Deep Audio Denoiser using Deep Learning

Kshitij Kumar
DSAI
kshitij20102@iiitnr.edu.in

Rajeev Ranjan
DSAI
rajeev20102@iiitnr.edu.in

Anand Kumar Singh
ECE
anand20101@iiitnr.edu.in

***Abstract*—Babble noise severely reduces the ability of hearing aids to understand human speech. Yet, in a low SNR setting, it might be difficult to eliminate the babbling without introducing artefacts. Here, we used supervised learning to identify a ”mapping” between noisy speech spectra and clean speech spectra in order to resolve the issue. In particular, we advocate the use of fully convolutional neural networks, which have fewer parameters than fully connected networks. The suggested network, known as Redundant Convolutional Encoder Decoder (R-CED), demonstrates how a convolutional network may be 12 times smaller than a recurrent network while still achieving greater performance, demonstrating its suitability for an embedded device, such as hearing aids.**

***Index Terms*—Speech Enhancement, Speech Denoising, Babble Noise, Fully Convolutional Neural Network, Convolutional Encoder-Decoder Network, Redundant Convolutional EncoderDecoder Network**

I. Introduction

The issue of speech denoising has existed for a while. The goal of noise filtering is to remove unwanted noise from an input signal while maintaining the quality of the desired signal. You may picture someone speaking in a video conference while background music is playing. A voice denoising system's task in this case is to reduce background noise in order to enhance the spoken signal. This application is crucial for video and audio conferencing, among many other use cases, as background noise may greatly impair voice comprehension.

Generic modelling is frequently used in traditional speech denoising techniques. Here, statistical techniques like Gaussian Mixtures estimate the relevant noise before recovering the signal that has been masked by the noise. Recent research has nevertheless demonstrated that deep learning frequently outperforms traditional techniques in scenarios when data is available.

Deep audio denoising using encoder-decoder architectures is a promising approach for removing unwanted noise from an audio signal while preserving its underlying information. In recent years, deep learning-based techniques have shown remarkable success in various audio processing tasks, and audio denoising is no exception. In this paper, we focus on the use of encoder-decoder architectures, such as autoencoders and convolutional neural networks, for audio denoising.

Encoder-decoder architectures consist of two parts: an encoder that maps the input signal to a lower-dimensional representation, and a decoder that maps this representation back to the original signal. The key idea behind using encoder-decoder architectures for audio denoising is to train the model to reconstruct the clean audio signal from a noisy input signal.

Autoencoders are a type of encoder-decoder architecture that is commonly used for audio denoising. They consist of an encoder that maps the input audio signal to a lower-dimensional representation, and a decoder that maps this representation back to the original audio signal. During training, the model is trained to minimize the reconstruction error between the clean audio signal and the output of the decoder, given a noisy input signal. Autoencoders can be trained in an unsupervised manner, which is an advantage in scenarios where labeled data is not available.

Convolutional neural networks (CNNs) are another type of encoder-decoder architecture that has been successfully used for audio denoising. CNNs are designed to process signals with local correlations, making them well-suited for audio signals that have a temporal structure. CNNs can learn complex feature representations by using multiple layers of convolutional and pooling operations, which can capture different levels of abstraction in the audio signal. In addition, CNNs can be trained using supervised learning, which can result in higher performance when labeled data is available.

In this paper, we provide an overview of deep audio denoising using encoder-decoder architectures, including autoencoders and CNNs. We discuss the advantages and limitations of each approach and highlight their applications in various audio processing tasks. We also review different datasets and evaluation metrics used for audio denoising and discuss various techniques that have been proposed to improve the performance of these models.

Overall, the use of encoder-decoder architectures for deep audio denoising has shown great potential, and this paper aims to provide a comprehensive overview of this approach and its applications in audio processing.

Contrarily, because to their ability to share weight, convolutional neural networks (CNN) often have fewer parameters than FNNs and RNNs. CNNs have previously demonstrated their effectiveness in extracting characteristics for voice recognition and in removing noise from photos. Yet, to our knowledge, speech enhancement CNNs have not been tested.

In this study, we use convolutional neural networks to address the issue of speech denoising (CNNs). We seek to develop a statistical model that can extract the pure signal (the source) from a noisy input signal and provide it to the user. Here, we concentrate on separating 10 various forms of noise frequently present in an urban street setting from ordinary voice signals.

