INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

SCP-GAN: Self-Correcting Discriminator Optimization for Training Consistency Preserving Metric GAN on Speech Enhancement Tasks

Vasily Zadorozhnyy[1,*], *Qiang Ye*[1,†], *Kazuhito Koishida*[2]

[1] Department of Mathematics, University of Kentucky, Lexington, USA
[2] Applied Sciences Group, Microsoft Corporation, Redmond, USA

vasily.zadorozhnyy@uky.edu, qye3@uky.edu, kazukoi@microsoft.com

Abstract

In recent years, Generative Adversarial Networks (GANs) have produced significantly improved speech enhancement (SE) task results. However, they are challenging to train. In this work, we introduce several improvements to GAN training schemes, which can be applied to most GAN-based SE models. We propose using consistency loss functions, which target the inconsistency in time and time-frequency domains caused by Fourier and Inverse Fourier Transforms. We also present self-correcting optimization for training a GAN discriminator on SE tasks which helps avoid "harmful" training directions for parts of the discriminator loss function. We have tested our proposed methods on several state-of-the-art GAN-based SE models and obtained consistent improvements, including new state-of-the-art results for the Voice Bank+DEMAND dataset.

Index Terms: Speech Enhancement, GAN, MetricGAN, Self-Correcting Optimization, STFT Consistency, Voice Bank+DEMAND

1. Introduction

Speech Enhancement (SE) is a process of making deteriorated speech signals more understandable and perceptually pleasing. The SE has been widely used for various applications, including mobile communication, speech recognition systems, hearing aids, etc. SE as an area of research interest has been around for several decades. Traditional SE techniques [1, 2] often use a heuristic or straightforward signal processing algorithm to estimate a gain function, which is then applied to the noisy input to produce improved speech. Recent developments in deep learning have inspired many Deep Neural Network (DNN)-based SE techniques [3, 4, 5, 6, 7] that outperform conventional signal processing-based methods. One particular DNN-based architecture, Generative Adversarial Network (GAN), has garnered much interest in the SE community for the past few years [5, 6, 8, 9]. In the applications of SE, GAN architecture is primarily employed to generate enhanced speech. One of the earliest works where GAN models were implemented on the SE domain is the SEGAN [5] model. It utilizes an adversarial framework to map the noisy waveform to a corresponding enhanced speech. Later, MetricGAN [6] introduced a metric score optimization scheme, where an evaluated metric was introduced into adversarial loss functions, replacing a traditional binary-classifier [5] and creating a new branch for SE GAN-based research. There have been several improvements to the MetricGAN model, e.g., MetricGAN+ [8], iMetricGAN [10], CMGAN [9], etc. More recently, with a rise of Transformers [11] and Conformers [12], models such as DB-AIAT [13], DPT-FSNet [14], SE-Conformer [15], CMGAN [9], etc. show significant improvements on SE tasks.

Despite much work, training of GAN-based models is prone to problems such as non-convergence, overfitting, and gradient instabilities. One common issue in GAN's discriminator training is a potentially "harmful" gradient direction [16] where parts of the model might train opposite to the desired direction. To overcome this problem, we propose a new method called Self-Correcting (SC) Discriminator Optimization. At the same time, the SE DNN-based models are subject to problems caused by the signal-processing tools, e.g., an inconsistency in the Short-Time Fourier Transform (STFT) and its inverse (iSTFT) [7, 17]. Inspired by [18], we adapt and introduce the consistency loss function as a part of Consistency Preserving (CP) Net into the GAN framework, where loss and architecture take into account the iSTFT effects. From our experiments, the combination of SC and CP methods improves the SE GAN-based models even further than either method; we call such a combination SCP-GAN.

The remainder of this paper is as follows. In section 2, we list earlier works pertinent to our current work. In section 3, we introduce improvements to current GAN-based SE models. We present and compare the SCP-GAN results on Voice Bank+DEMAND dataset [19] to the current state-of-the-art (SOTA) models in section 4. Then, in section 5, we provide an extensive ablation study to show the advantages of the proposed methods. Finally, in section 6, we highlight the methods' contributions to the field.

