HIGH SCHOOL EDITION
Journal of Student Research

Volume 11 Issue 3 (2022)

Denoising Speech Signals with Hifi-Coulomb-GANs

Anirudh Satheesh[1] and Karthick Muthu-Manivannan[#]

[1] Thomas Jefferson High School for Science and Technology, Alexandria, VA, USA
[#]Advisor

ABSTRACT

Recorded speech signals often contain noise that affects the quality of the signal and reduces intelligibility. Several studies have used Generative Adversarial Networks (GANs) to remove noise artifacts and improve speech intelligibility. However, GANs can suffer from gradient vanishing or gradient explosion that can reduce their effectiveness in denoising. To mitigate gradient vanishing, we applied the CoulombGAN architecture to speech denoising using a model structure similar to Hifi-GAN, the current state of the art speech denoiser. We call this new model Hifi-CoGAN. We used a WaveNet generator to denoise signals, a PostNet for general cleanup, and a Multi-Resolution Discriminator to evaluate the signal quality relative to the clean signal. Our results show that Hifi-CoGAN was able to outperform Hifi-GAN in many of the narrowband signals (signals with a limited range of frequencies) in terms of the Short-Term Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ) metrics. However, the model did not perform as well as Hifi-GAN with wideband noise signals (signals with a wider range of frequencies) such as white noise, so future work must be done to improve the model for these noise signals.

Introduction

Recorded speech signals are often corrupted by background noise, causing reductions in overall sound quality and intelligibility. With the prevalence of virtual meetings, there is an increased demand for speech denoising to remove noise artifacts from noisy speech signals. The majority of applications involve microphone recordings through communications software such as Zoom or WebEx, where noisy environments can hamper effective communication. Other uses are for speech recognition systems, where removing speech artifacts from speech datasets can improve a model's ability to classify the speaker and identify specific components of the signal [1]. Finally, speech enhancements can be applied to cochlear implants before passing in the signal into the speaker, improving intelligibility [2]. The goal of our work is to improve band-limited speech signals by removing noise artifacts and generating clean signals that are indicative of high-quality single-channel recordings.

One of the earliest forms of denoising speech systems was spectral subtraction, which estimated the noise spectrum from the signal and subtracted it from the noisy signal to create clean speech [3]. Spectral subtraction was effective at reducing noise to an extent but introduces additional speech artifacts [4]. Another method of denoising speech are Wiener filter algorithms, which filter out any noise to provide an estimate of the clean signal. However, the Wiener filter assumes that both the signal and noise are second-order stationary and that correlation properties are known, which may not be the case [5]. Overall, both Wiener filter algorithms and spectral subtraction result in only minor reductions in background noise and have been vastly outperformed by neural network-based models [6].

One of the most promising advances in speech enhancement in machine learning has been through Recurrent Neural Networks (RNNs), which have outperformed non-network based models in many objective metrics. For example, Huang et al. achieved a 4.32-5.42 GSIR dB gain in comparison to other models using their deep RNN [7]. Additionally, Long-Short Term Memory (LSTM) networks with auto-encoders have been applied to speech denoising and had similar success. Maas et. al used a Deep Recurrent Denoising Autoencoder with three hidden layers and three frames to consider larger temporal windows of the signal [8]. Pandey et. al used a variational auto-encoder with an RNN architecture and achieved improvements in monaural speech separation in the SDR, SIR, and SAR metrics [9]. However, both RNNs and LSTMs can create speech artifacts from a mismatching phase, reducing their own effectiveness at removing noise [10]. Most speech datasets are limited in size, which in turn hampers training RNNs, as they are limited in their size and complexity and may not achieve optimal results.

The most recent research in speech denoising through machine learning has been through Generative Adversarial Networks (GAN). One advantage of GANs over RNNs and LSTMs is the adversarial loss function of the discriminator GANS use. Instead of using a single loss function to increase the accuracy of the model, GANs use loss functions for both the generator and the discriminator. The discriminator is trained by this loss function to identify real and fake signals, improving the generator to better target the real dataset. Unlike RNNs that need labeled data, little to no training data is required for fully unsupervised GANs.

