Evaluation of Sound Enhancing GANs

Artur Kuramshin
Department of Computer Science
University of Toronto
Toronto, ON

Shuvam Das
Department of Computer Science
University of Toronto
Toronto, ON

Pranjal Bajaria
Department of Computer Science
University of Toronto
Toronto, ON

Abstract

Generative Adversarial Networks (GANs) use an adversarial process between two models(a discriminator and a generator) which are simultaneously trained to estimate a generative model. GANs have recently been shown to be efficient for speech enhancement. One such architecture, the Speech Enhancement GAN uses Least Squares GAN to perform speech enhancement on audio samples. We are examining the performances the SEGAN and its variations with the Departure From Normality(DFN) metric and Wasserstein loss(WSEGAN) on two data sets: one comprising of Human Speech and other comprising of instrumental music. We use Signal-to-Noise Ratio(SNR), Peak SNR and Signal-to-Distortion Ratio to quantitatively analyse the results of our architectures.

1 Introduction

Sound enhancement (SE) models have the aim of improving the quality of captured sound audio signals. We can define the problem formally: given a dataset $\mathcal{X} = \{(\mathbf{x}_1, \tilde{\mathbf{x}}_1), (\mathbf{x}_2, \tilde{\mathbf{x}}_2), ..., (\mathbf{x}_N, \tilde{\mathbf{x}}_N)\}$ which consists of N pairs of noisy $(\mathbf{x})$ and clean $(\tilde{\mathbf{x}})$ audio signals, find a mapping $f(\tilde{\mathbf{x}}) : \tilde{\mathbf{x}} \to \mathbf{x}$ [5].

Speech Enhancement GAN (SEGAN): The SEGAN [3] generator takes in the noisy signal $\tilde{\mathbf{x}}$ and a latent vector $\mathbf{z}$ to produce the clean signal $\hat{\mathbf{x}} = G(\tilde{\mathbf{x}}, \mathbf{z})$. Conversely, the discriminator D receives a pair of signals that it learns to classify as the real pair $(\mathbf{x}, \tilde{\mathbf{x}})$ or the fake pair $(\hat{\mathbf{x}}, \tilde{\mathbf{x}})$. SEGAN also employs the least-squares GAN (LSGAN) [2] to improve stability by replacing the traditional cross-entropy loss in the discriminator by the least-squares loss. Additionally, SEGAN includes an ℓ_1-norm term between the clean and generated signals to the generator's objective function to generate a more realistic result. The ℓ_1-norm term is regulated by the hyper-parameter λ. The SEGAN objective functions of D and G are written as follows:

$$\begin{aligned}\min\nolimits_D V_{LS}(D) = {} & \frac{1}{2}\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{\text{data}}(\mathbf{x},\tilde{\mathbf{x}})}(D(\mathbf{x},\tilde{\mathbf{x}})-1)^2 \\ & + \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{\text{data}}(\tilde{\mathbf{x}})} D(G(\tilde{\mathbf{x}},\mathbf{z}),\tilde{\mathbf{x}})^2 \qquad (1) \\ \min\nolimits_G V_{LS}(G) = {} & \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{\text{data}}(\tilde{\mathbf{x}})}(D(G(\tilde{\mathbf{x}},\mathbf{z}),\tilde{\mathbf{x}})-1)^2 \\ & + \lambda||G_n(\tilde{\mathbf{x}},\mathbf{z})-\mathbf{x}||_1 \qquad (2)\end{aligned}$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Wasserstein GANs (WGAN). In Wasserstein GANs [4] we replace the GAN discriminator with a critic that outputs predictions in the range of (-inf, inf) instead of [0, 1] in a vanilla GAN with cross entropy or least squares loss (enforcing the critic function to be 1-Lipschitz continuous by clipping the weights of the critic between training batches). We also use a different loss metric called the Wasserstein Loss which essentially estimates the Earth Mover(EM) distance metric [4]. We can then train the Wasserstein critic to convergence without worrying about vanishing gradients as the EM distance is continuous and differentiable everywhere, thus giving us useful gradients till we reach optimality.

