JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
ICASET
2024

Research Article
Collection: Intelligent Computing of Applied Sciences and Emerging Trends

Applied Weighted Parameters Approach for Noise Removal in Audio Processing Environment

Aleena Mumtaz[1], Sajid Ali[1*], Ghulam Irtaza[1], Muhammad Hassan Raza[1], Saif Ur Rehman Khan[2], and Muhammad Muzamil Aslam[3]

[1]Department of Information Sciences, University of Education, Lahore Pakistan.
[2]School of Computer Science and Engineering, Central South University, Changsha, China.
[3]School of Digital Science, Universiti Brunei Darussalam, Muara, Gadong, BE1410, Brunei Darussalam.
[*]Corresponding Author: Sajid Ali. Email: sajid.ali@ue.edu.pk

Academic Editor: Salman Qadri **Published:** April 01, 2024

Abstract: In the world of artificial intelligence and speech technology, it's becoming increasingly crucial to improve how we filter out background noise from audio, aiming for efficiency without unnecessary complexity. So, the challenge is to come up with a really effective algorithm for real-time noise reduction, ensuring optimal performance. In this study, we've delved into a deep learning approach using a convolutional neural network (CNN) to tackle noise in audio signals. We trained our model on a substantial dataset named "Edinburgh DataShare". Throughout the development of the CNN model, we incorporated Softmax and rectified linear unit as an activation functions, along with the ADAM optimization algorithm. To model evaluation, the model over 50 epochs showed a really low loss of 0.012. Hence, our findings affirm that the CNN network performs well in effectively mitigating noise from audio signals.

Keywords: Deep learning; noise removal of audio signal; Convolutional neural network; speech enhancement.

1. Introduction

Contemporary society has reached a juncture where the daily routine of every individual is inconceivable without the ubiquitous presence of electronic devices. From checking current weather conditions to maintaining connections with friends and family, these modern devices play a vital role. While these modern devices serve various purposes like entertainment, education, health, and advertising, they hold significant importance in the lives of individuals today. Communication, the dynamic process of information exchange from sender to receiver, unfolds through diverse mediums, encompassing traditional channels like newspapers and magazines, as well as contemporary electronic gadgets. This transmission of information assumes various forms, including textual, graphical, audio, and visual modalities.

When we talk about audio communication, we're referring to the kind of communication that happens through hearing rather than words. In this setup, the person receiving the message gets the information based on what they've heard. Humans can pick up sounds within a frequency range from roughly 20 Hz to 20 kHz [1]. When we send audio signals, it's not just the main message that gets through. All sorts of background noises like birds chirping, traffic noise, wind, children crying also add in original audio signal. This mixture of sounds can make communication uncertain. The person on the receiving end might struggle to pick up any useful information or only get a fuzzy version of it because of all the noise. To make sure the audio signal comes through crystal clear, we need to use a process that gets rid of the unwanted background noise. Getting rid of noise means taking out all those unwanted sounds from an audio signal. When you do this, the quality of the audio gets a major boost because a cleaned-up signal carries way more

useful information than a noisy one. This not only makes listening and learning more enjoyable since we can focus better on the clear audio, but it also improves communication. When the background noise is gone, the person on the other end can get the information they need without it getting all messed up by interference.

The need for reliable real-time communication and collaboration solutions has grown. In these situations, Maintaining consistent relationships and teamwork with others requires audio calls having remarkable audio worth. Many other kinds of background noises are readily noticeable to us, including traffic noise, electrical device noise, dog barking, unnecessary speaking, alarms, a baby wailing, kitchen noises, etc. Background noise has a significant impact on the clarity and understandability of speech that is heard, which leads to fatigue. Things like smart devices and hearing aids are problematic with background noise. Lack of materials for training is a common problem that arises when models employ deep learning methods for audio denoising. This is due to the fact that the training process requires both clear and noisy audio samples. However, obtaining the required training samples is challenging because real-world audio signals frequently contain noise that cannot be eliminated. Another issue is that most of the noisy audio samples used to train the model are artificial and cannot be assembled the same way as genuine noise. Natural generated noise obscures clearly discernible signal patterns, whereas synthetically made noise retains distinct signal patterns. Furthermore, although naturally created noise obscures discernible signal patterns, artificially generated noise retains distinct signal patterns. When recording or sending audio during online meetings, video conversations, digital updates, voice communications via the globe, and other similar scenarios, undesired signals, or noise, might cause problems. As a result, before utilizing several contemporary signal processing techniques like audio filtering, expansion, equalization, compression, coding, analysis, and recognition, the noise reduction approach should be used as a preparatory step. Researchers refer to this technique of eliminating unwanted sounds from the conversation as "Audio noise mitigation as well as enhancing." Given how frequently this audio degradation occurs, denoising is essential for improving communication between humans and machines as well as between humans and machines. Because audio processes are complex and the anti-data is unknown, the undefined scenario of only one-channel verbal reduction of noise is a very difficult but common variation of the task. The nature of the content—audio content has a high sample rate—adds to the complexity. Due to perceptual processes, minor errors can be detected by the average user in controlled human-to-human interaction [2]. The most disagreeable noise that skews each sample in the audio signal is called Gaussian noise. Consequently, several researchers worked very hard to get rid of this influence. Another important noise that could interfere with the audio stream is impulse noise [3]. Errors in the transmission equipment and clicks, bursts, and crackles are the causes.

Reverberation and additive noise impinging on a microphone may degrade the clarity and understandability of the speech recordings. The difficult issues of noisy signal cancellation, noise suppression, and dereverberation—and especially the combination of these tasks—are currently being researched, with multi-microphone-based techniques that make use of spatial variation drawing special attention [4]. Therefore, in scenarios when sufficient data is available, deep learning models have often outperformed previous methods in auditory processing, such as hidden Markov models, Gaussian mixture models and others. Although a lot of deep learning techniques are used on images, there are still important differences between the two domains that necessitate a deeper analysis of audio. Audio input consists of a one-dimensional (1D) array as opposed to the two-dimensional (2D) array produced by visual data. While sound waves need to be studied in chronological sequence, images are instantaneous pictures of a subject and are often assessed in an orderly or segmented manner. These features caused the emergence of audio-specific systems. [5].

Deep learning has proven useful in sound systems and minimizing noise via many kinds of factors. Another factor is the amount of data. Large training data sets are ideal for deep learning algorithms, and the audio signal processing and noise reduction domains contain an abundance of training data [6], which has been essential in their success. Furthermore, without the requirement for manually designed features, deep learning models are able to learn an endwise mapping from input audio signals to output signals. Deep learning models are becoming quite effective in applications like voice enhancement, where the goal is to improve the audio signal quality. In summary, these models are capable of non-linear processing, which is an essential function for simulating intricate and non-linear interactions seen in audio signals.

Because of this, deep learning has proven very useful for jobs like noise reduction, in which the objective is to remove undesired and distracting signals from the audio [7]. All of these factors work together to give deep learning an incredibly potent tool for noise reduction and signal processing in audio, which has in recent times resulted in major advances in these domains.

