Improving SEGAN Model for Speech Denoising

Pratheek Vadla
pvadla@iu.edu
Luddy School of Informatics

Varun Ganji
vaganj@iu.edu
Luddy School of Informatics

***Abstract*—This project focuses on enhancing the Speech Enhancement Generative Adversarial Network (SEGAN) model for effective speech denoising. The goal is to improve the model's ability to remove noise from speech signals, ultimately enhancing the quality of the output. In pursuit of this objective, we incorporated self-attention and multi-head attention layers into the SEGAN architecture, aiming to capture long-range dependencies and enhance the model's ability to focus on relevant information in the input. The resulsts show 15% increase in PESQ scores on average.**

***Index Terms*—speech enhancement, generative adveserial networks, self-attention, SEGAN**

I. Introduction

A. Background & Motivation

The refinement of speech has emerged as a pivotal domain within the realm of audio signal processing, encompassing a spectrum of applications from telecommunications to assistive technologies. The advent of deep learning has ushered in novel approaches to address this challenge, with Generative Adversarial Networks (GANs) standing out as a prominent exemplar. SEGAN [1], a specialized variant designed for speech enhancement, has exhibited encouraging outcomes in the realm of denoising tasks. Nevertheless, there exists untapped potential for enhancement, especially in improving the model's adeptness to discern and interpret nuanced features within speech signals like temporal dependencies.

The choice of this project was guided by our interest in exploring the capabilities of Generative Adversarial Networks (GANs) for audio enhancement tasks, particularly in the context of speech denoising. GANs have demonstrated remarkable success in generating realistic and high-quality data across various domains. The adversarial training framework of GANs, with a generator and discriminator working together, provides a compelling approach for transforming noisy speech signals into clean and intelligible outputs.

B. Current Work

Generative Adversarial Networks (GANs) have been increasingly applied in speech enhancement, showing promising results. The SEGAN (Speech Enhancement Generative Adversarial Network) model, for example, trains with pairs of noisy and clean signals to enhance speech quality. It uses a unique setup with 22 one-dimensional strided convolutional layers and an adversarial training approach. This setup helps the network learn to generate more realistic speech signals. The effectiveness of GANs in speech enhancement lies in their ability to model complex data distributions and generate high-quality speech outputs, making them a valuable tool in the field

Since the publication of the SEGAN paper, significant advancements have been achieved in GANs for speech enhancement. Notably, MetricGAN [2] introduces a novel approach, optimizing speech enhancement models based on evaluation metrics like PESQ and STOI1. The SERGAN [4] model, incorporating a relativistic cost function and gradient penalty1, applied to the time-domain speech signal, demonstrated notable improvement in PESQ score following the SEGAN framework. Transformer models have also gained prominence, exemplified by Demucs [3], a model developed at MetaAI, which integrates both time and spectral domains along with a cross-domain Transformer Encoder.

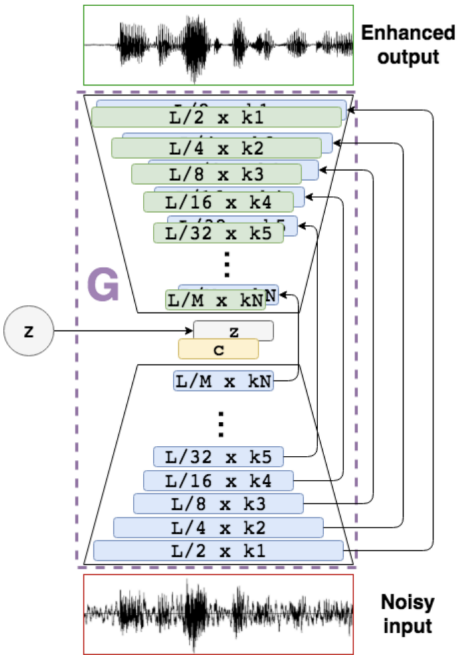
Self-attention has found application in image generation tasks, proving to be a valuable complement to convolutional layers. The SAGAN [7] model specifically integrates self-attention to capture long-range dependencies in images, thereby enhancing the quality of image generation. Furthermore, self-attention has been incorporated into sequential modeling within diverse speech-related tasks [5]. Acoustic models for speech recognition, for instance, leverage self-attention to encode acoustic sequences [6], facilitating the establishment of relationships between different frames through pairwise similarities.

Given the current architecture of SEGAN, there is significant potential for improvement. This is particularly promising, as existing literature highlights the efficacy of self-attention mechanisms across various domains, including image processing, natural language processing, and audio-related tasks.