Departure From Normality(DFN) metric[1]: In improving stability with LS GANs[2], the main motivation of the paper is to introduce a new loss metric for the Generator, this new similarity metric is in unitary space of Schur decomposition for 2D representation of audio and speech signals known as the DFN (Departure from Normality) metric[1].

2 Method

For the WSEGAN we change the discriminator loss and the generator loss in the SEGAN so that the new losses are:

$$\begin{aligned} V_{wasserstein}(D) &= \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{\text{data}}(\mathbf{x},\tilde{\mathbf{x}})}(D(\mathbf{x},\tilde{\mathbf{x}})) \\ &\quad - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{\text{data}}(\tilde{\mathbf{x}})} D(G(\tilde{\mathbf{x}},\mathbf{z}),\tilde{\mathbf{x}}) \qquad (1) \\ V_{wasserstein}(G) &= \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{\text{data}}(\tilde{\mathbf{x}})}(D(G(\tilde{\mathbf{x}},\mathbf{z}),\tilde{\mathbf{x}})) \\ &\quad + \lambda||G_n(\tilde{\mathbf{x}},\mathbf{z})-\mathbf{x}||_1 \qquad (2) \end{aligned}$$

Additionally we introduce gradient clipping with a lower limit -0.05 and an upper limit of 0.05 for the gradients. Notice here that in the generator loss we still have the same L1 loss (as vanilla SEGAN) along with the new adversarial loss. This is because the L1 norm encourages the generator to learn the structure of the audio data and output more realistic results. Our objective is to maximise the generator loss while minimising the discriminator loss.

For the third architecture, we use DFN metric on the SEGAN generator along with the loss function, the DFN metric is enforced as:

$$min\frac{1}{2}E_{z\sim p_z}[D(G(z)-1)^2] \ \ such \ \ that \ \ ||[E_{z \ p_z}\delta^2 G(z) - E_{x\sim p_r}\delta^2 G(x)]|| < \epsilon \text{ [1]}$$

δ^2 function is calculated the same way as mentioned by Mohammad Esmaeilpour[1]. The epsilon value chosen is the $\frac{max|V_x|}{min|V_g|}$ this is chosen from the paper. Here V_x eigen values of the real sample and V_g eigen values of the generated matrix. The idea of having it is because we don't want the loss to take too many large steps which may lead to skipping of adjacent sub spaces of p_r[1]

For training our models we initialise the loss arrays for the different types of losses we have: discriminator loss on fake(generated) samples, discriminator loss on real samples, adversarial generator loss on generated images and L1 generator loss. We then initialise the parameters like batch size and learning rates and train the models. For each model in each training loop, we first train the discriminator for one iteration and then train the generator and so on just like Pascual et al[]. We train both SEGANs(with and without DFN) for 5 epochs and WSEGANs for 15 epochs all with a batch size of 350.

3 Experiments

3.1 Datasets

In this section we will first introduce two datasets which we have chosen to evaluate the three architectures. The descriptions of the datasets used are as follows:

> **Speech** dataset presented by Valentini-Botinhao et al. (2016) for speech enhancement. The dataset contains audio from 30 speakers from the Voice Bank corpus [7] with the noisy

conditions created using artificial noise and noise stemmed from the Demand database [] at various signal-to-noise ratios (SNRs).

Music, a dataset that we created by applying noise to the original dataset presented by Mehri et al. (2016) for audio generation that consists of 10 hours of Beethoven sonatas. For the training data we used two noisy conditions (artificial white noise and one from the Demand database [8]) at an SNR of 20 dB. For the test data, we also used two noisy conditions of an artificial white noise and a noise from the Demand database (different noise than training) at an SNR of 25 dB.

