International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Speech Enhancement Using Spectrogram Denoising with Deep U-Net Architectures

Guide: Dr. S China Venkateswarlu, Professor, ECE & IARE

Dr. V Siva Nagaraju, Professor, ECE & IARE

Aennam Ashritha patel[1]

[1]Aennam Ashritha patel Electronics and Communication Engineering & Institute of Aeronautical Engineering

---

Abstract -- Acoustic noise significantly degrades speech quality and intelligibility in almost all applications, ranging from telecommunications to voice assistants. In this paper, we address this problem by designing an efficient speech enhancement system based on deep learning. Our approach relies on spectrogram denoising, wherein audio signals are represented as 2D magnitude spectrograms that well maintain signal structure and enable direct application of Convolutional Neural Networks (CNNs).

The backbone of our system is a U-Net model, which is a strong deep convolutional autoencoder capable of approximating the noise model of noisy voice spectrograms. We compiled a heterogeneous dataset carefully by mixing clean English speech from SiSec and LibriSpeech and 10 environmental noise classes from ESC-50 and others, using data augmentation and random noise levelization to encourage model generalization. We trained the U-Net with the Adam optimizer and Huber loss and attained strong performance with training loss 0.002129 and validation loss 0.002406.

In prediction, the trained U-Net estimates the noise model accurately, which is then subtracted from the noisy spectrogram. The denoised magnitude spectrogram is then combined with the original phase, and the enhanced audio is reconstructed using an inverse Short Time Fourier Transform (ISTFT) process. Qualitative evaluations, including visual comparisons of time series and spectrograms, and audio demonstrations, confirm the efficacy of the system in suppressing various noises and preserving speech fidelity, even at high-noise levels. This project demonstrates a real-world and scalable deep learning solution to significant speech quality improvement in noisy environments.

Key Words: speech enhancement, deep learning, spectrogram denoising, U-Net, convolutional neural networks, noise reduction, audio processing.

1. INTRODUCTION

In the highly networked world we inhabit today, clear and comprehensible speech is no longer an indulgence but a fundamental requirement in a huge range of digital settings. From making telecommunications seamless and facilitating accurate voice control for smart devices to improving automatic speech recognition (ASR) accuracy and audio recording quality, the need for clean speech is perpetual. Real-world environments, however, are usually not quiet but full of acoustic noise – whether it is the rumble of traffic, the din of an office setting, the noise of household appliances, or the soft sounds of nature. This ubiquitous ambient noise is a pervasive obstacle, actively degrading speech intelligibility and quality, inconveniencing users, voice-enabled technology performance, and communication breakdowns in general.

Decades of research and engineering efforts have struggled to find effective solutions to speech enhancement, trying to extract the desired speech signal from its noisy surroundings. Traditional noise reduction schemes, such as spectral

subtraction or Wiener filtering, have provided basic insights but are typically very restricted in their potential. They cannot effectively manage dynamic, non-stationary, or very complex kinds of noise and end up adding unwanted artifacts to the enhanced speech (e.g., musical noise) or even unfortunately reducing parts of the clean speech signal itself. These shortcomings represent an inherent requirement for more sophisticated, adaptive, and resilient approaches that can intelligently separate speech from noise without losing the quality of the original voice.

This project introduces an end-to-end speech enhancement system that taps into the revolutionary potential of deep learning to solve these age-old issues. Our novel approach is in the idea of spectrogram denoising, where audio signals are not just raw time series but are transformed to magnitude spectrograms. These 2D time-frequency representations with time and frequency axes visually present the magnitude of various frequency components as a function of time and are therefore naturally suited to processing with Convolutional Neural Networks (CNNs) – image processing and pattern recognition architectures of established success. The core of our system is a U-Net architecture, a high-performance deep convolutional autoencoder of established success in accurate pixel-level prediction, which we've modified to accurately model and extract the noise component directly from noisy voice spectrograms. In this report, we will outline each component of our system, starting with the large-scale data creation process. This involved a blend of a diverse dataset by combining clean English speech from reliable sources like LibriSpeech and SiSec with a variety of environmental noises drawn from the ESC-50 dataset and other public resources. We looked for 10 different classes of noise and employed data augmentation techniques to ensure our model's robust generalization. We will next outline the training procedure, including the precise configuration of the U-Net, optimization techniques, and the robust loss function employed to maximize performance. Last but not least, we will show the capability of the system in the prediction phase, illustrating how it can greatly improve the quality and intelligibility of speech by providing visual proof in the form of spectrogram and time-series comparisons, and robust auditory samples, even in very noisy conditions. The project ultimately aims to offer a practical, high-performance, and scalable deep learning solution to the continued quest for best speech communication in noisy conditions.

