applied sciences

MDPI

*Article*

# A Multi-Resolution Approach to GAN-Based Speech Enhancement

**Hyung Yong Kim , Ji Won Yoon , Sung Jun Cheon , Woo Hyun Kang and Nam Soo Kim ***

Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea; hykim@hi.snu.ac.kr (H.Y.K.); jwyoon@hi.snu.ac.kr (J.W.Y.); sjcheon@hi.snu.ac.kr (S.J.C.); whkang@hi.snu.ac.kr (W.H.K.)

* Correspondence: nkim@snu.ac.kr; Tel.: +82-2-880-8419

**Abstract:** Recently, generative adversarial networks (GANs) have been successfully applied to speech enhancement. However, there still remain two issues that need to be addressed: (1) GAN-based training is typically unstable due to its non-convex property, and (2) most of the conventional methods do not fully take advantage of the speech characteristics, which could result in a sub-optimal solution. In order to deal with these problems, we propose a progressive generator that can handle the speech in a multi-resolution fashion. Additionally, we propose a multi-scale discriminator that discriminates the real and generated speech at various sampling rates to stabilize GAN training. The proposed structure was compared with the conventional GAN-based speech enhancement algorithms using the VoiceBank-DEMAND dataset. Experimental results showed that the proposed approach can make the training faster and more stable, which improves the performance on various metrics for speech enhancement.

**Keywords:** speech enhancement; generative adversarial network; relativistic GAN; convolutional neural network

check for updates

**Citation:** Kim, H.Y.; Yoon, J.W.; Cheon, S.J.; Kang, W.H.; Kim, N.S. A Multi-Resolution Approach to GAN-Based Speech Enhancement. *Appl. Sci.* **2021**, *11*, 721. https://doi.org/10.3390/app11020721

Received: 2 December 2020
Accepted: 10 January 2021
Published: 13 January 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

Speech enhancement is essential for various speech applications such as robust speech recognition, hearing aids, and mobile communications [1–4]. The main objective of speech enhancement is to improve the quality and intelligibility of the noisy speech by suppressing the background noise or interferences.

In the early studies on speech enhancement, the minimum mean-square error (MMSE)-based spectral amplitude estimator algorithms [5,6] were popular producing enhanced signal with low residual noise. However, the MMSE-based methods have been reported ineffective in non-stationary noise environments due to their stationarity assumption on speech and noise. An effective way to deal with the non-stationary noise is to utilize a priori information extracted from a speech or noise database (DB), called the template-based speech enhancement techniques. One of the most well-known template-based schemes is the non-negative matrix factorization (NMF)-based speech enhancement technique [7,8]. NMF is a latent factor analysis technique to discover the underlying part-based non-negative representations of the given data. Since there is no strict assumption on the speech and noise distributions, the NMF-based speech enhancement technique shows robustness to non-stationary noise environments. Since, however, the NMF-based algorithm assumes that all data is described as a linear combination of finite bases, it is known to suffer from speech distortion not covered by this representational form.

In the past few years, deep neural network (DNN)-based speech enhancement has received tremendous attention due to its ability to model complex mappings [9–12]. These methods map the noisy spectrogram to the clean spectrogram via the neural networks such as the convolutional neural network (CNN) [11] or recurrent neural network

(RNN) [12]. Although the DNN-based speech enhancement techniques have shown promising performance, most of the techniques typically focus on modifying the magnitude spectra. This could cause a phase mismatch between the clean and enhanced speech since the DNN-based speech enhancement methods usually reuse the noisy phase for waveform reconstruction. For this reason, there has been growing interest in phase-aware speech enhancement [13–15] that exploits the phase information during the training and reconstruction. To circumvent the difficulty of the phase estimation, end-to-end (E2E) speech enhancement technique which directly enhances noisy speech waveform in the time domain has been developed [16–18]. Since the E2E speech enhancement techniques are performed in a waveform-to-waveform manner without any consideration of the spectra, their performance is not dependant on the accuracy of the phase estimation.

The E2E approaches, however, rely on a distance-based loss functions between the time-domain waveforms. Since these distance-based costs do not take human perception into account, the E2E approaches are not guaranteed to achieve good human-perception-related metrics, e.g., the perceptual evaluation of speech quality (PESQ) [19], short-time objective intelligibility (STOI) [20], and etc. Recently, generative adversarial network (GAN) [21]-based speech enhancement techniques have been developed to overcome the limitation of the distance-based costs [22–26]. Adversarial losses of GAN provide an alternative objective function to reflect the human auditory property, which can make the distribution of the enhanced speech close to that of the clean speech. To our knowledge, SEGAN [22] was the first attempt to apply GAN to the speech enhancement task, which used the noisy speech as a conditional information for a conditional GAN (cGAN) [27]. In [26], an approach to replace a vanilla GAN with advanced GAN, such as Wasserstein GAN (WGAN) [28] or relativistic standard GAN (RSGAN) [29] was proposed based on the SEGAN framework.

Even though the GAN-based speech enhancement techniques have been found successful, there still remain two important issues: (1) training instability and (2) a lack in considering various speech characteristics. Since GAN aims at finding the Nash equilibrium to solve a mini-max problem, it has been known that the training is usually unstable. A number of efforts have been devoted to stabilize the GAN training in image processing, by modifying the loss function [28] or the generator and discriminator structures [30,31]. However, in speech processing, this problem has not been extensively studied yet. Moreover, since most of the GAN-based speech enhancement techniques directly employ the models used in image generation, it is necessary to modify them to suit the inherent nature of speech. For instance, the GAN-based speech enhancement techniques [22,24,26] commonly used U-Net generator originated from image processing tasks. Since the U-net generator consisted of multiple CNN layers, it was insufficient to reflect the temporal information of speech signal. In regression-based speech enhancement, the modified U-net structure adding RNN layers to capture the temporal information of speech showed prominent performances [32]. In [33] for the speech synthesis, an alternative loss function depended on multiple sizes of window length and fast Fourier transform (FFT) was proposed and generated a good quality of speech, which also considered speech characteristics in frequency domain.

In this paper, we propose novel generator and discriminator structures for the GAN-based speech enhancement which reflect the speech characteristics while ensuring stable training. The conventional generator is trained to find a mapping function from the noisy speech to the clean speech by using sequential convolution layers, which is considered an ineffective approach especially for speech data. In contrast, the proposed generator progressively estimates the wide frequency range of the clean speech via a novel up-sampling layer.