3.2 Results

After training all of the models on both datasets, using the latest checkpoint we generated the enhanced audio samples of the noisy test sets. We then evaluated the models using three metrics, calculated on the enhanced test sets.

To evaluate the quality of the enhanced audio signals, we use three objective sound quality metrics. All three metrics are computed by comparing the enhanced signal with the corresponding clean signal. The three evaluation metrics we chose are: Segmental Signal-to-Noise Ratio (SSNR)[9, p. 41], Signal-to-Distortion Ratio (SDR)[10], and Peak Signal-to-Noise Ratio (PSNR). All three metrics are measured in dB ultimately comparing the power of the desired signal to the power of the noise signal. In all three metrics a higher more positive value is better.

Tables 1 and 2 show the results of these metrics on the speech and music datasets respectively. As a baseline for comparison, we also include the metrics values when calculated directly on the noisy signals which can be seen in the bottom rows of tables 1 and 2. From the results we can see that the additional of the DFN metric and the Wassertstein loss did not have a significant impact on evaluation metric performance when compared to the vanilla SEGAN.
In order to evaluate the stability of the three models, we plotted the generator adversarial loss over the training batches.

Table 1: Evaluation on the speech dataset

Model	Segmental SNR	SDR	PSNR
SEGAN	7.072027	39.720622	45.462866
SEGAN + DFN	3.993066	34.437286	39.509198
WSEGAN	8.088722	40.863101	46.304853
Noisy baseline	0.000265	31.805900	29.931449

Table 2: Evaluation on the music dataset

Model	Segmental SNR	SDR	PSNR
SEGAN	-6.924607934	2.408431	26.185589
SEGAN + DFN	-6.937759953	2.139467	26.091608
WSEGAN	-1.716425	-4.793059	33.101397
Noisy baseline	-10.0	-1.261633	6.005957

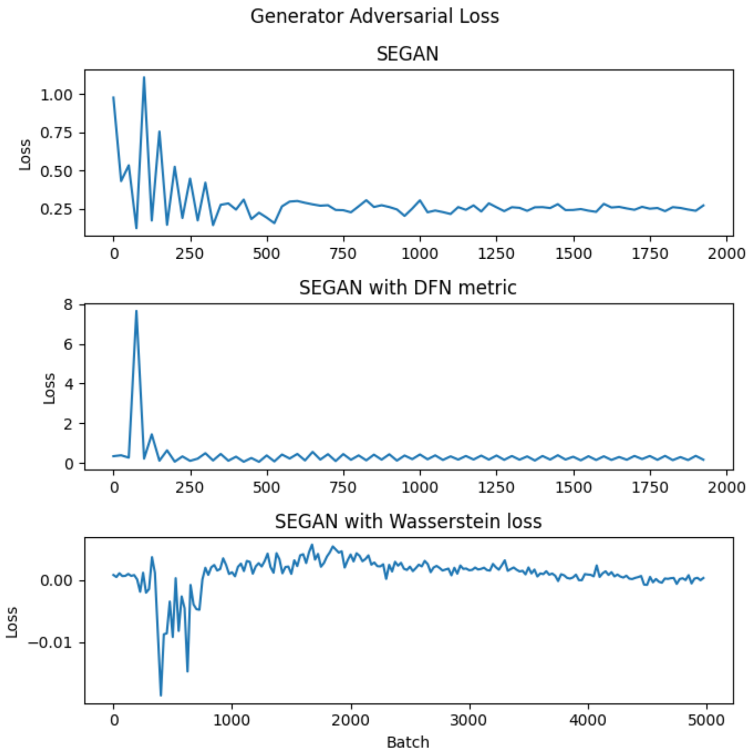

Figure 1: Generator adversarial loss plots of the three models.