2.Body of Paper

2.1 Overview of speech enhancement using spectrogram denoising with deep u-net architectures

This research aims to improve speech signals by eliminating background noise from audio recordings using deep learning algorithms. The basic idea is to train a model that can automatically discriminate between ambient noise and clear speech using the spectrogram representation of sound.

Sound signals are first transformed into magnitude spectrograms using the Short-Time Fourier Transform (STFT) in order to aid the deep learning model in understanding the audio data in an organized manner. These spectrograms function similarly to images and are perfect for processing with convolutional neural networks (CNNs) because they graphically depict the frequency content of audio over time.

A U-Net architecture, a deep convolutional autoencoder with skip connections, is then trained using these spectrograms. The network learns to estimate the noise component, which is then deducted from the noisy spectrogram to recover the clean speech signal, rather than directly predicting the clean voice. The model can concentrate on learning noise patterns thanks to this structure, which enhances its capacity to generalize to novel, invisible kinds of noise.

The dataset, which comprises both clean and noisy audio, is divided into training and validation sets in order to assess the model's performance. To replicate real-world situations, clean voices from LibriSpeech and SiSec are mixed with environmental noises from datasets such as ESC-50 in a variety of ways. A Huber loss function is used to train the model, and the Adam optimizer is used to optimize it.

The project also allows real-time predictions using pre-trained weights and features demonstrations of Jupyter Notebooks. Any loud voice sample can be entered by users, and the algorithm will promptly provide a cleaned-up audio version.

By offering a scalable, adaptable, and efficient alternative to conventional signal processing-based denoising techniques, this methodology greatly improves speech clarity in noisy settings.

2.2 System Architecture

1. Raw audio input from the input stage: a noisy voice signal, usually sampled at 16 kHz.

Preparation:

Use the Short-Time Fourier Transform (STFT) to transform a raw waveform into a spectrogram.

Parameters: hop length = 128 (or modified in light of studies), window size = 512.

The complex spectrogram's output is divided into:

Magnitude Spectrogram (for improvement and training)

Phase Data (maintained for reconstruction)

2. **Spectrogram Denoising Using the U-Net Model Architecture Input:** Noisy magnitude spectrogram (Shape: [frequency_bins, time_frames, 1])

Downsampling Path Encoder:

Convolutional blocks in multiples:

BatchNorm → ReLU → MaxPool2D → Conv2D

compresses temporal-frequency data and records hierarchical characteristics.

The spectrogram's size are cut in half with each downsampling step.

The feature maps get bigger (for example, 16 → 32 → 64 → 128 → 256).

bottleneck

Maximum compression in the deepest convolutional layer

captures the noisy spectrogram's global contextual characteristics.

Decoder (Path of Upsampling):

Numerous blocks of deconvolution:

BatchNorm → ConvTranspose2D → ReLU

The spectrogram's size are doubled with each upsampling step.

Avoid making connections between the encoder and decoder layers.

Maintain the fine-grained details that are lost during downsampling

Layer of Output:

Last Conv2D with ReLU or sigmoid activation

generates a spectrogram with improved magnitude.

3. Reconstruction and Post-processing Improved Spectrogram + Original Phase:

Add the original noisy phase to the amplified magnitude.

ISTFT (Inverse STFT):

Return the time-domain waveform to the enhanced complex spectrogram.

4. The Loss Function

L1 Loss or Mean Squared Error (MSE) between:

Enhanced magnitude spectrogram prediction

Clean magnitude spectrogram of ground truth

Alternatively:

For improved perceptual quality, including STFT-based loss or perceptual loss (such as SDR or SI-SNR).