In the early stage of GAN training, it is too easy for the conventional discriminator to differentiate real samples from fake samples for high-dimensional data. This often lets GAN fail to reach the equilibrium point due to vanishing gradient [30]. To address this issue, we propose a multi-scale discriminator that is composed of multiple sub-discriminators processing speech samples at different sampling rates. Even if the training is in the early stage, the sub-discriminators at low-sampling rates can not easily differentiate the real samples from the fake, which contributes to stabilize the training. Empirical results showed that the proposed generator and discriminator were successful in stabilizing GAN training and outperformed the conventional GAN-based speech enhancement techniques. The main contributions of this paper are summarized as follows:

- We propose a progressive generator to reflect the multi-resolution characteristics of speech.
- We propose a multi-scale discriminator to stabilize the GAN training without additional complex training techniques.
- The experimental results showed that the multi-scale structure is an effective solution for both deterministic and GAN-based models, outperforming the conventional GAN-based speech enhancement techniques.

The rest of the paper is organized as follows: In Section 2, we introduce GAN-based speech enhancement. In Section 3, we present the progressive generator and multi-scale discriminator. Section 4 describes the experimental settings and performance measurements. In Section 5, we analyze the experimental results. We draw conclusions in Section 6.

## 2. GAN-Based Speech Enhancement

An adversarial network models the complex distribution of the real data via a two-player mini-max game between a generator and a discriminator. Specifically, the generator takes a randomly sampled noise vector $z$ as input and produces a fake sample $G(z)$ to fool the discriminator. On the other hand, the discriminator is a binary classifier that decides whether an input sample is real or fake. In order to generate a realistic sample, the generator is trained to deceive the discriminator, while the discriminator is trained to distinguish between the real sample and $G(z)$. In an adversarial training process, the generator and the discriminator are alternatively trained to minimize their respective loss functions. The loss functions for the standard GAN can be defined as follows:

$$L_G = \mathbb{E}_{z\sim\mathbb{P}_z(z)}[log(1-D(G(z)))], \tag{1}$$

$$L_D = -\mathbb{E}_{x\sim\mathbb{P}_{clean}(x)}[log(D(x))] - \mathbb{E}_{z\sim\mathbb{P}_z(z)}[log(1-D(G(z)))] \tag{2}$$

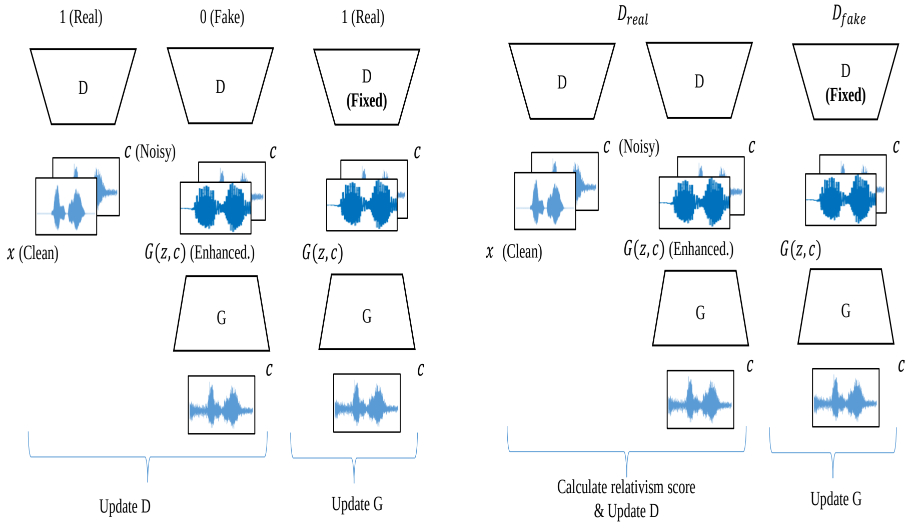
where $z$ is a randomly sampled vector from $\mathbb{P}_z(z)$ which is usually a normal distribution, and $\mathbb{P}_{clean}(x)$ is the distribution of the clean speech in the training dataset.

Since GAN was initially proposed for unconditional image generation that has no exact target, it is inadequate to directly apply GAN to speech enhancement which is a regression task to estimate the clean target corresponding to the noisy input. For this reason, GAN-based speech enhancement employs a conditional generation framework [27] where both the generator and discriminator are conditioned on the noisy waveform $c$ that has the clean waveform $x$ as the target. By concatenating the noisy waveform $c$ with the randomly sampled vector $z$, the generator $G$ can produce a sample that is closer to the clean waveform $x$. The training process of the cGAN-based speech enhancement is shown in Figure 1a, and the loss functions of the cGAN-based speech enhancement are

$$L_G = \mathbb{E}_{z\sim\mathbb{P}_z(z),c\sim\mathbb{P}_{noisy}(c)}[log(1-D(G(z,c),c))], \tag{3}$$

$$L_D = -\mathbb{E}_{x\sim\mathbb{P}_{clean}(x),c\sim\mathbb{P}_{noisy}(c)}[logD(x,c)] - \mathbb{E}_{z\sim\mathbb{P}_z(z),c\sim\mathbb{P}_{noisy}(c)}[log(1-D(G(z,c),c))] \tag{4}$$

where $\mathbb{P}_{clean}(x)$ and $\mathbb{P}_{noisy}(c)$ are respectively the distributions of the clean and noisy speech in the training dataset.



(**a**) cGAN-based speech enhancement (**b**) RSGAN-based speech enhancement

**Figure 1.** Illustration of the conventional GAN-based speech enhancements. In the training of cGAN-based speech enhancement, the updates for generator and discriminator are alternated over several epochs. During the update of the discriminator, the target of discriminator is 0 for the clean speech and 1 for the enhanced speech. For the update of the generator, the target of discriminator is 1 with freezing discriminator parameters. In contrast, the RSGAN-based speech enhancement trains the discriminator to measure a relativism score of the real sample $D_{real}$ and generator to increase that of the fake sample $D_{fake}$ with fixed discriminator parameters.

In the conventional training of the cGAN, both the probabilities that a sample is from the real data $D(x,c)$ and generated data $D(G(z,c),c)$ should reach the ideal equilibrium point 0.5. However, unlike the expected ideal equilibrium point, they both have a tendency to become 1 because the generator can not influence the probability of the real sample $D(x,c)$. In order to alleviate this problem, RSGAN [29] proposed a discriminator to estimate the probability that the real sample is more realistic than the generated sample. The proposed discriminator makes the probability of the generated sample $D(G(z,c),c)$ increase when that of the real sample $D(x,c)$ decreases so that both the probabilities could stably reach the Nash equilibrium state. In [26], the experimental results showed that, compared to other conventional GAN-based speech enhancements, the RSGAN-based speech enhancement technique improved the stability of training and enhanced the speech quality. The training process of the RSGAN-based speech enhancement is given in Figure 1b, and the loss functions of RSGAN-based speech enhancement can be written as:

$$L_G = -\mathbb{E}_{(x_r,x_f)\sim(\mathbb{P}_r,\mathbb{P}_f)}[log(\sigma(C(x_f) - C(x_r)))], \tag{5}$$

$$L_D = -\mathbb{E}_{(x_r,x_f)\sim(\mathbb{P}_r,\mathbb{P}_f)}[log(\sigma(C(x_r) - C(x_f)))] \tag{6}$$

where the real and fake data-pairs are defined as $x_r \triangleq (x,c) \sim \mathbb{P}_r$ and $x_f \triangleq (G(z,c),c) \sim \mathbb{P}_f$, and $C(x)$ is the output of the last layer in discriminator before the sigmoid activation function $\sigma(\cdot)$, i.e., $D(x) = \sigma(C(x))$.