Researchers have used GANs for speech denoising with both spectral features and waveforms. For example, Donahue et. al used a spectral feature mapping approach in both a Speech Enhancement GAN model (SEGAN) and Frequency-domain Speech Enhancement GAN model (FSEGAN) in the context of Automatic Speech Recognition (ASR) [11]. Both models produced an improvement in their ASR model in comparison to no speech enhancement model. Fu et. al built on the work of Donahue et. al by introducing MetricGAN, which has a L_p loss in the SEGAN model to ensure that the discriminator and generator are trained based on objective metrics such as PESQ and STOI [12]. For waveform models, Phan et al. created an iterative SEGAN model and deep SEGAN model to test the impact of multiple generators and performed well in objective metrics [13]. One of the current state-of-the-art models, Hifi-GAN, combines both these approaches by using deep feature matching and waveform processing [14].

In this paper, we extend the Hifi-GAN based approach by applying the CoulombGAN architecture to speech denoising, which can prevent gradient vanishing and better target the clean distribution. We call this new model Hifi-CoGAN. The underlying idea behind CoulombGAN is that the generated and true samples create an electric potential field, and the model learns by minimizing the potential between a generated and true sample. Our goal is to improve upon the existing GAN-based speech denoising techniques in terms of both objective metrics and removal of speech artifacts. CoulombGAN has been previously used for images, and to the best of our knowledge, we are the first to adopt CoulombGAN for processing 1D data and speech signal denoising. Our results improve over the current state of the art results of Hifi-GAN.

The rest of the paper is outlined as follows: we provide an overview of GANs (Section 2), we introduce the proposed method in more detail (Section 3) and the setup of the experiment (Section 4), we report the results (Section 5), we discuss the results (Section 6), and finally, we reach a conclusion (Section 7).

Generative Adversarial Networks

Generative Adversarial Networks work by mapping a sample z from an initial distribution to a sample x in a target distribution, which could be speech signals, images, etc. GANs accomplish this with two separate models, the generator G and the discriminator D [15]. The generator takes in the sample z, and produce a fake sample output that imitates the real sample x. Then, both the fake sample output and x are inputted into the discriminator, which outputs a probability of the fake sample being part of the target distribution. When the model is training, D learns to find more features in the target distribution, and G updates its parameters to imitate more on the target distribution. This is done through an adversarial minimax game, represented by the equation

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D\big(G(z)\big)\right)\right] \quad (1)$$

where z is a random noise vector, $G(z)$ is the fake sample, $D(x)$ is the discriminator's estimate of whether the real sample x is part of the target distribution, and $D(G(z))$ is the discriminator's estimate of whether the fake sample is part of the target distribution. Through this process, the model will be able to understand the underlying structure of the target distribution and effectively reproduce samples based on this structure.

However, one problem with GANs is that they can suffer from vanishing gradients and not converge fully to model the target distribution [16]. Gradient vanishing happens when the gradients of the loss functions of the discriminator and generator become zero at local minimums and do not continue to learn (in the adversarial sense, this is equivalent to a local Nash Equilibrium [18]). The main workaround for gradient vanishing was made by Goodfellow et. al [17] by adding a non-saturating loss of $-\log\,(D(G(z))$, which is used today in GANs, but still doesn't entirely fix the issue. If $D(G(z))$ approaches 0, then the non-saturating loss will increase tremendously and cause gradient explosion, which is as detrimental to the model training as gradient vanishing [16].