Noise incursion can dramatically lower the perceived quality and understandability of voice communication. Related tasks including automatic speech recognition, voice-based input devices, interactive voice systems (IVS), and vocal biometrics can all be severely impacted by noise interference. Recently, promising results in voice augmentation have been obtained with deep learning (DL) techniques, especially when managing non-stationary sounds in challenging scenarios. Deep learning works well for augmenting audio data in both multi- and monaural (single-channel) formats [8]. Based on artificial neural networks (ANNs), deep learning is a technique that processes and computes massive amounts of data using numerous layers of ANNs (impacted by how the human brain functions and works). Popular algorithms used for deep learning applications include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short Term Memory Networks (LSTMs), Generative Adversarial Networks (GANs), etc.

Deep learning techniques for denoising and voice augmentation improve audio quality, particularly for critical speech perception applications such as audio communication systems, hearing aids, and speech recognition. By reducing noise and interference, these algorithms enhance performance and dependability .Speech-Denoising algorithms are used in a variety of real-world applications, including voice assistants, hearing aids, automatic speech recognition (ASR), security, and surveillance. An improvement in speech signal quality during video conferences and phone calls is beneficial to telecommunication. Speech-Denoising algorithms can be used by those who have hearing loss to improve speech intelligibility and clarity in noisy settings. Speech denoising algorithms are used by voice assistants like Alexa, Siri, and Google Assistant to improve speech recognition accuracy even in noisy surroundings. Speech denoising methods are used by ASR systems to enhance transcription accuracy and signal quality. By enhancing and recovering speech signals from chaotic audio recordings, speech denoising algorithms also aid in security and surveillance by making it simpler to recognize speakers and follow their discussions.

Designing a Convolutional Neural Network (CNN) for noise removal in audio data involves several key components. Below is a generalized structure for a CNN model in Figure 1. This architecture is a supervised learning approach, where you have pairs of clean and noisy audio samples for training. Noise removal in audio involves a combination of preprocessing, deep learning with CNNs, and post-processing techniques. The goal is to produce high-quality, denoised audio that preserves the integrity of the original content. Continuous improvement through fine-tuning and adaptation to specific noise scenarios ensures robust performance in various audio environments.

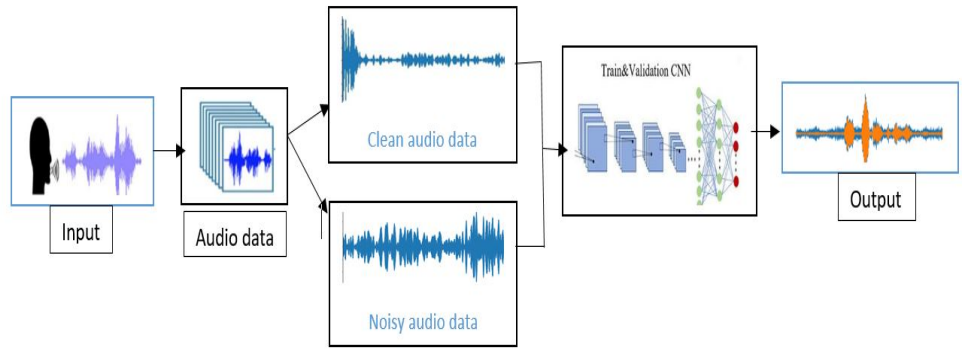

Figure 1. Schema of the proposed CNN based noise removal architecture.

Input Layer: Accepts audio waveform in pairs of clean and noisy audio samples.
CNN: Convolutional layers capture spatial dependencies, and pooling layers help in feature extraction. Layers with filters, kernel sizes, and activation functions.
Output Layer: Provides denoised audio output.

The suggested technique uses a deep features loss function in conjunction with a fully convolutional neural network to suppress noise signals in audio input. By changing the weights, they were able to attain extraordinarily good results, which optimized the suggested approach.

2. Literature Review

Awad et al.'s study [9] discovered that impulse noise in the audio stream could be identified and recovered. Cascade staging was used to assess a noisy input signal. A number of experiments had been carried out to assess the effectiveness of the suggested methodology. It was discovered that the suggested study successfully achieves audio denoising while preserving the original audio signal, requiring less processing power and being simple to execute.

A study utilizing the Kalman filter and Sidelobe cancelling technique was proposed for the reduction of audio noise, cancellation of interfering signals, and speech reverberation [10]. Two methods, multi-channel linear prediction and extend Sidelobe cancellation (SLC), were combined: spatial filtration and deconvolutions. Recognition, classification, and audio signal processing are areas where deep learning has shown promise. Deep learning-based audio signal processing has shown promise for application across multiple disciplines. After being used in several studies for image processing, deep learning became well-liked for processing audio signals, music, and ambient noise [11].

Sequence classification is the process of identifying a label to one category, like a speaker's, musically, vocabulary, or acoustical situation. On the other hand, when there are multiple classes, any one of them might represent desired label. In situations when classes overlap, multi-label classification may perform better. On the other hand, sequence regression forecasts the next audio sample or estimates musical tempo by predicting a value within a continuous range [12]. It is important to remember that discretizing the output range might convert regression difficulties into classification problems. For instance, by quantizing audio samples into 256 classes, audio sample prediction can be approached as a classification problem. A set amount of sound clips could be included in every phase when predicting a label when using an 8-bit audio sample; this yields an expected sequence size which represents a percentage of the original sequence length. In regression per time step, continuous predictions are generated, e.g., the pitch of a voice, the separation between sources, or the distance to a moving sound source. Noise interference has the potential to drastically lower the perceived quality and understandability of speech communication. Lately, DL methods have yielded promising results in voice enhancement, especially in challenging situations involving non-stationary sounds. Deep learning can enhance both single- and multi-channel voice augmentation, depending on the application.

GANs have been used to improve audio signals and lower noise levels in those signals. This proposed solution has the advantage of working end-to-end with decoded audio signals without requiring the deployment of a front-end application [13]. Furthermore, it can enhance the quality of low-bit-rate audio signals without requiring any changes to the current encoders. A dual-path RNN was proposed as a method to denoise underwater audio sources [14]. In essence, the algorithm detects the unwanted noise signal in real-time and then creates signal that has the same intensity and opposite aspect as the original noise signal. For years, scholars have been investigating audio denoising through the application of both traditional and deep learning-based methods. These techniques can only be used with artificially generated noise or low-quality denoised audio, though. The author [15] gathered 14,120 actual bird sound recordings and suggested a denoising method based on image segmentation. The outcomes show how well this method works to lower noise and improve audio quality. Wiener filtering is a statistical method that calculates a weighting factor for each frequency bin in the data spectrum by estimating the signal-to-noise ratio (SNR). A clean output is obtained by filtering the noisy signal using the weighting factor [16]. Using the wavelet transform, this method divides the audio signal into several frequency bands for wavelet denoising.

3. Methodology

The suggested approach is predicated on the advancement of deep learning algorithms designed to reduce noise in audio signals.

The following steps provide an explanation of the methodology:

- Dataset loading and visualization.
- Pre-processing
- Dataset creation for pipeline
- Deep learning model development
- Model training.