2. Related Work

2.1. Adaptively Weighted GAN (awGAN)

The discriminator plays a very important role in training GAN-based models. However, optimizing the discriminator loss function(s) has been a challenge [16]. In the image generation domain, most discriminator loss functions have a form of two equally weighted parts, where one of these parts only relies on the original dataset. The second part depends on the generator network, its output, and not the original data, [16] calls them 'real' and 'fake' parts, respectively. However, the training with an equally weighted discriminator loss function is not performed equally on the real and fake parts, but it depends on the angle between real and fake gradients and their magnitudes. Under such conditions, the actual training direction might end up in the opposite direction to either real or fake gradients, which is undesirable as it can cause issues with convergence and stability [16]. To solve such issue [16] proposed the method of adaptive weights for the discriminator loss function and the algorithm for choosing such weights on image generating tasks.

* Supported by Microsoft during internship.
† Supported by NSF grants DMS-2208314 and DMS-1821144.

10.21437/Interspeech.2023-456

2.2. STFT Consistencies in SE DNN models

The short-time Fourier transform (STFT) is one of the most fundamental and widely used methods in audio signal processing. Most DNN-based SE models [7, 9, 17] use a complex-valued STFTs generator to suppress noise and preserve speech. However, using STFT methods has its issues. One of those issues is the STFT consistency. This is an issue when a loss function does not consider iSTFT signal reconstruction. Several works have been done to resolve this issue. [17] presented an algorithm for a phase reconstruction based on a local approximation of the consistency constraints. Adding simple differentiable projection layers to the enhancement DNN to solve the issue was proposed by [7]. More recently, [18] introduced the iSTFT into back-propagation methods for SE DNN-based models.

3. SCP-GAN

We propose the following two innovative learning strategies to enhance the performance of SE GAN-based models.

3.1. Self-Correcting Discriminator Optimization

Notation: The angle between two gradients $\nabla\mathcal{L}_\alpha$ and $\nabla\mathcal{L}_\beta$ is defined as $\angle_2\left(\nabla\mathcal{L}_\alpha, \nabla\mathcal{L}_\beta\right) = \cos^{-1}\left(\frac{\langle\nabla\mathcal{L}_\alpha, \nabla\mathcal{L}_\beta\rangle_2}{||\nabla\mathcal{L}_\alpha||_2||\nabla\mathcal{L}_\beta||_2}\right)$, where $\langle\cdot,\cdot\rangle_2$ and $||\cdot||_2$ denote the Euclidean inner product and the Euclidean 2-norm, respectively.

We introduce the Self-Correcting (SC) Discriminator Optimization method, a generalization of the method from [16] to the SE domain. A large number of existing SE GAN-based models have the discriminator loss function consisting of either two [6, 9] or three [8] equally weighted parts:

$$\mathcal{L}_D = \mathcal{L}_C + \mathcal{L}_E \tag{1}$$

$$\mathcal{L}_D = \mathcal{L}_C + \mathcal{L}_E + \mathcal{L}_N, \tag{2}$$

where $\mathcal{L}_C$, $\mathcal{L}_E$, and $\mathcal{L}_N$ exclusively rely on clean, enhanced, and noisy datasets, respectively. For example, MetricGAN [6] has a two-part discriminator loss (i.e. Eq. (1)) with

$$\mathcal{L}_C = \mathbb{E}_y\left(D(y,y) - Q(y,y)\right)^2 \tag{3}$$

$$\mathcal{L}_E = \mathbb{E}_{x,y}\left(D(G(x),y) - Q(G(x),y)\right)^2, \tag{4}$$

and, MetricGAN+ [8] also uses the $\mathcal{L}_N$ (i.e. Eq. (2)) such as

$$\mathcal{L}_N = \mathbb{E}_{x,y}\left(D(x,y) - Q(x,y)\right)^2. \tag{5}$$

Above, x is a noisy signal, y is its corresponding clean version, and $D(\cdot,\cdot)$, $G(\cdot)$, and $Q(\cdot,\cdot)$ are the discriminative model, generative model, and evaluation metric function, respectively. Moreover, notations $\mathbb{E}_y$ and $\mathbb{E}_{x,y}$ denote the expectation over $\{y\}$ and $\{(x,y)\}$, respectively. In such a setup as [6], G only takes the noisy signal while both D and Q take two inputs, either (y,y) (as in (3)) or $(G(x),y)$ (as in (4)) for training D to approximate Q on clean and enhanced signals, respectively.