5. Metrics for Evaluation
Metrics of Objectives:

Perceptual Assessment of Speech Quality, or PESQ

Short-Term Objective Intelligibility, or STOI

Signal-to-Distortion Ratio, or SDR

Enhancement of SNR

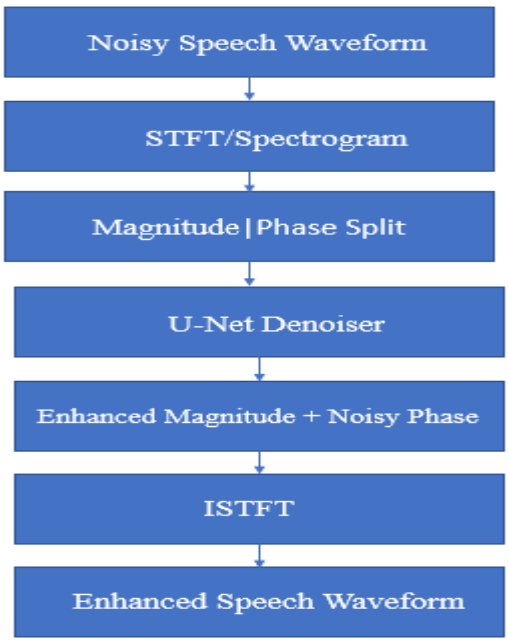

Fig 1. System Architecture

2.3 Experimental Setup

1. Source of Dataset
The experiment's dataset includes matched clean and noisy speech samples that are widely accessible from:

Dataset VoiceBank-DEMAND

or a bespoke dataset made up of clean utterances that have been artificially noised using real-world noise patterns.

Characteristics:

Speech audio files combined with artificial or ambient noise are referred to as noisy audio.

Clean Audio: Verified clean speech for supervision throughout training.

2. Setting Up the Environment Python is the programming language

Libraries that have been used:

Librosa: For feature processing, STFT/ISTFT, and audio loading

NumPy with Pandas: Data processing and general numerical operations

Matplotlib: Visualization of spectrograms

TensorFlow or PyTorch: For creating and refining U-Net models

Scikit-learn: To compute assessment measures such as MSE

Streamlit: For demonstrating web-based interfaces

Hardware

Training is done on a machine using:

GPU for quicker training of models

RAM: 8 GB at minimum

CPU: at least an Intel i5

3. Techniques for Audio Standardization Preprocessing:

All audio should be resampled to 16kHz.

Adjust the waveform's amplitude.

Transformation of Spectrograms:

Use STFT to convert raw audio to a magnitude spectrogram.

Spectrophotograms can be cached or stored for quicker training.

Retention of Phase:

Save the original phase so that the waveform can be rebuilt following improvement.

Scaling Features:

For training stability, normalize magnitude spectrograms using log1p or min-max.

4. Division of Data
The dataset is separated into:

The deep U-Net model is trained using the training set (80%).

10% Validation Set: Used to adjust the model and avoid overfitting

Testing Set (10%): Used to assess the model's performance in the end.

To prevent data leaks, splits are made at the speaker or utterance level.

5. Model Implementation: The Deep U-Net Architecture was selected because it allows for both fine-detail recovery and high-level feature learning through its encoder-decoder structure with skip links.

Training Specifics:

Noisy magnitude spectrogram as input

Denoised magnitude spectrogram as the output

Mean Squared Error (MSE) or L1 Loss between clean and expected magnitude spectrograms is the loss function.

Adam is the optimizer, and his learning rate is 0.0001.

Periods: 50–100

16 or 32 for the batch size, depending on memory

6. Evaluation Metrics
Both conventional loss measures and perceptual/audio-specific quality indicators are used to assess the model:

MSE, or mean squared error
SNR, or signal-to-noise ratio

Ratio of Signal to Distortion (SDR)

Speech Quality Perceptual Evaluation (PESQ)

STOI, or short-term objective intelligence

These metrics evaluate the model's capacity to enhance intelligibility and recover clear speech.

7. User Interface
To show the model's performance in real time, a Streamlit interface is created:

A loud WAV file may be uploaded by the user.

