e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

IMPROVING SPEECH ENHANCEMENT USING GENERATIVE ADVERSARIAL NETWORKS (SEGAN) BY USING MULTISTAGE-ENHANCEMENT

Iyengar Vijay*[1], Himanshu Banwari*[2], Gagandeep Saluja*[3], Prof. Ajay Khatri*[4]

*[1,2,3]Student, Department Of Computer Science & Engineering, Acropolis Institute Of Technology & Research, Indore, Madhya Pradesh, India.

*[4]Professor, Department Of Computer Science & Engineering, Acropolis Institute Of Technology & Research, Indore, Madhya Pradesh, India.

ABSTRACT

We have been using Generative adversarial networks (GANs) for speech enhancement methods as they have proven to be efficient. However, we have various types of existing SEGANs that use a single generator to perform one-stage enhancement mapping. To do this particular job, it's better to use multiple generators that will efficiently perform multi-stage enhancement mapping, which eventually refines the noisy input signals in a stage-wise fashion. Furthermore, we have to analyze some specific cases: (1) the parameters will be shared by the generators, (2) there are independent parameters. There is a difference here between the former and the latter methods. The former restrains the generators from learning a somewhat common mapping which is used at all stages of enhancement and the product is a footprint of a small model. Whereas, in the other case, the generators are allowed to learn different mappings at various networking levels, but resulting in the increase of model size. We are trying to prove that our advanced multi-stage enhancement approach performs more efficiently and accurately as compared to the one-stage SEGAN baseline, where the independent generators lead to more preferable results than the tied generators.

Keywords: Speech Enhancement; Gans; Distill Knowledge; Convoluted Neural Network; Isegan.

I. INTRODUCTION

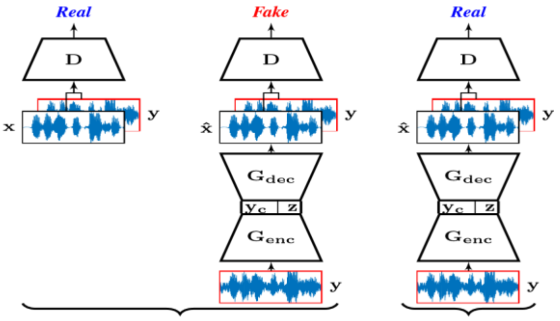
Speech enhancement's main job is to improve mainly two aspects: quality and understandability of speech which are polluted by background noise. Speech enhancement can definitely serve as a forefront platform for increasing the efficiency of the automatic speech recognition system. It also plays an important role in applications like communication systems, hearing aids, and cochlear implants in which contaminated speech gets enhanced prior to signal amplification to reduce discomfort. Significant progress is happening on this topic using several DNN based techniques. Various betterments on performance of the enhancement methods are being shown by these methods unlike some other general ones, like Wiener filtering, spectral subtraction etc. Generative adversarial network is combination of two individually functioning networks. These two networks are called Generator(G)and discriminator(D). Generator neural network creates artificial data and mixes it with real data, then it sends the data to the discriminator. The Discriminator then tries to differentiate between the original data and the artificial data. These Two network are actually two varied types of CNNs, that means network will get better and better every time which means they will run on basis of feedback.

Here, we will provide a large dataset that has noise and other mixture of sounds in it. Now, the model with the GAN infrastructure will receive the data. Now the data is passed on to the generator neural network. The generator creates artificial copies of the original dataset and then mixes that and the real data together. Then it passes on the data to the discriminator neural network. The discriminator will now try to differentiate between the real and artificial data. During this process both the network suffer losses. This goes around as feedback as both are types of CNNs. With every passing round the generator and discriminator will improve in accuracy and will become more efficient in recognizing the data. When maximum accuracy is achieved then they transfer the data as the output.