However, gradient descent training with $\nabla\mathcal{L}_D$ is not performed equally on clean and enhanced parts; its effect depends on the angle between $\nabla\mathcal{L}_C$, $\nabla\mathcal{L}_E$, and $\nabla\mathcal{L}_N$ (if used) and their magnitudes. For example in (1), if the angle between $\nabla\mathcal{L}_C$ and $\nabla\mathcal{L}_E$ is a large obtuse angle and $||\nabla\mathcal{L}_C||_2 >> ||\nabla\mathcal{L}_E||_2$, then $\nabla\mathcal{L}_D$ would make an obtuse angle with $\nabla\mathcal{L}_E$ and thus training along $\nabla\mathcal{L}_D$ would increase the loss $\mathcal{L}_E$, which would be undesirable (or even harmful) to the enhanced part of the model.

To address this issue, as in [16] for GAN, we introduce weights into the two-part discriminator loss in Eq. (1),

$$\mathcal{L}_D^{SC} = w_C\mathcal{L}_C + w_E\mathcal{L}_E \tag{6}$$

and call it SC_2 - Self-Correcting two terms discrimination loss. Moreover, we introduce weights into the three-part discriminator in Eq. (2)

$$\mathcal{L}_D^{SC} = w_C\mathcal{L}_C + w_E\mathcal{L}_E + w_N\mathcal{L}_N \tag{7}$$

and call it SC_3 - Self-Correcting three terms discrimination loss. We choose the weights so that $\nabla\mathcal{L}_D^{SC}$ does not make an obtuse angle with any of $\nabla\mathcal{L}_C$, $\nabla\mathcal{L}_E$, or $\nabla\mathcal{L}_N$. While for the SC_2 method, weights can be easily generalized from the aw-GAN algorithm in [16]; for the SC_3 method, determination of the weights is much more complicated involving many cases. We have analyzed all possible cases and derived corresponding formulas for each scenario. The precise mathematical statement with proof is provided in *Theorem 1* in the *Supplementary Material*. There are a total of 7 general cases (up to symmetry and a pick of direction(s) that we want to prioritize); however, these cases can be categorized into four groups:

1. all angles are acute
2. one obtuse angle
3. two obtuse angles
4. all angles are obtuse

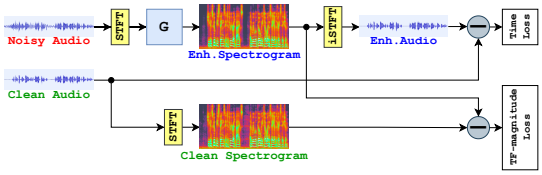
Some of the above cases require non-trivial projections into desirable subspaces with substantial computations.

It is important to note that the MetricGAN discriminator loss function(s) has little in common with traditional loss functions [16] (e.g., approximation of the metric vs. binary classification). Moreover, SE GAN's behavior differs from image-generating GAN since one takes a noisy counterpart of clean speech and the other takes a random sample from a simple distribution. These differences affect their behavior and must be considered when designing algorithms for adaptive decision-making regarding which parts of the loss to prioritize, how to rotate gradients, etc. With that in mind, we propose Algorithm 1, which determines weights for (6) and (7) using $\nabla\mathcal{L}_C$, $\nabla\mathcal{L}_E$, or $\nabla\mathcal{L}_N$, resulting in a self-correcting discriminator gradient. Similar to [16], the goal of Algorithm 1 is to minimize the potential harm to parts of the model by ensuring the training gradient $\nabla\mathcal{L}_D^{SC}$ does not go obtuse to $\nabla\mathcal{L}_C$, $\nabla\mathcal{L}_E$, or $\nabla\mathcal{L}_N$.

