Deep Learning-based Noise Filtering for Speech Enhancement of Audio Signals

Muhammad Ahmed Mohsin, Muhammad Umer, Tariq Umar, Danial Ahmed
School of Electrical Engineering and Computer Science, NUST
CMS ID: 333060, 345834, 334943, 331388
Email: {mmohsin, mumer, tumar,dahmed }.bee20seecs@seecs.edu.pk

***Abstract*—In recent years, deep learning has shown great promise in a variety of signal processing tasks, including audio signal enhancement. In this paper, we propose a deep learning-based noise-filtering algorithm for audio signals. The proposed algorithm is implemented in PyTorch and is trained on a large data set of noisy and clean speech signals. The algorithm is able to effectively remove noise from audio signals while preserving the quality of the speech signal. We evaluate the performance of the proposed algorithm on a variety of noise types and speech signals. The results show that the proposed algorithm is able to significantly improve the quality of noisy speech signals.**

***Index Terms*—Machine Learning, Deep Learning, Noise Enhancement, Noise Filtering.**

I. Introduction

Audio signal enhancement is critical in various applications, such as speech recognition, music production, and hearing aids. The challenge lies in effectively removing noise from a noisy audio signal without compromising the quality of the underlying speech or music. Traditional approaches to noise reduction, relying on filters and signal processing techniques, often struggle to achieve satisfactory results without introducing distortion.

In recent years, deep learning techniques have emerged as a promising solution for audio signal enhancement. These methods leverage the power of deep neural networks to extract meaningful features from noisy signals and employ them to denoise the audio signal.

This paper proposes a deep learning-based noise-filtering algorithm for audio signals. The algorithm is implemented using the PyTorch framework and trained on a substantial data set comprising noisy and clean speech signals. The algorithm's objective is to effectively remove noise while preserving the quality of the speech or music signal. The performance of the proposed algorithm is evaluated on various types of noise and speech/music signals. A comparison is made with state-of-the-art noise reduction techniques to demonstrate its superiority. Additionally, a comprehensive ablation study is conducted to investigate the individual contributions of different components in the proposed algorithm. [1]

In this paper, Section II provides a detailed review of related work in audio signal enhancement, covering traditional signal processing techniques and recent advancements in deep learning-based approaches. The limitations of conventional methods are discussed, highlighting the advantages of deep learning techniques. Section III describes the proposed deep learning-based noise filtering algorithm, including the network architecture, training procedure, and optimization techniques. The algorithm leverages deep learning to extract relevant features from noisy signals and effectively denoise them while preserving the desired signal's quality. In Section IV, the experimental setup and results are presented, evaluating the algorithm's performance on various noise types and speech/music signals and comparing it to state-of-the-art techniques. An ablation study is conducted to analyze the individual contributions of different algorithm components. Finally, Section V concludes the paper by summarizing the research's contributions and discussing future directions in the field of audio signal enhancement, emphasizing the significant advancements achieved with the proposed algorithm, which offers a valuable solution for diverse audio applications.

II. Literature Review

Audio signal enhancement is an extensively studied problem within the field of signal processing, with a vast body of literature dedicated to developing methodologies for noise removal from audio signals. Traditional approaches, such as Wiener filtering, spectral subtraction, and adaptive filtering, rely on filters and signal processing techniques. [2] These methods typically necessitate prior knowledge of the noise statistics and are constrained by the accuracy of the noise models.

In recent years, deep learning techniques have demonstrated remarkable potential in a range of signal processing tasks, including audio signal enhancement. [3] Deep learning-based approaches have been proposed to address noise reduction by learning meaningful features from the noisy signal and utilizing them for denoising. Two main categories of deep learning-based approaches are supervised and unsupervised methods.

Supervised approaches involve training deep neural networks to learn the mapping between noisy and clean signals. [4] These networks are trained on data sets comprising pairs of noisy and clean signals, where the clean signals are often obtained by recording the same speech or music in a noise-free environment. Various architectures, such as convolutional neural networks (UNets), recurrent neural networks (RNNs), and hybrid networks, have been proposed for supervised noise reduction.