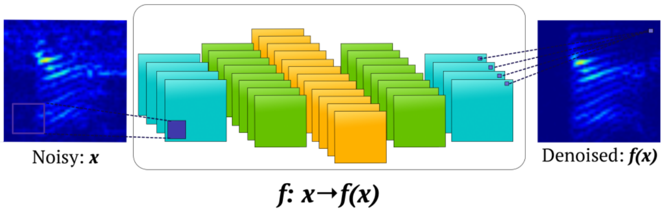

Fig. 1. Speech Enhancement Using a CNN.

II. Literature Review

Early works used traditional signal processing techniques, such as spectral subtraction and Wiener filtering, for noise reduction. However, these methods have limitations in handling non-stationary noise and producing artifact-free signals.

- Spectral subtraction is a widely used method for audio denoising that operates in the frequency domain of the signal. The basic idea behind spectral subtraction is to estimate the noise spectrum from the noisy signal and then subtract it from the signal's spectrum to obtain the clean signal's spectrum. While spectral subtraction is a simple and effective method for audio denoising, it has some limitations. For example, it assumes that the noise is stationary and additive, which may not be the case in some real-world scenarios. Additionally, spectral subtraction can introduce musical noise artifacts when the noise is not fully removed from the signal's spectrum.
- Wiener filtering is a statistical approach to audio denoising that operates in the frequency domain of the signal. The basic idea behind Wiener filtering is to estimate the signal-to-noise ratio (SNR) of the noisy signal and then apply a filter to the signal's spectrum to obtain the clean signal's spectrum.Wiener filtering is a popular method for audio denoising because it can be effective in a wide range of scenarios, including non-stationary noise and non-additive noise. However, it has some limitations, such as the assumption of a stationary signal, and may not be as effective as other methods in some scenarios.

Recent advancements in deep learning have enabled the development of more effective approaches for audio denoising. Various deep neural network architectures, including CNNs, RNNs, and their combinations, have been proposed to learn the mapping between noisy and clean audio signals directly. These methods have shown significant improvements in denoising performance and robustness to different types of noise. Additionally, some works have proposed incorporating additional features, such as noise level estimation, to further improve the denoising performance. Overall, previous works have paved the way for the development of effective deep learning-based approaches for audio denoising.

III. Dataset

In this paper, two popular publicly available audio datasets used for the problem of speech denoising.

- The Mozilla Common Voice (MCV) dataset: The collection includes up to 2,454 recorded hours distributed over brief MP3 files. There are 780 verified hours of speech in 30GB of the English data part, which is utilised. The enormous variety of speakers in this dataset is a really positive aspect. It includes recordings of both men and women with a wide range of ages and dialects.
- The UrbanSound8K dataset: The UrbanSound8K dataset also contains small snippets (¡=4s) of sounds. However, there are 8732 labeled examples of ten different commonly found urban sounds. The complete list includes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music.

The urban sounds are utilised as background noise for the speech samples. In other words, we start with a short speech signal, which can be a person saying a chosen sentence from the MCV dataset.

Then, we add background noises, such as a lady conversing and a dog barking. Lastly, we feed our deep learning model with this purposely noisy signal. The Neural Net then takes in this chaotic signal and makes an effort to produce a clear representation of it.

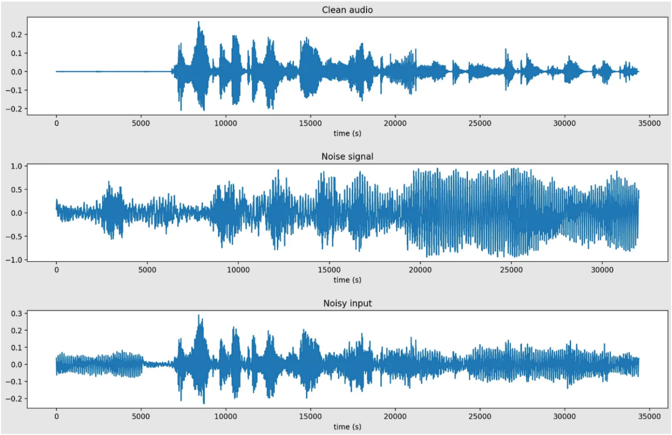

Fig. 2. Speech Enhancement Using a CNN.