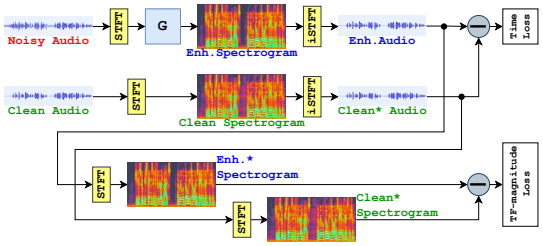
Algorithm 1: Self-Correcting Discriminator Method

Compute: $\nabla\mathcal{L}_C, \nabla\mathcal{L}_E, \nabla\mathcal{L}_N$ **if** $\mathcal{L}_N$ is used
if $\angle_2\left(\nabla\mathcal{L}_C, \nabla\mathcal{L}_E\right) < 90^\circ$ **then**
  $w_C = 1$ and $w_E = 1$
  if $\mathcal{L}_N$ *is used* **and** $\angle_2\left(w_C\nabla\mathcal{L}_C + w_E\nabla\mathcal{L}_E, \nabla\mathcal{L}_N\right) < 90^\circ$ **then**
    $w_N = 1$
  else
    $w_N = -\frac{\langle\nabla\mathcal{L}_C, \nabla\mathcal{L}_N\rangle_2}{||\nabla\mathcal{L}_N||_2^2} - \frac{\langle\nabla\mathcal{L}_E, \nabla\mathcal{L}_N\rangle_2}{||\nabla\mathcal{L}_N||_2^2}$
  end
else
  $w_C = 1$ and $w_E = -\frac{\langle\nabla\mathcal{L}_C, \nabla\mathcal{L}_E\rangle_2}{||\nabla\mathcal{L}_E||_2^2}$
  if $\mathcal{L}_N$ *is used* **and** $\angle_2\left(w_C\nabla\mathcal{L}_C + w_E\nabla\mathcal{L}_E, \nabla\mathcal{L}_N\right) < 90^\circ$ **then**
    $w_N = 1$
  else
    $w_N = -\frac{\langle\nabla\mathcal{L}_C, \nabla\mathcal{L}_N\rangle_2}{||\nabla\mathcal{L}_N||_2^2} + \frac{\langle\nabla\mathcal{L}_C, \nabla\mathcal{L}_E\rangle_2\langle\nabla\mathcal{L}_E, \nabla\mathcal{L}_N\rangle_2}{||\nabla\mathcal{L}_E||_2^2||\nabla\mathcal{L}_N||_2^2}$
  end
end

Algorithm 1 is designed using Theorem 1 to ensure that the angle between $\nabla\mathcal{L}_C$ and $\nabla\mathcal{L}_E$ never becomes obtuse while prioritizing the $\nabla\mathcal{L}_C$ direction, which is particularly important in earlier training since the model has not seen much clean data. Towards the end of the training, $\nabla\mathcal{L}_E$ leans towards $\nabla\mathcal{L}_C$ (in both direction and magnitude) as enhanced examples become less noisy; therefore, prioritization of direction would not be necessary. With the add-on of the noisy part, we want to ensure that $\nabla\mathcal{L}_N$ never goes in the opposite direction to $w_C\nabla\mathcal{L}_C +$



(a) *Process of computing Time and TF-magnitude losses inside the GAN-based SE model Generator (G)*



(b) ***Consistency Preserving (CP) Net:*** *Depiction of the process for computing Time and TF-magnitude losses with CP method inside the GAN-based SE model Generator (G)*

Figure 1: *Traditional vs. Consistency Preserving SE GAN-based models*

$w_E \nabla \mathcal{L}_E$ (after the rotation), which is already a strong direction since the 2-term loss model performs at a good level. Figure 2 depicts the aforementioned behavior of $\angle_2 (\nabla \mathcal{L}_N, \nabla \mathcal{L}_E)$ and $\angle_2 (\nabla \mathcal{L}_C, \nabla \mathcal{L}_E)$ during training.