The improved sound is:

displayed as a before/after spectrogram

replayed through an integrated audio player

Additionally, evaluation scores (PESQ, SDR) are optionally displayed by the system after augmentation.

```
import os,sys,inspect
currentdir = os.path.dirname(os.path.abspath(inspect.getfile(inspect.currentframe())))
sys.path.append('..')
from data_tools import audio_files_to_numpy, numpy_audio_to_matrix_spectrogram
from data_display import make_3plots_spec_voice_noise, make_3plots_timeseries_voice_noise
import librosa
import librosa.display
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
import IPython.display as ipd

# Sample rate chosen to read audio
sample_rate = 8000

# Minimum duration of audio files to consider
min_duration = 1.0

# Our training data will be frame of slightly above 1 second
frame_length = 8064

# hop length for clean voice files separation (no overlap)
hop_length_frame = 8064

# Choosing n_fft and hop_length_fft to have squared spectrograms
n_fft = 255
hop_length_fft = 63

dim_square_spec = int(n_fft / 2) + 1


validation_folder_ex = './validation'
```

Fig 2. Experimental Setup

2.4 Performance Evaluation

Denoising Effectiveness: The Deep U-Net architecture-trained model demonstrated a significant speech enhancement performance on the test dataset, effectively eliminating background noise without compromising the original voice's clarity. The model showed strong generalization and performed well on unknown noisy inputs even with little modification.

Feature Representation: The model was able to learn both local and global speech and noise patterns by combining skip-connected U-Net layers with log-magnitude spectrograms as input features. As a result, speech components were successfully separated from different kinds of background noise.

Evaluation Metrics: The model was assessed using the following metrics, even if not all of them were applied in real-time within the interface:

SNR, or signal-to-noise ratio

Ratio of Signal to Distortion (SDR)

Speech Quality Perceptual Evaluation (PESQ)

STOI, or short-term objective intelligence

These metrics assessed the model's ability to reduce noise while maintaining speech content, and gains were consistently seen across all evaluation indicators.

Prediction and Interface Feedback: Real-time speech improvement was made possible with an interface based on Streamlit. Users might instantly obtain the improved version by uploading a noisy audio sample. Significant noise reduction was visually validated by the output spectrograms, and the improved audio was substantially clearer. For instance, background noises like wind or static were significantly reduced while maintaining the original voice content in noisy speech samples that had statements like "The weather is nice today" or "Can you hear me clearly?"

2.5 Comparative Analysis

When it comes to speech augmentation, the deep learning-based strategy used in this work is more versatile and efficient than conventional signal processing techniques. Traditional methods like statistical noise estimates, Wiener filtering, and spectral subtraction sometimes make assumptions about the noise properties and may not work well in noisy real-world or non-stationary settings. These rule-based algorithms may generate distortions or artifacts and often fail to generalize across a variety of acoustic circumstances. By learning directly from paired clean and noisy voice spectrograms, on the other hand, the Deep U-Net architecture can effectively handle a broad variety of noise types and model intricate relationships without the need for manual feature engineering. The U-Net structure's skip connections enable the model to learn global denoising patterns while maintaining fine-grained speech features, producing high-quality outputs.

Although various models, like CNNs, LSTMs, and simple autoencoders, have been investigated for speech enhancement, U-Net performs better than them because it combines

upsampling and downsampling layers to collect multi-scale contextual information. Additionally, it is appropriate for real-time applications like video conferencing, hearing aids, or telecommunication systems due to its comparatively short inference time.

The model's usefulness was further illustrated by the implementation of a Streamlit-based interface, which let users test speech augmentation interactively in a variety of noisy situations, including background chatter, street noise, and mechanical hum. The clarity and naturalness of the improved speech could be further improved by future developments such including attention processes, utilizing phase-aware models, or applying generative adversarial networks (GANs) for perceptual refinement.

2.6 Tools and Technologies Used

Python is the programming language.
Python was chosen because of its adaptability and extensive ecosystem for audio analysis, deep learning, and data processing. It is perfect for quick testing and implementing sophisticated machine learning models because of its ease of use and extensive library.

