AI Based Noise-Cancellation(Using DCU-Net Model)

UG PROJECT

MEMBERS

Soojal Singh	-	21095115
Avnesh Kumar	-	21095029
Ravi Kumar Meena	-	21095093

Under the supervision of:

Dr. Kishor P. Sarawadekar

DEPARTMENT OF ELECTRONICS ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI

CERTIFICATE

This is to certify that the UG Project entitled "**AI Based Noise-Cancellation(Using DCU-Net Model)**" submitted by Avnesh Kumar(21095029), Soojal Singh(21095115) and Ravi Kumar Meena(21095093), to the Department of Electronics Engineering, Indian Institute of Technology (Banaras Hindu University) Varanasi, in partial fulfilment of the requirements for the award of the degree "Bachelor of Technology" in Electronics Engineering is an authentic work carried out at Department of Electronics Engineering, Indian Institute of Technology (Banaras Hindu University) Varanasi under my supervision and guidance on the concept vide project grant as acknowledged.

Dr. Kishor P. Sarawadekar

Associate Professor

Department of Electronics Engineering,

Indian Institute of Technology (BHU) Varanasi

DECLARATION

I hereby declare that the work presented in this project titled "**AI Based Noise-Cancellation(Using DCU-Net Model)**" is an authentic record of our own work carried out at the Department of Electronics Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi as requirement for the award of degree of Bachelors of Technology in Electronics Engineering, submitted in the Indian Institute of Technology (Banaras Hindu University) Varanasi under the supervision of **Dr. Kishor P. Sarawadekar,** Department of Electronics Engineering, Indian Institute of Technology (Banaras Hindu University) Varanasi. It does not contain any part of the work, which has been submitted for the award of any degree either in this Institute or in other University/Deemed University without proper citation.

Avnesh kumar
(21095029)

Ravi Kumar Meena
(21095093)

Soojal Singh
(21095115)

ABSTRACT

This paper tackles the problem of the heavy dependence of clean speech data required by deep learning based speech enhancement methods by showing that it is possible to train deep speech enhancement networks using only noisy speech samples. Conventional wisdom dictates that in order to achieve good speech enhancement performance, there is a requirement for a large quantity of both noisy speech samples and perfectly clean speech samples, resulting in a need for expensive audio recording equipment and extremely controlled soundproof recording studios. These requirements pose significant challenges in data collection, especially in economically disadvantaged regions and for low resource languages. This work shows that speech enhancement deep neural networks can be successfully trained utilizing only noisy training audio. Furthermore, such training regimes achieve superior denoising performance over conventional training regimes utilizing clean training audio targets.

Contents

Chapter 1
INTRODUCTION

Speech enhancement is one of the most important and challenging tasks in speech applications where the goal is to separate clean speech from noise when noisy speech is given as an input. As a fundamental component for speech-related systems, the applications of speech enhancement vary from speech recognition front-end modules to hearing aid systems for the hearing-impaired.

Due to recent advances in deep learning, the speech enhancement task has been able to reach high levels in performance through significant improvements. When using audio signals with deep learning models, it has been a common practice to transform a time-domain waveform to a time-frequency (TF) representation (i.e. spectrograms) via short-time-Fourier-transform (STFT). Spectrograms are represented as complex matrices, which are normally decomposed into magnitude and phase components to be used in real-valued networks. In tasks involving audio signal reconstruction, such as speech enhancement, it is ideal to perform correct estimation of both components. Unfortunately, complex-valued phase has been often neglected due to the difficulty of its estimation. This has led to the situation where most approaches focus only on the estimation of a magnitude spectrogram while reusing noisy phase information

However, reusing phase from noisy speech has clear limitations, particularly under extremely noisy conditions, in other words, when signal-to-noise ratio (SNR) is low. This can be easily verified by simply using the magnitude spectrogram of clean speech with the phase spectrogram of noisy speech to reconstruct clean speech.

