN (DSEGAN). In the ISEGAN system, parameters of its generator are fixed, constraining ISEGAN's generators to apply the same mapping iteratively, as its name implies. DSEGAN's generators have their own independent parameters, allowing them to learn different mappings flexibly. However, parameters of DSEGAN's generators are N_G times that of ISEGAN's generators. N_G is the number of generators.

Afterwards, Phan et al. [30] revealed that the existing class of GANs for speech enhancement solely relies on the convolution operation, which may obscure temporal dependencies across the sequence input. To remedy this issue, they propose a self-attention layer adapted from non-local attention, coupled with the convolutional and deconvolutional layers of the SEGAN, referred to as SASEGAN. Furthermore, they empirically studied the effect of placing the self-attention layer at the (de)convolutional layers with varying layer indices, including all layers as long as memory allows.

As Pascual et al. [14] state, they open the exploration of generative architectures for speech enhancement to progressively incorporate further speech-centric design choices for performance improvement. This study aims to further optimize SEGAN, especially its variant with a self-attention mechanism. Unlike [37], we preserve the single generator architecture to maintain the light-weight parameter scale. The authors of [30] focused on coupling only one self-attention layer to one convolutional layer in the encoder. Namely, a maximum of three layers of SEGAN are equipped with the self-attention mechanism each time: one convolutional layer of the encoder, one deconvolutional layer of the decoder, and one convolutional layer of the discriminator. Although they also experimented with the performance of SASEGAN-all, i.e., simply coupling self-attention layers to all (de)convolutional layers, we query whether there are more optimized coupling combinations. (Actually, ref. [30] only couples the self-attention layer to the 3rd–11th layers in the encoder, decoder, and the discriminator because of the memory limitation, although they refer it to SASEGAN-all.) For example, can coupling the self-attention mechanism to the 10th and 11th (de)convolutional layers outperform SASEGAN-all with even smaller parameters? In addition, inspired by [32,33], we explore the feasibility of substituting self-attention layers with (de)convolutional layers completely, namely SEGAN with stand-alone self-attention layers. Moreover, to take full advantage of the self-attention layer's flexibility of modeling of both long-range and local dependencies, we introduce the parameter-free locality modeling [34] of the self-attention mechanism in SEGAN. To our best knowledge, the three following explorations: stand-alone self-attention layer, the locality modeling on self-attention layers, and optimized combination of coupling self-attention layers with convolutional layers, were never executed by previous works in the SEGAN class.

3. Self-Attention Mechanism

Self-attention [20] relates the information over different positions of the entire input sequence for computing the attention distribution using scaled dot product attention:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}. \tag{1}$$

$\boldsymbol{Q} \in \mathbb{R}^{t_q \times d_q}$, $\boldsymbol{K} \in \mathbb{R}^{t_k \times d_k}$, and $\boldsymbol{V} \in \mathbb{R}^{t_v \times d_v}$ are three inputs of the self-attention layer: queries, keys, and values, where t_q, t_k, and t_v are the element numbers in different inputs and d_q, d_k, and d_v denote the corresponding element dimensions. The scalar $\frac{1}{\sqrt{d_k}}$ prevents the softmax function from falling into regions with tiny gradients. One query's output is computed as a weighted sum of the values, where each weight of the value is computed by a designated function of the query with the homologous key.

4. Self-Attention Speech Enhancement GANs

4.1. Speech Enhancement GANs

Given a dataset $\mathcal{X} = \{(\boldsymbol{x}_1^*, \tilde{\boldsymbol{x}}_1), (\boldsymbol{x}_2^*, \tilde{\boldsymbol{x}}_2), \cdots, (\boldsymbol{x}_N^*, \tilde{\boldsymbol{x}}_N)\}$ consisting of N pairs of raw signals: clean speech signal $\boldsymbol{x}^*$ and noisy speech signal $\tilde{\boldsymbol{x}}$, speech enhancement aims to find a mapping $f_\theta(\tilde{\boldsymbol{x}}) : \tilde{\boldsymbol{x}} \to \hat{\boldsymbol{x}}$ to transform the raw noisy signal $\tilde{\boldsymbol{x}}$ to the enhanced signal $\hat{\boldsymbol{x}}$. θ contains the parameters of the enhancement network.

SEGANs designate the generator G for the enhancement mapping, i.e., $\hat{\boldsymbol{x}} = G(\tilde{\boldsymbol{x}})$, while designating the discriminator D to guide the training of G by classifying $(\boldsymbol{x}^*, \tilde{\boldsymbol{x}})$ as real and $(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ as fake. Eventually, G learns to produce enhanced signals $\hat{\boldsymbol{x}}$ good enough to fool D such that D classifies $(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ as real.

4.2. Stand-Alone Self-Attention Speech Enhancement GANs

In this section, we demonstrate the self-attention layer adapted in GANs [31], which enables both the generator and the discriminator to efficiently model relations between widely separated spatial regions. Given the feature map $\boldsymbol{F} \in \mathbb{R}^{L \times C}$ as the input of the self-attention layer, where L is the time dimension and C is the number of channels, the query matrix $\boldsymbol{Q}$, the key matrix $\boldsymbol{K}$, and the value matrix $\boldsymbol{V}$ are obtained via transformations:

$$\boldsymbol{Q} = \boldsymbol{F}\boldsymbol{W}^Q, \boldsymbol{K} = \boldsymbol{F}\boldsymbol{W}^K, \boldsymbol{V} = \boldsymbol{F}\boldsymbol{W}^V, \tag{2}$$

where $\boldsymbol{W}^Q \in \mathbb{R}^{C \times \frac{C}{b}}$, $\boldsymbol{W}^K \in \mathbb{R}^{C \times \frac{C}{b}}$, and $\boldsymbol{W}^V \in \mathbb{R}^{C \times \frac{C}{b}}$ denote the learnt weight matrices of the 1×1 convolutional layer. b is a factor for reducing the channel numbers. Additionally, a max pooling layer with filter width and stride size of p is deployed to reduce the number of keys and values for memory efficiency. Therefore, the dimensions of the matrices are $\boldsymbol{Q} \in \mathbb{R}^{L \times \frac{C}{b}}$, $\boldsymbol{K} \in \mathbb{R}^{\frac{L}{p} \times \frac{C}{b}}$, and $\boldsymbol{V} \in \mathbb{R}^{\frac{L}{p} \times \frac{C}{b}}$. The attention map $\boldsymbol{A}$ is then computed as

$$\boldsymbol{A} = \mathrm{softmax}(\boldsymbol{Q}\boldsymbol{K}^T), \quad \boldsymbol{A} \in \mathbb{R}^{L \times \frac{L}{p}}, \tag{3}$$

$$a_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{L} \exp(s_{ij})}, \quad \text{where } s_{ij} = \boldsymbol{Q}(\boldsymbol{x_i})\boldsymbol{K}(\boldsymbol{x_j})^T. \tag{4}$$