Existing SEGAN models have a common aspect – they achieve the mapping (enhancement through a single-stage by a single generator(G), which might not prove to be the best one. Here, our goal is to segregate the enhancement process into various stages and achieve it via multiple enhancement mappings, one on every stage. Each mapping is then understood by one generator, and chaining in the generators helps to enhance a noisy input signal eventually, step by step, to produce a single enhanced signal. By doing so, a generator has the task of refining the result produced in the previous stage. We have proposed to test out some new enhancement architectures, which go by the names, iterated and deep SEGANs or simply iSEGAN and dSEGAN to study two scenarios: (1) use a common mapping for all the enhancement stages, (2) using independent mappings at various enhancement stages. In these frameworks, the GAN architectures include multitudes of generator networks are connected with each other for success of the mapping (multiple-stage variation) which eventually processes the noisy data input in a stage-wise fashion. In iSEGANs, the generators are made to share their parameters for learning iterative mapping. Whereas, in dSEGANs, the generators do share a common structure but they have independent parameters and limitations; due to this case, the learning of different mappings happens at multiple different stages.

iSEGAN (Iterated SEGAN):

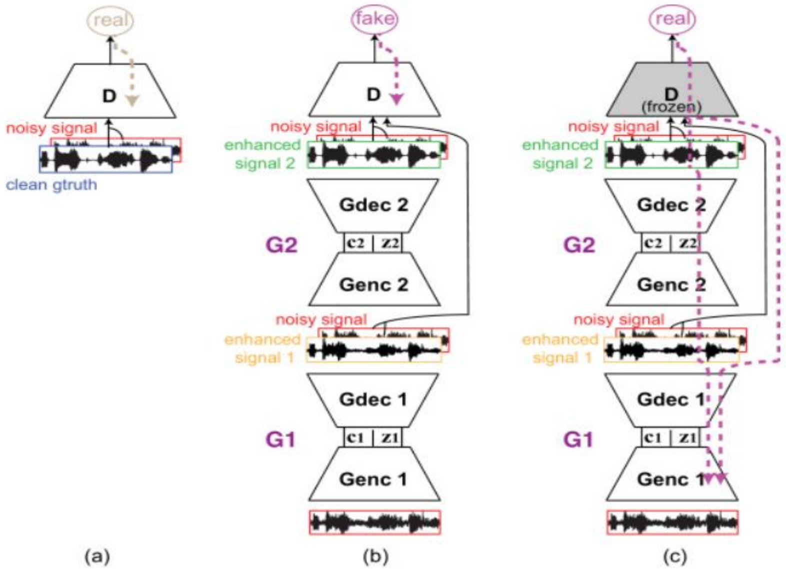
Popular deep learning-based speech enhancement methodologies work on the magnitude spectrogram and ignore the discrepancy between the noisy and clean speech signals. Conditional GANs or cGANs can properly address the phase-mismatch discrepancy by direct mapping of the noisy speech signals to the clean versions of speech signal. However, stabilization and training of the cGAN systems is quite hard and they still fall short when it comes to the performance achieved by the spectral enhancement approaches. This paper tries to investigate whether different normalization techniques and one-sided label smoothing can further stabilize the cGAN-based speech enhancement model. The outcome of the simulation shows proper improvement in the speech enhancement performance of cGAN systems and also yield improved stability and reduced computational effort.



The above image is about training a cGAN based speech enhancement system

DSEGAN (Deep SEGAN):

The one striking aspect that differentiates both iterative and deep SEGAN is that dSEGAN uses an independent generator that has a level N depth. Whereas, iSEGAN has multiple generators from G_1 to G_N, and 'N' here is the amount of iterations. In iSEGAN, the chain of generators produce an enhanced result by interpreting the results on various checkpoints. However, the last generator's outcome is considered as the final one.



The above image shows the training mechanism of both 'i' and 'd' SEGANS.

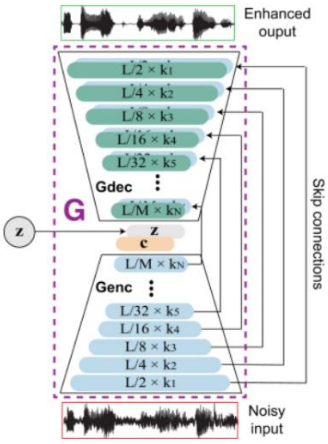
II. NETWORK ARCHITECTURE

GENERATOR(G):

The generators of both the SEGANs use encoder-decoder type fully convolutional layers. Each one of the generator network gets approximately 16kHz of speech input. The encoder comprises of 11 one dimensional convolutional layers (strided) with a common filter dimension of 2*31. The filters are designed specifically such that they increase with the encoder depth to make them more adjustable for smaller outputs. The output sized that we receive are of dimensions:

8192×16, 4096×32, 2048×32, 1024×64, 512×64, 256×128, 128×128, 64×256, 32×256, 16×512, 8×1024

The decoder part in the generator is an exact mirror or opposite of the encoder structure.