CoulombGAN

To combat gradient vanishing, CoulombGANs represent the real and fake samples as a negative charge and positive charge respectively in an electric field, where the loss function is represented by the electric potential between the samples [18]. In an electric field, the potential is minimized when two charges are at a distance zero away from one another, and there are no local minimums, so in the CoulombGAN, there would be no local Nash Equilibrium and no gradient vanishing. Unterthiner et al. [18] defines $p_x(\boldsymbol{a})$ as a model density and $p_y(\boldsymbol{a})$ as a target density for sample $\boldsymbol{a}$ and as the difference in these densities. Analogous to the electric potential function $\frac{Q}{4r\pi\varepsilon_0}$, they also proposed the function that calculates the potential between sample $\boldsymbol{a}$ and sample $\boldsymbol{b}$ as follows:

$$\Phi(\boldsymbol{b}) = \int \rho(\boldsymbol{a})k(\boldsymbol{a},\boldsymbol{b})d\boldsymbol{b}, k(\boldsymbol{a},\boldsymbol{b}) = \frac{1}{\sqrt{(||\boldsymbol{a}-\boldsymbol{b}||)^2+\epsilon^2}^{\,d}} \quad (2)$$

where $k(\boldsymbol{a},\boldsymbol{b})$ is the kernel function analogous to r of the electric field potential function, ϵ is to introduce nonlinearity into the model to achieve more accurate results, and d defines the dimensionality of the kernel.

In the proposed method we use the CoulombGAN architecture and apply it to speech denoising by applying the following loss functions for the generator and discriminator respectively:

$$\mathcal{L}_D(D;G) = \frac{1}{2}E_{p_a}\big((D(\boldsymbol{a}) - \hat{\Phi}(\boldsymbol{a}))^2\big) \quad (3)$$

$$\mathcal{L}_G(G;D) = -\frac{1}{2}E_{p_z}\Big(D\big(G(\boldsymbol{z})\big)\Big) \quad (4)$$

where $\hat{\Phi}(\boldsymbol{a})$ is the batched potential function. CoulombGANs work by using batches to compute a batched potential function, so that over time, the average of these potential functions converges to the overall potential function $\Phi(\boldsymbol{a})$ for sample $\boldsymbol{a}$.

Method

Model Structure

Our model structure was inspired by Hifi-GAN and builds on previous work involving generative adversarial networks in denoising speech systems. However, there are two key differences between the models: Hifi-GAN uses a Mel-Spectrogram Discriminator, while our model does not, and our model uses CoulombGAN loss function. This is because we found that the Mel-Spectrogram discriminator did not have a significant impact on the denoising effort but increased training time for the model. The model begins with the generator, which is the enhancement network that imitates the clean speech target distribution. The generator consists of a WaveNet, which has shown success in removing speech artifacts in previous works. In the WaveNet, several causal filters and dilated convolutions allow for receptive fields to grow exponentially, which yields better removal of artifacts [19]. Next, we pass the generated sample into a PostNet, which has been able to improve speech quality before samples are passed into the discriminator [20]. Our PostNet involves six 1D convolutional layers, each with 256 layers, size 64 kernel, and a sigmoid as an activation function. PostNet allows the generator to generate signals without worrying about creating any potential noise artifacts. Finally, we pass the signal into the discriminator, which contains two 1D grouped convolutional layers to reduce the number of parameters, and a mean pool to determine whether the inputted signal is real or fake. The discriminator we used is a waveform discriminator, because previous research has shown that waveform discriminators have been successful at removing background noise [14]. The output of the discriminator is passed into the CoulombGAN loss functions for both the generator and discriminator, which calculates the batched potential and updates the parameters of each component.