$a_{j,i}$ denotes the extent to which the model attends to the ith location when synthesizing the jth column $\boldsymbol{v_j}$ of $\boldsymbol{V}$. The output of the attention layer $\boldsymbol{O}$ is then computed as

$$\boldsymbol{O} = (\boldsymbol{A}\boldsymbol{V})\boldsymbol{W}^O, \quad \boldsymbol{W}^O \in \mathbb{R}^{\frac{C}{b} \times C}. \tag{5}$$

With the weight matrix $\boldsymbol{W}^O$ realized by a 1 × 1 convolution layer of C filters, the shape of $\boldsymbol{O}$ is restored to the original shape $L \times C$. Eventually, there is a learnable scalar β weighing the output of the attention layer in the final output as

$$\tilde{\boldsymbol{F}} = \beta\boldsymbol{O} + \boldsymbol{F}. \tag{6}$$

Various loss functions have been proposed to improve the training of GANs, e.g., Wasserstein loss [38], relativistic loss [39], metric loss [38], and least-squares loss [40]. In our work, the least-squares loss [40] with binary coding is utilized instead of the cross-entropy loss. Due to the effectiveness of the L_1 norm in the image manipulation domain [41], it is deployed in G to gain more fine-grained and realistic results. The scalar λ controls the magnitude of the L_1 norm. Consequently, the loss functions of G and D are

$$\begin{aligned} \min_{D} \mathcal{L}(D) = &\frac{1}{2}\mathbb{E}_{\boldsymbol{x}^*,\tilde{\boldsymbol{x}} \sim p_{data}(\boldsymbol{x}^*,\tilde{\boldsymbol{x}})}[D(\boldsymbol{x}^*,\tilde{\boldsymbol{x}}) - 1]^2 + \\ &\frac{1}{2}\mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z}),\tilde{\boldsymbol{x}} \sim p_{data}(\tilde{\boldsymbol{x}})}[D(G(\boldsymbol{z},\tilde{\boldsymbol{x}}),\tilde{\boldsymbol{x}})]^2, \end{aligned} \tag{7}$$

$$\begin{aligned} \min_{G} \mathcal{L}(G) = &\frac{1}{2}\mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z}),\tilde{\boldsymbol{x}} \sim p_{data}(\tilde{\boldsymbol{x}})}[D(G(\boldsymbol{z},\tilde{\boldsymbol{x}}),\tilde{\boldsymbol{x}}) - 1]^2 \\ &+ \lambda\|G(\boldsymbol{z},\tilde{\boldsymbol{x}}) - \boldsymbol{x}^*\|_1. \end{aligned} \tag{8}$$

We illustrate the diagram of a simplified self-attention layer with $L = 9$, $C = 6$, $p = 3$, and $b = 2$ in Figure 1.

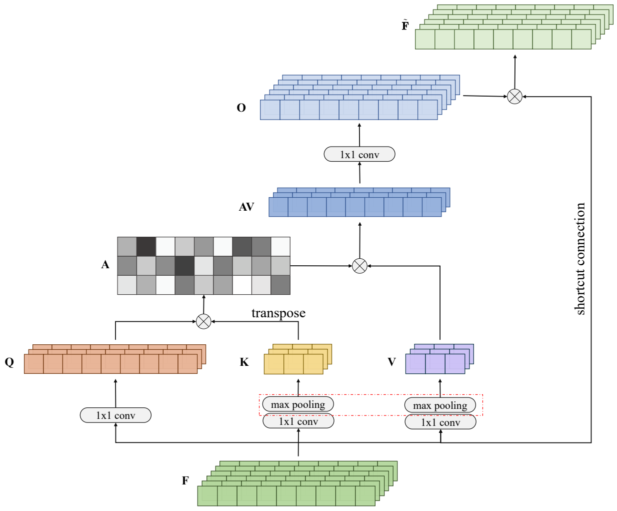


Figure 1. Illustration of the stand-alone self-attention layer with $L = 9$, $C = 6$, $p = 3$, and $b = 2$. The max pooling layers in the red frame are discarded for matrix $\boldsymbol{K}$ and $\boldsymbol{V}$ when modeling locality of the stand-alone self-attention layers.

4.3. Locality Modeling for Stand-Alone Self-Attention Layers

As illustrated in Figure 2, for the query $\boldsymbol{Q}$, we restrict its attention region (e.g., $\boldsymbol{K} = \{\boldsymbol{k}_1, \cdots, \boldsymbol{k}_l, \cdots, \boldsymbol{k}_L\}$) to a local scope with a fixed window size $M+1$ ($M \leq L$) centered at the position l as

$$\hat{\boldsymbol{K}} = \{\boldsymbol{k}_{l-\frac{M}{2}}, \cdots, \boldsymbol{k}_l, \cdots, \boldsymbol{k}_{l+\frac{M}{2}}\}, \quad \hat{\boldsymbol{K}} \in \mathbb{R}^{(M+1)\times\frac{c}{b}}, \tag{9}$$

$$\hat{\boldsymbol{V}} = \{\boldsymbol{v}_{l-\frac{M}{2}}, \cdots, \boldsymbol{v}_l, \cdots, \boldsymbol{v}_{l+\frac{M}{2}}\}, \quad \hat{\boldsymbol{V}} \in \mathbb{R}^{(M+1)\times\frac{c}{b}}. \tag{10}$$

