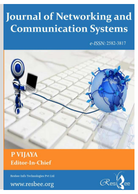
Vol.6 No.1 Jan 2023

Journal of Networking and Communication Systems

Received 27 September, Revised 29 November, Accepted 18 December


Journal of Networking and Communication Systems
e-ISSN: 2582-3817
P VIJAYA
Editor-In-Chief
Resbee Info Technologies Pvt Ltd
www.resbee.org

FATFA-based DRN: Feedback Artificial Tree Firefly Algorithm-Enabled Deep Residual Network for Pathological Speech Enhancement

Srinivas Katkam
Department of CSE
Geethanjali College of Engineering and Technology, Medchal, Telangana, India.

Abstract: The voice is a most important tool for the communication among the people in their day-to-day life. However, any slight change in the voice production system may affect the quality of the voice. For the past years, researchers worked to develop an effective automatic system for the clinicians to perform a preventive diagnosis for detecting the voice pathologies in an early stage. In this paper, the Feedback Artificial Tree Firefly Algorithm (FATFA)-based Deep Residual Network (DRN) is developed for the pathological speech enhancement. Here, the Hanning window is employed for extracting the frames from the speech signals. The multiband spectral subtraction technique is utilized for improving the extracted frames from the speech signals. Additionally, the DRN classifier is used for enhancing the pathological speech signals where the employed classifier is trained by the proposed optimization algorithm, called FATFA. Here, the developed FATFA-based DRN is the integration of Feedback Artificial Tree (FAT) and Firefly Algorithm (FA). However, the developed pathological speech enhancement method achieved efficient performance in terms of Perceptual Evaluation of Speech Quality (PESQ), and Root Mean Square Error (RMSE) with a higher PESQ of 3.907, and lesser RMSE of 30.64, respectively.

Keywords: Deep Residual Network, Feedback Artificial Tree, Firefly algorithm, Hanning Window, Speech enhancement.

1. Introduction

Voice is commonly considered as the fundamental mode of communication for humans. The pathological voice detection becomes a challenging task that should be robustly implemented [12] [13]. With respect to the perspective on the field of health diagnosis, pathologic voice detection depends on the phonetic indication or physiological defects of the voice scheme. In the medical field, the most commonly used diagnostic methods are Stroboscopy, Laryngoscopy, and Endoscopy, which is considered as a time consuming, persistent, and an invasive method [14] [15]. These techniques require skilled medical experts, well specialized clinical equipment's, and are time consuming and expensive. Hence, a robust and effective automatic voice diagnosis technique is more helpful for the human voice health. Compared with conventional health related techniques, computer-aided detection technique is efficient, noninvasive, limited cost, real time, and robust [7]. In addition, around 25% of the people undergo voice related problems because of the unhealthy standard of living and voice abuse [16]. People who make use of their voices extremely, and whose jobs entail them to speak loud. For instance, lawyers, singers, actors, teachers, etc. are mainly at danger of being affected by numerous types of voice issues [8]. Automatic recognition of voice pathology enables an objective evaluation and an early involvement for the analysis. A typical system based on the voice pathology detection comprises of two phases: The initial phase is the illustration of the acoustic speech signal input called feature extraction and the second phase is the classifier phase for making the normal and pathological decision. The feature sets utilized for the detection of the voice pathology can be generally categorized into 3 types [17] [18], namely perturbation measures, spectral and cepstral measures, and complexity measures.