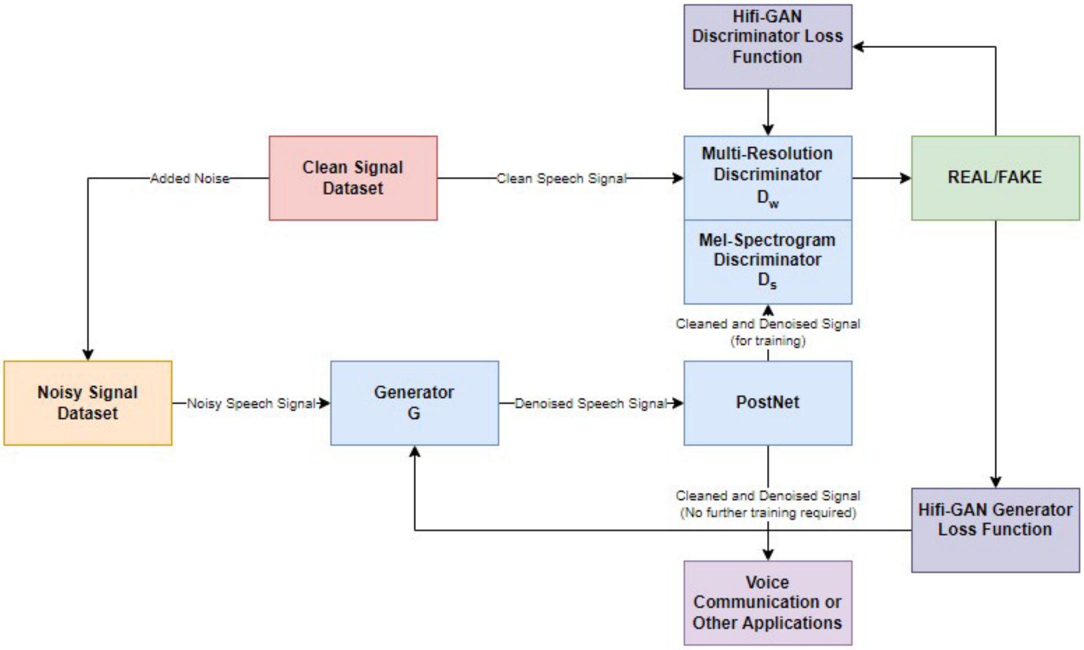
Hifi-GAN Discriminator Loss Function
Clean Signal Dataset
Added Noise
Clean Speech Signal
Multi-Resolution Discriminator D_w
REAL/FAKE
Mel-Spectrogram Discriminator D_s
Cleaned and Denoised Signal (for training)
Noisy Signal Dataset
Noisy Speech Signal
Generator G
Denoised Speech Signal
PostNet
Cleaned and Denoised Signal (No further training required)
Hifi-GAN Generator Loss Function
Voice Communication or Other Applications

Figure 1. Block Diagram of the Hifi-GAN model.

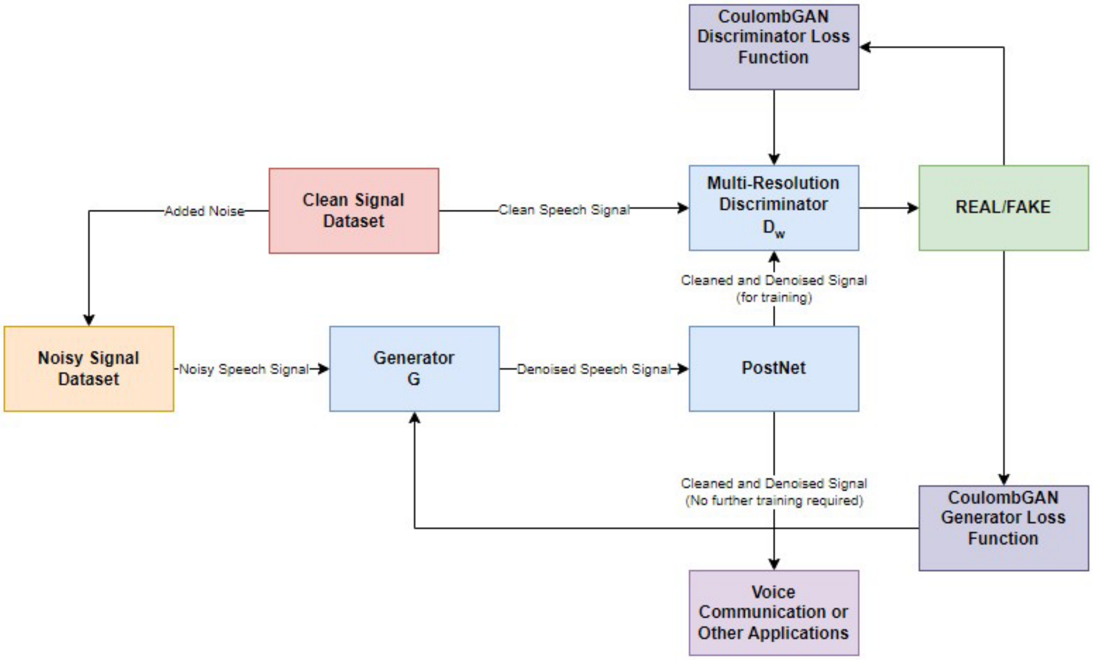
CoulombGAN Discriminator Loss Function
Added Noise
Clean Signal Dataset
Clean Speech Signal
Multi-Resolution Discriminator D_w
REAL/FAKE
Cleaned and Denoised Signal (for training)
Noisy Signal Dataset
Noisy Speech Signal
Generator G
Denoised Speech Signal
PostNet
Cleaned and Denoised Signal (No further training required)
CoulombGAN Generator Loss Function
Voice Communication or Other Applications