When we apply the locality modeling to the self-attention layer, the factor p should be discarded to help preserve the original neighborhood for the centered position. Accordingly, the local attention map and the output of the attention layer are modified as

$$\hat{\boldsymbol{A}} = \text{softmax}(\boldsymbol{Q}\hat{\boldsymbol{K}}^T), \quad \hat{\boldsymbol{A}} \in \mathbb{R}^{L\times(M+1)}, \tag{11}$$

$$\hat{\boldsymbol{O}} = (\hat{\boldsymbol{A}}\hat{\boldsymbol{V}})\boldsymbol{W}^O, \quad \boldsymbol{W}^O \in \mathbb{R}^{\frac{C}{b}\times C}. \tag{12}$$

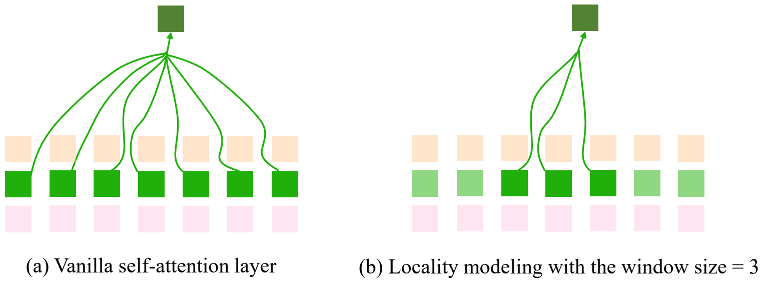

(a) Vanilla self-attention layer
(b) Locality modeling with the window size = 3

Figure 2. Illustration of (**a**) vanilla self-attention layer; (**b**) locality modeling with the window size = 3. Semi-transparent colors represent masked tokens that are invisible to the self-attention layer.

4.4. Attention augmented convolutional SEGAN

We implement an attention augmented convolutional SEGAN by coupling the self-attention layer with the (de)convolutional layer(s). We assume this proposed architecture possesses an advantage, where the distance-aware information from the convolutional layer and the distance-agnostic dependencies modeled by the self-attention layer are supplementary and complementary to each other.

To this end, we introduce two learnable parameters, κ and γ, to weigh the input feature map $\boldsymbol{F}$ and the output feature map $\boldsymbol{O}$ of the coupled layer (with self-attention mechanism and convolution) in the augmented output $\tilde{\boldsymbol{F}}$ as

$$\tilde{\boldsymbol{F}} = \kappa\boldsymbol{O} + \gamma\boldsymbol{F}. \tag{13}$$

5. Experimental Setups

We systematically evaluate the effectiveness of (i) the stand-alone self-attention layer on SEGAN, (ii) the locality modeling on the stand-alone self-attention layer, and (iii) the attention augmented convolutional SEGAN.

5.1. Dataset

We conduct all experiments on the publicly available dataset introduced in [42]. The dataset includes 30 speakers from the Voice Bank corpus [43], with 28 speakers selected for the training set and 2 for the test set. The noisy training set contains 40 different conditions, obtained by 10 types of noise (2 artificial and 8 from the Demand database [44]) with 4 signal-to-noise ratios (SNRs) (15, 10, 5, and 0 dB) each. There are approximately 10 sentences in each condition per training speaker. The noisy test set contains 20 different conditions, obtained by 5 types of noise with 4 SNRs (17.5, 12.5, 7.5, and 2.5 dB) each. There are around 20 sentences in each condition per test speaker. Notably, all the speakers and conditions included in the training and test set are totally different from each other, i.e., the test set is entirely unseen by the training set.

5.2. Evaluation Criteria

All the proposed architectures are evaluated with the following classic objective criteria for speech enhancement (the higher the better):

- SSNR: Segmental SNR [45] (in the range of $[0, +\infty)$);
- STOI: Short-time objective intelligibility [46] (in the range of $[0, 100]$);
- CBAK: Mean opinion score (MOS) prediction of the intrusiveness of background noises [47] (in the range of $[1, 5]$);
- CSIG: MOS prediction of the signal distortion attending only to the speech signal [47] (in the range of $[1, 5]$);
- COVL: MOS prediction of the overall effect [47] (in the range of $[1, 5]$);
- PESQ: Perceptual evaluation of speech quality, using the wide-band version recommended in ITU-T P.862.2 [48] (in the range of $[-0.5, 4.5]$).

All results have been computed using the implementation demonstrated in [1], and is available on the publisher's website: https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip (accessed on 15 October 2020).

5.3. Network Architecture

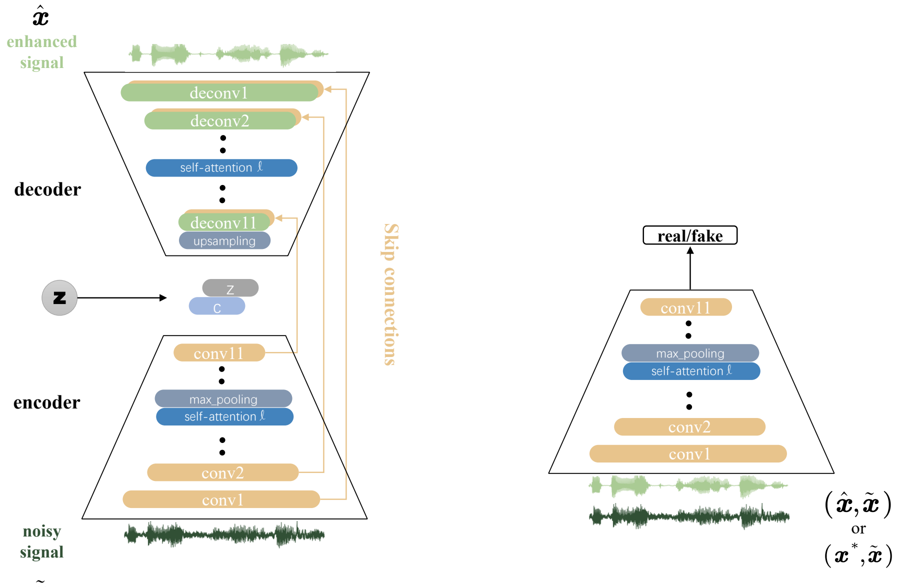
We first introduce the architecture of SEGAN, as all three variants, namely (i) SEGAN with stand-alone self-attention layers, (ii) stand-alone self-attention layers with locality modeling, and (iii) attention augmented convolutional SEGAN, are based on it.