The area of speech enhancement (SE) deals with enhancing the speech signals that have been corrupted by noise [19]. Speech enhancement is commonly utilized in the systems based on Automatic Speech Recognition (ASR) as a preprocessing phase for enhancing the accuracy of the system in noisy surroundings [20]. In recent times, investigation into this purpose has been developed, resulting in a considerable performance increase in the ASR systems. This accomplishment has also lead to a new

https://doi.org/10.46253/jnacs.v6i1.a1

interest in the speech enhancement applications for human listeners where the objective is to improve the speech intelligibility and speech quality by making the speech easier to recognize and making more comfortable for the listeners [21] [22]. The latter application is particularly most important in the areas of hearing assistive technology and telecommunications [1]. The monaural speech enhancement is a difficult task as it contains only the single channel information. However, it is significant in the applications based on the speech signal processing, namely mobile communication, robust speech recognition approach, and hearing aids. The main objective of the speech enhancement is to enhance the intelligibility and quality under the interfering noise situations. Speech enhancement methods are introduced and deployed to improve the superiority and the lucidity of the noisy speech signals. Conventional speech enhancement approaches involves the filtering methods, namely spatial filtering, and spectral filtering. In the earlier approaches, the interfering noises were mitigated by considering the spatial properties, whereas the interfering voice signals captured by a particular microphone sensor is processed in the latter method.

Deep neural network (DNN) has been developed to model the data in several applications [23]. With the expansive growth in the machine learning techniques, and big data approaches empowered by the health services of the common population [24], it is necessary to re-investigate the computerized voice classification systems [11]. The fundamental basis of the pathological voice analysis using the deep learning strategies is to distantly evaluate the health of the voice and review voice treatment processes. In recent times, deep learning techniques are widely employed in the speech recognition, and the image processing areas [25]. Nowadays, deep learning has a progressive growth in the quantity and superiority of pathological voice judgment [26]. Hence, deep learning techniques have the capability for solving the pathological voice classification problems [7]. In most of the fields, DNN approaches have gained extensive development because of their capability to enhance the various existing approaches. Techniques based on DNN [27] are currently receiving a lot of interest due to their potential to outperform earlier techniques. For the ASR systems, the system performance is calculated by the ability of the SE system to minimize the error rate. Based on the subjective evaluation of speech intelligibility and quality [19], the performance is preferably determined for the human listeners. In general, these voice enhancement-based tests compares the evaluation results of the interference sound before and after the enhancement process to quantify the effects of various mechanisms [1]. Compuer-aided pathological voice detection techniques are well-organized for the initial screening of pathological voice, and have established high intellectual and medical consideration [7].

This research work is focused on designing the proposed FATFA-based DRN algorithm for the pathological speech enhancement. The developed FATFA-based DRN includes several stages, such as framing, initial improvement, and pathological speech enhancement. Initially, the pathological voice signals are utilized for extracting the frames from the speech signals where the process is carried out using the Hanning window. The extracted frames are presented to the initial improvement stage to improve the voice signals using the multiband spectral subtraction. After that, the pathological voice enhancement is performed using the DRN classifier where the classifier is trained by the FATFA algorithm. However, the developed FATFA is derived by the integration of the FAT and FA respectively.

The major contribution of the research work is described as follows:

- ❖ **Proposed FATFA-based DRN:** An effective pathological speech enhancement approach is devised using developed FATFA-based DRN classifier. Here, the Deep Residual network classifier is employed for improving the pathological speech enhancement procedure where the classifier is trained by the developed optimization algorithm, named FATFA approach.

The research paper is organized as follows: section 2 describes the various pathological speech enhancement approaches, section 3 portrays the proposed method for pathological speech enhancement, section 4 explains the results and discussion, and finally the conclusion of the paper is illustrated in section 5.

2. Motivation

In this section, some of the existing pathological speech enhancement methods are reviewed along with their advantages and drawbacks that motivate the researchers to develop the proposed mechanism for enhancing the pathological voice.