A popular approach to speech enhancement is to optimize a mask which produces a spectrogram of clean speech when applied to noisy input audio. One of the first mask-based attempts to perform the task by incorporating phase information was the proposal of the phase-sensitive mask. Since the performance of PSM was limited because of reusing noisy phase, later studies proposed using complex-valued ratio mask (cRM) to directly optimize on

complex values. We found this direction promising for phase estimation because it has been shown that a complex ideal ratio mask (cIRM) is guaranteed to give the best performance out of other ideal masks such as ideal binary masks, ideal ratio masks, or PSMs. Moreover, this approach jointly estimates magnitude and phase, removing the need of separate models. To estimate a complex-valued mask, a natural desire would be to use an architecture which can handle complex-domain operations. Recent work gives a solution to this by providing deep learning building blocks adapted to complex arithmetic.

Chapter 2

Detailed Literature Survey of Existing Technologies

With the rise of urbanization and technological advancements, noise pollution has become a prevalent issue in our modern society. However, technology has also provided us with solutions to tackle this problem. Two commonly used approaches are Active Noise Cancellation (ANC) technology, primarily used to clean the sound from the surrounding environment for hearing better (think headphones on airplanes), and AI-based noise cancellation algorithms, which excel at filtering noise from microphones (outbound stream) and headphones (inbound stream) for real-time communications. These two solutions differ significantly. ANC is directly perceived in the ear, while microphone noise canceling allows filtering out noise from the surrounding environment which passes through a microphone, enabling clear voice communications during calls or recordings.

ANC typically involves physical mechanisms to block or reduce external noises. These mechanisms can be found in various devices such as noise-canceling headphones, earplugs, and even architectural designs of buildings. They utilize microphones to capture external sounds and generate sound waves that are 180 degrees out of phase with the incoming noise, effectively canceling out the unwanted sounds.

It’s important to note that ANC often require specific hardware devices, while microphone noise-cancellation algorithms are software-based, making them adaptable to a variety of digital devices and applications. AI-based algorithms, in particular, are highly efficient, increasing their scalability and accessibility across different devices like wearables, smart speakers and smartphones.

ANC is a technique based on the principle of superposition, i.e., an antinoise with the same amplitude and opposite phase is generated and combined with an unwanted noise, thus resulting in the cancellation of both noises. However, ANC is still not widely used owing to the effectiveness of control algorithms, and to the physical and economical constraints of practical applications.

#Our Approach Using AI Based Model

The approach is to use an end-to-end model which takes noisy audio as raw waveform inputs. First, the time domain waveform is converted into the time-frequency domain using the Short Time Fourier Transform (STFT). This transform outputs a linearly scaled, complex matrix spectrogram, factorizable into a real-valued phase component and a complex-valued magnitude component. The STFT is computed with a FFT size of 3072, number of bins equaling 1536, and hop size of 16ms. Normalization is then carried out to ensure compliance with Parseval's energy-conservation property, meaning that the energy in the spectrogram equals the energy in the original time domain waveform. The output frequency-time domain waveform is converted again into time domain throught ISTFT(Inverse Short Time Fourier Transform).

#Project Flow

2.1 Datasets used and Dataset Generation

Due to the lack of a pre-existing benchmark dataset containing noise in both the input and target, a collection of datasets was generated in order to perform Noise2Noise training. We have used “UrbanSoundk8k” for real-world noise samples and “Voice Bank + Demand” for speech samples . All 10 noise categories of the UrbanSound8K dataset were used. This dataset was chosen for its collection of samples from numerous real-world noise categories. Separate training and testing datasets are created for each UrbanSound8K noise category N. For each noise type N, the input training audio file is generated by overlaying a random noise sample from N with repetition on top of a clean audio file. The noise is then overlapped over the clean audio using PyDub, which truncates or repeats the noise such that it covers the entire speech segment. Next, a corresponding target training audio file is generated using the same underlying clean audio file, and a random noise sample from a category that is not N. The Mixed category dataset was created by picking a random noise category for the input file, while picking another random noise category for the target file, ensuring both don’t use the same noise category N. The White noise category dataset was generated by using random additive white gaussian noise with SNR scaled randomly in the range 0 to 10, on both the input and target training files. The testing dataset was generated in the same fashion. The testing input is the noisy audio file, whereas the testing reference is the underlying clean audio file.