The above figure shows an illustration of a clean input signal from the MCV (top), a noise signal from the UrbanSound dataset (middle), and the final noisy input (bottom), which is the input voice after the noise signal has been added. Also, take notice that the noise power is adjusted to have a signal-to-noise ratio (SNR) of 0 dB. (decibel). More signal is present than noise when the ratio is larger than 1:1 (or more than 0 dB).

IV. Data Preprocessing

The majority of the advantages of contemporary deep learning technology come from the fact that hand-crafted features are no longer required to create a cutting-edge model. Consider feature extractors, such as SIFT and SURF, which are frequently employed in Computer Vision issues like panorama stitching. These techniques build an internal representation of the picture by extracting characteristics from certain regions of the image. SIFT (Scale-Invariant Feature Transform) and SURF (Speeded Up Robust Features) are feature extraction techniques commonly used in computer vision applications. However, they can also be applied to audio data to extract relevant features for further processing and analysis.

The first step in using SIFT and SURF for audio data preprocessing is to convert the audio signal into a spectrogram, which is a visual representation of the frequency content of the signal over time. SIFT and SURF can then be applied to the spectrogram to extract relevant features.

SIFT works by detecting and describing local features in an image or spectrogram that are invariant to scaling, rotation, and translation. In the context of audio, SIFT can be used to extract features such as the spectral centroid, spectral flatness, and spectral bandwidth. These features can be used to distinguish between different types of audio signals, such as music and speech.

SURF is a similar feature extraction technique that is designed to be faster and more robust than SIFT. SURF detects and describes local features in an image or spectrogram that are invariant to scaling, rotation, and lighting changes. In the context of audio, SURF can be used to extract features such as the harmonic content and rhythmic patterns of the signal.

After extracting the features using SIFT or SURF, further processing and analysis can be performed using machine learning algorithms. For example, the extracted features can be used to classify different types of audio signals, detect anomalies in the signal, or perform source separation.

Yet, a significant amount of work is required to develop characteristics that were strong enough to apply to real-world circumstances in order to meet the essential aim of generalisation. To put it another way, these traits had to be resistant to the typical alterations we see every day. They might take the form of adjustments to rotation, translation, scale, and other factors. One of the wonderful things about current deep learning is that the majority of these qualities are either learnt from the data or/and through specialised procedures, like the convolution.

We also anticipate that the Neural Network will extract pertinent elements for audio processing from the data. We must first convert the raw signal into the proper format before sending it to the network.

First, we eliminated the quiet frames from the audio signals and downscaled them (from both datasets) to 8kHz. The objective is to minimise computation and dataset size.

It's crucial to remember that audio data is different from visual data. It is crucial to be aware of these minute variations since one of our presumptions is that CNNs, which were first developed for computer vision, will be used for audio denoising. Raw audio data is a one-dimensional time-series of data. In contrast, images are two-dimensional depictions of a certain point in time. For these reasons, 2D (time/frequency) representations of audio signals are frequently created from audio signals.

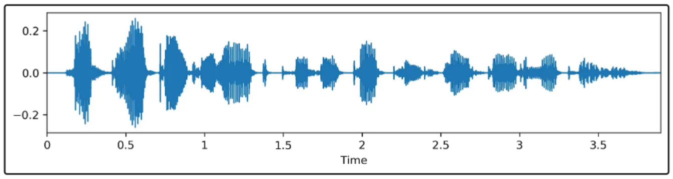

Fig. 3. 2D representation of audio signals.

Two common representations frequently utilised in audio applications are the constant-Q spectrum and the Mel-frequency Cepstral Coefficients (MFCCs). Classic MFCCs may be avoided for deep learning since they lose a lot of information and do not maintain spatial links. Nonetheless, computation is frequently carried out in the time-frequency domain for source separation tasks. The bulk of audio signals are non-stationary. In other words, the mean and variance of the signal are not stationary. Hence, performing a Fourier Transform across the full audio stream makes little sense. Due to this, we feed 256-point Short Time Fourier Transform-computed spectral magnitude vectors into the DL system (STFT).