2.1 Literature Survey

Some of the existing pathological speech enhancement methods are reviewed in this section.Femke B. Gelderblom *et al.* [1] developed a Deep Neural Networks (DNN). This method achieved higher accuracy to improve the intelligibility and the quality of the speech. However this method was very complex to recognize the speech signals in babble noise. R.Saravana Ram *et al*. [2] introduced a Generative Adversarial Networks (GAN) for improving the speech signals. This generator improved the noisy data input, whereas the discriminator embedded with this generator classified the out generated and the database clean contents. This method minimized the computational issues. However, this method failed to examine the fake classification. QizhengHuang *et al*. [3] devised a speech enhancement method using Multi-Band Excitation (MBE) model for the speech enhancement. This model comprises of two phases, namely training phase, and an enhancement phase. In the training phase, the DNN networks were trained, whereas in the enhancement phase, the noisy speech was initially pre-processed using the multi-band spectral subtraction technique, and finally MBE parameters were utilized for enhancing the noisy speech. This model enhanced the quality of the speech. However, this approach failed to estimate the pitch accurately at low SNR levels for the non-stationary noise. KouheiSekiguchi *et al*. [4] designed a Multichannel Non-negative Matrix Factorization (MNMF) and Independent Low Rank Matrix Analysis (ILRMA) for enhancing the speech signals. This technique utilized only the noise free data, and this method can work without spatial information under the single channel circumstances. However, this method failed to perform dereverberation and source localization

Mazin Abed Mohammed *et al*. [5] developed a Convolutional Neural Networks (CNN) for the detection and classification of the voice pathology. This approach comprises of four phase, namely dataset preparation, learning phase, training and testing, and finally the interference processing. This method achieved accurate detection of pathological voice. However, this method failed to test various types of CNN and training models in order to enhance the detection performance. Sudarsana Reddy Kadiri and Paavo Alku [6] introduced a Quasi-Closed Phase (QCP) for the automatic analysis of voice pathology. Here, the glottal features were derived from the glottal flows, estimated source signals, and auditory voice signals. This method provided better discrimination results. This method failed to estimate the severity level in pathological voice. Lili Chen and Junjiang Chen [7] introduced a Mel Frequency Cepstral Coefficients (MFCC) in order to classify the pathological voice from healthy voice. In this approach, DNN was constructed with the two-layer stacked sparse auto encoders,and a softmax layer. This method achieved higher specificity, sensitivity and accuracy. However, this method failed to use more data in order to train the DNN. I.Hammami *et al*. [8] devised an Empirical Mode Decomposition-Discrete Wavelet Transform (EMD-DWT) for improving the classification performance of voice pathology. In this method, discrete wavelet features were extracted, and then tested. This method achieved higher classification accuracy. However, this method failed to consider deep learning technologies to improve the identification accuracy.

In 2023, Costantini et al. [32] have used CNN and ML throughout the process. Initially, the voice recording was collected from PD patients. Then the voice recordings were divided into groups namely recently identified PD patients, patients not started taking any medication, and medicine intaking persons. Following that, pipelines of ML were compared with the above groups. Finally, CNN architecture was used along with deep learning to get the patient's details accurately. This method was more effective in grouping different categorized PD patients however there were many features to improve the efficiencies that were not explored.

In 2023, Liu et al. [33] have implemented a soft-adaptive threshold spectral enhancement (SATSE) and Spectral Cross-domain neural network (SCDNN). Initially, Input Data were gathered from ResNet18. Then the gathered information was sent to Fast Fourier Transformation to gather data such as spectral domains and extraction time through SATSE blocks. Then the adoptive average pooling layer was followed by an adoptive max pooling layer and it was further refined and passed to a layer that was fully connected. Finally, SCDNN was tested on different task-implemented datasets. The result was more efficient than another existing method however it required a high computational task.

In 2023, Ksibi et al. [34] have ensembled Deep learning technologies. Initially, the datasets were collected from open-source platforms and sorted into healthy and diseased samples then using a deep learning model the audio of the extracted samples was classified. Finally, with the help of a two-level cascaded model they classified the sample as healthy and diseased. Thus the result provide more accurate results in early diagnosis but they focused only on single vowels.

In 2023, Kim et al. [35] have implemented a deep convolution neural network. Initially, samples from the diseased patients and the healthy people were recorded on a mobile phone. Where the text was pre-given to the people and they have to read the text. Here they used three methods to detect depressed people and healthy people. They are CNN based on features, classification through a log-Mel spectrogram, and training and classification through a log-Mel spectrogram by using known pre-trained networks. In the end, acoustic characteristics were used to predict depression automatically. The predicted result showed an accurate result however it was tested with only fewer samples.