The classic enhancement systems are based on the short-time Fourier analysis/synthesis framework [1]. They assume that the short-time phase is not important for speech enhancement [49], and they only process the spectrum magnitude. However, further studies [50] show the intensive relation between the clean phase spectrum and the speech quality. Therefore, we use raw inputs for SASEGAN. We extract approximately one-second waveform chunks (~16,384 samples) with a sliding window every 500 ms. G makes use of an encoder–decoder structure. The encoder is composed of 11 1-dim strided convolutional layers of filter width 31 and stride 2, followed by parametric rectified linear units (PReLUs) [51]. Along with the depth, the number of filters per layer increases while the signal duration shrinks. To compensate for the smaller and smaller convolutional output, the number of filters increases along the encoder's depth $\{16, 32, 32, 64, 64, 128, 128, 256, 256, 512, 1024\}$, resulting in dimensions per layer of $\{8192 \times 16, 4096 \times 32, 2048 \times 32, 1024 \times 64, 512 \times 64, 256 \times 128, 128 \times 128, 64 \times 256, 32 \times 256, 16 \times 512, 8 \times 1024\}$. At the 11th layer of the encoder, the encoding vector $\boldsymbol{c} \in \mathbb{R}^{8\times1024}$ is stacked with the noise $\boldsymbol{z} \in \mathbb{R}^{8\times1024}$, sampled from the distribution $\mathcal{N}(0, I)$, and presented to the decoder. The decoder mirrors the encoder architecture entirely to reverse the encoding process by means of the transposed convolution, termed as deconvolution. There are skip connections connecting each convolutional layer to its homologous deconvolutional layer. They bypass the compression performed in the middle of the model and allow the fine-grained details (e.g., phase, alignment) of speech signals to flow into the decoding stage directly. This is done because if we force all information flow through bottleneck structures, much useful low-level information could be lost through the compression. In addition, skip connections offer a better training behavior, as the gradients can flow more deeply through the whole structure [52]. Notably, skip connections and the addition of the latent vector double the number of feature maps in every layer.

The discriminator D resembles the encoder's structure with the following differences: (i) it receives a pair of raw audio chunks as the input, i.e., $(\boldsymbol{x}^*, \tilde{\boldsymbol{x}})$ or $(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$; (ii) it utilizes virtual batch-norm before LeakyReLU [53] activation with $\alpha = 0.3$; (iii) in the last activation layer, there is a 1×1 convolution layer to reduce the feature required for the final classification neuron from 8×1024 to 8.

5.3.1. SEGAN with Stand-Alone Self-Attention Layers

We substitute the self-attention layer, illustrated in Section 4.2, with the (de)convolutional layers of both G and D. Figure 3a,b demonstrate an example of substituting the self-attention layer with the lth (de)convolutional layers.



(a) Generator (b) Discriminator

Figure 3. Illustration of the architecture of speech enhancement GAN (SEGAN) with stand-alone self-attention layers. (**a**) The generator component. (**b**) The discriminator component.

When the stand-alone self-attention layer is deployed on the lth convolutional layer of the encoder, the mirroring lth deconvolutional layer of the decoder and the lth convolutional layer in the discriminator are also replaced with it. To keep the dimension of the feature map per layer in accordance with that of SEGAN, a max pooling layer with a kernel size of 2 and a stride length of 2 follows every stand-alone self-attention layer in the encoder. Accordingly, the upsampling needs to be deployed before the 11th deconvolutional layer in the decoder to ensure the same feature dimensions flowing in the skip connections. We experiment with two interpolation methods: nearest and bilinear. The results suggest that the bilinear interpolation outperforms the nearest one, so we utilize the bilinear interpolation for upsampling in all our experiments. Theoretically, the stand-alone self-attention layer can be placed in any number, even all, of the (de)convolutional layers. We empirically study the effect of placing the stand-alone self-attention layer at (de)convolutional layers with lower or higher layer indices as well as layer combinations, provided that memory allows.

Similarly, the D component follows the same structure as G's encoder stage. We set $b = 8$, $p = 4$ for the self-attention layer. λ is set to be 100, and we initialize $\beta = 0$ for the self-attention layer.

5.3.2. Stand-Alone Self-Attention Layer with Locality Modeling

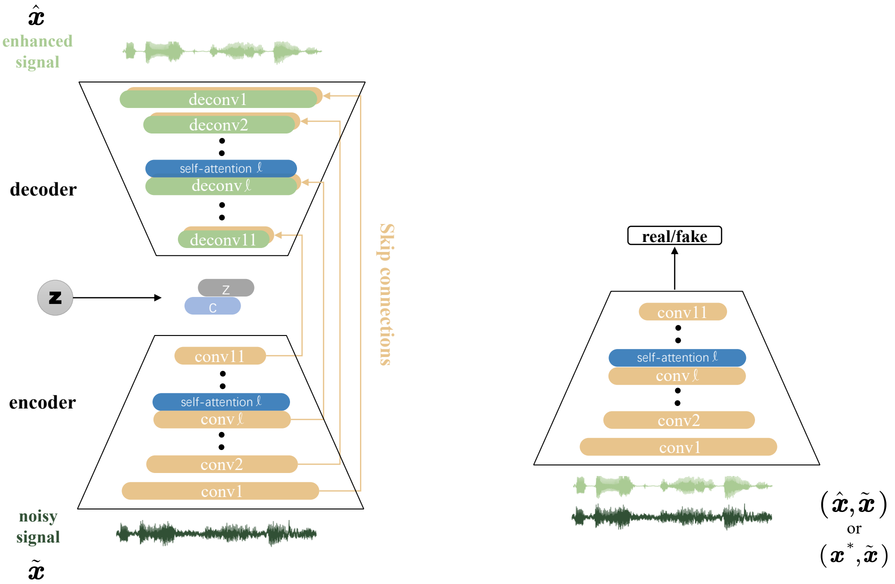
The general architecture remains the same as the case of Section 5.3.1. However, the factor p is eliminated to help preserve the original neighborhood for the centered position. Namely, the max pooling layers in Figure 1 are discarded for matrix $\boldsymbol{K}$ and $\boldsymbol{V}$. The factor b remains as 8. We conduct ablation tests on the impacts of the window size $(M + 1)$ and the placement of the window on the system performance.