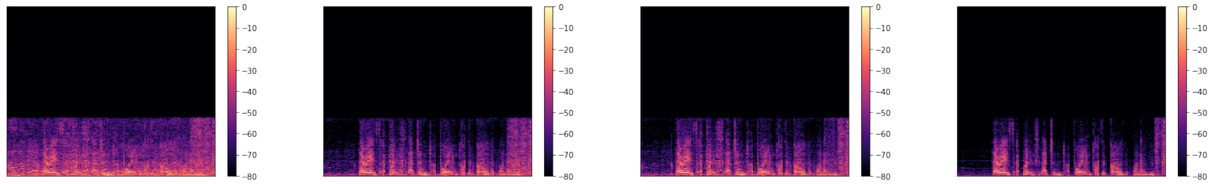

Figure 2: Noisy Spectrogram, SEGAN Spectrogram, SEGAN with DFN spectrogram, SEGAN with Wassterstein loss spectrogram

We observe that the addition of the DFN metric To the SEGAN makes the loss curve more stable this is because DFN converges earlier and has less variance. WSEGAN training seems very unstable but it performs as expected as we are trying to maximise the adversarial loss. From the spectrograms we observe that there is clear reduction of noise in all three models but the WSEGAN has more defined peaks than SEGAN with DFN metric and SEGAN. This may be because we train WSEGAN for longer and it tries to solve the vanishing gradients problem that we would see in an LSGAN [2].

4 Conclusion

This paper examined the performances of three different GAN architectures on audio enhancement. Using the SEGAN model architecture as the backbone, we evaluated the performance of the DFN metric and Wassertstein loss on objective audio quality measures and training loss stability. We did not find a significant improvement in the quality measures over the original SEGAN by the other two models. We did see evidence of the DFN metric stabilizing the generator adversarial loss during training.

5 References

[1] Esmaeilpour, Mohammad and Sallo, Raymel Alfonso and St-Georges, Olivier and Cardinal, Patrick and Koerich, Alessandro Lameiras *Improving Stability of LS-GANs for Audio and Speech Signals, 2020* https://arxiv.org/abs/2008.05454

[2] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, *Least squares generative adversarial networks,* in IEEE Intl Conf onComputer Vision (ICCV), 2017, pp. 2813–2821.

[3] Santiago Pascual1 , Antonio Bonafonte1 , Joan Serra *SEGAN: Speech Enhancement Generative Adversarial Network*, 2017

[4] Martin Arjovsky , Soumith Chintala , and Leon Bottou *Wasserstein GAN* https://arxiv.org/pdf/1701.07875.pdf

[5] Huy Phan and Ian V. McLoughlin and Lam Pham and Oliver Y. Chen and Philipp Koch and Maarten De Vos and Alfred Mertins, "Improving GANs for Speech Enhancement," in *IEEE Signal Processing Letters* , vol. 27, pp. 1700-1704, 2020.

[6] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noiserobust text-to-speech," in *9th ISCA Speech Synthesis Workshop*, pp. 146–152.

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

[8] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," in Journal of the Acoustical Society of America, vol. 133, no. 5, pp. 3591–3591, 2013.

[9] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, Measures of Speech Quality. Englewood Cliffs, NJ: PrenticeHall, 1988.

[10] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis, "mir_eval: A Transparent Implementation of Common MIR Metrics", of the 15th International Conference on Music Information Retrieval, 2014.

6 Contributions

Shuvam: Contributed in deciding the topic for the project. Implemented the DFN metric. Trained the DFN metric. Analyzed the DFN metric and loss plots and spectrograms. Planned on the evaluation metrics for speech and music. Helped formatting plots. Contributed in Method, Experiment and qualitative evaluation of the output.

Artur: Helped put out ideas about possible research topics for the project. Created the datasets used for training. Created the evaluation pipeline for all three evaluation metrics. Helped in implementation of the DFN metric.

Pranjal: Helped in background research for choosing the topic; Implemented and trained the WSEGAN. Contributed to the the abstract, introduction and results. Helped in formatting the report.