The above diagram shows the generator (G) architecture. Encoder-decoder structure that features U-shaped skip-connections that are employed for enhancing speech. The arrows represent the inter-connections in the layers. The output shapes alongside the input signals are provided which are designated as L and K_N. y_c is the encoder output in the diagram. Z are the distributions from the prior sample sets.

DISCRIMINATOR(D):

The next component i.e., discriminator (D) uses a structure that is very similar to the part which has the encoder of the generator. The one exception here is that it uses a two-channel input mechanism and also it has a 1-D convolutional layer which has only one filter and that too with a width of one unit (i.e. 1 × 1) to minimize the last output dimension from 8 × 1024 to mere 8 features.

III. DATASET

The dataset for this research work is from research section of University of Cambridge. The data consists the voice recordings of about 35 speakers, out of which we used 30 to train the models and the remaining 5 for testing. We used about 50 distinct types of background noise aberrations, to make a dataset of noisy audios. We gave the model about 16 kHz of audio on a single iteration. The dataset includes an array of audios ranging from 4 to 17 dBs.

IV. RESULTS OBTAINED

As we see in the result evaluation, as predicted, SEGAN enhances and works on the noisy signals to give speech signals that are better in quality and are intelligible, the proof is its better results across the objective metrics compared to those measured in the noisy data. When compared to SEGAN, iSEGAN performs relatively better in the aspects of speech-quality metrics, slightly surpassing the baseline in evaluation of perception (PESQ), rate of distortment in speech (CBAK), and ratio of signal and the noise (SSNR) (i.e. with N = 2, 3 and 4) but marginally underperforming in the other shown cases. Whereas, dSEGAN obtains the optimal results, consistently outperforming both SEGAN and iSEGAN across all the speech quality metrics. The outcomes are portraited in a graphical format below:

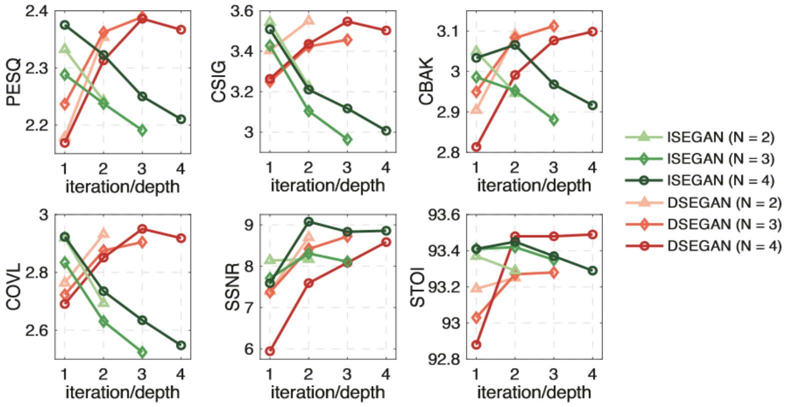


Figure: The image above shows the graphical diagram of the outcomes that were observed with both iSEGAN and dSEGAN algorithms.

V. CONCLUSION

In this work, we have presented a GAN method with multiple generators to rectify the problems in speech enhancement. We used multiple generators in a chained format with the aim to learn multiple enhancement mappings, to achieve a multiple-stage enhancement structure. We proposed two new types of GAN architectures: 'i' and 'd' SEGANs. Since, iSEGAN's generators share parameters, they are restricted to learn only a common mapping for all enhancement stages. Whereas, dSEGAN has different independent generators that can allow them to learn several mappings at every individual stage. The tests conducted clearly represent that the newly proposed GANs architectures perform comfortably and that they perform much better than SEGAN on speech-quality metrics and this shows that learning independent mappings results in better performances than a common mapping. Also, both the proposed architectures achieve much more favorable results as compared to SEGAN in speech-intelligibility metric as well as the perceptual test.

VI. REFERENCES

[1] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays, HSCMA*, Mar 2017, pp. 136– 140.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, Dec 2014, pp. 2672–2680.

[3] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *CoRR*, vol. abs/1711.05747, 2017.

[4] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *IEEE International Conference*

[5] *on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2015, pp. 4624-4628.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep 2011.

[7] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, 2017.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[9] B. Li C. Donahue and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. ICASSP,*, 2018, pp. 5024–5028.

[10] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017, pp. 2813–2821.

[11] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: design, collection and data analysis of a large regional accent speech database," in *Proc. 2013 International Conference Oriental COCOSDA*, 2013, pp. 1–4.