2.2 Pre-processing of data

The original raw waveforms were first downsampled from 48kHz to 16kHz. For the actual model input, complex-valued spectrograms were obtained from the downsampled waveforms via STFT with a 64ms sized Hann window and 16ms hop length.

2.3 Phase Aware Speech Enhancement

In this section we will provide details on our approach, starting with our proposed model Deep Complex U-Net, followed by the masking framework based on the model. Finally, we will introduce a new loss function to optimize our model, which takes a critical role for proper phase estimation.

The notations used in the report are defined in this section.The input mixture signal $x(n) = y(n) + z(n) \in R$ is assumed to be a linear sum of the clean speech signal $y(n) \in R$ and noise $z(n) \in R$, where estimated speech is denoted as $\hat{y}(n) \in R$. Each of the corresponding time-frequency (t, f) representations computed by STFT is denoted as $X_{t,f} \in C$, $Y_{t,f} \in C$, $Z_{t,f} \in C$, and $\hat{Y}_{t,f} \in C$. The ground truth mask cIRM is denoted as $M_{t,f} \in C$ and the estimated cRM is denoted as $\hat{M}_{t,f} \in C$, where $M_{t,f} = Y_{t,f} / X_{t,f}$.

->DEEP COMPLEX U-NET

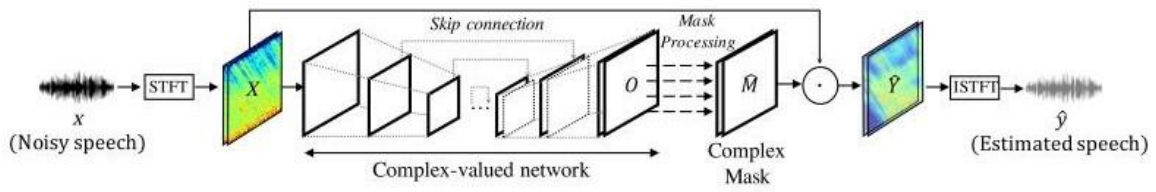


Figure 2: Illustration of speech enhancement framework with DCUnet.

U-Net is an architecture for semantic segmentation. It consists of a contracting path and an expansive path. The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step we double the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total the network has 23 convolutional layers.

Deep Complex U-Net (DCUnet) is an extended U-Net, refined specifically to explicitly handle complex domain operations. In this section, we will describe how U-Net is modified using the complex building blocks.

Our network consists of two parts: the **encoder** which extracts relevant features from images, and the **decoder** part which takes the extracted features and reconstructs a segmentation mask.

In this manner, high-resolution features from the encoder path are combined and reused with the upsampled output.