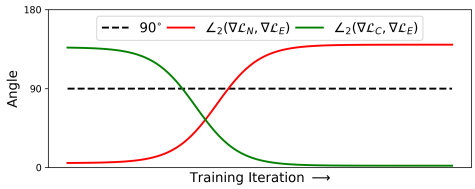


Figure 2: *The behavior of $\angle_2 (\nabla \mathcal{L}_N, \nabla \mathcal{L}_E)$ and $\angle_2 (\nabla \mathcal{L}_C, \nabla \mathcal{L}_E)$ during the SE MetricGAN's training*

3.2. Consistency Preserving Network

Most GAN-based SE models [6, 9, 13, 14] have a generator (G) that accepts the STFT spectrogram of a noisy waveform as input. The G's output is an enhanced spectrogram that later uses iSTFT to produce the enhanced waveform; Figure 1a illustrates the process. The G is then updated using a combination of various loss functions, e.g., Time Loss [20], TF-magnitude Loss [21], Adversarial Metric Loss [6], etc. For example, the TF-magnitude Loss [21] is computed between the enhanced and clean spectrograms; see Figure 1a. However, such loss and architectural setup do not consider the effect of the iSTFT reconstruction, which causes inconsistencies between signals.

We incorporate the idea from [18] to SE GAN-based models by modifying architecture and loss function(s) such that any input into a loss function (including the Adversarial Loss) undergoes the same process, taking into consideration the effects of signal reconstruction from the spectrogram; we call such process and loss function a Consistency Preserving (CP) Network and a consistency loss, respectively. Particularly, the CP Net ensures that the same number of STFT and iSTFT transforms are applied on clean, enhanced, and noisy (if used) signals and avoids distortion(s) that could happen on the ends of the audio segments, where the edge regions have insufficient data to reconstruct a signal from the spectrogram with the overlap-add operation. Our approach addresses this issue using the same STFT-iSTFT process and avoids such unexpected behavior. Figure 1b depicts the process of computing Time and TF-magnitude losses using the proposed CP method by ensuring that the same transforms are applied on enhanced, clean, and noisy (if used) signals.

Note: The Clean Audio to the Clean* Audio process inside the CP Net (2$^{\text{nd}}$ row from the top of Figure 1b) can be performed at the data preprocessing stage.

4. Experiments

4.1. Dataset

We use the publicly accessible Voice Bank+DEMAND [19] dataset to evaluate and compare our proposed SCP-GAN method. The training set of the Voice Bank+DEMAND dataset consists of 11,572 individual recordings of 28 speakers from the Voice Bank corpus [24], which are mixed with DEMAND [25] database and some artificial background noises at the signal-to-noise ratios (SNRs) of 0, 5, 10, and 15 dB. The test set has 824 utterances of two speakers from the Voice Bank corpus, which are mixed with unseen DEMAND noises at the SNRs of 2.5, 7.5, 12.5, and 17.5 dB. All utterances were resampled to 16kHz; in addition, for all the experiments, the frame length was set to 100-sample and the frame rate of the STFT was 160, following setups from [8, 9].

4.2. Evaluation Metrics

To assess the speech quality, we select a set of widely used metrics, including the Perceptual Evaluation of Speech Quality (PESQ) [26] (ranging between -0.5 and 4.5), the Segmental Signal-to-Noise Ratio (SSNR), the Short-Time Objective Intelligibility (STOI) [27] (with a range 0 to 1), and three Mean Opinion Score (MOS) [28] based metrics: the MOS prediction of the signal distortion (CSIG), the MOS prediction of the intrusiveness of background noise (CBAK), and MOS prediction of the overall effect (COVL) (MOS metrics range between 1 and 5). For all metrics, higher numbers denote better performance.

4.3. Experimental Results

We have applied our proposed methods to two baseline models: a widely-used MetricGAN+ [8] model and a current SOTA model - CMGAN [9]. Our SCP method shows a consistent improvement over the compared baseline. On the MetricGAN+ model, our SCP-MetricGAN+ improved by 0.04, 0.06, 0.04, and 0.01 on the PESQ, CSIG, CBAK, and COVL metrics, respectively. Our improvements with the SCP-CMGAN model are 0.11, 0.12, 0.03, and 0.13 on the same scores. Moreover, we have compared our method to other recent SOTA models which can be seen in Table 1.

Note: Results provided in Table 1 for MetricGAN+ and CMGAN models are quoted from the original papers; however, to verify them, we have obtained our results: MetricGAN+ (repro.) and CMGAN (repro.). The results for the CMGAN (repro.) model are very similar to the results from [9] with one exception - the SSNR metric, where our result is lower: 10.61 (ours) vs. 11.10 [9]. Furthermore, the results for MetricGAN+ (repro.) model are slightly lower than the ones provided in [8].