In 2022, Weise et al. [36] have applied voice conversion technology and auto-encoding architecture. Initially, the speech information was decomposed into different components such as content, pitch/intonation, timing/rhythm, and timbre. The encoding and decoding process was used to compare the healthy references using the Mean squared error loss function. It contained three encoders to obtain the latent codes from the log-Mel spectrogram. In the decoding process, the audio from the original input was based on codes and with an encoder speaker id. In addition to this, the Random Resampling technique was used to disturb the temporal structure during encoding data points. The result showed significant improvement over the existing method however it was limited to only English Language.

In 2022, Yang, M., et al. [37] have practiced extended Geneva Acoustic Parameter Set (eGeMAPS), Deep learning technology to maximize quality and openSMILE. It contains two processes. Initially, eGeMAPS creates differential estimators to finely tune the voice of the existing system. Then the obtained results were evaluated using Deep noise suppression. This approach reduced the variance between the initial speech and the enhanced speech however it was difficult to apply experimentally using limited resources.

In 2022, Miller et al. [38] have adopted the autoencoder technique to detect Parkinson's disease patients' speech signals. Initially, the speech signals from the patients were extracted. Using various parameters, the speech signals were analyzed. Then auto-encoder and multi-spectral fusion were used to identify various pathologies of the patient. Finally, the results were evaluated to understand the severity of the patient’s disease.

In 2022, Zakariah et al. [39] have tried AI, DNN, and different types of feature analysis. Initially, audio from the patients was collected using DNN and those data were sent to three features for analysis to separate the health and affected sounds. Then AI analyses the voice and DNN collects accurate data and sends it to the examiner. Here it provides absolute details and also the gender information.

2.2. Challenges

The various challenges faced during the pathological speech enhancement are as follows,

- In [1], DNN is developed for the speech signal enhancement. However, this method failed to introduce the objective measures that correspond well with the speech quality, and speech intelligibility for the difficult nonlinear degradation channels processing with the DNN-based SE system that impedes the progress for DNN-based SE systems, like telecommunication and hearing assistive devices.
- GAN is developed for enhancing the speech signals, but this approach failed to optimize the various circumstances selected for extracting the samples in order to provide more efficient speech enhancement system [2].
- In [4], MNMF and ILRMA are developed for the enhancement of the speech. However, these methods failed to consider the online expansion of the full-rank representation for the real-time applications in order to improve the performance of the system.
- QCP method is designed for the pathological voice detection and classification [6]. However, the challenge lies in utilizing the glottal source features for the pathology classification, and also for estimating the level of pathology as mild, medium, high and very high to enhance the diagnosis.
- In [7], MFCC method is developed for the automatic recognition of pathological voices. However this method does not investigate the DNN applications in different experiments and clinical practices, and also failed to utilize more amounts of data to verify the performance of DNN.

3. Proposed FATFA-based DRN for Pathological Speech Enhancement

This section explains the design and development of a pathological speech enhancement approach using a novel optimization algorithm, named FATFA. This developed pathological speech enhancement method mainly includes the following three phase, namely, framing, initial improvement, and pathological speech enhancement. Initially, the pathological voice signals is given to the framing phase in order to perform segmentation for extracting the frames from the speech signal, which is carried out using the Hanning window. Once the frames are formed, they undergo initial improvement. The improvement of the pathological voice signals is done based on the multi-band spectral subtraction in order to improve the quality of the speech signals. Once the initial improvement is done, the improved speech signals are subjected to the final phase, called pathological speech enhancement phase. The enhancement of the pathological voice signals is carried out using the DRN [31], which is trained using the proposed FATFA optimization algorithm. The proposed FATFA algorithm is designed by the integration of the FAT [9] and FA [10], respectively. Fig.1 illustrates the schematic diagram of the proposed FATFA-based DRN for the pathological speech enhancement. Fig.1 illustrates the schematic diagram of the proposed FATFA-based DRN for the pathological speech enhancement.