In order to stabilize GAN training, there are two penalties commonly used: A gradient penalty for discriminator [28] and $L_1$ loss penalty for generator [24]. First, the gradient penalty regularization for discriminator is used to prevent exploding or vanishing gradients. This regularization penalizes the model if the $L_2$ norm of the discriminator gradient moves away from 1 to satisfy the Lipschitz constraint. The modified discriminator loss functions with the gradient penalty are as follows:

$$L_{GP}(D) = \mathbb{E}_{\tilde{x},c\sim\tilde{\mathbb{P}}}\Big[(||\nabla_{\tilde{x},c}C(\tilde{x},c))||_2 - 1)^2\Big], \tag{7}$$

$$L_{D-GP}(D) = -\mathbb{E}_{(x_r,x_f)\sim(\mathbb{P}_r,\mathbb{P}_f)}[log(\sigma(C(x_r) - C(x_f)))] + \lambda_{GP}L_{GP}(D) \tag{8}$$

where $\tilde{\mathbb{P}}$ is the joint distribution of $c$ and $\tilde{x} = \epsilon x + (1-\epsilon)\hat{x}$, $\epsilon$ is sampled from a uniform distribution in $[0,1]$, and $\hat{x}$ is the sample from $G(z,c)$. $\lambda_{GP}$ is the hyper-parameter that controls the gradient penalty loss and the adversarial loss of the discriminator.

Second, several prior studies [22–24] found that it is effective to use an additional loss term that minimizes the $L_1$ loss between the clean speech $x$ and the generated speech $G(z,c)$ for the generator training. The modified generator loss with the $L_1$ loss is defined as

$$L_1(G) = \|G(z,c) - x\|_1, \tag{9}$$

$$L_{G-L_1}(G) = -\mathbb{E}_{(x_r,x_f)\sim(\mathbb{P}_r,\mathbb{P}_f)}[log(\sigma(C(x_f) - C(x_r)))] + \lambda_{L_1}L_1(G) \tag{10}$$

where $\|\cdot\|_1$ is $L_1$ norm, and $\lambda_{L_1}$ is a hyper-parameter for balancing the $L_1$ loss and the adversarial loss of the generator.

## 3. Multi-Resolution Approach for Speech Enhancement

In this section, we propose a novel GAN-based speech enhancement model which consists of a progressive generator and a multi-scale discriminator. The overall architecture of the proposed model is shown in Figure 2, and the details of the progressive generator and the multi-scale discriminator are given in Figure 3.

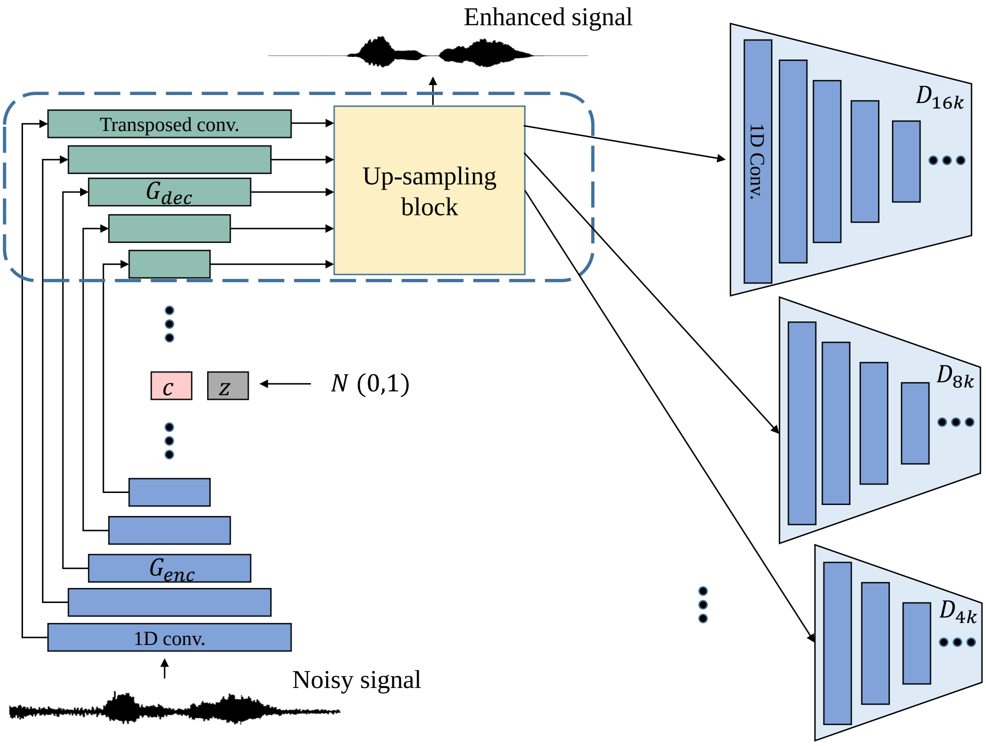


**Figure 2.** Overall architecture of the proposed GAN-based speech enhancement. The up-sampling block and the multiple discriminators $D_n$ are newly added, and the rest of the architecture is the same as that of [26]. The components within the dashed line will be addressed in Figure 3.

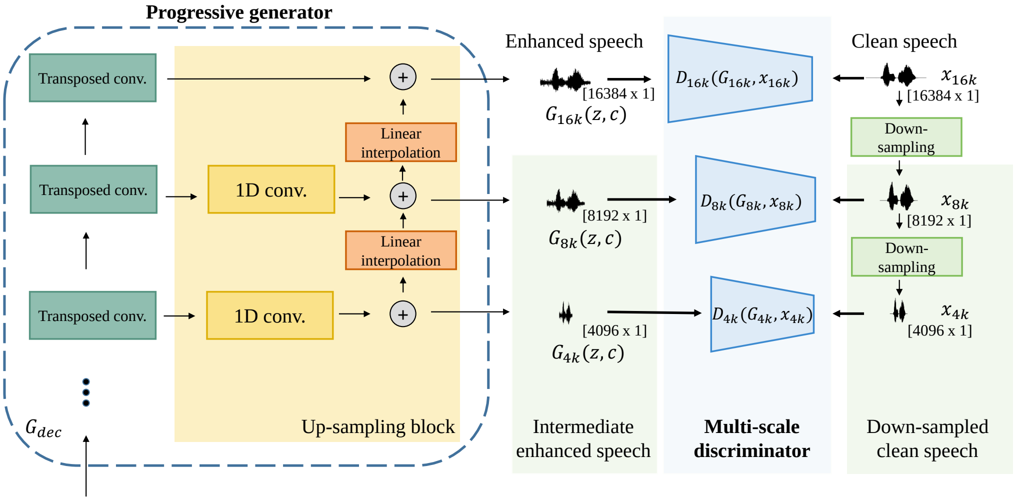


**Figure 3.** Illustration of the progressive generator and the multi-scale discriminator. Sub-discriminators calculate the relativism score $D_n(G_n, x_n) = \sigma(C_n(x_{r_n}) - C_n(x_{f_n}))$ at each layer. The figure is the case when $p, q = 4k$, but it can be extended for any $p$ and $q$. In our experiment, we covered that $p$ and $q$ were from $1k$ to $16k$.