- Optimization of model
- Inference of TFLite algorithm
- Another example of the suggested process is shown in block diagram Figure 2.

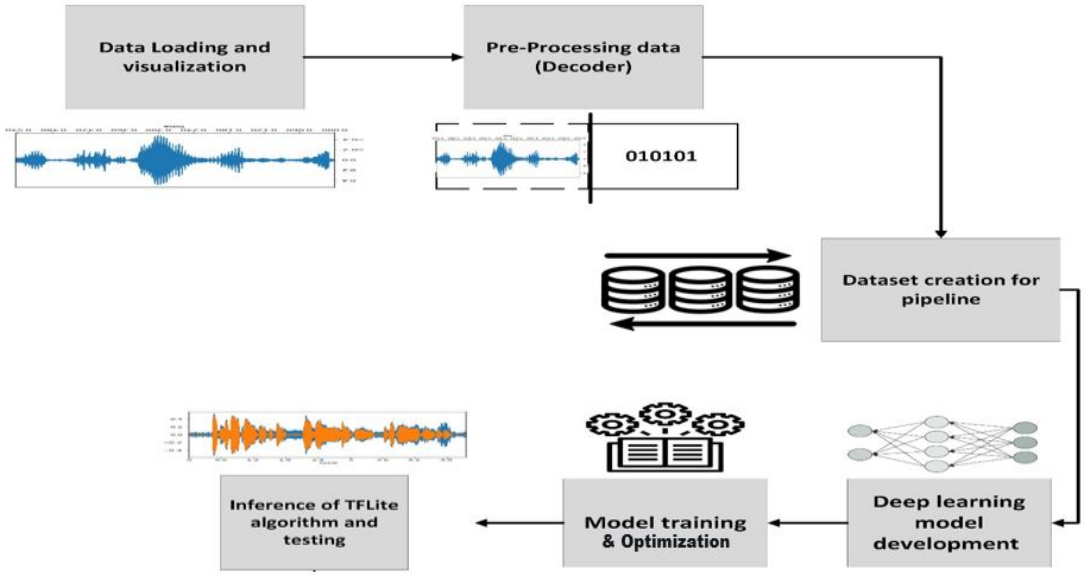

Figure 2. Block diagram of proposed methodology

3.1 Dataset loading and visualization

An initial batch of data was in a compressed file. Unzipping it and loading it into Google Colab's main directory was the first step. The dataset, which included both clear and noisy audio waves, was read and loaded using the TensorFlow library. See Figure 3 for the mixed signal. As a Tensor Shape, the input data has the form ([97445791, 1]), where 1 represents the size and 97445791 represents the depth of the audio signal. Signals were represented graphically using the TQDM loader.

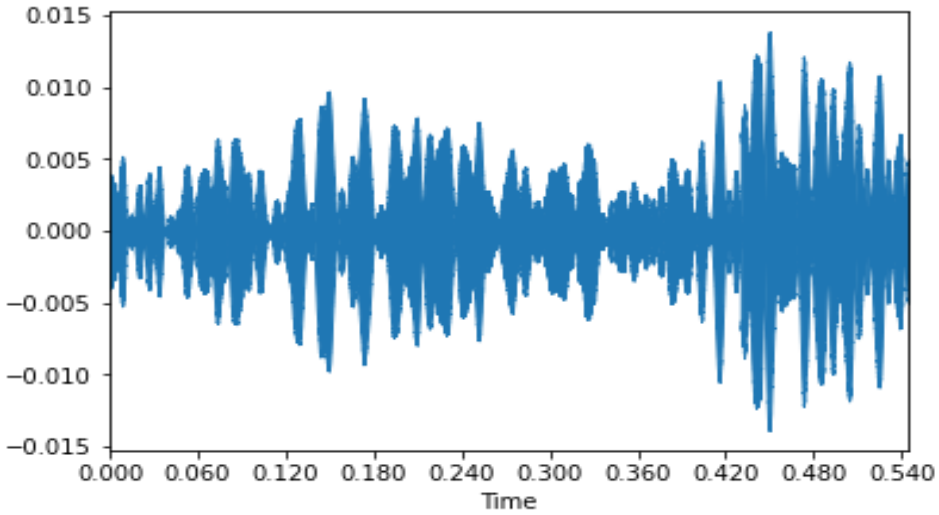

Figure 3. Visual form of mixed signal

3.2 Preprocessing of Signals

Following dataset loading, pre-processing is performed, where a batching size of 12,000 was applied. The audio signals, both clean and loud, were supplemented with the batching size that indicates there are 12000 data points to form batches from the supplied data. After that, the neural network is sequentially trained on each batch to teach it how to filter out noise from the given data.

Also configure the individual audio files using tf.audio.decode-wav () and concatenate them to get two tensors named clean_sounds_list and noisy_sounds_list. The visualization of audio signals are represented graphically as noisy and clean signals in Figure 4 and Figure 5.

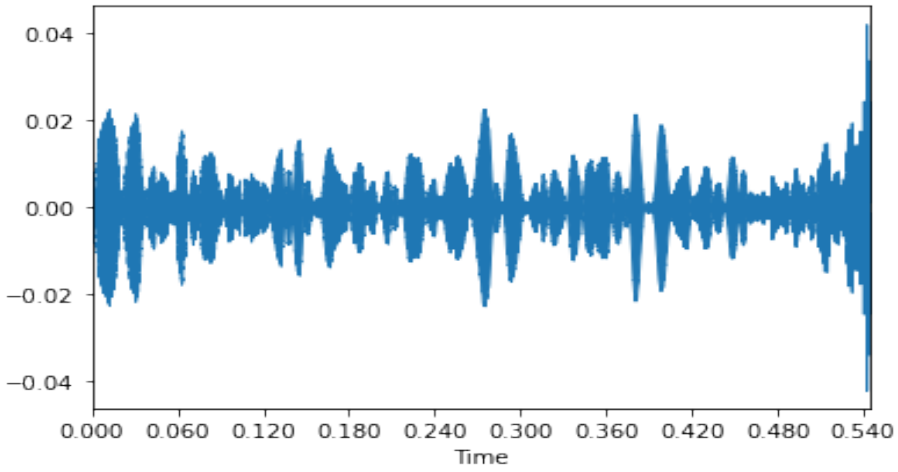

Figure 4. Visual waveform of clean signal

On plotting the data as seen below, we can see that the noise is quite visible is like small fluctuations.