Peters et al. [54] and Raganato and Tiedemann [55] indicated that higher layers of the system tend to learn semantic information while lower layers capture more surface and lexical information. Therefore, we centrally apply locality modeling to the lower layers, in line with the configuration in [56,57]. Then, we expect that the representations are learned in a hierarchical fashion. Namely, the distance-aware and local information extracted by the lower layers can complement distance-agnostic and global information captured by the higher layers.

5.3.3. Attention Augmented Convolutional SEGAN

Instead of replacement (Section 5.3.1), we couple stand-alone self-attention layers with (de)convolutional layers of both the generator and discriminator. As illustrated in Figure 4, when the stand-alone self-attention layer is coupled with the lth convolutional layer of the encoder, the mirror lth deconvolutional layer of the decoder and the lth convolutional layer in the discriminator are also coupled with it. Spectral normalization is applied to all the (de)convolutional layers. For scalars γ and κ, we experiment with three initialization pairs: $\gamma = 0$ and $\kappa = 0$, $\gamma = 0.75$ and $\kappa = 0.25$ (inspired by the results provided by [58]), and $\gamma = 0.25$ and $\kappa = 0.25$, finding that best results are acquired when both γ and κ are initialized as 0.25. We empirically study the effect of coupling the self-attention layer with lower and higher layer indices as well as different layer combinations. Two introduced factors (introduced in Section 4.2) are set to be $p = 4$ and $b = 8$.

Since coupling the self-attention layer with a single convolutional layer has already been studied by [30] in detail, this study focuses the more optimized couple combination for this topic.



(a) Generator (b) Discriminator

Figure 4. Illustration of the attention augmented convolutional SEGAN architecture. (**a**) The generator component. (**b**) The discriminator component.

5.4. Baseline Systems

For comparison, we take the seminal work [14], and other SEGAN variants [30,37] that we introduced in Section 2 as baseline systems. From [37], we choose the results of ISEGAN with two shared generators and DSEGAN with two independent generators as baseline results (the situation of $N_G = 2$) for two reasons. On one hand, the number of generators leads to an exponential parameter increment. On the other hand, Phan et al. [37] indicated that marginal impacts of ISEGAN's number of iterations and DSEGAN's depth larger than $N_G = 2$ with no significant performance improvements are seen. The authors of [30] present detailed results of the influence of the self-attention layer placement in the generator and the discriminator. We choose the average result of coupling the self-attention layer with a single (de)convolutional layer (referred to as *SASEGAN-avg*), and the result of coupling self-attention layers with all (de)concolutional layers (referred to as SASEGAN-all) to ensure a fair comparison. It is worth noting that it is stated in [30] that compared to

SASEGAN-avg, results of SASEGAN-all are slightly more boosted, but these gains are achieved at the cost of increased computation time and memory requirements.

5.5. Configurations

Networks are trained with RMSprop [59] for 100 epochs with a minibatch size of 50. A high-frequency preemphasis filter of coefficient 0.95 is applied to both training and test samples. In the test stage, we slide the window through the whole duration of our test utterance without overlap, and the enhanced outputs are deemphasized and concatenated at the end of the stream.

6. Results

We systematically evaluate the effectiveness of (i) the stand-alone self-attention layer on SEGAN, (ii) the locality modeling on the stand-alone self-attention layer, and (iii) attention augmented convolutional SEGAN. Table 1 exhibits the performance of SEGAN equipped with the stand-alone self-attention layer. The upper part displays the results of baseline systems, while the lower part displays the results of our work.

As shown in Table 1, when we replace the 6th and 10th (de)convolutional layers with the stand-alone self-attention layer, the system overtakes all baselines across all metrics and achieves the best SSNR result, with 47% fewer (compared to DSEGAN) or 12% fewer (compared to SASEGAN-all) parameters. Furthermore, when we adopt the stand-alone self-attention layer at the 9th to 11th (de)convolutional layers, it still yields comparable or even better (STOI) results, with parameters plunging drastically to merely 5% (95% fewer) of DESEGAN or 9% (91% fewer) of SASEGAN-all. When $l = 4$, the parameter scale of the proposed system is closest to that of the baseline systems. Under such circumstances, it outperforms baseline systems in PESQ, CBAK, COVL, and STOI, and achieves the best results in PESQ, CBAK, and COVL. When we substitute the self-attention layer with lth ($6 \leq l \leq 11$) (de)concolutional layers, the system performance plummets as the parameters of the whole system are only 1% of SASEGAN-all [14] and 0.6% of DSEGAN [37]. The results reveal that the stand-alone layer can be a powerful and light-weight primitive for speech enhancement.

Table 1. Effects of the stand-alone self-attention layer(s) on speech enhancement GANs (SEGANs). We denote the proposed architecture with the stand-alone self-attention layer(s) at the lth (de)convolutional layer(s) as *standalone-l*. Values that overtake all baseline systems are in bold. Values with an asterisk are the best ones achieved for each metric.