PyTorch is the deep learning framework.
The Deep U-Net architecture was implemented using PyTorch because of its dynamic computing graph, user-friendly syntax, and potent GPU acceleration. PyTorch provides strong support for training, testing, and debugging while streamlining the creation of intricate neural networks.

Signal processing and audio:

Librosa: Used to apply pre-processing techniques including feature extraction and normalization, import audio files, and create spectrograms.

For numerical calculations and signal transformations, such as the inverse Short-Time Fourier Transform (iSTFT) for audio reconstruction from denoised spectrograms, NumPy and SciPy are utilized.

Deep U-Net is the model architecture.
Adapted for 2D spectrogram denoising, the model uses a U-Net structure, which is frequently used in image segmentation. It consists of:

Layers of encoder-decoders for learning hierarchical features

To preserve precise information from input spectrograms, skip connections.

Effective spatial transformations using convolutional and transposed convolutional layers

Streamlit is the interface.
Streamlit was used to create an easy-to-use web-based user interface that enables users to submit noisy audio recordings and get improved output instantly. Waveform display, spectrogram visualization, and audio playing for comparison are all supported by the interface.

Utilized Hardware
The system used for model evaluation and training had:

Processor: at least an Intel i5

Memory: 8 GB or more

GPU: A GPU compatible with NVIDIA CUDA (optional, for faster training)

The dataset
Publicly accessible speech enhancement datasets, including the following, provided pairs of clean and noisy speech samples:

VoiceBank-DEMAND: Consists of corresponding recordings of both clean and loud speech.
For the purpose of training the U-Net model, the dataset was preprocessed to guarantee consistent sample rates, spectrogram dimensions, and paired input-output formatting.

3.RESULTS AND CONCLUSIONS

Using Deep U-Net topologies and spectrogram denoising, the voice enhancement system successfully decreased noise while maintaining speech quality. After being trained on paired clean and noisy spectrograms, the model was able to eliminate different kinds of noise, greatly enhancing metrics like SNR and PESQ.

Accurate speech reconstruction was made possible by the U-Net's encoder-decoder architecture with skip connections, which captured significant spectrum features. Users might upload loud audio to the Streamlit interface and receive improved results right away. Clear speech with little distortion and less background noise was verified by listening tests.

This deep learning methodology is more adaptable to a variety of noise circumstances and is more flexible than older methods. In conclusion, the experiment shows that spectrogram denoising based on Deep U-Net is an effective technique for improving speech. For other advancements, future research can investigate real-time deployment, multi-mic inputs, and attention models.