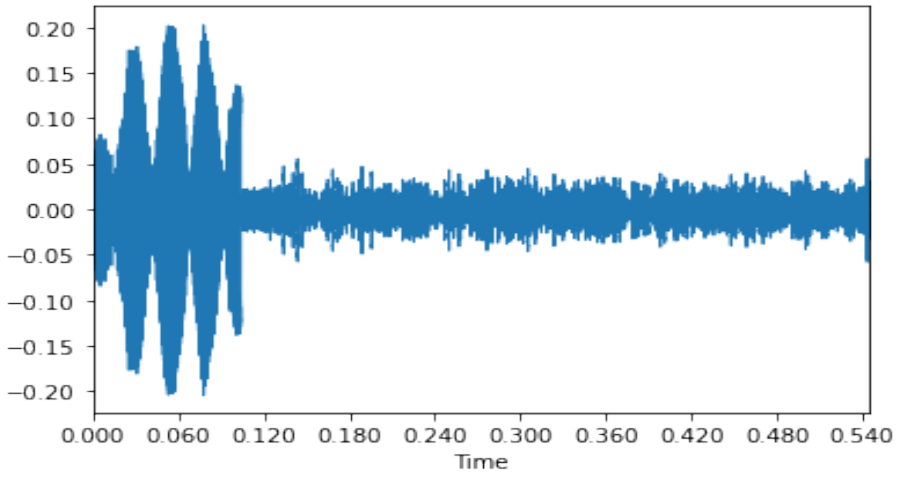

Figure 5. Visual waveform of noisy signal

The creation of the tf.dataset was the next step in pipelining the steps of model training, testing, and development. A technique for training a model in deep learning frameworks, whole dataset is divided into fixed batches (64). When "drop_remainder=True" is specified, any leftover dataset components are removed instead of added to final batches of reduced volume. This method is popular in deep learning because it allows for better memory consumption and processing on devices like GPUs, which can handle data more quickly in parallel and in larger batches.

The official Edinburg Data Share website provided the dataset that was used in this study. Table 3.1 contains a description. The dataset makes it possible to train deep learning models, and developers can use it to assess the models' performance. For the purpose of developing models, the dataset is essential. The dataset was split into two subsets: the test set, which included both clear, noise speech the train set.

Table 1. Dataset Description

No.	Name	Description	Size
1	Train set clean speech	28 natural English speakers are represented by clean speech waveforms at 48 kHz, apiece with about 400 phrases.	821.6MB

2	Train set noisy speech	28 natural English speakers are represented by noisy speech waveforms at 48 kHz, typically with about 400 phrases.	912.7MB
3	Test set clean speech	Two fluent English speakers are heard in clear voice at a rate of 48 kHz, speaking about 400 phrases each.	147.1MB
4	Test set noisy speech	Two native English speakers are heard in two noisy speech 48 kHz waveforms, each speaking about 400 phrases.	162.6MB

The dataset is divided into train and test data which corresponds to 80:20 % split. The dataset division has been shown in Figure 6**.**

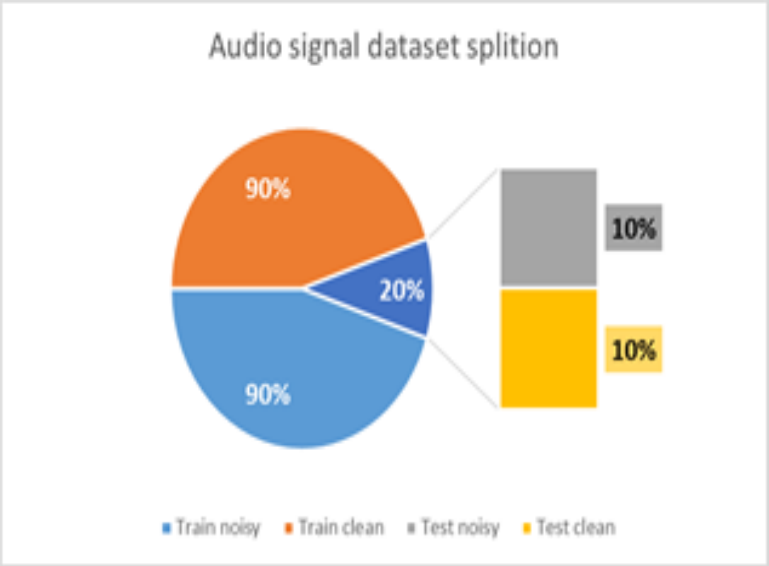

Figure 6. Dataset division

To assess the noise suppression algorithm's performance, a high-quality dataset is essential. A large dataset makes sure the model is evaluated with a variety of inputs, which helps identify the model's advantages and disadvantages.

3.3. Model Development and Training

The creation of a deep learning model came next, and a 1D convolutional neural network model was created during this stage. 1D CNNs are a kind of deep learning method that have shown promise in signal processing and image identification, among other uses. 1D CNNs are capable of detecting and removing unwanted noise patterns from signals when they are utilized for noise reduction. Typically, 1D CNNs are used to reduce noise by training the network on a dataset that contains both clean and noisy signals. During training, the CNN detects unwanted signals in the signal being input by applying a sequence of convolutional filters, which results in the production of a noise-free output signal. Signal interference primarily comes in two forms. Arbitrary and fixed value signal inference. The only two numbers in fixed-value impulse sound are 0 and 1. The value in random value signal interference may lie between the maximum and minimum values. Let us assume that p is the probability of unwanted speech, given speech signal is x and the corresponding inference position inside the sound signal is $\boldsymbol{xij}$, where background noise occurs at position ij. Additionally, let p represents the probability of background speech signal or it could be 1-p, can be represented by $\boldsymbol{nij}$, which can indicate the relationship between the unwanted signal and its position in the speech signal.

$$\boldsymbol{x_{ij}} = \begin{cases} x_{ij} \\ n_{ij} \end{cases} \quad (1)$$

Where $\boldsymbol{n_{ij}}$ is with probability p, and $\boldsymbol{x_{ij}}$ with ***1-p*** probability.

3.3.1 Input layer

An input layer is the first layer, which takes in an audio signal with the shape size (None, 12000, 1), where 1 denotes the number of filters, 12000 is the number of features in the input sample, and none indicates the batch size—in our instance, 64. The output of this layer was sent to a convolutional 1D block.

3.3.2 Convolutional Block:

There are five convolutional layers in the convolutional block. Applying a convolution function to audio signals was the fundamental purpose of the convolutional layer. The fundamental building element of any CNN model development is a convolutional layer. The preceding layer's output feeds into a convolutional block, which processes and extracts features from it. To the following layers is supplied the output (None, 375, 32). Given by (2-5), this convolution function.

C= conv1D (x, filters, kernel_size (F), padding (p), activation) (2)

Sliding a filter over the input data and conducting element-wise multiplication and summing are the steps involved in the convolution action.

Given an input signal x and a filter k:

The mathematics operation can be represented as:

$$C = F + p \quad (3)$$

$$C = 1^* k + p \quad (4)$$

$$Conv\ (1,k) = ø_\alpha\ (1 *k + p) \quad (5)$$

At each step, an elements-wise multiplication are performed between filter & an input signal.