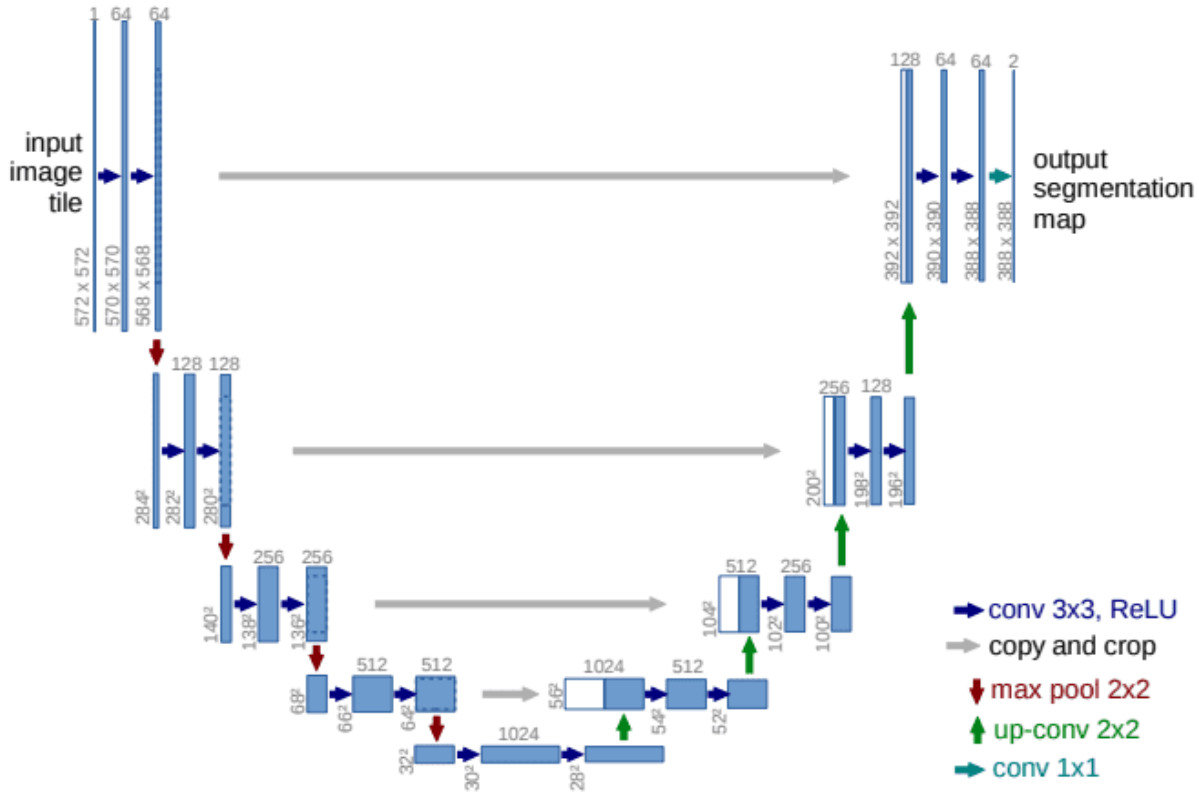


Fig - DCUNet Model

Encoder :

It consists of the repeated application of two 3x3 convolutions. Each conv is followed by a ReLU and batch normalization. Then a 2x2 max pooling operation is applied to reduce the spatial dimensions. Again, at each downsampling step, we double the number of feature channels, while we cut in half the spatial dimensions.

The encoder finally extracts relevant features from images.

Decoder :

Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 transpose convolution, which halves the number of feature channels. We also have a concatenation with the corresponding feature map from the contracting path, and usually a 3x3 convolutional (each followed by a ReLU). At the final layer, a 1x1 convolution is used to map the channels to the desired number of classes.

Moreover, padding is used to keep the size of the feature maps the same after convolution operations.

Thereby, decoder part takes extracted features and reconstructs a segmentation mask.

Complex-valued Building Blocks : Given a complex-valued convolutional filter $W = A+iB$ with real-valued matrices A and B, the complex convolution operation on complex vector $h = x + iy$ with W is done by,

$$W * h = (A*x - B*y) + i(B*x + A*y).$$

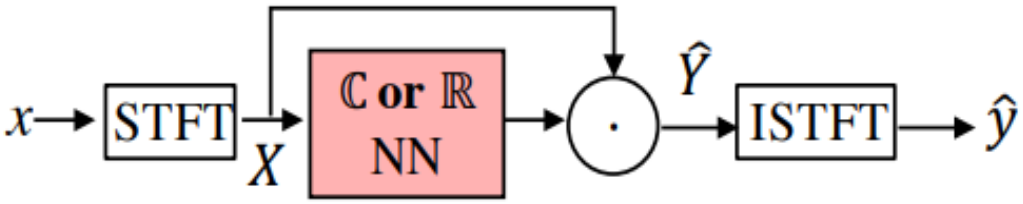
In practice, complex convolutions can be implemented as two different real-valued convolution operations with shared real-valued convolution filters.

Activation functions like ReLU were also adapted to the complex domain.CReLU, an activation function which applies ReLU on both real and imaginary values, is used to produce the best results.

Modifying U-Net :

The proposed Deep Complex U-Net is a refined U-Net architecture applied in STFT-domain. Modifications done to the original U-Net are as follows. Convolutional layers of UNet are all replaced to complex convolutional layers. Here, the convolution kernels are set to be independent to each other by initializing the weight tensors as unitary matrices for better generalization and fast learning. Complex batch normalization is implemented on every convolutional layer except the last layer of the network. In the encoding stage, max pooling operations are replaced with strided complex convolutional layers to prevent spatial information loss. In the decoding stage, strided complex deconvolutional operations are used to restore the size of input. For the activation function, we modified the previously suggested CReLU into leaky CReLU, where we simply replace ReLU into leaky ReLU,, making training more stable.