Table 1: ***Performance comparison on Voice Bank+DEMAND dataset [19]:*** *"-" denotes the results not provided in the original paper; † - quoted from [22]; repro. - our reproduction of experiments*

Model	# of Param.	PESQ	CSIG	CBAK	COVL	SSNR	STOI
Noisy Data	n/a	1.97	3.35	2.44	2.63	1.68	0.91
MANNER [23]	-	3.21	4.53	3.65	3.91	-	-
DB-AIAT [13]	2.81M	3.31	4.61	3.75	3.96	10.79	0.96
DPT-FSNet [14]	0.91M	3.33	4.58	3.72	4.00	-	0.96
PCS [22]	-	3.35	4.43	-	3.92	-	0.95
MetricGAN+ [8]	2.6M	3.15	4.14	3.16	3.64	-	0.93†
MetricGAN+ (repro.)	2.6M	3.08	4.05	3.01	3.60	-	0.92
SCP-MetricGAN+ (ours)	2.6M	3.19	4.20	3.20	3.65	-	0.93
CMGAN [9]	1.83M	3.41	4.63	3.94	4.12	**11.10**	0.96
CMGAN (repro.)	1.83M	3.39	4.62	3.93	4.13	10.61	0.96
SCP-CMGAN (ours)	1.83M	**3.52**	**4.75**	**3.97**	**4.25**	10.82	**0.96**

Finally, to demonstrate the significance of our best model, we ran a paired T-test between SCP-CMGAN and CMGAN(repro.) (as our baseline) models (`scipy.ttest_rel(scp-cmgan,cmgan(repro.))`) using a VoiceBank-DEMAND test dataset, see Table 2 for results. The test confirmed that our results are better than the baseline, with a p-value less than 0.05 for every metric.

Table 2: ***T-test Statistics and p-value:*** *Results of a paired T-test between SCP-CMGAN and CMGAN(repro.) models on Voice Bank+DEMAND test dataset [19], a p-value less than 0.05 indicates statistically significant results.*

Metric	Statistic	p-value	Metric	Statistic	p-value
PESQ	14.889	$1.419 \cdot 10^{-44}$	COVL	19.515	$5.362 \cdot 10^{-70}$
CSIG	18.402	$1.326 \cdot 10^{-63}$	SSNR	4.830	$1.630 \cdot 10^{-6}$
CBAK	14.565	$6.322 \cdot 10^{-43}$	STOI	2.511	$1.222 \cdot 10^{-2}$

5. Ablation Study

We have conducted an ablation study to demonstrate the importance of our methods. We have chosen the CMGAN [9] model as the base model due to its SOTA performance at the time of this study. Table 3 shows the average results of each model's best performance over three randomly chosen seeds.

First, we have retrained the CMGAN [9] model to verify the results from [9]. The results obtained from our experiments are relatively close to the results stated in [9], except for the SSNR metric where we have obtained slightly lower results, i.e., SSNR of 10.61 (ours) vs. SSNR of 11.10 [9].

Next, we have added Noisy Data (ND) to the CMGAN model discriminator training ('+ ND' in Table 3); however, such an addition had slight improvement over the baseline. Following it, we have added our SC method to the baseline model ('+ SC_2' in Table 3) and nothing else. With SC_2, we saw some improvements in PESQ, COVL, and SSNR metrics. Furthermore, we have analyzed the advantages of the CP method ('+ CP' in Table 3) without any add-ons. The CP method shows significant improvements in PESQ, CSIG, and COVL metrics and is comparable in the others.

Then, we combined ND and SC_3 methods ('+ ND, SC_3' in Table 3). Such a setup further improves baseline as well as single methods, particularly in COVL and SSNR metrics. A combination of ND and CP methods ('+ ND, CP' in Table 3) has the same nature of improvements, producing better results in PESQ, CSIG, and COVL metrics. The last combination of SC_2 and CP methods ('+ SC_2, CP' in Table 3) demonstrates that together both proposed methods achieve significant improvements on the SE task. Moreover, this particular model achieved the highest SSNR result of 10.91.

Finally, we have combined ND, SC_3, and CP methods in the model we call SCP-CMGAN in Table 1. Adding ND and switching from SC_2 to SC_3 further improves the 'CMGAN + SC_2, CP' model, achieving new SOTA results.