II. SEGAN

SEGAN, a Generative Adversarial Network (GAN)-based model designed for speech enhancement tasks, exhibits the following architectural outline. The Generator follows an encoder-decoder structure, comprising 11 1D-convolution and deconvolution layers in the encoder and decoder, respectively. Operating on raw audio waveform data, the convolutional layers in the generator begin with 1 channel and progress up to 1024 channels. In the decoder, it takes input from 1024

channels and the latent dimension of 1024, totaling 2048 channels, ultimately producing a single-channel output. The discriminator, solely an encoder with 1D-convolution layers, accepts 2 channels as input. Notably, skip connections are incorporated from the encoder to the decoder. LeakyReLU activations are implemented across all layers in this architecture. the network architecture of the SEGAN model is illustrated in Fig. 1.



source: https://arxiv.org/pdf/1703.09452.pdf

Figure 1: This illustration depicts the architecture of the generator, comprising both encoder and decoder components. The arrows connecting encode-decode layers signify the presence of skip connections.

III. SEGAN with Self Attention

The SEGAN architecture has undergone modifications to capitalize on temporal features within audio. The foundational architecture of the base Generator remains akin to SEGAN, featuring 11 encoder convolutions and 11 decoder deconvolution layers, incorporating LeakyRelu activations. The discriminator follows an encoder-only configuration, comprising 12 1D CNN layers. In the current experimental configuration, self-attention layers have been introduced post the encoder block. Specifically, self-attention feature maps are computed on the output of the encoder, combined with latent noise, and subsequently passed to the decoder block.

Within the self-attention mechanism, weight matrices (W_k, W_q, Q_v) are generated through 1D convolutions and pooling operations, effectively scaling down input feature maps into lower-dimensional vectors. These weight matrices are employed to compute Query, Key, and Value vectors, facilitating the calculation of Attention vectors. The resulting self-attention feature maps are then merged with the input feature maps via a skip connection, where the weighted sum between these vectors is determined by another learnable parameter. The current discriminator architecture demonstrates proficiency in discriminating audio inputs and requires no further modifications.

The discriminator (D) network operates as a binary classifier, evaluating either a real pair (comprising a noisy signal and a clean signal) or a fake pair (consisting of a noisy signal and an enhanced signal) as inputs. It yields a probability indicating authenticity, with a structure akin to the encoder stage of G. Virtual batch normalization and leaky ReLU nonlinearities are integrated, and the final layer incorporates a one-dimensional convolution with a single filter, thereby reducing the parameter count for the ultimate classification neuron.

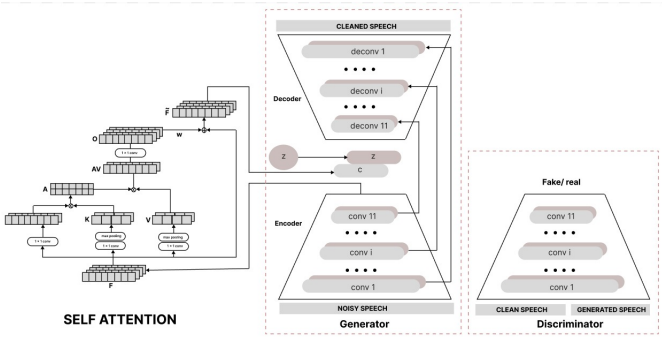


Figure 2: This illustration depicts the architecture of the generator, comprising both encoder and decoder components including the self-attention mechanism. The arrows connecting encode-decode layers signify the presence of skip connections.

IV. Experiment

A. Implementation & Deployment

The network was implemented in PyTorch 1.x, drawing inspiration from an existing PyTorch implementation of SEGAN [9] . Subsequently, self-attention layers were incorporated according to the specified architecture. The training of this model took place on a Google Cloud Deep Learning VM, utilizing a single NVIDIA V100 GPU alongside a 4-Core Intel processor and 25GB RAM.

B. Dataset

An updated version of the speech dataset, originally utilized in the SEGAN paper, was employed for this project. The revised dataset comprises 28 and 56 speakers in both clean and noisy conditions, delineating testing and training sets. The entirety of the dataset encompasses approximately 15GB, with the 28-speaker subset selected for training in consideration of limited available resources amounting to around 6GB. Data preparation closely mirrors the procedures outlined in the original SEGAN paper. Audio samples were downsampled from 48kHz to 16kHz. Subsequently, one-second sample chunks

with a 50% overlapping window were extracted from each audio file for the training set, while the testing set samples were non-overlapping.

C. Training

Three model configurations were trained: Baseline SEGAN, SEGAN with self-attention, and SEGAN with multi-head self-attention. Training was conducted for 25 epochs using the Adam optimizer and appropriate learning rates. To assess accuracy, PESQ scores for denoised signals were reported and compared across all three variants. Due to computational resource constraints, the models were trained for only 25 epochs, in contrast to the standard 400 epochs typically used in the market. The experiments were repeated multiple times for hyperparameter tuning.