### *3.1. Progressive Generator*

Conventionally, GAN-based speech enhancement systems adopt U-Net generator [22] which is composed of two components: An encoder $G_{enc}$ and a decoder $G_{dec}$. The encoder $G_{enc}$ consists of repeated convolutional layers to produce compressed latent vectors from a noisy speech, and the decoder $G_{dec}$ contains multiple transposed convolutional layers to restore the clean speech from the compressed latent vectors. These transposed convolutional layers in $G_{dec}$ are known to be able to generate low-resolution data from the compressed latent vectors, however, the capability to generate a high-resolution data is severely limited [30]. Especially in the case of speech data, it is difficult for the transposed convolutional layers to generate the speech with a high-sampling rate because it should cover a wide frequency range.

Motivated from the progressive GAN, which starts with generating low-resolution images and then progressively increases the resolution [30,31], we propose a novel generator that can incrementally widen the frequency band of the speech by applying an up-sampling block to the decoder $G_{dec}$. As shown in Figure 3, the proposed up-sampling block consists of 1D-convolution layers, element-wise addition, and liner interpolation layers. The up-sampling block yields the intermediate enhanced speech $G_n(z, c)$ at each layer through the 1D convolution layer and element-wise addition so that the wide frequency band of the clean speech is progressively estimated. Since a sampling rate is increased through the linear interpolation layer, it is possible to generate the intermediate enhanced speech at the higher layer while maintaining estimated frequency components at the lower layer. This incremental process is repeated until the sampling rate reaches the target sampling rate which is $16kHz$ in our experiment. Finally, we exploit the down-sampled clean speech $x_n$ processed by low-pass filtering and decimation as the target for each layer to provide multi-resolution loss functions. We define the real and fake data-pairs at different sampling rates as $x_{r_n} \triangleq (x_n, c_n) \sim \mathbb{P}_{r_n}$ and $x_{f_n} \triangleq (G_n(z,c), c_n) \sim \mathbb{P}_{f_n}$, and the proposed multi-resolution loss functions with $L_1$ loss are given as follows:

$$\begin{aligned} L_G(p) &= \sum_{\substack{n \geq p \\ n \in N_G}} L_{G_n} + \lambda_{L_1} L_1(G_n), N_G \in \{1k, 2k, 4k, 8k, 16k\}, \\ &= \sum_{\substack{n \geq p \\ n \in N_G}} -\mathbb{E}_{(x_{r_n}, x_{f_n}) \sim (\mathbb{P}_{r_n}, \mathbb{P}_{f_n})}[log(\sigma(C_n(x_{f_n}) - C_n(x_{r_n})))] + \lambda_{L_1} \|G_n(z,c) - x_n\|_1 \end{aligned} \tag{11}$$

where $N_G$ is the possible set of $n$ for the proposed generator, and $p$ is the sampling rate at which the intermediate enhanced speech is firstly obtained.

### 3.2. *Multi-Scale Discriminator*

When generating high-resolution image and speech data in the early stage of training, it is hard for the generator to produce a realistic sample due to the insufficient model capacity. Therefore, the discriminator can easily differentiate the generated samples from the real samples, which means that the real and fake data distributions do not have substantial overlap. This problem often causes training instability and even mode collapses [30]. For the stabilization of the training, we propose a multi-scale discriminator that consists of multiple sub-discriminators treating speech samples at different sampling rates.

As presented in Figure 3, the intermediate enhanced speech $G_n(z,c)$ at each layer restores the down-sampled clean speech $x_n$. Based on this, we can utilize the intermediate enhanced speech and down-sampled clean speech as the input to each sub-discriminator $D_n$. Since each sub-discriminator can only access limited frequency information depending on the sampling rate, we can make each sub-discriminator solve different levels of discrimination tasks. For example, discriminating the real from the generated speech is more difficult at the lower sampling rate than at the higher rate. The sub-discriminator at a lower sampling rate plays an important role in stabilizing the early stage of the training. As the training progresses, the role shifts upwards to the sub-discriminators at higher sampling rates. Finally, the proposed multi-scale loss for discriminator with gradient penalty is given by

$$\begin{aligned} L_D(q) &= \sum_{\substack{n \geq q \\ n \in N_D}} L_{D_n} + \lambda_{GP} L_{GP}(D_n), N_D \in \{1k, 2k, 4k, 8k, 16k\}, \\ &= \sum_{\substack{n \geq q \\ n \in N_D}} -\mathbb{E}_{(x_{r_n}, x_{f_n}) \sim (\mathbb{P}_{r_n}, \mathbb{P}_{f_n})}[log(\sigma(C_n(x_{r_n}) - C(x_{f_n})))] + \lambda_{GP} \mathbb{E}_{\widetilde{x_n}, c_n \sim \widetilde{\mathbb{P}_n}}[(|| \nabla_{\widetilde{x_n}, c_n} C(\widetilde{x_n}, c_n))||_2 - 1)^2] \end{aligned} \tag{12}$$

where $\widetilde{\mathbb{P}_n}$ is the joint distribution of the down-sampled noisy speech $c_n$ and $\widetilde{x_n} = \epsilon x_n + (1-\epsilon)\hat{x_n}$, $\epsilon$ is sampled from a uniform distribution in $[0,1]$, $x_n$ is the down-sampled clean speech, and $\hat{x_n}$ is the sample from $G_n(z,c)$. $N_D$ is the possible set of $n$ for the proposed discriminator, and $q$ is the minimum sampling rate at which the intermediate enhanced output was utilized as the input to a sub-discriminator for the first time. The adversarial losses $L_{D_n}$ are equally weighted.

## 4. Experimental Settings

### 4.1. *Dataset*

We used a publicly available dataset in [34] for evaluating the performance of the proposed speech enhancement technique. The dataset consists of 30 speakers from the Voice Bank corpus [35], and used 28 speakers (14 male and 14 female) for the training set (11572 utterances) and 2 speakers (one male and one female) for the test set (824 utterances). The training set simulated a total of 40 noisy conditions with 10 different noise sources (2 artificial and 8 from the DEMAND database [36]) at signal-to-noise ratios (SNRs) of 0, 5, 10, and 15 dB. The test set was created using 5 noise sources (living room, office, bus, cafeteria, and public square noise from the DEMAND database), which were different from the training noises, added at SNRs 2.5, 7.5, 12.5, and 17.5 dB. The training and test sets were down-sampled from 48 kHz to 16 kHz.