Architecture	**Params (M)**	**Metric**					
		PESQ	**CSIG**	**CBAK**	**COVL**	**SSNR**	**STOI**
Noisy	-	1.97	3.35	2.43	2.63	1.69	92.10
SEGAN [14]	294	2.16	3.48	2.94	2.79	7.66	93.12
ISEGAN [37]	294	2.24	3.23	2.93	2.68	8.19	93.29
DSEGAN [37]	513	2.35	3.56	3.10	2.94	8.70	93.25
SASEGAN-avg [30]	295	2.33	3.52	3.05	2.90	8.08	93.33
SASEGAN-all [30]	310	2.35	3.55	3.10	2.91	8.30	93.49
standalone-4	293	2.49 *	3.54	3.62 *	3.11 *	7.70	93.72
standalone-6	292	2.41	3.54	3.07	2.96	8.10	93.63
standalone-11	103	2.43	3.74 *	3.01	3.07	7.11	93.55
standalone-6,10	274	2.39	3.59	3.12	2.98	8.71 *	93.66
standalone-10,11	51	2.37	3.57	3.00	2.95	7.71	93.54
standalone-4,6,10	275	2.45	3.61	3.10	3.01	8.30	93.73
standalone-9,10,11	28	2.43	3.50	2.97	2.88	8.51	93.90 *
standalone-6–11	3	2.01	3.36	2.62	2.64	6.98	93.32

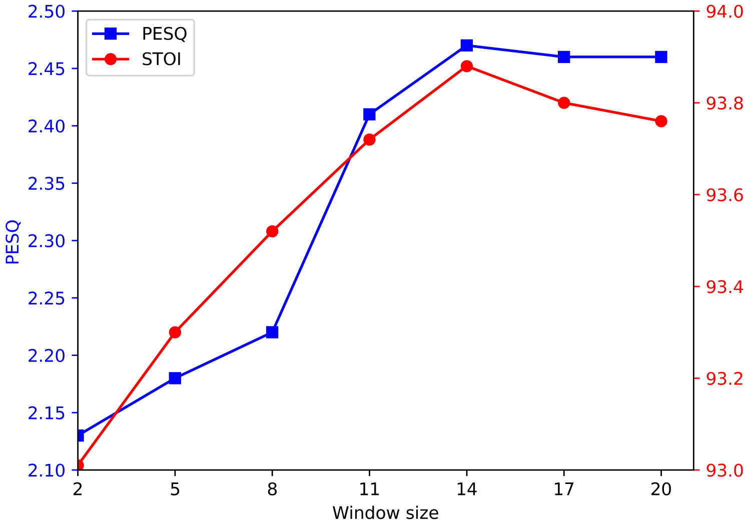


Figure 5. Effects of the window size on the self-attention layers.

Table 2. Effects of the locality modeling on the stand-alone self-attention layer. The layer numbers with curly braces represent the employment of locality modeling on the current layer. Values that overtake all baseline systems are in bold. Values with an † are the ones that overtake their best counterparts on the same metric in Table 1.

Architecture	Params (M)	Metric					
		PESQ	CSIG	CBAK	COVL	SSNR	STOI
standalone-{4}	293	**2.50** †	3.55	**3.64** †	**3.13** †	8.10	**93.83**
standalone-{6}	292	**2.41**	3.54	3.08	**2.96**	7.90	**93.68**
standalone-{6},10	274	**2.40**	**3.59**	**3.12**	**2.99**	**8.70**	**93.76**
standalone-{4},{6},10	275	**2.45**	**3.58**	**3.11**	**3.03**	8.30	**93.88**

Next, we investigate the effects of locality modeling and its window size. Prior studies [56,60] indicate that lower layers usually extract lower-level features, so they should attend to the local field more. Additionally, they prove empirically that modeling locality on lower layers achieves better performances. Therefore, we only apply the locality modeling on layers not deeper than the 6th one. As plotted in Figure 5, the tiny window size limits the receptive field too much, and hence degrades the performance due to the deprivation of the ability of modeling long-range and multi-level dependencies. It appears that a window size of 14 is superior to other settings, approximately consistent with [34] on machine translation tasks. When the window size continues to increase, the performance tends to be the same without windows, which is self-explanatory. Completed results on six criteria are exhibited in Table 2. Compared to Table 1, employing locality modeling on the 4th layer yields the most significant improvement, in accordance with the conclusion in [56,60]. It also achieves the best or comparable results across all criteria, which demonstrate the functionality of the locality modeling without further computational cost. An explanation of the undesirable SSNR is that the suboptimal upsampling method introduces speech distortion, which is also manifested on CSIG. Importantly, a fixed-size window is not the state-of-the-art approach in the field of locality modeling of the self-attention mechanism. We choose it as it is parameter free, corresponding to our goal of a light-weight system. It is worth noting that the proposed SEGAN with stand-alone self-attention layers is general enough to combine other more advanced locality modeling approaches [61,62] in cases where the computation complexity is secondary.

Table 3. Performance of the attention augmented convolutional SEGAN. We denote the proposed architecture where the self-attention couples the lth (de)convolutional layer(s) as *augmentation-l*. Values that overtake all baseline systems are in bold. Values with an ‡ are the ones that overtake their best counterparts on the same metric in Tables 1 and 2.

Architecture	Params (M)	Metric					
		PESQ	CSIG	CBAK	COVL	SSNR	STOI
augmentation-4,6	293	**2.43**	**3.64**	**3.11**	**3.02**	8.20	**93.99** ‡
augmentation-6,10	299	**2.47**	**3.58**	**3.15**	**3.01**	**8.87** ‡	**93.61**
augmentation-4,10	298	**2.45**	**3.58**	**3.10**	**3.00**	7.86	**93.61**
augmentation-4,6,10	299	**2.39**	3.50	3.08	2.92	8.41	**93.68**

Lastly, we investigate the functionality of the attention augmented convolutional networks according to Equation (13). We choose the combination from lower-to-middle layers (augmentation-4,6), middle-to-higher layers (augmentation-6,10), and all layer ranges (augmentation-4,10 and augmentation-4,6,10). As displayed in Table 3, coupling the self-attention layer on the 4th and 6th layers is more competitive on CSIG, COVL, and STOI (the best results in Table 3), and it achieves the best STOI performance. In contrast, adding the self-attention layer on the 6th and 10th layers overtakes all baseline systems across all metrics, and it gives the best result on SSNR. The combination of the 4th and 10th layers still outperforms baseline systems, except for SSNR. However, the combination of the 4th, 6th, and 10th layers only outperforms baseline systems on PESQ and STOI, although it still yields decent results on other metrics. These results demonstrate the efficiency of the attention augmentation for the convolutional SEGAN. Nevertheless, it is worth noting that system parameters inevitably increase when coupling the self-attention layer to (de)convolutional layers.