Figure 2. Block Diagram of the Hifi-CoGAN model.

Dataset

To train the CoulombGAN model, one needs both a noisy signal and its corresponding clean signal, as the GAN must compare both to identify the structure of the clean signal and generate accurate samples. However, large amounts of training data with clean and noisy signals are difficult to obtain. Instead, we obtained the clean signal through single-channel speech recordings and artificially added noise to create the noisy signal pair. This allowed us to add different kinds of noise to the clean signal and measure the extent of denoising by the model.

We trained the CoulombGAN model using the Microsoft Scalable Noisy Speech Dataset (MS-SNSD), which contains 27500 clean signals and 26 different background noise types (air conditioning, babble, munching etc.) sampled at 16 kHz. [21]. We chose 16kHz mono speech signals, as they represent the most used recording format for voice communication applications. We removed all noise types that had less than 10 recordings because we wanted to ensure that the model can generalize to all recordings of that specific noise type, reducing the number of noise types to 10. We also included additive white noise as a noise type to test the model's ability to denoise not only narrowband noise (such as the Air Conditioner and Copy Machine noise types) but also broadband noise. As a result, our dataset contained 155 noise signals comprising 11 noise types. To create the dataset of clean and noisy signal pairs, we batched 155 clean signals and combined each with a randomly chosen noise signal. This was done at three levels of SNR: 10 dB, 20 dB, and 30 dB, resulting in a dataset with 82,500 signal pairs. If the noise component was longer than the clean signal, then we truncated the noise component to end where the clean signal ends.

Training

Using the created dataset, we passed in each batch of clean and noisy signal pairs into the model. For the first 100K iterations, we trained the generator to generate clean signals that at least resemble the target distribution at a learning rate of 0.0002. This was done to avoid gradient diminishing, because if we trained the discriminator at the same time as the generator, there is a possibility that the discriminator becomes too successful and discriminates all fake samples before the generator learns the structure of clean signals. Next, we used the PostNet to remove any speech artifacts that could have been created during generation while also cleaning up the signal for another 100K iterations at a learning rate of 0.0001. Finally, we used the discriminator to fine tune the generator for 500K iterations at a learning rate of 0.0001.

During the training process, the discriminator and generator learned by Equation 3 and Equation 4 respectively, which relies on the kernel function to measure the influence between the clean and noisy samples. We decided to use the kernel function with d=3, which has been shown to train smoothly without oscillations and achieve accurate results [18].

Results

To measure the accuracy of our model on the signals from the MS-SNSD dataset, we use the Short-Term Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ) metrics. These are the most used metrics in denoising speech systems. We measured the STOI and PESQ of signals at different SNR values to evaluate the model's ability to adapt to different noise levels. Table 1 shows the STOI and PESQ values for all signals at SNRs of 10 dB, 20 dB, and 30 dB (higher is better). Our model has been able to achieve better denoising results for higher SNR, but there are still improvements to be made for lower SNRs.

Table 1. Objective metrics for all generated signals from the MS-SNSD dataset.