### *4.2. Network Structure*

The configuration of the proposed generator is described in Table 1. We used the U-Net structure with 11 convolutional layers for the encoder $G_{enc}$ and the decoder $G_{dec}$ as in [22,26]. Output shapes at each layer were represented by the number of temporal dimensions and feature maps. Conv1D in the encoder denotes a one-dimensional convolutional layer, and TrConv in the decoder means a transposed convolutional layer. We used approximately 1 s of speech (16384 samples) as the input to the encoder. The last output of the encoder was concatenated with a noise which had the shape of 8 × 1024 randomly sampled from the standard normal distribution $N(0,1)$. In [27], it was reported that the generator usually learns to ignore the noise prior $z$ in the CGAN, and we also observed a similar tendency in our experiments. For this reason, we removed the noise from the input, and the shape of the latent vector became 8 × 1024. The architecture of $G_{dec}$ was a mirroring of $G_{enc}$ with the same number and width of the filters per layer. However, skip connections from $G_{enc}$ made the number of feature maps in every layer to be doubled. The proposed up-sampling block $G_{up}$ consisted of 1D convolution layers, element-wise addition operations, and linear interpolation layers.

**Table 1.** Architecture of the proposed generator. Output shape represented temporal dimension and feature maps.

| Block | Operation | | Output Shape | |
|---|---|---|---|---|
| | Input | | 16,384 × 1 | |
| Encoder | Conv1D<br>(filterlength = 31, stide = 2) | | 8192 × 16 | |
| | | | 4096 × 32 | |
| | | | 2048 × 32 | |
| | | | 1024 × 64 | |
| | | | 512 × 64 | |
| | | | 256 × 128 | |
| | | | 128 × 128 | |
| | | | 64 × 256 | |
| | | | 32 × 256 | |
| | | | 16 × 512 | |
| | Latent vector | | 8 × 1024 | |
| Decoder | Trconv<br>(filterlength = 31, stide = 2) | | 16 × 1024 | |
| | | | 32 × 512 | |
| | | | 64 × 512 | |
| | | | 128 × 256 | |
| | | | 256 × 256 | |
| | | | 512 × 128 | |
| | Trconv<br>(filterlength = 31, stide = 2) | Conv1D<br>(filterlength = 17, stide = 1)<br>Element-wise addition<br>Linear interpolation layer | 1024 × 128 | 1024 × 1 |
| | | | 2048 × 64 | 2048 × 1 |
| | | | 4096 × 64 | 4096 × 1 |
| | | | 8192 × 32 | 8192 × 1 |
| | | | 16,384 × 1 | |

In this experiment, the proposed discriminator had the same serial convolutional layers as $G_{enc}$. The input to the discriminator had two channels of 16,384 samples, which were the clean speech and enhanced speech. The rest of the temporal dimension and feature-maps were the same as those of $G_{enc}$. In addition, we used LeakyReLU activation function without a normalization technique. After the last convolutional layers, there were a $1 \times 1$ convolution, and its output was fed to a fully-connected layer. To construct the proposed multi-scale discriminator, we used 5 different sub-discriminators, which were $D_{16k}, D_{8k}, D_{4k}, D_{2k}, and D_{1k}$ trained according to in Equation (12). Each sub-discriminator had a different input dimension depending on the sampling rate.

The model was trained using the Adam optimizer [37] for 80 epochs with 0.0002 learning rate for both the generator and discriminator. The batch size was 50 with 1-s audio signals that were sliced using windows of length 16,384 with 8192 overlaps. We also applied a pre-emphasis filter with impulse response $[-0.95, 1]$ to all training samples. For inference, the enhanced signals were reconstructed through overlap-add. The hyper-parameters to balance the penalty terms were set as $\lambda_{L_1}$ = 200 and $\lambda_{GP}$ = 10 such that they could match the dynamic range of magnitude with respect to the generator and discriminator losses. Note that we gave the same weight to the adversarial losses, $L_{G_n}$ and $L_{D_n}$, for all $n \in \{1k, 2k, 4k, 8k, 16k\}$. We implemented all the networks using Keras with Tensorflow [38] back-end using the public code (The SERGAN framework is available at https://github.com/deepakbaby/se_relativisticgan). All training was performed on single Titan RTX 24 GB GPU, and it took around 2 days.

### *4.3. Evaluation Methods*

#### 4.3.1. Objective Evaluation

The quality of the enhanced speech was evaluated using the following objective metrics:

- PESQ: Perceptual evaluation of speech quality defined in the ITU-T P.862 standard [19] (from $-0.5$ to 4.5),
- STOI: Short-time objective intelligibility [20] (from 0 to 1),
- CSIG: Mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal [39] (from 1 to 5),
- CBAK: MOS prediction of the intrusiveness of background noise [39] (from 1 to 5),
- COVL: MOS prediction of the overall effect [39] (from 1 to 5).

#### 4.3.2. Subjective Evaluation

To compare the subjective quality of the enhanced speech by baseline and proposed methods, we conducted two pairs of AB preference tests: AECNN versus the progressive generator and SERGAN versus the progressive generator with the multi-scale discriminator. Two speech in each pair were given in arbitrary order. For each listening test, 14 listeners participated, and 50 pairs of the speech were randomly selected. Listeners could listen to the speech pairs as many times as they wanted and were instructed to choose the speech with better perceptual quality. If the quality of the two samples was indistinguishable, listeners could select no preference.

## 5. Experiments and Results

In order to investigate the individual effect of the proposed generator and discriminator, we experimented on the progressive generator with and without the multi-scale discriminator. Furthermore, we plotted $L_1$ losses at each layer $L_1(G_n)$ to show that the proposed model makes training fast and stable. Finally, the performance of the proposed model is compared with that of the other GAN-based speech enhancement techniques.