7. Discussion

In general, the biggest advantage of applying the stand-alone self-attention layer in SEGAN is that it simultaneously outperforms baseline systems, and decreases the model complexity drastically. In particular, when applying the stand-alone self-attention layer as the 6th and 10th layers of the system, the resultant system overtakes all baselines across all metrics and achieves the best SSNR results with only $\sim$50% parameters (compared to DSEGAN [37]). In addition, locality modeling can be an effective auxiliary to stand-alone self-attention layers, which further improves their performance without any extra parameter increment. Notably, locality modeling on a lower self-attention layer delivers more perceptible performance improvements, consistent with [34,56,57]. For the self-attention augmented SEGAN, it performs modestly better. Although it is less light-weight than those two approaches, it still has 42% fewer parameters compared to DSEGAN.

However, different placements of the stand-alone self-attention layer or the coupled self-attention layer lead to different performance improvements, and the compromise between system performance and system complexity is always ineluctable. We only present the achieved performance and the homologous model complexity for representative placements, which readers can take for reference according to the desired application.

8. Conclusions

We integrate the self-attention mechanism with SEGAN to improve its flexibility of both long-range and local dependency modeling for speech enhancement in three methods, namely, applying the stand-alone self-attention layer, modeling locality on the stand-alone self-attention layer, and coupling the self-attention layer with the (de)convolutional layer. The proposed systems deliver consistent performance improvements. The main merit of the stand-alone self-attention layer is its low model complexity, and it can perform even better when equipped with the locality modeling. In contrast, the self-attention augmented convolutional SEGAN delivers more stable improvements, whereas it increases the model complexity.

Importantly, the locality modeling method utilized in this study is basic. We choose it to achieve the goal of light weight, but more advanced locality modeling approaches can be applied simply. Moreover, all proposed approaches described in this paper are generic enough to be applied to existing SEGAN models for further performance improvements. We leave these topics for future studies.

Author Contributions: Conceptualization, L.L.; methodology, L.L.; software, L.L. and Z.L.; validation, L.L. and Z.L.; formal analysis, L.L. and Z.L.; investigation, L.L. and Z.L.; resources, L.L. and Z.L.; data curation, Z.L.; writing—original draft preparation, L.L. and Z.L.; writing—review and editing, L.L., Z.L., T.W., L.K. and G.R.; visualization, L.L., Z.L., T.W., L.K. and G.R.; supervision, G.R. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by the Technical University of Munich Publishing Fund.

Data Availability Statement: The dataset utilized in this study is publicly available on the publisher's webpage: https://datashare.ed.ac.uk/handle/10283/1942, (accessed on 29 June 2021).

Acknowledgments: The authors wish to thank the editors and reviewers for their valuable comments.

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

Abbreviations

The following abbreviations are used in this manuscript:

DNN	deep neural network
CNN	convolutional neural network
RNN	recurrent neural network
GAN	generative adversarial network
SEGAN	speech enhancement generative adversarial network
G	generator
D	discriminator
SNR	signal-to-noise ratio
MOS	mean opinion score
SSNR	segmental SNR
STOI	short-time objective intelligibility
CBAK	MOS prediction of the intrusiveness of background noises
CSIG	MOS prediction of the signal distortion attending only to the speech signal
COVL	MOS prediction of the overall effect
PESQ	perceptual evaluation of speech quality

References

1. Loizou, P.C. *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2013.
2. Weninger, F.; Erdogan, H.; Watanabe, S.; Vincent, E.; Le Roux, J.; Hershey, J.R.; Schuller, B. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International Conference on Latent Variable Analysis and Signal Separation*; Springer: Berlin, Germany, 2015; pp. 91–99.
3. Taherian, H.; Wang, Z.Q.; Chang, J.; Wang, D. Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1293–1302. [CrossRef]
4. Avila, A.R.; Alam, M.J.; O'Shaughnessy, D.D.; Falk, T.H. Investigating Speech Enhancement and Perceptual Quality for Speech Emotion Recognition. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 3663–3667.
5. Lai, Y.H.; Zheng, W.Z. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users. *Biomed. Signal Process. Control.* **2019**, *48*, 35–45. [CrossRef]
6. Wang, D.; Hansen, J.H. Speech enhancement for cochlear implant recipients. *J. Acoust. Soc. Am.* **2018**, *143*, 2244–2254. [CrossRef]
7. Lim, J.; Oppenheim, A. All-pole modeling of degraded speech. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 197–210. [CrossRef]
8. Nie, S.; Liang, S.; Xue, W.; Zhang, X.; Liu, W. Two-stage multi-target joint learning for monaural speech separation. In Proceedings of the INTERSPEECH, Dresden, Germany, 6–10 September 2015; pp. 1503–1507.