3.3.3 Transpose Block:

The transposing block that follows has seven 1D convolutional transpose layers with seven concatenation layers (seven concatenations) between each layer. Let C be the feature map that was produced using a convolutional block, where N, X, and L stand for the batch size, number of channels, and feature map length, respectively. (6) yields the transpose layer's output feature map,

Y = Conv1DTranspose (C, f, k_st, p) (6)

Above equation 6 Y[i] represents the output of the proposed method and W[c] are the filter weight used in this neural network. At each step, an element-wise multiplication is performed between the filter and a section of the input signal. Denoising auto encoder was used as transpose layer, a pair of encoders and decoders. The encoder processes the input signal to extract high-level features, which the decoder uses to reconstruct the original signal. The transpose layer of the decoder enlarges the feature maps by interpolating zeros between them in order to boost the output signal's resolution. Through the process of up sampling, the model is able to recover any information that were lost during the encoder's down sampling and restore the input signal to its original shape. This method works especially well for removing noise from audio, where a high-precision output signal is crucial. Along the channel dimension, feature maps from several earlier layers are combined in the concatenation layer. Feature maps from each layer are stacked in this procedure to produce a new feature map with more channels. The concatenation layer is typically employed in skip or residual connections to enable a network to simultaneously acquire the highest-level and lowest-level information. In these connections, the result of one layer merges with the outputs of the layer preceding it, typically with fewer channels or lower resolution. As a result, the network can bypass some layers while keeping crucial characteristics that would have been lost in the down sampling process. This block's output form was (12000, 1).

3.3.4 Output layer

The processed audio signal with the shape (12000, 1) is contained in the output layer. The ability of 1D CNNs to recognize complex patterns found in the input data that could be difficult to discern using traditional methods for signal processing is one of the main benefits of employing them for noise reduction. Furthermore, 1D CNNs are a flexible tool for a variety of applications since they can be taught to reduce

noise from a wide span of the input signals. Studies have demonstrated that 1D CNNs are a powerful tool for noise reduction; their versatility in learning complex patterns and generalizing to a wide range of input signals makes them a popular choice for noise reduction jobs in many different fields. In Figure 7, the suggested layered design is displayed.

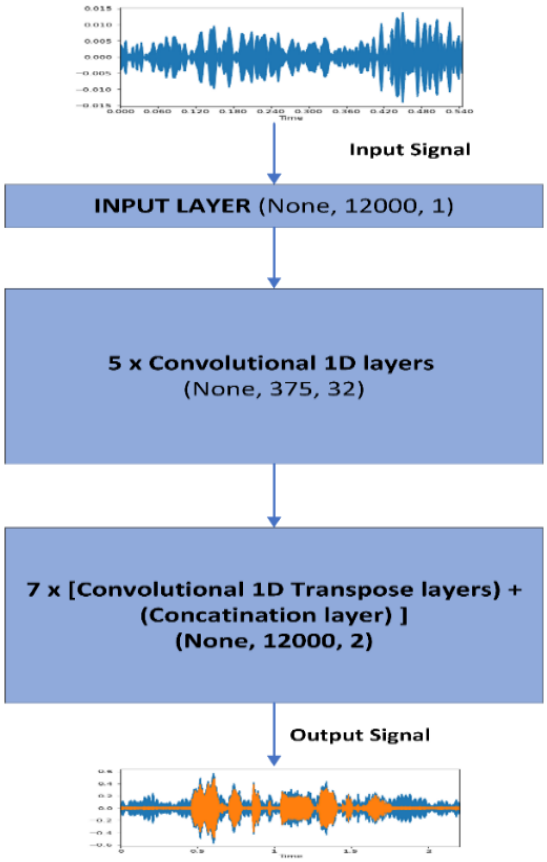

Figure 7. High-level abstraction of proposed CNN architecture

3.3.5 Model training and optimization

In next step, after creating CNN was to train model and optimize for better noise reduction outcomes. Through deep learning, deep neural network learns to analyze given dataset and make assumptions regarding its meaning. Before the network is able to make deductions based on the desired result, it must go through a great deal of trial and error. During training, the CNN feeds a large number of tagged audio signals with the class labels that correspond to the neural network. Each signal is processed by the CNN network using randomly assigned values, and it then compares the output to the class label of the input audio signal.

The applied CNN has the padding set to "same". Convolutional neural networks (CNNs) use padding, which is the process of adding extra values—usually zeros—to the input's boundaries volumes. This is done to guarantee that, after the convolution procedure, the input volume's spatial dimensions are maintained. The convolutional layer's output volume size is equal to the input volume size if the padding parameter is set to "same". In Figure 8, the entire network of these layers is displayed.

3.3.6 Inference of TFLite algorithm and testing

In the last phase, model receives an unwanted speech input, extracts high-level features using convolutional layers. Subsequent layers process these features in order to produce an output signal that has been denoised. Because of its compact size and streamlined architecture, the TFLite variant can be used on mobile and edge devices. This enables it to carry out real-time noise reduction on devices—like voice recognition or real-time audio processing applications—where low latency and high accuracy are crucial.