2.4 Complex-valued Masking on Polar Coordinates



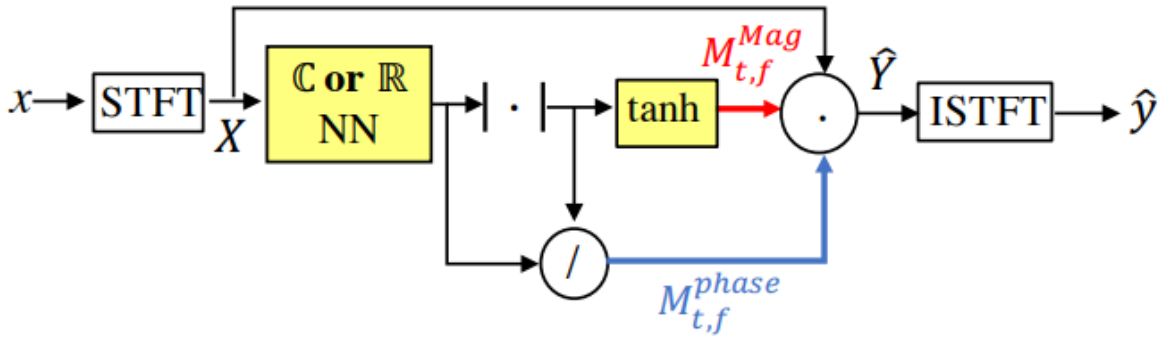
1. Unbounded Mask

As our proposed model can handle complex values, we aim to estimate cRM(Complex Ratio Mask) for speech enhancement. Although it is possible to directly estimate the spectrogram of a clean source signal, but better performance can be achieved by applying a weighting mask to the mixture spectrogram. One thing to note is that real-valued ratio masks (RM) only change the scale of the magnitude without changing phase, resulting in irreducible errors. On the other hand, cRM also perform a rotation on the polar coordinates,

allowing to correct phase errors. In other words, the estimated speech spectrogram Yˆ (t,f) is computed by multiplying the estimated mask Mˆ (t,f) on the input spectrogram X(t,f) as follows:

$$\hat{Y}_{t,f} = \hat{M}_{t,f} \cdot X_{t,f} = \left|\hat{M}_{t,f}\right| \cdot \left|X_{t,f}\right| \cdot e^{i(\theta_{\hat{M}_{t,f}} + \theta_{X_{t,f}})}$$

In this state, the real and imaginary values of the estimated cRM is unbounded. Although estimating an unbounded mask makes the problem well-posed, we can imagine the difficulty of optimizing from an infinite search space compared to a bounded one.



2. Bounded (tanh) Mask

To tackle wih the problems encountered in the unbounded cRM mask, we used a polar-coordinate-wise cRM method that imposes non-linearity on the magnitude part and not on the phase part. More specifically, we use a hyperbolic tangent non-linearity to bound the range of magnitude part of the cRM be [0, 1) which makes the mask bounded in an unit-circle in complex space. The

corresponding phase mask is naturally obtained by dividing the output of the model with the magnitude of it.