### *5.1. Performance of Progressive Generator*

#### 5.1.1. Objective Results

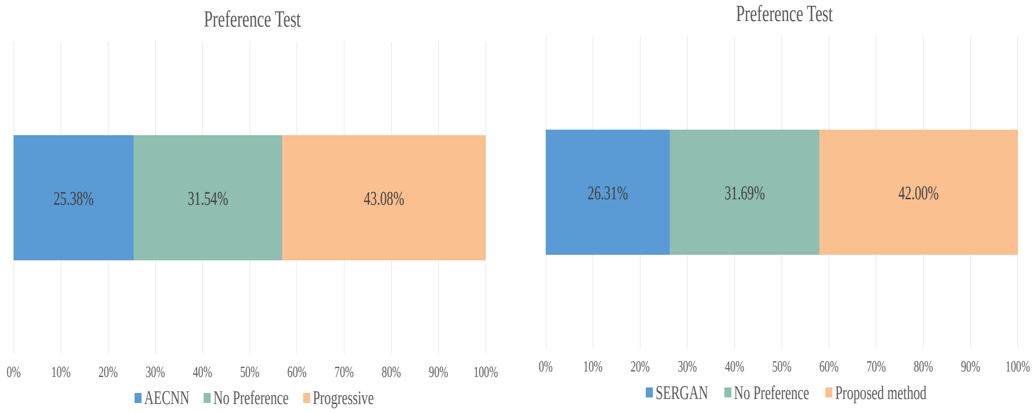
The purpose of these experiments is to show the effectiveness of the progressive generator. Table 2 presents the performance of the proposed generator when we minimized only the $L_1(G_n)$ in Equation (11). In order to better understand the influence of the progressive structure on the PESQ score, we conducted an ablation study with different $p$ in $\sum_{n \geq p} L_1(G_n)$. As illustrated in Table 2, compared to the auto-encoder CNN (AECNN) [26] that is the conventional U-net generator minimizing the $L_1$ loss only, the PESQ score of the progressive generator improved from 2.5873 to 2.6516. Furthermore, for the smaller $p$, we got a better PESQ score, and the best PESQ score was achieved when $p$ was the lowest, i.e., $1k$. For enhancing high-resolution speech, we verified that it is very useful to progressively generate intermediate enhanced speech while maintaining the estimated information obtained at lower sampling rate. We used the best generator $p = 1k$ in Table 2 for the subsequent experiments.

**Table 2.** Comparison of results between different architectures of the progressive generator. The best model is shown in bold type.

| **Model** | $\sum_{n \geq p} L_1(G_n)$ | **PESQ** |
|---|---|---|
| AECNN [26] | $p = 16k$ | 2.5873 |
| **Proposed** | $p = 8k$ | 2.6257 |
| | $p = 4k$ | 2.6335 |
| | $p = 2k$ | 2.6407 |
| | $\boldsymbol{p = 1k}$ | **2.6516** |

### 5.1.2. Subjective Results

The preference score of AECNN and the progressive generator was shown in Figure 4a. The progressive generator was preferred to AECNN in 43.08% of the cases, while the opposite preference was 25.38% (no preference in 31.54% of the cases). From the results, we verified that the proposed generator could produce the speech with not only higher objective measurements but also better perceptual quality.



(**a**) AECNN versus Progressive generator

(**b**) SERGAN versus Proposed method

**Figure 4.** Results of AB preference test. A subset of test samples used in the evaluation is accessible on a webpage https://multi-resolution-SE-example.github.io.

## 5.2. *Performance of Multi-Scale Discriminator*

### 5.2.1. Objective Results

The goal of these experiments is to show the efficiency of the multi-scale discriminator compared to the conventional single discriminator. As shown in Table 3, we evaluated the performance of the multi-scale discriminator while varying $q$ of the multi-scale loss $L_D(q)$ in Equation (12), which means varying the number of sub-discriminators. Compared to the baseline proposed in [26], the progressive generator with the single discriminator showed an improved PESQ score from 2.5898 to 2.6514. The multi-scale discriminators outperformed the single discriminators, and the best PESQ score of 2.7077 was obtained when $q = 4k$. Interestingly, we could observe that the performance was degraded when the $q$ became below $4k$. One possible explanation for this phenomenon would be that since the progressive generator faithfully generated the speech below the 4 kHz sampling rate, it was difficult for the discriminator to differentiate the fake from the real speech. This let the additional sub-discriminators a little bit useless for performance improvement.

**Table 3.** Comparison of results between different architectures of the multi-scale discriminator. Except for the SERGAN, the generator of all architectures used the best model in Table 2. The best model is shown in bold type.

| **Model** | **Generator** | **Discriminator** | $L_D(q)$ | **PESQ** | **RTF** |
|---|---|---|---|---|---|
| SERGAN [26] | U-net | Single | $q = 16k$ | 2.5898 | 0.008 |
| **Proposed** | Progressive | Single | $q = 16k$ | 2.6514 | 0.010 |
| | **Progressive** | **Multi-scale** | $q = 8k$ | 2.6541 | |
| | | | $\boldsymbol{q = 4k}$ | **2.7077** | |
| | | | $q = 2k$ | 2.6664 | |
| | | | $q = 1k$ | 2.6700 | |

### 5.2.2. Subjective Results

The preference scores of SERGAN and the progressive generator with multi-scale discriminator were shown in Figure 4b. The proposed method was preferred over SERGAN in 42.00% of the cases, while SERGAN was preferred in 26.31% of the cases (no preference in 31.69% of the cases). These results showed that the proposed method could enhance the speech with better objective metrics and subjective perceptual scores.

### 5.2.3. Real-Time Feasibility

SERGAN and the proposed method were evaluated in terms of the real-time factor(RTF) to verify the real-time feasibility, which is defined as the ratio of the time taken to enhance the speech to the duration of the speech (small factors indicate faster processing). CPU and graphic card used for the experiment were Intel Xeon Silver 4214 CPU 2.20 GHz and single Nvidia Titan RTX 24 GB. Since the generator of AECNN and SERGAN is the same, their RTF has the same value. Therefore, we only compared the RTF of SERGAN and the proposed method in Table 3. As the input window length was about 1 s of speech (16,384 samples), and the overlap was 0.5 s of speech (8192 samples), the total processing delay of all models can be computed by the sum of the 0.5 s and the actual processing time of the algorithm. In Table 3, we observed that the RTF of SERGAN and the proposed model was small enough for the semi-real-time applications. The similar value of the RTF for SEGAN and the proposed model also verified that adding the up-sampling network did not significantly increase the computational complexity.

## *5.3. Analysis and Comparison of Spectorgrams*

An example of the spectrograms of clean speech, noisy speech, and the enhanced speech by different models are shown in Figure 5. First, we focused on the black box to verify the effectiveness of the progressive generator. Before 0.6 s, a non-speech period, we could observe that the noise containing wide-band frequencies was considerably reduced since the progressive generator incrementally estimated the wide frequency range of the clean speech. Second, when we compared spectrograms of the multi-scale discriminator and that of the single discriminator, the different pattern was presented in the red box. The multi-scale discriminator was able to suppress more noise than the single discriminator in the non-speech period. We could confirm that the multi-scale discriminator selectively reduced high-frequency noise in a speech period as the sub-discriminators in multi-scale discriminator differentiate the real and fake speech at the different sampling rates.