Type of Noise Signal	STOI – Hifi-GAN	STOI – Hifi-CoGAN	PESQ – Hifi-GAN	PESQ – Hifi-CoGAN
Clean	1.00	1.00	4.36	4.36
Noisy	0.927	0.927	1.92	1.92
10 dB	0.935	0.937	2.06	2.09
20 dB	0.945	0.954	2.14	2.25
30 dB	0.957	0.965	2.39	2.44

We also evaluated the model over the different types of noise added to each clean signal through the objective metrics. Table 2 shows the STOI and PESQ values for signals of each noise type. The model was best at removing noise from the copying machine noise type, as it had the highest STOI and PESQ. On the other hand, the model was unable to significantly remove white noise, with the lowest STOI and PESQ, showing that there are still improvements to be made for denoising broadband signals (such as white noise).

Table 2. Objective Metrics for each noise type.

Noise Type	STOI – Hifi-GAN	STOI – Hifi-CoGAN	PESQ – Hifi-GAN	PESQ – Hifi-CoGAN
Air Conditioner	0.960	0.966	2.35	2.38
Airport Announcements	0.954	0.959	2.24	2.28
Babble	0.948	0.961	2.27	2.33
Copy Machine	0.951	0.968	2.2	2.42
Munching	0.945	0.963	2.22	2.35
Neighbor Speaking	0.945	0.958	2.21	2.31
Shutting Door	0.952	0.944	2.14	2.08
Squeaky Chair	0.951	0.965	2.16	2.38
Typing	0.951	0.949	2.15	2.13
Washing Machine	0.945	0.951	2.19	2.24
White Noise	0.900	0.888	2.03	1.96

Discussion

The observed performance increase over Hifi-GAN can be attributed to a lack of local Nash Equilibriums, which prevents gradient vanishing. As such, the use of the CoulombGAN loss function helped better represent the target distribution of clean signals and achieve better results. From the objective metrics, the model denoised best on narrowband signals such as the Copy Machine and Air Conditioner. This is most likely because both types of signals have limited range of frequencies and unique noise signatures, so both the generator and discriminator recognize them during training. On the other hand, the model did not perform as well on white noise because, both the generator and discriminator may not recognize which frequencies contain white noise and which are part of the clean signal, so denoising broadband signals such as white noise has been much more difficult for the model. Additionally, the model achieved better results on higher SNR values, with a 3% improvement in STOI for 30 dB compared to 10 dB. This is to be expected, as a 10 dB SNR has much more noise in the overall signal compared to a 30 dB SNR, so the frequencies in the clean signal are much less pronounced. As a result, it becomes harder for the generator and discriminator to separate the two and generate clean signals.

One way to improve the model would be to add more components, such as an autoencoder, converting the model into an Energy Based GAN (EBGAN) or a Boundary Equilibrium GAN (BEGAN). EBGANs work by using an autoencoder as the discriminator, where the discriminator loss is represented instead with a reconstruction loss that allows for multiple targets in the target distribution [22]. In the speech denoising context, this would equate to a better representation of the clean signals and better results in terms of the objective metrics. BEGANs are similar to EBGANs, but they use a Wasserstein loss to measure the convergence of the model to the target distribution, which could further improve generated signals' representation of clean signals [23]. Finally, all the signals used to train the model have been at a sampling rate of 16 kHz, whereas high-fidelity speech can have a sampling rate of 44.1 kHz or 48kHz.

Conclusion

The results from objective metrics such as STOI and PESQ show that using our Hifi-CoGAN model has outperformed Hifi-GAN which is considered state of the art with a narrowband noise. However, this architecture does not significantly denoise white noise, so further research must be done for similar broadband signals. One area to target in the future could be real-time speech denoising, which could have significant applications in communication systems.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, 1996, pp. 929-932 vol.2, doi: 10.1109/ICSLP.1996.607754.

[2] Yang LP, Fu QJ. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. J Acoust Soc Am. 2005 Mar;117(3 Pt 1):1001-4. doi: 10.1121/1.1852873. PMID: 15806989.

[3] Upadhyay, Navneet & Karmakar, Abhijit. (2015). Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study. Procedia Computer Science. 54. 574-584. 10.1016/j.procs.2015.06.066.