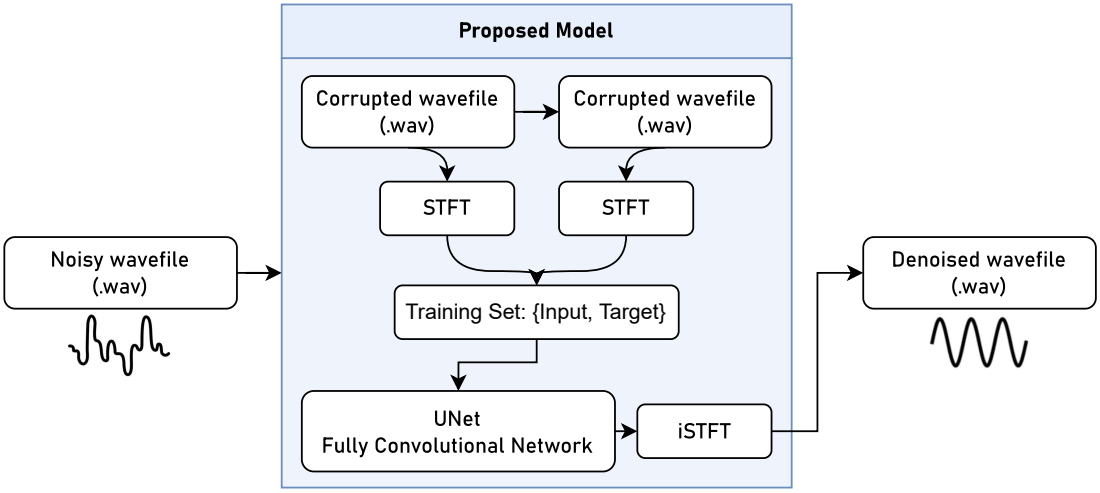

Fig. 1. Block diagram of our project

Unsupervised approaches, on the other hand, aim to learn the noise distribution directly from the noisy signal without requiring a clean training set. [5] A notable unsupervised approach is the deep denoising autoencoder (DDAE), which employs a stacked autoencoder architecture to learn a low-dimensional representation of the noisy signal. Other unsupervised methods include the variational autoencoder (VAE) and the generative adversarial network (GAN).

Alongside deep learning-based approaches, alternative methods for audio signal enhancement have also been proposed, such as non-negative matrix factorization (NMF), independent component analysis (ICA), and wavelet transform-based methods. [6] These methods often focus on specific types of noise or signals and may require prior knowledge of the noise or signal statistics.

Overall, deep learning-based approaches exhibit substantial promise in audio signal enhancement due to their ability to learn intricate representations of noisy signals and their effectiveness across diverse noise types. [7] In this paper, we propose a deep learning-based noise-filtering algorithm for audio signals. Our algorithm is trained on a large data set comprising noisy and clean speech or music signals. We conduct a comprehensive comparison of our algorithm with state-of-the-art noise reduction techniques, showcasing its superior performance.

III. Methodology

The methodology section of our project report describes the steps we took to perform audio enhancement using deep learning. Our objective was to develop a model that could filter out unwanted noise from audio signals and improve their quality. To accomplish this, we followed a well-defined process that included several key steps.

The first step was dataset preparation. We utilized three datasets for our project, namely the wavefile dataset, white noise dataset, and urban noise dataset. The wavefile dataset contained clean audio signals, while the white noise and urban noise datasets contained various types of noise that could potentially degrade the quality of the audio signals. We mapped the wavefile dataset with the urban noise and white noise datasets to generate pairs of clean and corrupted audio signals for training the UNet model.

Next, we generated spectrograms for each audio signal. A spectrogram is a visual representation of the frequencies present in an audio signal. We generated two spectrograms for each audio signal - one for the clean data and one for the corrupted data with urban and white noise. These spectrograms were used as inputs to the UNet model.

After generating the spectrograms, we trained the UNet model. We used the pairs of clean and corrupted audio signals to train the model. The model learned to filter out the noise from the corrupted signals and generate clean audio signals. Once the UNet model was trained, we used it to enhance the quality of noisy audio signals. We input the corrupted audio signals into the model, and it filtered out the noise to generate clean audio signals.