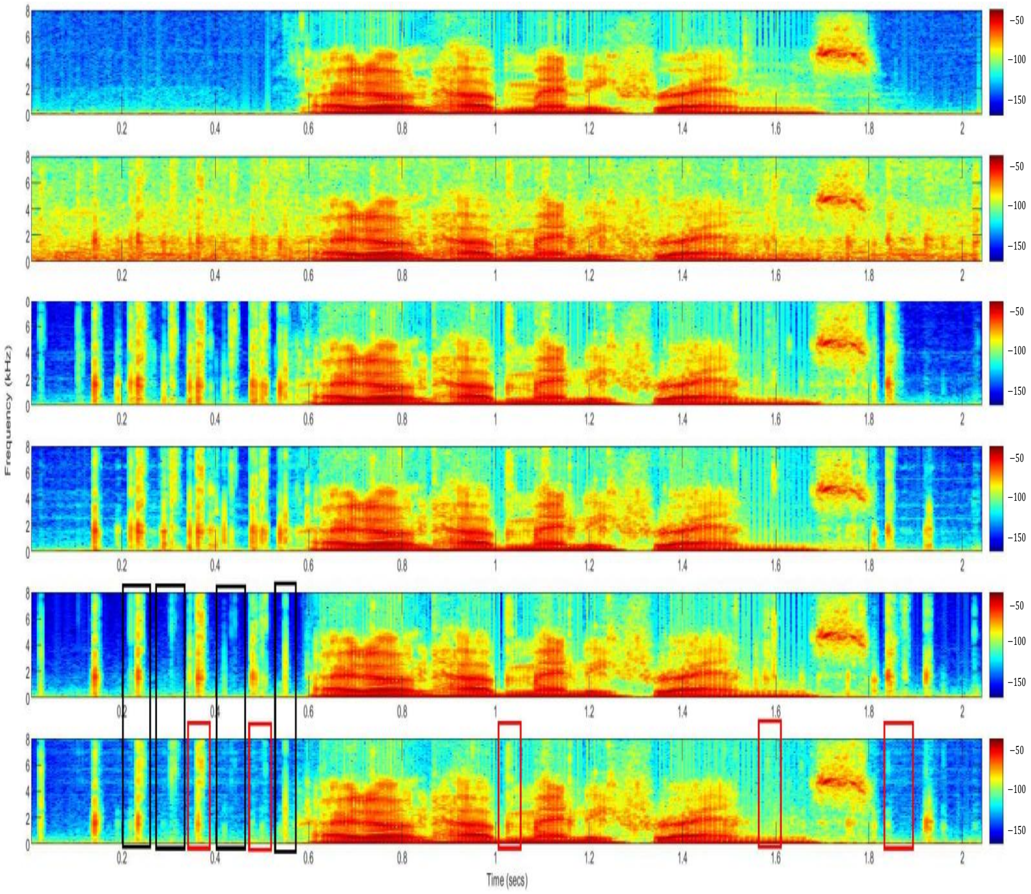


**Figure 5.** Spectrograms from the top to the bottom correspond to clean speech, noisy speech, enhanced speech by AECNN, SERGAN, progressive generator, progressive generator with multi-scale discriminator, respectively.

### 5.4. Fast and Stable Training of Proposed Model

To analyze the learning behavior of the proposed model in more depth, we plotted $L_1(G_n)$ in Equation (11) obtained from the best model in Table 3 and SERGAN [26] during the whole training periods. As the clean speech was progressively estimated by the intermediate enhanced speech, the stable convergence behavior of $L_1(G_n)$ was shown in Figure 6. With the help of $L_1(G_n)$ at low layers ($n = 1, 2, 4, 8$), $L_1(G_{16k})$ for the proposed model decreased faster and more stable than that of SERGAN. From the results, we can convince that the proposed model accelerates and stabilizes the GAN training.

### 5.5. Comparison with Conventional GAN-Based Speech Enhancement Techniques

Table 4 shows the comparison with other GAN-based speech enhancement methods that have the E2E structure. The GAN-based enhancement techniques which were evaluated in this experiment are as follows: **SEGAN** [22] has the U-net structure with conditional GAN. Similar to the structure of SEGAN, **AECNN** [26] is trained to only minimize $L_1$ loss, and **SERGAN** [26] is based on relativistic GAN. **CP-GAN** [40] has modified the generator and discriminator of SERGAN to utilize contextual information of the speech. The progressive generator without adversarial training even showed better results than CP-GAN on PESQ and CBAK. Finally, the progressive generator with the multi-scale discriminator outperformed the other GAN-based speech enhancement methods for three metrics.

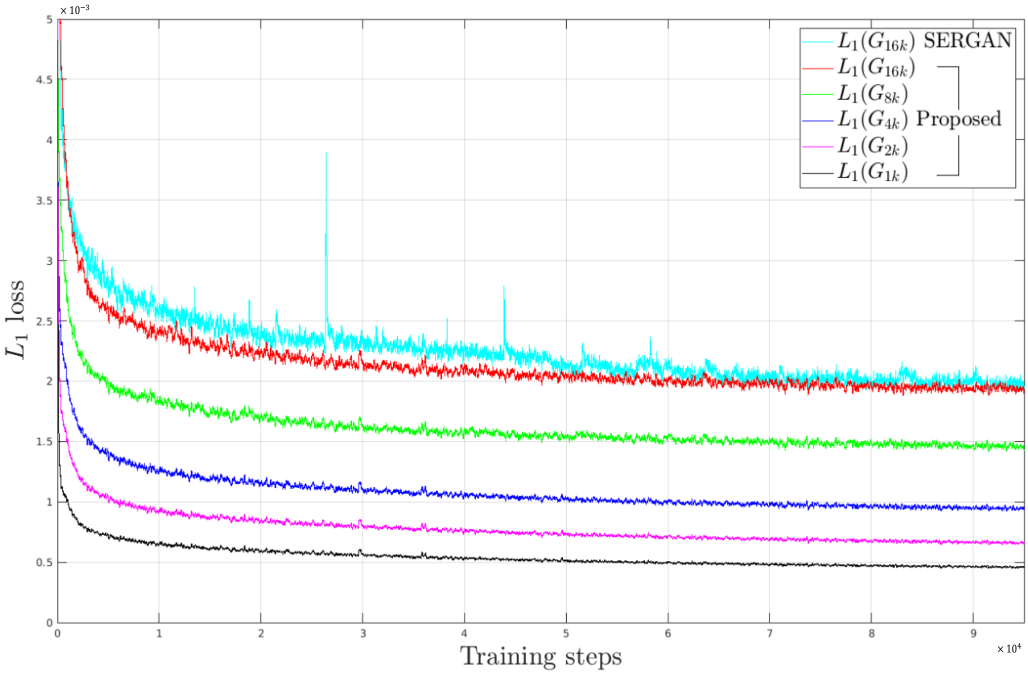


**Figure 6.** Illustration of $L_1(G_n)$ as a function of training steps.

**Table 4.** Comparison of results between different GAN-based speech enhancement Techniques. The best result is highlighted in bold type.

| Model | PESQ | CSIG | CBAK | COVL | STOI |
|---|---|---|---|---|---|
| Noisy | 1.97 | 3.35 | 2.44 | 2.63 | 0.91 |
| SEGAN [24] | 2.16 | 3.48 | 2.68 | 2.67 | 0.93 |
| AECNN [26] | 2.59 | 3.82 | **3.30** | 3.20 | 0.94 |
| SERGAN [26] | 2.59 | 3.82 | 3.28 | 3.20 | 0.94 |
| CP-GAN [38] | 2.64 | 3.93 | 3.29 | 3.28 | 0.94 |
| The progressive generator without adversarial training | 2.65 | 3.90 | **3.30** | 3.27 | 0.94 |
| The progressive generator with the multi-scale discriminator | **2.71** | **3.97** | 3.26 | **3.33** | 0.94 |

## 6. Conclusions

In this paper, we proposed a novel GAN-based speech enhancement technique utilizing the progressive generator and multi-scale discriminator. In order to reflect the speech characteristic, we introduced a progressive generator which can progressively estimate the wide frequency range of the speech by incorporating an up-sampling layer. Furthermore, for accelerating and stabilizing the training, we proposed a multi-scale discriminator which consists of a number of sub-discriminators operating at different sampling rates.