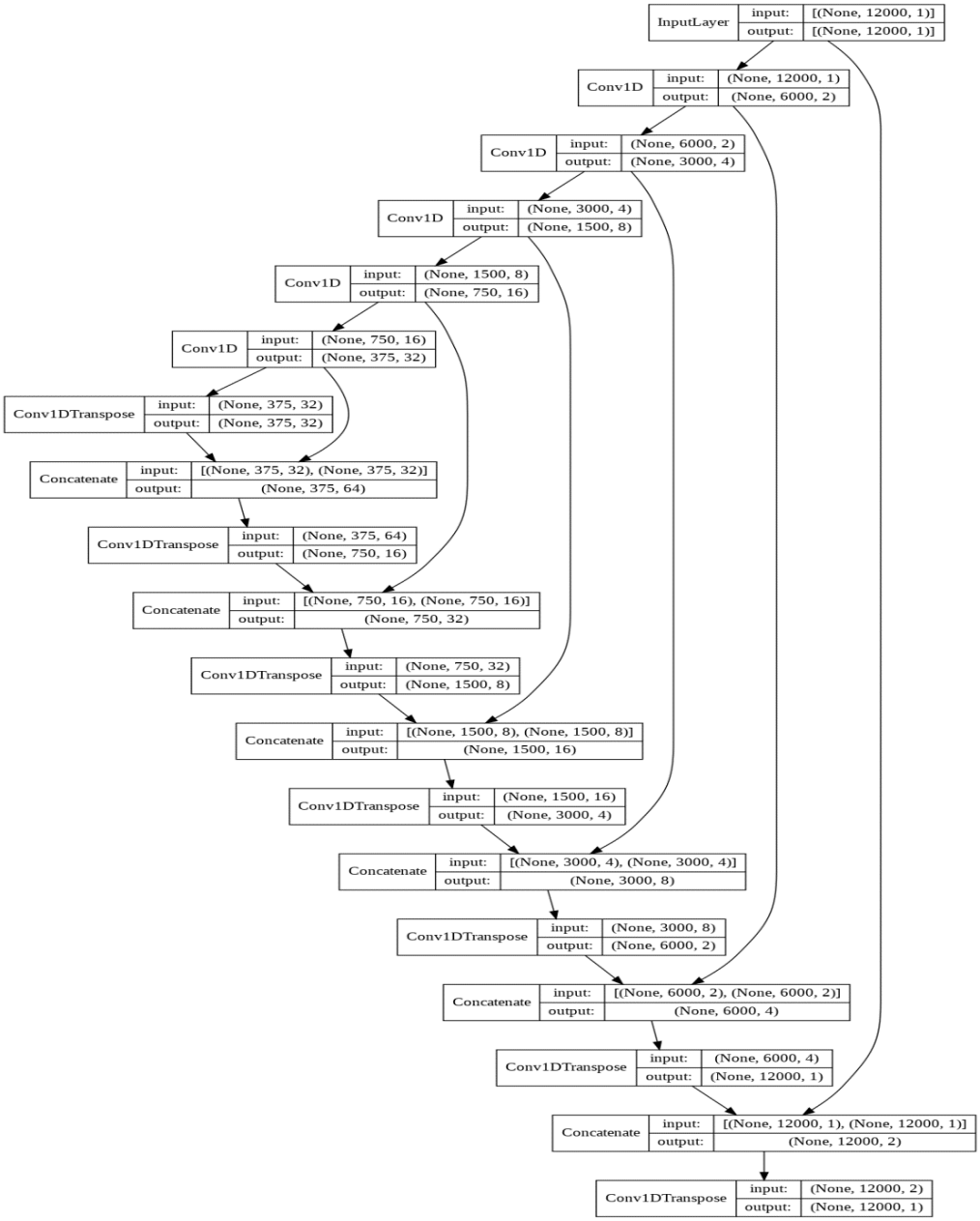

Figure 8. Layered architecture of proposed neural network

4. Results and Analysis

The experiment results have been discussed and provided in this section. The variables influencing the suggested CNN model's efficacy, dependability, and performance have also been discussed. To ensure that our model is accurate, we ran a number of tests comparing the effects of the suggested denoising technique. When discussing network performance, there are numerous things to consider. Apart from the obvioan anti-us aspects like assessment scores, we also look at how easy it was to train the model and the training process itself.

The hyper-parameters are shown in table 2, which are set before training and to optimize results.

Table 2. Model Parameters

Parameter	Value/type
Epochs	50
Batch size	64
Learning rate	0.001

Activation function	ReLU
Optimizer	Adam
Padding	Same
Activation type	Linear

The loss assessment measure is used to evaluate model. It combines the features of the mean squared error (MSE) function with the absolute value function.

As can be seen in Figure 9, the loss decreases quickly in the first few epochs, develops gradually, and then continues to decrease. This indicates that the model is learning efficiently. The loss function has a start value of 0.011 or an ending value of 0.010.

There is a slight training loss in the model, as indicated by the loss function of 0.011

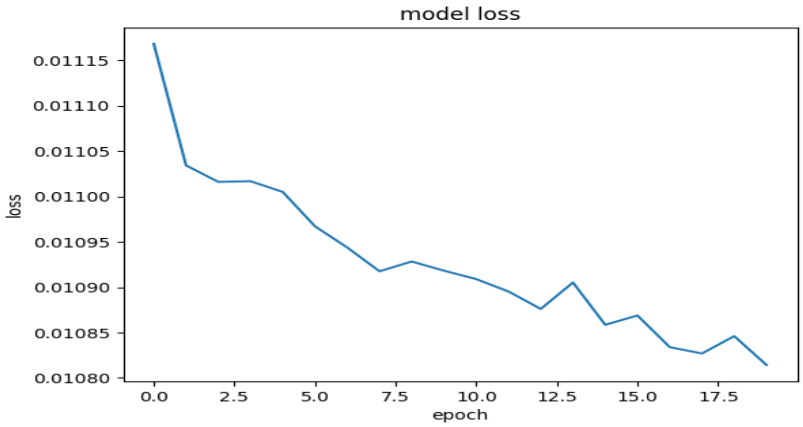

Figure 9. Model evaluation training loss

Following trained model optimization, the model's mean loss value decreases, with the loss function's start value being 0.026 and final value being 0.012. Deep learning and machine learning aim to minimize loss; Figure 10 illustrates how increasing the period size decreased the model loss by 0.012.

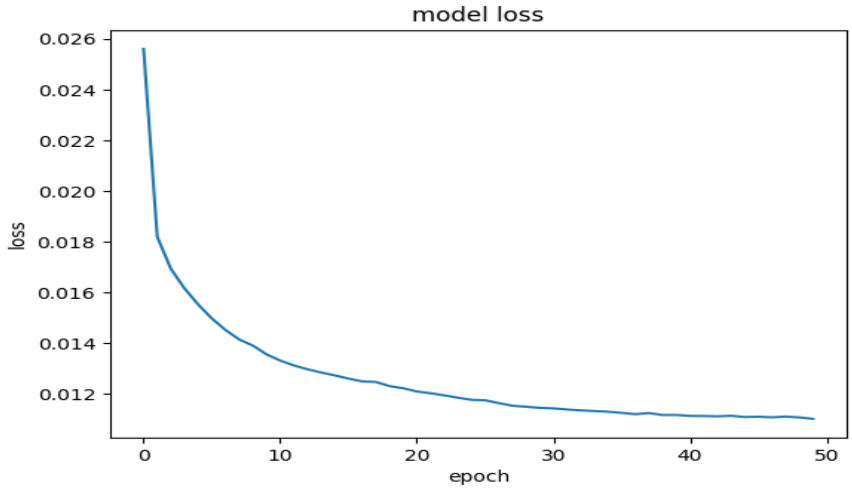

Figure 10. Model evaluation training loss

4.1 Testing phase

Following hyper-parameter optimization, a great deal of trial and error has been involved in trying to train the model. The constructed model must next be evaluated using test data. The following graphs and pictures show the outcomes of the model's successful testing.

Three images are used in each test to depict sound with noise added (orange), original sound (blue), and the denoised audio signal (orange). The suggested approach creates sound pressure, which varies with the pressure of the atmosphere to make sound. Sound intensity is measured in terms of magnitude. The

speech waveform represents the test audio inside the time area. It displays the evolution of the magnitude throughout time. Figure 11 displays a plot of sample A, the original audio signal.

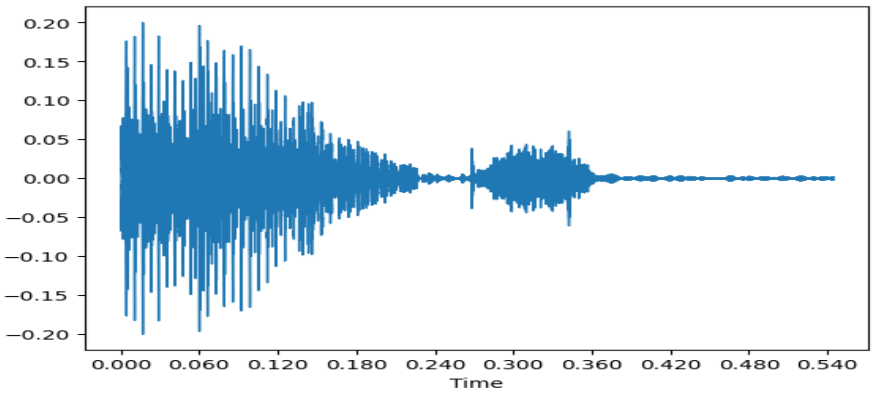

Figure 11. Waveform of original signal

As seen in Figure 12, the highest degree of distortion occurs when a certain range of frequencies is added to the original audio, altering the speed at which each frequency segment propagates.