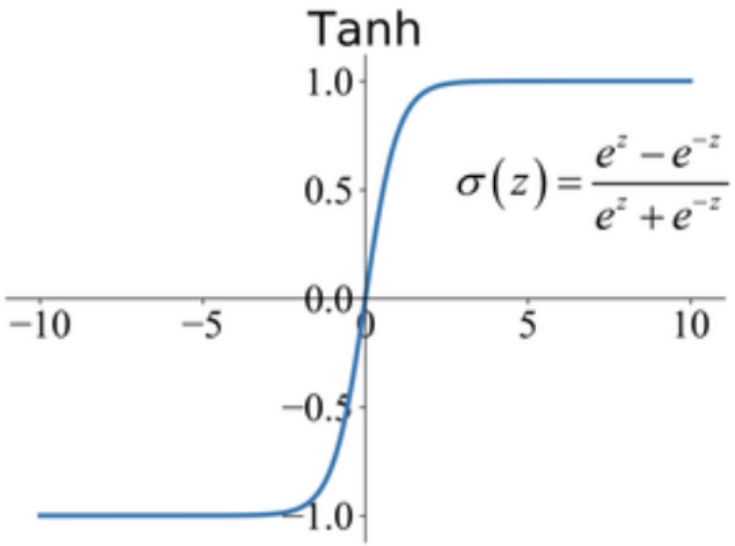


Fig - tan hyperbolic function

More formally, let g(·) be our neural network and the output of it be Ot,f = g(Xt,f). The proposed complex-valued mask Mˆ t,f is estimated as follows:

$$\hat{M}_{t,f} = \left|\hat{M}_{t,f}\right| \cdot e^{i\theta_{\hat{M}_{t,f}}} = \hat{M}_{t,f}^{mag} \cdot \hat{M}_{t,f}^{phase}$$

$$\hat{M}_{t,f}^{mag} = \begin{cases} tanh(|O_{t,f}|) & \text{(bounded cond.)} \\ |O_{t,f}| & \text{(unbounded cond.)} \end{cases}, \qquad \hat{M}_{t,f}^{phase} = O_{t,f}/|O_{t,f}|$$

Chapter 3

Preliminary Results and Insights

We have trained our model with the Noise2Noise approach, for all 10 (numbered 0-9) UrbanSound noise classes. The output spectrograms obtained for each of the stages is shown below:-