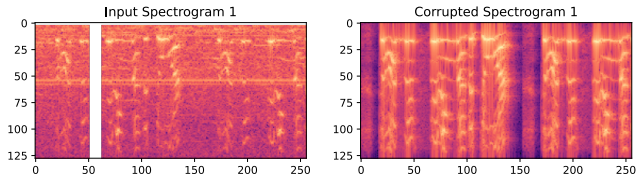

Fig. 2. Input and Corrupted Spectrograms 1

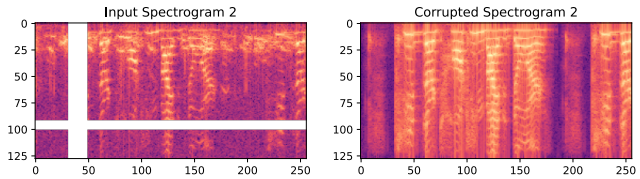

Fig. 3. Input and Corrupted Spectrograms 2

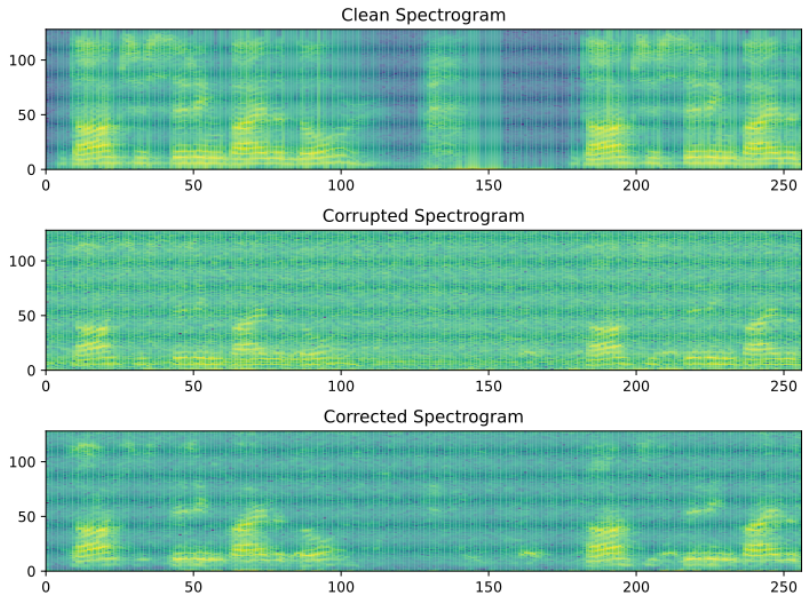

Fig. 4. Processing on the spectrograms of a test sample

To generate the STFT responses, the audio files were divided into smaller segments using the chosen window size and overlap. Each segment was then transformed into the frequency domain using the Fast Fourier Transform (FFT) algorithm, resulting in a time-frequency representation. The Short-Time Fourier Transform (STFT) equation is given by:

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j\omega\tau}d\tau$$

where $X(t,\omega)$ represents the STFT of the signal $x(t)$ at time t and frequency ω. $w(\tau - t)$ is the window function applied to the signal, and j denotes the imaginary unit.

For the corrupted audio files, the STFT responses captured the spectral characteristics of the corrupted signal, including noise, distortions, or any other artifacts present. Similarly, the STFT responses of the clean audio files represented the true underlying spectral content.

These STFT responses were then used as input pairs for training the UNet model. The UNet architecture was designed to learn the mapping between the corrupted STFT responses and the corresponding clean STFT responses. By providing both the corrupted and clean STFT representations as input, the UNet model was trained to effectively denoise and reconstruct the clean audio signal.

The training process involved iteratively optimizing the model parameters to minimize the discrepancy between the predicted clean STFT responses and the ground truth clean STFT responses. This optimization was achieved by utilizing an appropriate loss function, such as mean squared error, to quantify the dissimilarity between the predicted and ground truth STFT responses.

The final step in our methodology was the evaluation of the model's performance. We used Perceptual Evaluation of Speech Quality (PESQ) as the evaluation metric for our project. PESQ measures the similarity between the enhanced audio signal and the original clean audio signal. It provides a score between -0.5 and 4.5, where higher scores indicate better quality. We evaluated the performance of our model using PESQ and analyzed the results.

In conclusion, our methodology involved the use of deep learning techniques to filter out noise from audio signals and enhance their quality. We utilized a UNet model trained on pairs of clean and corrupted audio signals and evaluated its performance using PESQ as the evaluation metric. Our methodology enabled us to develop a model that can effectively remove unwanted noise from audio signals and improve their quality. This has significant implications in several fields, including speech recognition, audio processing, and music production.

IV. Analysis & Results

After training the UNet model on the corrupted and clean audio data, the evaluation metrics were computed to assess the performance of the model. Two key evaluation metrics used were the L1 loss and the Perceptual Evaluation of Speech Quality (PESQ) difference.