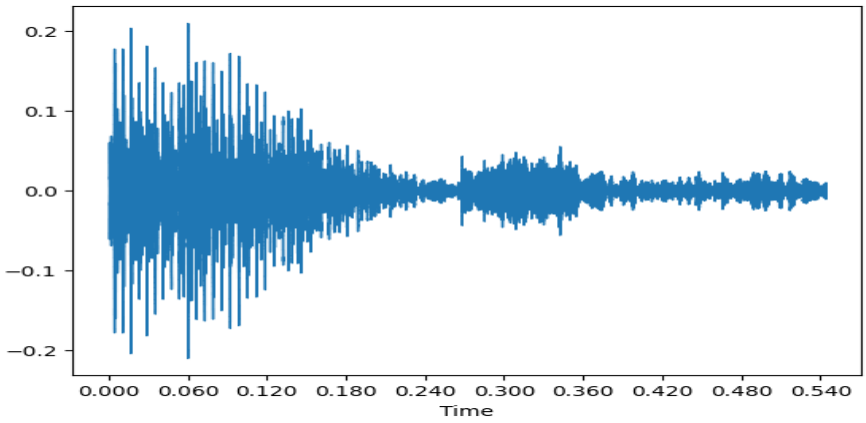

Figure 12. Waveform of noisy signal

Figure 13 displays the denoised signal that results from the estimation and reduction of the noise samples from the frames. The final output of the stage is an orange-colored denoised signal that is used as an input for the next speech enhancement method.

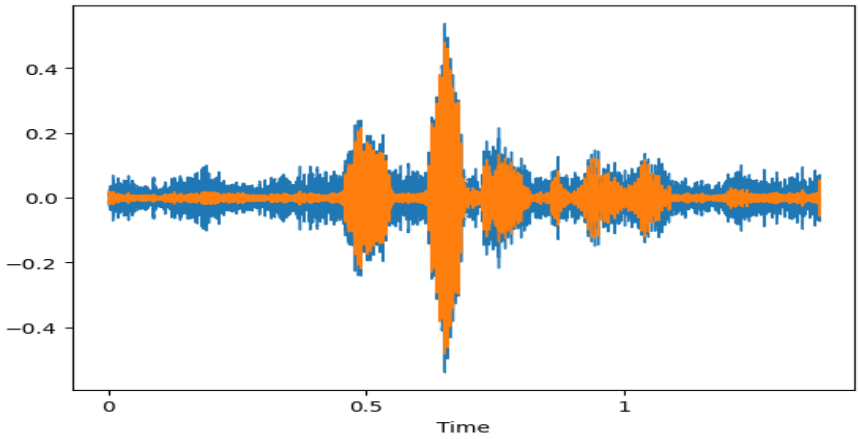

Figure 13. Denoised signal produced by proposed system

The test audio is shown as a speech waveform in the subsequent speech sample, which is situated inside a temporal domain. It displays the evolution of the magnitude throughout time. In Figure 14, an original audio signal plot is displayed.

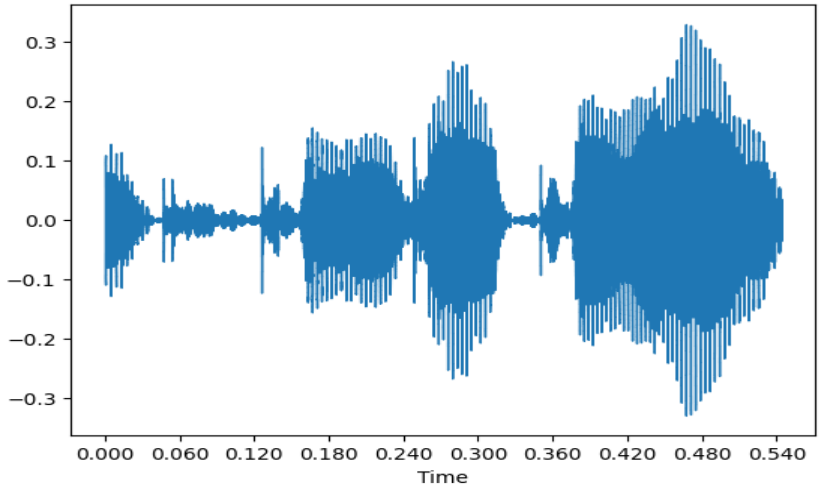

Figure 14. Waveform of original signal

A higher degree of distortion exists at a distinct frequency band, as seen in Figure 15, distortion is introduced to the original sound to alter each frequency segment's velocity of propagation.

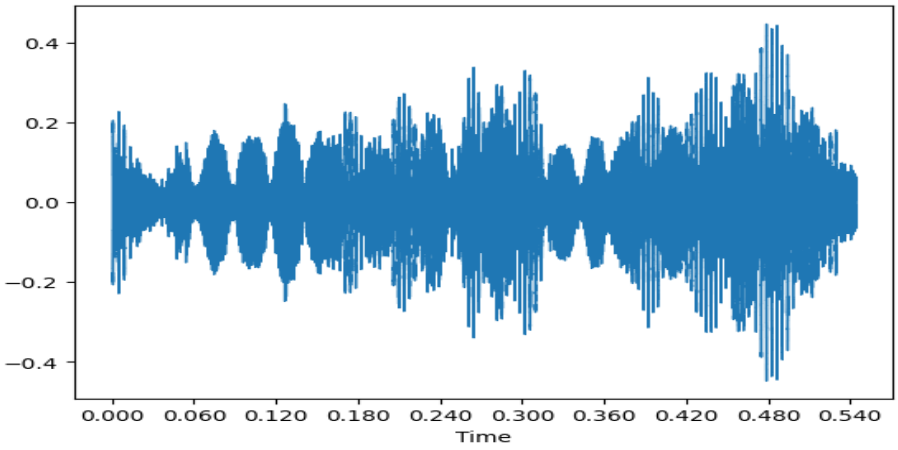

Figure 15. Waveform of noisy signal

Figure 16 displays the denoised signal that results from the estimation and reduction of the noise samples from the frames .The final output of the stage is an orange-colored denoised signal that is used as an input for the next speech method.