For performance evaluation of the proposed methods, we conducted a set of speech enhancement experiments using the VoiceBank-DEMAND dataset. From the results, it was shown that the proposed technique provides a more stable GAN training while showing consistent performance improvement on objective and subjective measures for speech enhancement. We also checked the semi-real-time feasibility by observing a small increment of RTF between the baseline generator and the progressive generator.

As the proposed network mainly focused on the multi-resolution attribute of speech in the time domain, one possible future study is to expand the proposed network to utilize the multi-scale attribute of speech in the frequency domain. Since the progressive generator and multi-scale discriminator can also be applied to the GAN-based speech reconstruction models such as neural vocoder for speech synthesis and codec, we will study the effects of the proposed methods.

**Author Contributions:** Conceptualization, H.Y.K. and N.S.K.; methodology, H.Y.K. and J.W.Y.; software, H.Y.K. and J.W.Y.; validation, H.Y.K. and N.S.K.; formal analysis, H.Y.K.; investigation, H.Y.K. and S.J.C.; resources, H.Y.K. and N.S.K.; data curation, H.Y.K. and W.H.K.; writing—original draft preparation, H.Y.K.; writing—review and editing, J.W.Y., W.H.K., S.J.C., and N.S.K.; visualization, H.Y.K.; supervision, N.S.K.; project administration, N.S.K.; funding acquisition, N.S.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFCIT1701-04.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. Benesty, J.; Makino, S.; Chen, J.D. *Speech Enhancement*; Springer; New York, NY, USA, 2007.
2. Boll, S.F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
3. Lim, J.S.; Oppenheim, A.V. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE* **1979**, *67*, 1586–1604. [CrossRef]
4. Scalart, P. Speech enhancement based on a priori signal to noise estimation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, GA, USA, 7–10 May 1996; pp. 629–632.
5. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1109–1121. [CrossRef]
6. Kim, N.S.; Chang, J.H. Spectral enhancement based on global soft decision. *IEEE Signal Process. Lett.* **2000**, *7*, 108–110.
7. Kwon, K.; Shin, J.W.; Kim, N.S. NMF-based speech enhancement using bases update. *IEEE Signal Process. Lett.* **2015**, *22*, 450–454. [CrossRef]
8. Wilson, K.; Raj, B.; Smaragdis, P.; Divakaran, A. Speech denoising using nonnegative matrix factorization with priors. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV, USA, 30 March–4 April 2008; pp. 4029–4032.
9. Xu, Y.; Du, J.; Dai, L.R.; Lee, C.H. A regression approach to speech enhancement based on deep neural networks. *IEEE Trans. Audio Speech Lang. Process.* **2015**, *23*, 7–19. [CrossRef]
10. Grais, E.M.; Sen, M.U.; Erdogan, H. Deep neural networks for single channel source separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3734–3738.
11. Zhao, H.; Zarar, S.; Tashev, I.; Lee, C.H. Convolutional-recurrent neural networks for speech enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2401–2405.
12. Huang, P.S.; Kim, M.; Hasegawa. J.M.; Smaragdis, P. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE Trans. Audio Speech Lang. Process.* **2015**, *23*, 2136–2147. [CrossRef]
13. Roux, J.L.; Wichern, G.; Watanabe, S.; Sarroff, A.; Hershey, J. The Phasebook: Building complex masks via discrete representations for source separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 66–70.
14. Wang, Z.; Tan, K.; Wang, D. Deep Learning based phase reconstruction for speaker separation: A trigonometric perspective. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 71–75.
15. Wang, Z.; Roux, J.L.; Wang, D.; Hershey, J. End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction. *arXiv* **2018**, arXiv:1804.10204.
16. Pandey, A.; Wang, D. A new framework for supervised speech enhancement in the time domain. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 1136–1140.
17. Stoller, D.; Ewert, S.; Dixon, S. Wave-U-net: A multi-scale neural network for end-to-end audio source separation. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 23–27 September 2018; pp. 334–340.
18. Rethage, D.; Pons, J.; Xavier, S. A wavenet for speech denoising. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5069–5073.

19. ITU-T. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. Rec. ITU-T P. 862; 2000. Available online: https://www.itu.int/rec/T-REC-P.862 (accessed on 18 February 2019).
20. Jensen, J.; Taal, C.H. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE Trans. Audio Speech Lang. Process.* **2016**, *24*, 2009–2022. [CrossRef]
21. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
22. Pascual, S.; Bonafonte, A.; Serrà, J. SEGAN: Speech enhancement generative adversarial network. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 3642–3646.
23. Soni, M.H.; Shah, N.; Patil, H.A. Time-frequency masking-based speech enhancement using generative adversarial network. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5039–5043.
24. Pandey, A.; Wang, D. On adversarial training and loss functions for speech enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5414–5418
25. Fu, S.-W.; Liao, C.-F.; Yu, T.; 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 September 2019.
26. Baby, D.; Verhulst, S. Sergan: speech enhancement using relativistic generative adversarial networks with gradient penalty. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 106–110.
27. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1125–1134.
28. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans, In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5769–5779.
29. Jolicoeur-Martineau, A. The Relativistic Discriminator: A Key Element Missing from Standard GAN. *arXiv* **2018**, arxiv:1807.00734.
30. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* **2018**, arxiv:1710.10196.
31. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of styleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8107–8116.
32. Alexandre, D.; Gabriel, S.; Yossi, A. Real time speech enhancement in the Waveform domain. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 3291–3295.
33. Yamamoto, R.; Song, E.; Kim, J.-M. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.
34. 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 International Symposium on Computer Architecture, Seoul, Korea, 18–22 June 2016; pp. 146–152.
35. 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 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.
36. 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.
37. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
38. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467.
39. Hu, Y.; Loizou, P.C. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 229–238. [CrossRef]
40. Liu, G.; Gong, K.; Liang, X.; Chen, Z. CP-GAN: Context pyramid generative adversarial network for speech enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6624–6628.