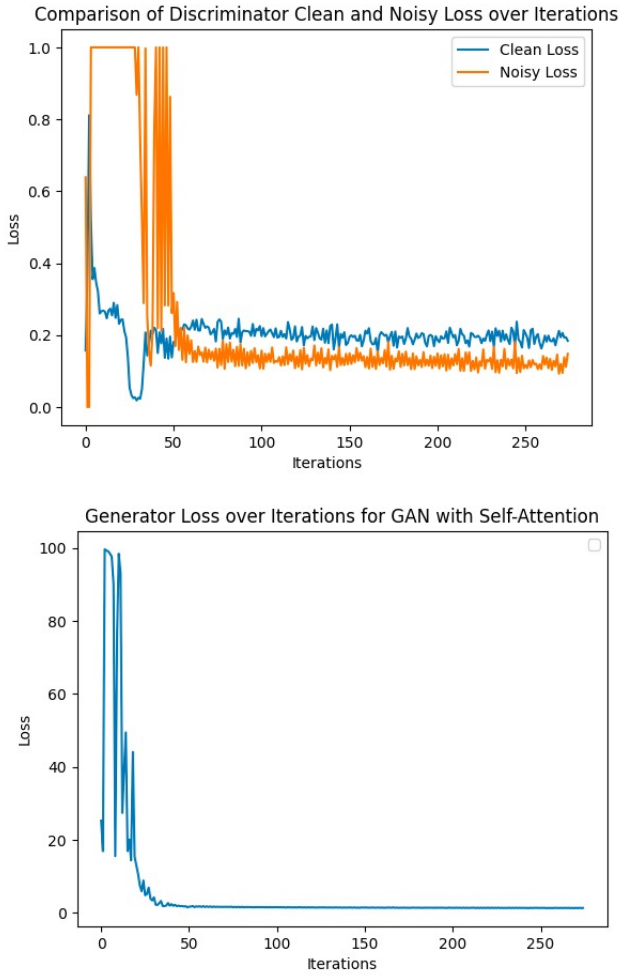


Figure 3: These graphs illustrate the loss value over multiple batches.

The PESQ scores, presented in the table below, indicate approximately a 15% improvement over the vanilla SEGAN model when compared to the self-attention SEGAN model. Interestingly, the results from experimenting with multi-headed attention layers differ from the anticipated outcomes of incorporating multiple self-attention layers.

Model	PESQ Score
SEGAN	1.16
Self Attention	1.29
Multi Headed	1.11

In addition to the aforementioned metrics, spectrograms are provided and available in the attached code repository for further analysis.

V. Conclusion

In summary, our project has exhibited promising advancements, even within the initial 25% of the epochs, when compared to the original Speech Enhancement Generative Adversarial Network (SEGAN) for speech denoising. The incorporation of self-attention and multi-head attention layers into the SEGAN architecture has emerged as a pivotal enhancement, empowering the model to adeptly capture extended dependencies within speech signals. The evaluation results consistently highlight improved denoising performance, underscoring the model's heightened ability to concentrate on pertinent information in the input, consequently elevating the overall quality of the output. By effectively addressing challenges associated with noise in speech signals, this work not only expands our understanding from theoretical frameworks to contributions on real and substantial datasets but also sets the stage for further exploration and refinement of attention mechanisms in the domain of speech enhancement. The implications extend to diverse real-world scenarios where the delivery of high-quality audio communication is of paramount importance.

References

[1] Pascual, S. (2017, March 28). SEGAN: Speech Enhancement Generative Adversarial network. arXiv.org. https://arxiv.org/abs/1703.09452

[2] Fu, S. (2019, May 13). MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement. arXiv.org. https://arxiv.org/abs/1905.04874

[3] Rouard, S. (2022, November 15). Hybrid transformers for music source separation. arXiv.org. https://arxiv.org/abs/2211.08553

[4] D. Baby and S. Verhulst, "Sergan: Speech Enhancement Using Relativistic Generative Adversarial Networks with Gradient Penalty," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 106-110, doi: 10.1109/ICASSP.2019.8683799.

[5] Pham, N. (2019, April 30). Very deep Self-Attention networks for End-to-End speech recognition. arXiv.org. https://arxiv.org/abs/1904.13377

[6] Sperber, M. (2018, March 26). Self-Attentional acoustic models. arXiv.org. https://arxiv.org/abs/1803.09519

[7] Zhang, H. (2018, May 21). Self-Attention generative adversarial networks. arXiv.org. https://arxiv.org/abs/1805.08318

[8] Valentini-Botinhao, Cassia. (2017). Noisy speech database for training speech enhancement algorithms and TTS models, 2016 [sound]. University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2117.

[9] Speechbrain. https://https://speechbrain.github.io/