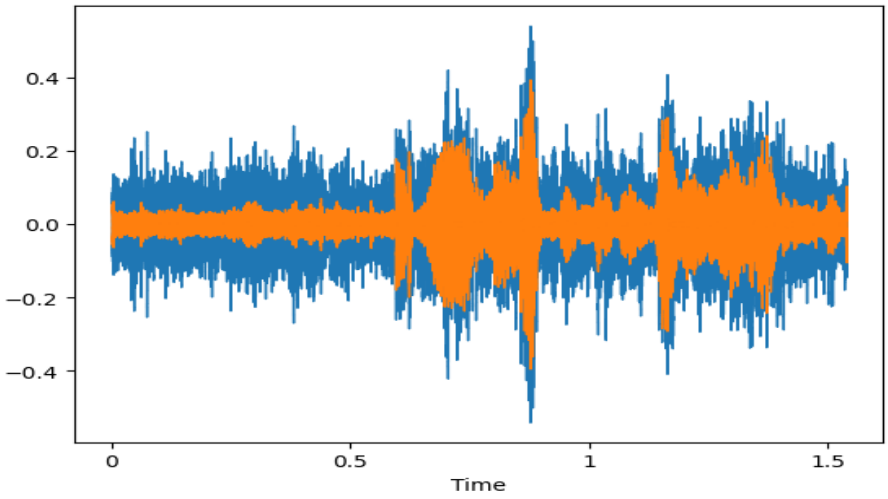

Figure 16. Denoised signal produced by proposed system

In the following table 3, proposed methodology is presented along with comparison with state-of-the-art algorithms from literature.

Table 3. Result Comparison with State of Art Algorithms

Sr.no	Reference	Metrics	Results
1	(Sun et al., 2021)	Loss function	2.64
2	(Bhat et al., 2019)	Loss value	1.26
3	(Park et al., 2020)	Loss evaluation	0.16
		Training Loss	2.60
4	(Dogra et al., 2021)	Loss function	0.13
			1.02
5	(Ghimire et al., 2022)	Huber loss	0.0205
6	**Proposed method**	**Loss function**	**0.012**

In the following Figure 17 shows the results in graph form. In this, proposed method has minimum training loss value which is 0.012 that is showed in the last column of this graph via green color bar.

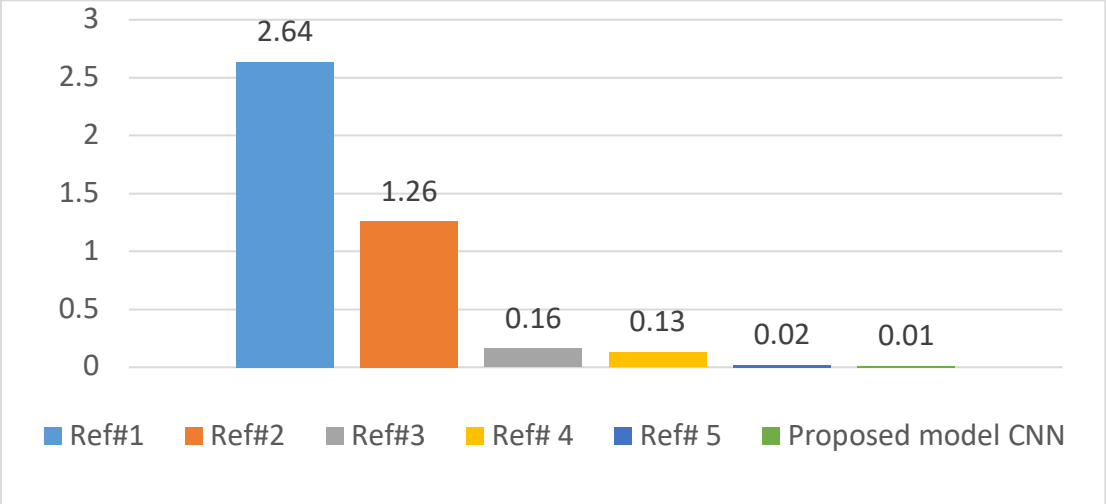

Figure 17. Result comparison graph with state of art algorithms

5. Conclusions

Deep learning breakthroughs have improved speech quality and reduced noise interference, enabling more efficient noise reduction and audio signal processing. As a result, deep learning has emerged as a potent instrument in numerous for-profit applications involving human-to-human communication. Using a deep convolutional neural network (CNN), we have created an all-inclusive audio noise reduction method. As a result, deep learning has emerged as a potent instrument in numerous commercial applications pertaining to human-to-human communication. Our method involves training the CNN on a large variety of noisy and clean audio data, which enables it to recognize intricate correlations and patterns within audio signals as well as evaluate speech structure at various scales. As a result, we can conclude from the results that the CNN network was effectively used to reduce audio signal noise, and that its training methodology improves the suggested system's performance without adding to its complexity. Through the utilization of deep CNNs and the integration of an extensive training plan, we have made significant progress in this area.

References

1. Kristoff Fluyt Wouter Tirry Maximilian Strake, Bruno Defraene and Tim Fingscheidt. Speech enhancement by lstm-based noise suppression followed by cnn-based speech restoration. 2020.
2. Germain, F. G., Chen, Q., & Koltun, V. (2018). Speech Denoising with Deep Feature Losses. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019-September, 2723–2727. https://doi.org/10.48550/arxiv.1806.10522
3. Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion, 49, 69–78.
4. Dietzen, T., Doclo, S., Moonen, M., & van Waterschoot, T. (2020). Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 740–754
5. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., & Sainath, T. (2019b). Deep Learning for Audio Signal Processing. IEEE Journal of Selected Topics in Signal Processing, 13(2), 206–219. https://doi.org/10.1109/JSTSP.2019.2908700
6. Saleem, N., & Khattak, M. I. (2020). Deep neural networks for speech enhancement in complex-noisy environments.
7. Ouyang, Z., Yu, H., Zhu, W.-P., & Champagne, B. (2019). A fully convolutional neural network for complex spectrogram processing in speech enhancement. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5756–5760.
8. Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. ArXiv Preprint ArXiv:2008.00264.
9. Awad, A. (2019). Impulse noise reduction in audio signal through multi-stage technique. Engineering Science and Technology, an International Journal, 22(2), 629–636
10. Bai, M. R., & Kung, F. J. (2022). Speech Enhancement by Denoising and Dereverberation Using a Generalized Sidelobe Canceller-Based Multichannel Wiener Filter. Journal of the Audio Engineering Society, 70(3), 140–155. https://doi.org/10.17743/JAES.2021.0059
11. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., & Sainath, T. (2019a). Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 13(2), 206–219.
12. Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261
13. Biswas, A., & Jia, D. (2020). Audio Codec Enhancement with Generative Adversarial Networks. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May, 356–360. https://doi.org/10.1109/ICASSP40776.2020.9053113
14. Song, Y., Liu, F., & Shen, T. (2023). A novel noise reduction technique for underwater acoustic signals based on dual-path recurrent neural network. IET Communications, 17(2), 135–144. https://doi.org/10.1049/CMU2.12518
15. Zhang, Y., & Li, J. (2023). BirdSoundsDenoising: Deep Visual Audio Denoising for Bird Sounds (pp. 2248–2257).
16. Bai, M. R., & Kung, F. J. (2022). Speech Enhancement by Denoising and Dereverberation Using a Generalized Sidelobe Canceller-Based Multichannel Wiener Filter. Journal of the Audio Engineering Society, 70(3), 140–155. https://doi.org/10.17743/JAES.2021.0059.
17. Bhat, G. S., Shankar, N., Reddy, C. K. A., & Panahi, I. M. S. (2019). A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone. IEEE Access, 7, 78421–78433. https://doi.org/10.1109/ACCESS.2019.2922370