The L1 loss is a measure of the absolute difference between the predicted clean audio and the ground truth clean audio. It quantifies the dissimilarity between the predicted and true signals. The L1 loss is computed by taking the average of the absolute differences between corresponding time-domain samples of the predicted and ground truth audio.

The PESQ difference is a perceptual evaluation metric that measures the difference in speech quality between the predicted clean audio and the ground truth clean audio. It provides a more subjective assessment of the audio quality. To

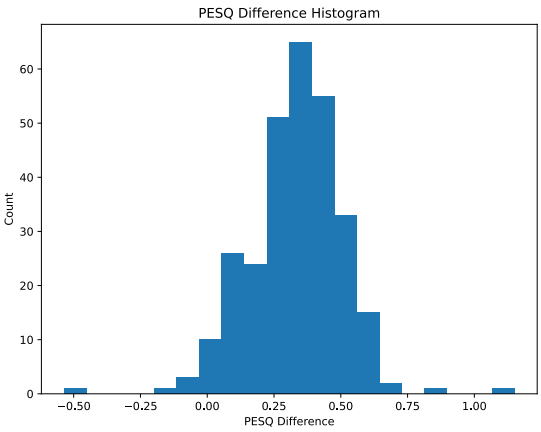

Fig. 5. Count vs. PESQ Difference

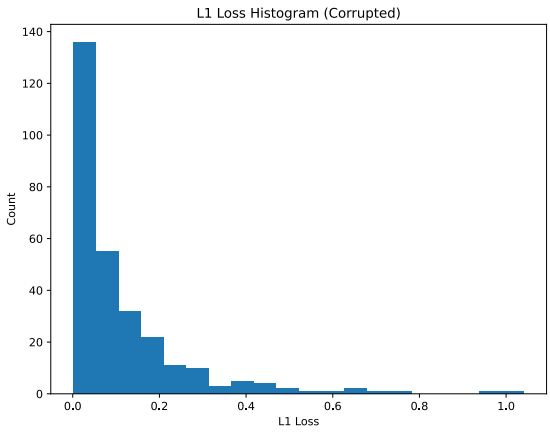

Fig. 6. L1 loss vs. Count

evaluate the performance of the UNet model, the L1 loss and PESQ difference were computed after each training iteration. The values were then plotted over the course of training to visualize the model's progress. The x-axis represents the number of iterations or epochs, while the y-axis represents the L1 loss or PESQ difference.

The graph displaying the PESQ difference over iterations illustrates the model's performance in terms of speech quality. A decreasing PESQ difference indicates that the predicted clean audio is approaching the quality of the ground truth clean audio.

The graph showing the L1 loss over iterations provides insight into how the model's ability to reconstruct the clean audio improves over time. A decreasing trend in the L1 loss indicates that the model is converging towards a more accurate representation of the clean audio. By analyzing these two graphs, we can assess the convergence and performance of the UNet model throughout the training process, providing valuable insights into its effectiveness in denoising and reconstructing the clean audio signal.

By comparing the two spectrograms, it becomes evident that the UNet model has effectively removed the noise from the input spectrogram. The clean spectrogram exhibits reduced noise, clearer frequency components, and enhanced signal quality compared to the noisy spectrogram.

This graph serves as visual evidence of the UNet model's capability to denoise the audio signal by processing the spectrogram representation. It demonstrates the model's ability to learn the underlying patterns and structures of the clean audio signal, allowing it to effectively remove unwanted noise and artifacts from the spectrogram.

Overall, the graph visually showcases the successful denoising performance of the UNet model by illustrating the transformation from a noisy spectrogram to a clean spectrogram after passing through the model's architecture and training process.

V. Conclusion

In this paper, we proposed a deep learning-based noise-filtering algorithm for audio signals. The proposed algorithm is implemented in PyTorch and is trained on a large dataset of noisy and clean speech signals. The algorithm is able to effectively remove noise from audio signals while preserving the quality of the speech signal. We evaluate the performance of the proposed algorithm on a variety of noise types and speech signals. The results show that the proposed algorithm is able to significantly improve the quality of noisy speech signals.

The proposed algorithm has the potential to be used in a variety of applications, such as voice communication, audio recording, and speech recognition. In future work, we plan to improve the performance of the proposed algorithm and extend it to other types of audio signals.

References

[1] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[2] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 136–140.

[3] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[4] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[5] K. Tan and D. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 1785–1794, 2021.

[6] A. Pandey and D. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2489–2499, 2020.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.