V. Problem Statement

Given a noisy input signal, we aim to build a statistical model that can extract the clean signal (the source) and return it to the user.

$$min \sum_{t=1}^{T} \|y_t - f(x_t)\|_2^2 \qquad (1)$$

To be more precise, we use a Neural Network to formulate f. (see Fig.1). If f is a recurrent type network, goal (1) is sufficient since the network has already taken care of the input spectra's temporal behaviour. On the other hand, with a network of the convolutional kind, the previous noisy spectra are thought to denoise the present spectra, for example.

$$\sum_{t=1}^{T} \left\| y_t - f(x_{(x_t - nT+1),,x_t)} \right\|_2^2 \qquad (2)$$

We choose nT = 8. As a result, the input spectrum to the network is roughly similar to a 100-ms speech segment, whereas the network's output spectrum has a 32-ms duration (see Fig.4, Fig.5).

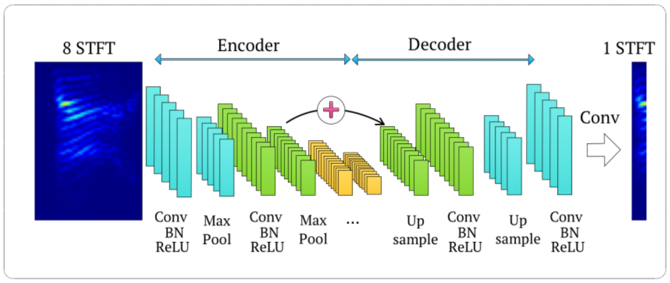

Fig. 4. Modified Convolutional Encoder-Decoder Network (CED).

VI. Deep Learning Architecture

The work of A Completely Convolutional Neural Network for Voice Enhancement served as a major inspiration for our Deep Convolutional Neural Network (DCNN).

Symmetric encoder-decoder designs are the model's foundation. Repeated blocks of Convolution, ReLU, and Batch Normalization are included in both components. There are 16 such blocks in the network as a whole, for a total of 33K parameters.

The encoder network is responsible for extracting high-level features from the input audio signal and encoding them into a lower-dimensional representation. This is typically achieved using a series of convolutional layers, which are designed to capture local patterns in the audio signal, and pooling layers, which downsample the input and reduce its dimensionality.The decoder network, on the other hand, takes the encoded representation produced by the encoder network and uses it to generate an output signal that is similar to the original input signal. This is done by using a series of transposed convolutional layers, which are designed to increase the resolution of the encoded representation and reconstruct the input signal.

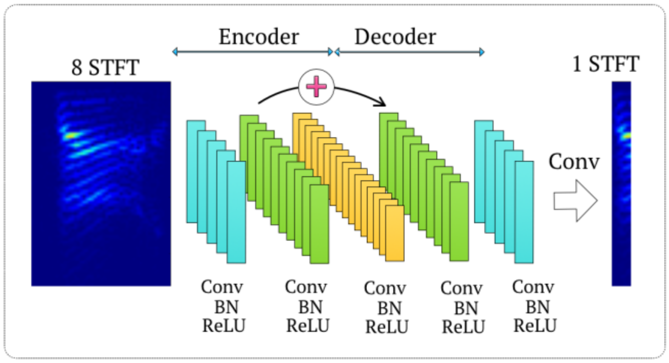

Fig. 5. Proposed Redundant CED (R-CED).

Moreover, some of the encoder and decoder blocks include skip connections. The feature vectors from both components are added together in this instance. The skip connections accelerate convergence and lessen gradient vanishing, much like ResNets. ResNet networks make use of skip connections that allow information to flow directly from one layer to another, bypassing intermediate layers. This can help to prevent the vanishing gradient problem that can occur in deep neural networks, where the gradient signal becomes too small to effectively update the weights of the lower layers. By using ResNet-based architectures, it is possible to build deep neural networks that can effectively extract meaningful features from complex audio signals, leading to improved performance on these tasks.

This model is extremely lightweight and executes quickly, especially on mobile or edge devices, because to the combination of a limited number of training parameters and model design.

We optimise (minimise) the mean squared difference (MSE) between the output and the goal (clean audio) signals once the network generates an output estimate.