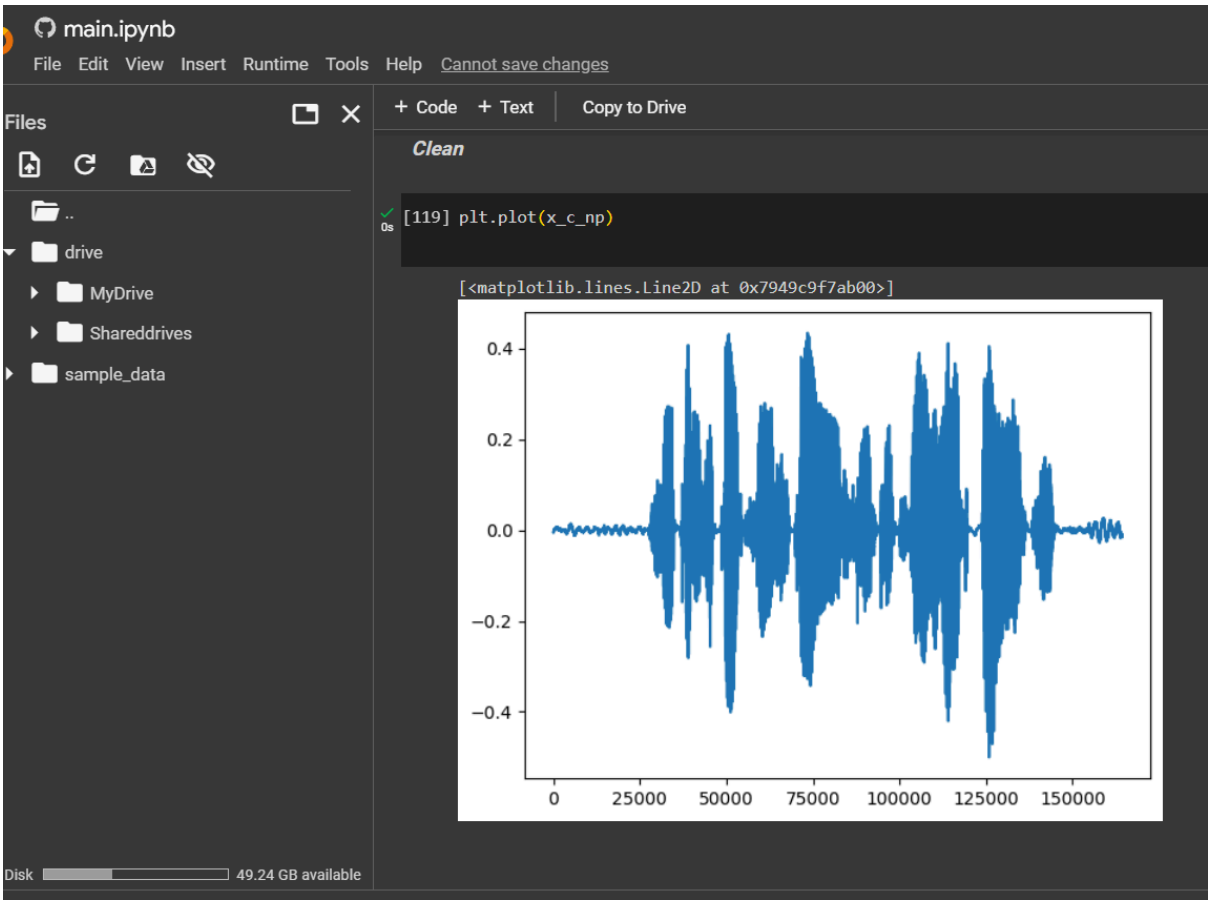


Fig - Spectrogram of Clean speech sample obtained from the dataset itself

Inference - The predicted spectrogram is close to the original clean speech sample but not accurate. This is because we trained only 1000 files out of 11000 files dataset as it required a very high end GPU but we were short of it.