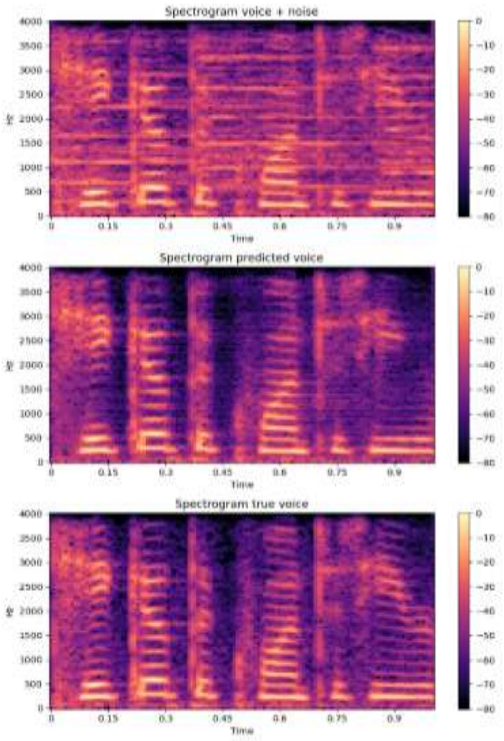

Fig 3. Bells Example

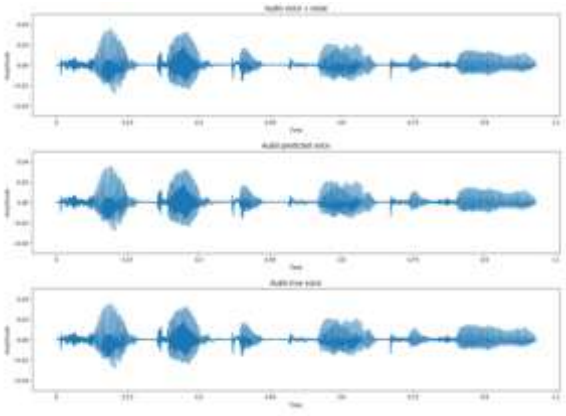

Fig 4. Bells Spectrogram

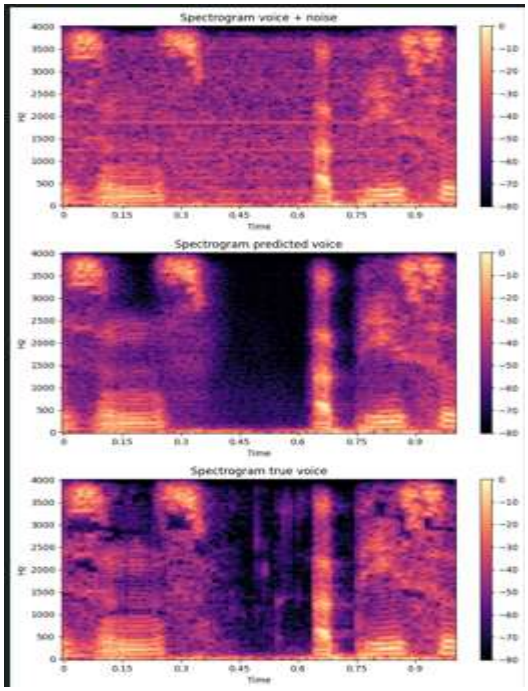

Fig 5. Vacuum Cleaner Example

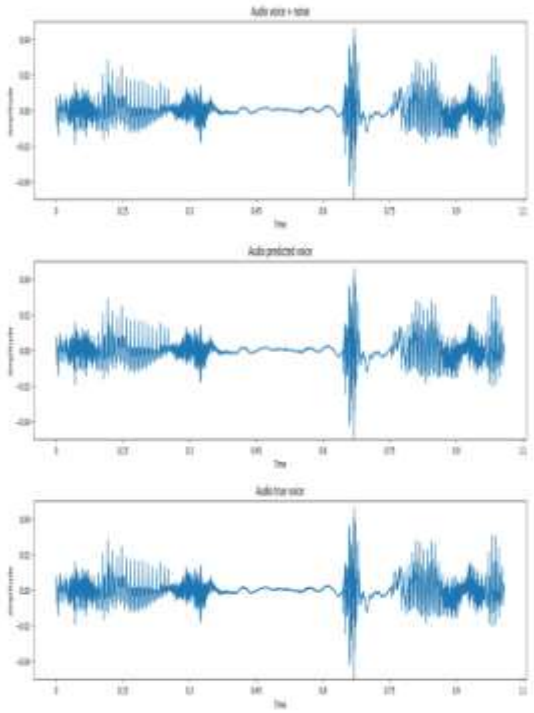

Fig 6. Vacuum Cleaner Spectrogram

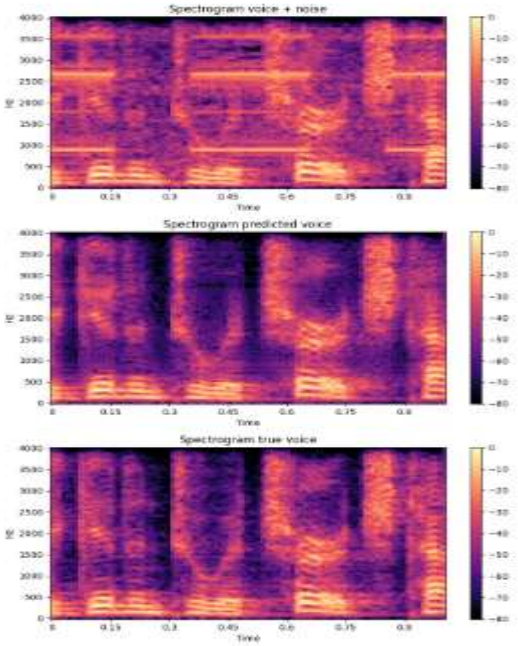

Fig 7. Alarm Example

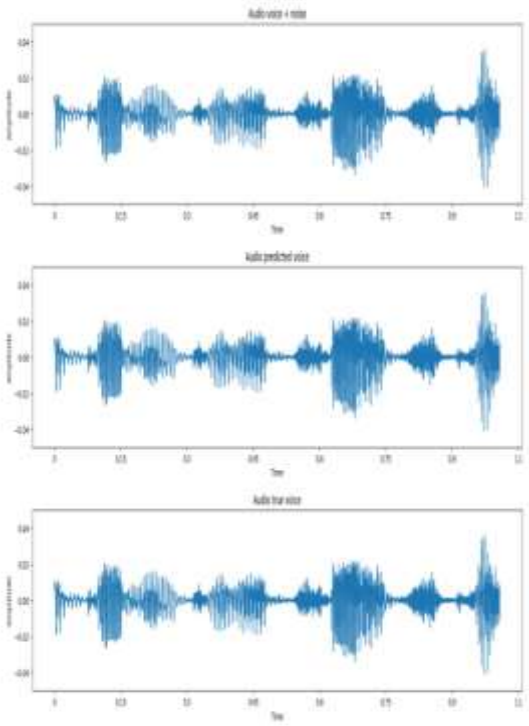

Fig 8. Alarm Spectrogram

ACKNOWLEDGEMENT

The author sincerely acknowledges the invaluable guidance, continuous support, and constructive feedback provided by Dr. S. China Venkateswarlu and Dr. V. Siva Nagaraju faculty members of the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). Their expert advice and encouragement have been instrumental throughout the entire course of this research.

Special thanks are also extended to the faculty and staff of the Institute for providing a conducive academic environment and essential resources that greatly facilitated the successful completion of this work. The author appreciates the support and collaboration of peers and colleagues who contributed their time and expertise.

REFERENCES

[1] Luo, Y., & Mesgarani, N. (2019). **Conv-TasNet: Surpassing Ideal Time–Frequency Masking for Speech Separation**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8), 1256-1266.

[2] Fu, S. W., Tsao, Y., Lu, X., & Kawai, H. (2017). **Raw waveform-based speech enhancement by fully convolutional networks**. *Interspeech 2017*, 1993-1997.

[3] Pandey, A., & Wang, D. (2019). **TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain**. *ICASSP 2019*, 6875-6879.

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015). **U-Net: Convolutional networks for biomedical image segmentation**. *MICCAI 2015*, 234-241.