Note that all of the above models were trained under the same conditions without changing the hyperparameters and with identical software and hardware settings: Python 3.8.13, PyTorch 1.10, and CUDA 11 on NVIDIA Tesla V100 GPUs.

Table 3: ***Ablation Study on Voice Bank + DEMAND:*** *STOI results are equal to 0.96 for all the tests; † - results from our tests; ND - Noisy Data, CP - Consistency Preserving Generator, SC_2 - SC with $\mathcal{L}_C$ and $\mathcal{L}_E$, SC_3 - SC with $\mathcal{L}_C$, $\mathcal{L}_E$, and $\mathcal{L}_N$.*

Model	PESQ	CSIG	CBAK	COVL	SSNR
CMGAN (repro.)†	3.39	4.62	3.93	4.13	10.61
+ ND	3.41	4.65	3.92	4.13	10.68
+ SC_2	3.44	4.65	3.92	4.17	10.70
+ CP	3.47	4.71	3.93	4.20	10.54
+ ND, SC_3	3.43	4.64	3.93	4.18	10.76
+ ND, CP	3.47	4.73	3.93	4.22	10.53
+ SC_2, CP	3.49	4.72	3.96	4.24	**10.91**
+ ND, SC_3, CP	**3.52**	**4.75**	**3.97**	**4.25**	10.82

6. Conclusion

This paper presents several improvements to SE GAN-based models. The proposed method of Consistency Preservation reconciles the issue with Fourier and Inverse-Fourier transforms inside the generative models. At the same time, the Self-Correcting Discriminator Optimization method helps with training the discriminator by avoiding gradient directions that are potentially harmful to the training. Our experiments demonstrate the proposed methods' advantages, including new SOTA results for the Voice Bank+DEMAND dataset.

7. Acknowledgements

We thank the University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for using the Lipscomb Compute Cluster. We also thank Rebecca Zadorozhnyy for reading the manuscript and providing us with many valuable comments and suggestions.

8. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[3] A. E. Bulut and K. Koishida, "Low-latency single channel speech enhancement using u-net convolutional neural networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6214–6218.

[4] D. N. Tran and K. Koishida, "Single-channel speech enhancement by subspace affinity minimization." in *INTERSPEECH*, 2020, pp. 2447–2451.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," 2017. [Online]. Available: https://arxiv.org/abs/1703.09452

[6] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2031–2041.

[7] S. Wisdom, J. R. Hershey, K. W. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 900–904, 2019.

[8] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv preprint arXiv:2104.03538*, 2021.

[9] R. Cao, S. Abdulatif, and B. Yang, "Cmgan: Conformer-based metric gan for speech enhancement," *arXiv preprint arXiv:2203.15149*, 2022.

[10] H. Li, S.-W. Fu, Y. Tsao, and J. Yamagishi, "imetricgan: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning," *ArXiv*, vol. abs/2004.00932, 2020.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *ArXiv*, vol. abs/2005.08100, 2020.

[13] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7847–7851.

[14] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6857–6861.

[15] E. Kim and H. Seo, "SE-Conformer: Time-Domain Speech Enhancement Using Conformer," in *Proc. Interspeech 2021*, 2021, pp. 2736–2740.

[16] V. Zadorozhnyy, Q. Cheng, and Q. Ye, "Adaptive weighted discriminator for training generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 4781–4790.

[17] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency," in *Proc. DAFx*, vol. 10, 2010, pp. 397–403.

[18] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 656–660, 2021.

[19] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *SSW*, 2016.

[20] S. Abdulatif, K. Armanious, J. Thaiparambil Sajeev, K. Guirguis, and B. Yang, "Investigating cross-domain losses for speech enhancement," 08 2021, pp. 411–415.

[21] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 72–76, 2021.

[22] R. Chao, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Perceptual contrast stretching on target feature for speech enhancement," *Proc. of INTERSPEECH*, 2022.

[23] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "Manner: Multi-view attention network for noise erasure," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7842–7846.

[24] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 035081, 2013. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.4799597

[26] A. W. Rix, J. G. Beerends, M. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, 2001.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.