9. Erdogan, H.; Hershey, J.R.; Watanabe, S.; Le Roux, J. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 708–712.
10. Nie, S.; Liang, S.; Liu, W.; Zhang, X.; Tao, J. Deep learning based speech separation via nmf-style reconstructions. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 2043–2055. [CrossRef]
11. Xu, Y.; Du, J.; Dai, L.R.; Lee, C.H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 7–19. [CrossRef]
12. Park, S.R.; Lee, J. A fully convolutional neural network for speech enhancement. *arXiv* **2016**, arXiv:1609.07132.
13. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661.
14. Pascual, S.; Bonafonte, A.; Serra, J. SEGAN: Speech enhancement generative adversarial network. *arXiv* **2017**, arXiv:1703.09452.
15. Pascual, S.; Serrà, J.; Bonafonte, A. Towards generalized speech enhancement with generative adversarial networks. *arXiv* **2019**, arXiv:1904.03418.
16. Soni, M.H.; Shah, N.; Patil, H.A. Time-Frequency Masking-Based Speech Enhancement Using Generative Adversarial Network. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5039–5043. [CrossRef]
17. Michelsanti, D.; Tan, Z.H. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. *arXiv* **2017**, arXiv:1709.01703.
18. Li, P.; Jiang, Z.; Yin, S.; Song, D.; Ouyang, P.; Liu, L.; Wei, S. PAGAN: A Phase-Adapted Generative Adversarial Networks for Speech Enhancement. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6234–6238. [CrossRef]
19. Higuchi, T.; Kinoshita, K.; Delcroix, M.; Nakatani, T. Adversarial training for data-driven speech enhancement without parallel corpus. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 40–47.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
21. Dong, L.; Xu, S.; Xu, B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5884–5888.
22. Pham, N.Q.; Nguyen, T.S.; Niehues, J.; Müller, M.; Stüker, S.; Waibel, A. Very deep self-attention networks for end-to-end speech recognition. *arXiv* **2019**, arXiv:1904.13377.
23. Sperber, M.; Niehues, J.; Neubig, G.; Stüker, S.; Waibel, A. Self-attentional acoustic models. *arXiv* **2018**, arXiv:1803.09519.
24. Katona, J. Examination and comparison of the EEG based Attention Test with CPT and TOVA. In Proceedings of the 2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, 19–21 November 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 117–120.
25. Katona, J.; Ujbanyi, T.; Sziladi, G.; Kovari, A. Examine the effect of different web-based media on human brain waves. In Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Debrecen, Hungary, 11–14 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 000407–000412.
26. Katona, J.; Kovari, A. The evaluation of BCI and PEBL-based attention tests. *Acta Polytech. Hung.* **2018**, *15*, 225–249.
27. Cheng, J.; Liang, R.; Zhao, L. DNN-based speech enhancement with self-attention on feature dimension. *Multimed. Tools Appl.* **2020**, *79*, 32449–32470. [CrossRef]
28. Koizumi, Y.; Yaiabe, K.; Delcroix, M.; Maxuxama, Y.; Takeuchi, D. Speech enhancement using self-adaptation and multi-head self-attention. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 181–185.
29. Pandey, A.; Wang, D. Dense CNN with self-attention for time-domain speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1270–1279. [CrossRef] [PubMed]
30. Phan, H.; Nguyen, H.L.; Chén, O.Y.; Koch, P.; Duong, N.Q.; McLoughlin, I.; Mertins, A. Self-Attention Generative Adversarial Network for Speech Enhancement. *arXiv* **2020**, arXiv:2010.09132.
31. 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.
32. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *arXiv* **2019**, arXiv:1906.05909.
33. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the Relationship between Self-Attention and Convolutional Layers. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
34. Yang, B.; Wang, L.; Wong, D.F.; Chao, L.S.; Tu, Z. Convolutional Self-Attention Networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4040–4045.
35. Guo, Q.; Qiu, X.; Xue, X.; Zhang, Z. Low-Rank and Locality Constrained Self-Attention for Sequence Modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2213–2222. [CrossRef]

36. Xu, M.; Wong, D.F.; Yang, B.; Zhang, Y.; Chao, L.S. Leveraging local and global patterns for self-attention networks. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3069–3075.
37. 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]
38. 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. In Proceedings of the Interspeech 2020, Shanghai, China, 14–18 September 2020; pp. 3266–3270.
39. Baby, D.; Verhulst, S. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 106–110.
40. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
41. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
42. 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 SSW, Sunnyvale, CA, USA, 13–15 September 2016; pp. 146–152.
43. 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. [CrossRef]
44. 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, Acoustical Society of America, Montreal, QC, Canada, 2–7 June 2013; Volume 19, p. 035081.
45. Quackenbush, S.R. Objective Measures of Speech Quality. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 1995.
46. 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]
47. Hu, Y.; Loizou, P.C. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *16*, 229–238. [CrossRef]
48. International Telecommunication Union. P. 862.2: Wideband Extension to Recommendation P. 862 for the Assessment of Wideband Telephone Networks and Speech Codecs. 2005. Available online: https://www.itu.int/rec/T-REC-P.862.2 (accessed on 1 May 2021).
49. Wang, D.; Lim, J. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust. Speech Signal Process.* **1982**, *30*, 679–681. [CrossRef]
50. Paliwal, K.; Wójcicki, K.; Shannon, B. The importance of phase in speech enhancement. *Speech Commun.* **2011**, *53*, 465–494. [CrossRef]
51. 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.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the ICML, Atlanta, GA, USA, 16–21 June 2013; Volume 30, p. 3.
54. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 2227–2237.
55. Raganato, A.; Tiedemann, J. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*; The Association for Computational Linguistics: New Orleans, LA, USA, 2018.
56. Yu, A.W.; Dohan, D.; Luong, M.T.; Zhao, R.; Chen, K.; Norouzi, M.; Le, Q.V. QANet: Combining local convolution with global self-attention for reading comprehension. *arXiv* **2018**, arXiv:1804.09541.
57. Yang, B.; Tu, Z.; Wong, D.F.; Meng, F.; Chao, L.S.; Zhang, T. Modeling localness for self-attention networks. *arXiv* **2018**, arXiv:1810.10182.
58. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3286–3295.
59. 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.
60. Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Zhang, C. Bi-directional block self-attention for fast and memory-efficient sequence modeling. *arXiv* **2018**, arXiv:1804.00857.

61. Yu, A.W.; Dohan, D.; Le, Q.; Luong, T.; Zhao, R.; Chen, K. Fast and Accurate Reading Comprehension by Combining Self-Attention and Convolution. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
62. Vaswani, A.; Ramachandran, P.; Srinivas, A.; Parmar, N.; Hechtman, B.; Shlens, J. Scaling local self-attention for parameter efficient visual backbones. In Proceedings of the CVPR, Virtual, 19–25 June 2021. Available online: https://openaccess.thecvf.com/content/CVPR2021/papers/Vaswani_Scaling_Local_Self-Attention_for_Parameter_Efficient_Visual_Backbones_CVPR_2021_paper.pdf (accessed on 1 May 2021).