[4] T. Biswas, C. Pal, S. B. Mandal and A. Chakrabarti, "Audio de-noising by spectral subtraction technique implemented on reconfigurable hardware," 2014 Seventh International Conference on Contemporary Computing (IC3), 2014, pp. 236-241, doi: 10.1109/IC3.2014.6897179.

[5] Wahlberg, P., & Schreier, P. J. (2010). On wiener filtering of certain locally stationary stochastic processes. Signal Processing, 90(3), 885-890. https://doi.org/10.1016/j.sigpro.2009.09.013

[6] M. Coto-Jimenez, J. Goddard-Close, L. Di Persia and H. Leonardo Rufiner, "Hybrid Speech Enhancement with Wiener filters and Deep LSTM Denoising Autoencoders," 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 2018, pp. 1-8, doi: 10.1109/IWOBI.2018.8464132.

[7] Huang, Po-Sen & Kim, Minje & Hasegawa-Johnson, Mark & Smaragdis, Paris. (2015). Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation. Audio, Speech, and Language Processing, IEEE/ACM Transactions on. 23. 10.1109/TASLP.2015.2468583.

[8] Maas, A.L., Le, Q.V., O'Neil, T.M., Vinyals, O., Nguyen, P., Ng, A.Y. (2012) Recurrent neural networks for noise reduction in robust ASR. Proc. Interspeech 2012, 22-25, doi: 10.21437/Interspeech.2012-6

[9] Pandey, L., Kumar, A., Namboodiri, V. (2018) Monoaural Audio Source Separation Using Variational Autoencoders. Proc. Interspeech 2018, 3489-3493, DOI: 10.21437/Interspeech.2018-1140.

[10] K. Osako, R. Singh and B. Raj, "Complex recurrent neural networks for denoising speech signals," 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1-5, doi: 10.1109/WASPAA.2015.7336896.

[11] Donahue, C., Li, B., & Prabhavalkar, R. (n.d.). Exploring speech enhancement with generative adversarial networks for robust speech recognition. ICASSP 2018. https://doi.org/10.48550/arXiv.1711.05747

[12] Fu, S., Liao, C., Tsao, Y., & Lin, S. (2019). MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement. ArXiv, abs/1905.04874.

[13] H. Phan et al., "Improving GANs for Speech Enhancement," in IEEE Signal Processing Letters, vol. 27, pp. 1700-1704, 2020, doi: 10.1109/LSP.2020.3025020.

[14] Su, J., Jin, Z., & Finkelstein, A. (2020). HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. doi:10.48550/ARXIV.2006.05694

[15] Pascual, Santiago & Bonafonte, Antonio & Serrà, Joan. (2017). SEGAN: Speech Enhancement Generative Adversarial Network. 3642-3646. 10.21437/Interspeech.2017-1428.

[16] Wiatrak, M., Albrecht, S. V., & Nystrom, A. (2019). Stabilizing Generative Adversarial Networks: A Survey. doi:10.48550/ARXIV.1910.00927

[17] Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622.

[18] Unterthiner, T., Nessler, B., Klambauer, G., Heusel, M., Ramsauer, H., & Hochreiter, S. (2018). Coulomb GANs: Provably Optimal Nash Equilibria via Potential Fields. ArXiv, abs/1708.08819.

[19] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., … Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. doi:10.48550/ARXIV.1609.03499

[20] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[21] Reddy, Chandan & Beyrami, Ebrahim & Pool, Jamie & Cutler, Ross & Srinivasan, Sriram & Gehrke, Johannes. (2019). A Scalable Noisy Speech Dataset and Online Subjective Test Framework. 1816-1820. 10.21437/Interspeech.2019-3087.

[22] Zhao, Junbo, et al. Energy-Based Generative Adversarial Network. arXiv, 6 Mar. 2017. arXiv.org, https://doi.org/10.48550/arXiv.1609.03126.

[23] Berthelot, David, et al. BEGAN: Boundary Equilibrium Generative Adversarial Networks. arXiv, 31 May 2017. arXiv.org, https://doi.org/10.48550/arXiv.1703.10717.