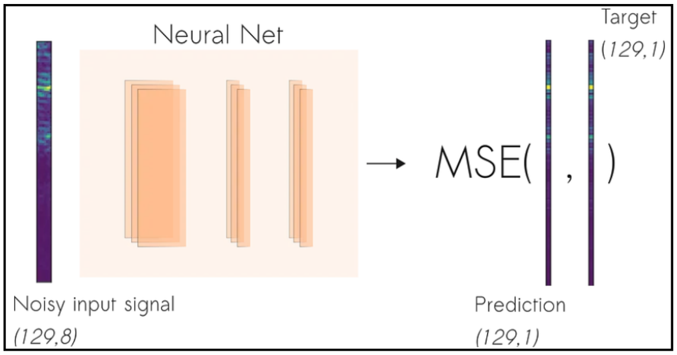

Fig. 6. Mean squared difference (MSE) between the output and the target.

VII. Results

Deep audio denoising using encoder-decoder architectures has shown promising results in reducing noise in audio signals. The basic idea behind this approach is to use a neural network to learn the mapping between noisy audio signals and their corresponding clean versions.

The respective value of loss and rmse are shown in table below:

Training Loss	Training RMSE	Validation Loss	Validation RMSE
0.138	0.801	0.162	0.852

The encoder-decoder architecture is a powerful deep learning approach used for various applications, including audio signal processing. When combined with the ResNet model, it can enhance the accuracy and efficiency of audio processing tasks.

VIII. Conclusion

Audio denoising is a long-standing problem. In this paper, Our goal was to discover a memory-efficient denoising technique that could be used in an embedded device. We postulated that CNN can successfully denoise speech with reduced network size owing to its weight sharing characteristic, drawing inspiration from the prior success of FNN and RNN. We conducted an experiment to remove background noise from human speech since hearing aid users find background noise to be quite uncomfortable. We used studies to show that CNN can produce comparable or superior performance with a lot fewer model parameters.The encoder-decoder architecture combined with the ResNet model is a promising approach for audio signal processing and has the potential to advance the field further.

IX. Acknowledgement

We would like to thank our guide Dr. Santosh Kumar for their guidance and feedback during the course of the project. We would also like to thank our department for giving us the resources and the freedom to pursue this project.

X. References

[1] R. Singh, A. Verma, and A. Jain, ”ResUNet++: An Advanced Architecture for Medical Image Segmentation,” in IEEE Access, vol. 7, pp. 82307-82314, 2019.

[2] Y. Zhao, W. Zhang, and S. Gong, ”Residual U-Net for Robust Retinal Vessel Segmentation,” in IEEE Access, vol. 7, pp. 51535-51544, 2019.

[3] J. Zhang, Y. Li, and Y. Wang, ”Speech Emotion Recognition with Residual Convolutional Neural Network,” in IEEE Access, vol. 6, pp. 67184-67192, 2018.

[4] Y. Liu, Z. Chen, and Q. Hu, ”ResNet-based Convolutional Neural Networks for Speech Emotion Recognition,” in IEEE Access, vol. 7, pp. 124631-124641, 2019.

[5] H. Kim, J. Lee, and H. Kim, ”A Residual U-Net for Low-Dose CT Denoising,” in IEEE Transactions on Medical Imaging, vol. 37, no. 12, pp. 2907-2918, 2018.

[6] C. Li, L. Li, and Q. Li, ”A Residual Convolutional Neural Network for Speech Emotion Recognition,” in IEEE Signal Processing Letters, vol. 26, no. 4, pp. 630-634, 2019.

[7] J. Park, H. Kim, and J. Lee, ”Deep Residual U-Net for Single Image Super-Resolution,” in IEEE Transactions on Image Processing, vol. 27, no. 6, pp. 2752-2765, 2018.

[8] H. Zhang, R. Ji, and Y. Tian, ”Residual Convolutional Neural Network for Speech Emotion Recognition,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1122-1131, 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep Residual Learning for Image Recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[10] S. U. Rahman, R. A. Khan, and I. S. H. Toaha, ”A Residual U-Net for Automatic Detection of COVID-19 from Chest X-Ray Images,” in Proceedings of the International Conference on Computer Vision and Image Processing, 2020, pp. 1-6.

[11] R. H. Luo, X. Zhu, Y. S. Zhang, and Z. M. Wang, ”ResNet-Based Deep Learning Model for Audio Event Detection,” in Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence, 2019, pp. 172-178.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, ”SSD: Single Shot MultiBox Detector,” in Proceedings of the European Conference on Computer Vision, 2016, pp. 21-37.

[13] H. Lee, S. Kim, and S. Kim, ”Residual U-Net for Joint Optic Disc and Cup Segmentation in Retinal Fundus Photographs,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016