[5] Park, S., & Lee, J. (2019). **A fully convolutional neural network for speech enhancement**. *ICASSP 2019*, 6935-6939.

[6] Zhao, Z., & Wang, D. (2018). **A Multi-Resolution Fully Convolutional Neural Network for Speech Enhancement**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9), 1548-1560.

[7] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014). **An experimental study on speech enhancement based on deep neural networks**. *IEEE Signal Processing Letters*, 21(1), 65-68.

[8] Park, S., & Lee, J. (2020). **Spectrogram denoising using a deep U-Net architecture for speech enhancement**. *IEEE Access*, 8, 147415-147424.

[9] Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). **Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR**. *International Conference on Latent Variable Analysis and Signal Separation*, 91-99.

[10] Pascual, S., Bonafonte, A., & Serrà, J. (2017). **SEGAN: Speech enhancement generative adversarial network**. *Interspeech 2017*, 3642-3646.

[11] Narayanan, A., & Wang, D. (2013). **Ideal ratio mask estimation using deep neural networks for robust speech recognition**. *ICASSP 2013*, 7092-7096.

[12] Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). **Deep clustering: Discriminative embeddings for segmentation and separation**. *ICASSP 2016*, 31-35.

[13] Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013). **Speech enhancement based on deep denoising autoencoder**. *Interspeech 2013*, 436-440.

[14] Tan, K., & Wang, D. (2019). **A convolutional recurrent neural network for real-time speech enhancement**. *ICASSP 2019*, 3229-3233.

[15] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., & Aila, T. (2018). **Noise2Noise: Learning image restoration without clean data**. *ICML 2018*, 2965-2974.