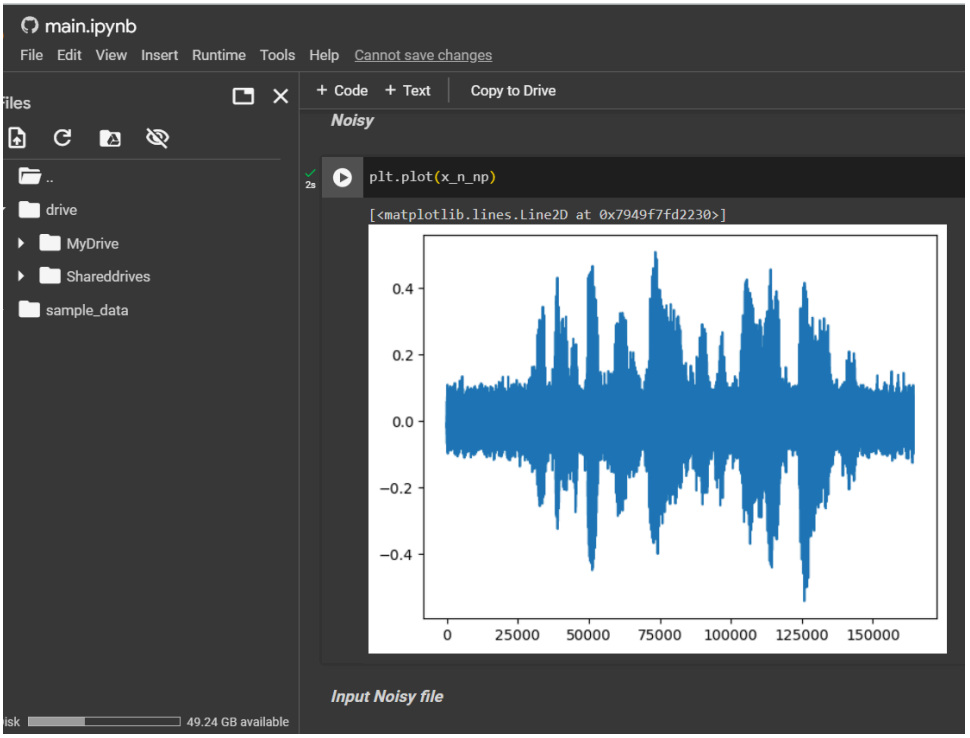


Fig - Spectrogram of input noise signal

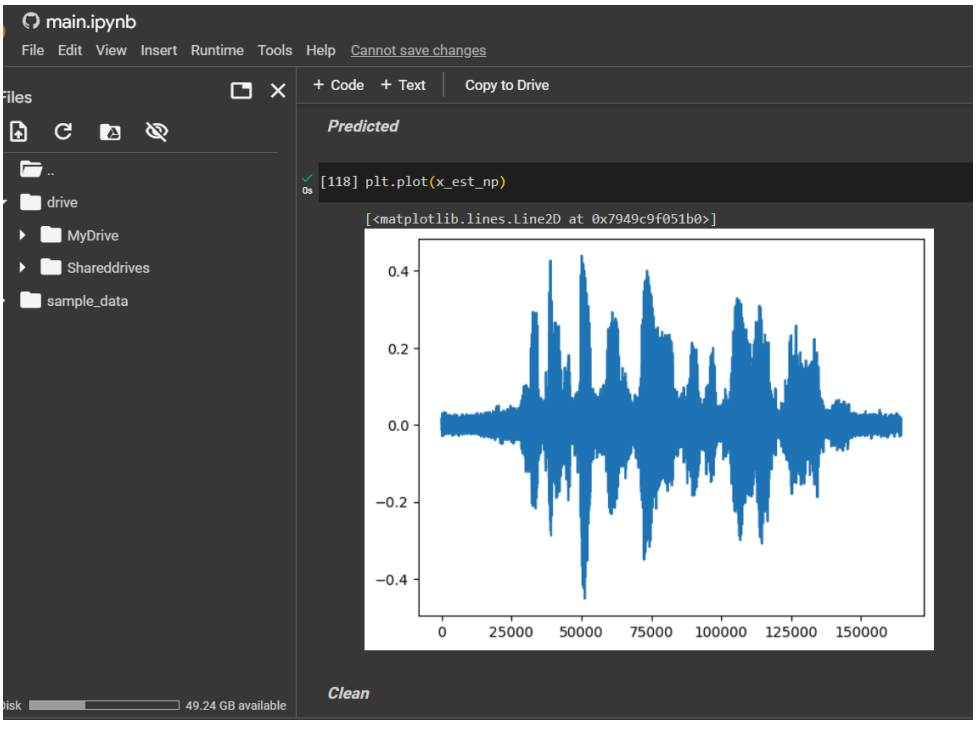


Fig - Spectrogram of our predicted speech signal output

Chapter 4

Conclusions and Future Scope

This work proves that deep neural networks can be trained to denoise audio by employing a technique that uses only noisy audio samples as both the input as well as the target to the network, subject to the noise distributions being zero mean and uncorrelated. This is demonstrated by using the DCUnet-20 model to denoise real-world UrbanSound8K noise categories. Furthermore we see that our proposed Noise2Noise approach in the speech domain produces superior denoising performanc for low SNR UrbanSound8K noise categories. This is a general conclusion seen across all noise categories and from the UrbanSound8K dataset. A limitation of this approach is the fact that the noisy training input and target pairs need to have the same underlying clean speech. Although this type of data collection is still practical - for example having multiple microphones in various spatial locations to the noisy speech source.

There's potential for integrating AI-based noise cancellation algorithms, such as those based on DCU-Net architecture, into hardware devices and consumer electronics. This could lead to the development of smarter noise-canceling headphones, audio recording equipment, and communication devices that can effectively remove unwanted noise in real-time. The mobile companies have already started providing AI features for their handsets. For instance, recently Samsung have released an option for all their flagship devices of live noise cancellation while a person is on a phone call.

Chapter 5

Bibliography

Research Papers :

[1]Wenling Shan.Kihyuk Sohn.2016.Concatenated ReLU.

https://arxiv.org/abs/1603.05201

[2]Hyeong-Seok Choi.2019.Phase aware speech enhancement with Deep Complex U-net.

https://paperswithcode.com/paper/phase-aware-speech-enhancement-with-deep-1

[3]Yoshinobu Kajikawa.2013. Active Noise Control.

https://www.researchgate.net/publication/261047508_Recent_applications_and_challenges_on_active_noise_control

Online resources :

[1] Minh Tran.2022.Understanding U-net.
https://towardsdatascience.com/understanding-u-net-61276b10f360

[2]Nikolas Adologlou.2021.An overview of Unet architectures for semantic segmentation and biomedical image segmentation.
https://theaisummer.com/unet-architectures/

[3]Hongwei Ding, Leiyang Chen.2021. DCU-Net.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8359769/