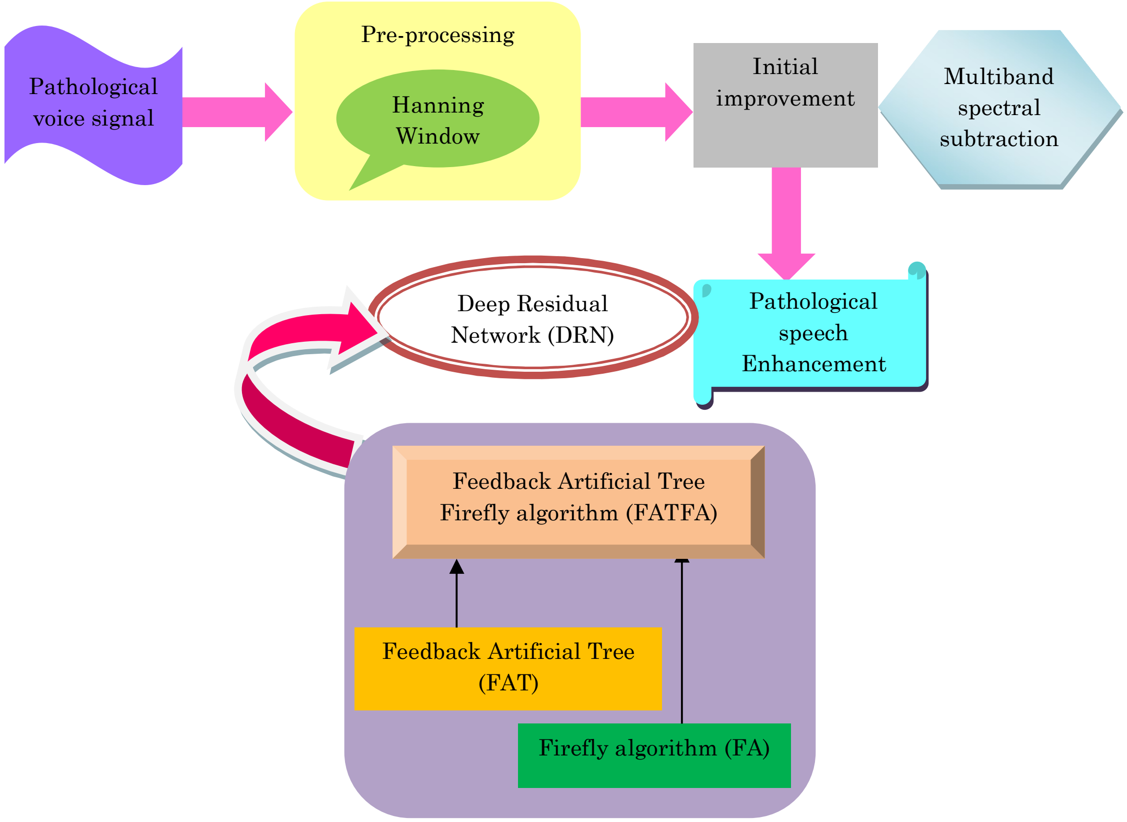

Pathological voice signal
Pre-processing
Hanning Window
Initial improvement
Multiband spectral subtraction
Pathological speech Enhancement
Deep Residual Network (DRN)
Feedback Artificial Tree Firefly algorithm (FATFA)
Feedback Artificial Tree (FAT)
Firefly algorithm (FA)

Fig.1. *Schematic diagram of the proposed FATFA-based DRN for t pathological speech enhancement.*

3.1 Input Pathological Voice Signal Acquisition

Initially, the input pathological voice signal is collected from the dataset, and various noises, such as airport noise, babble noise, car noise, train noise and exhibition noise is involved in order to generate noisy pathological voice signal. Let us consider the dataset D with f noisy pathological speech signal, and it is expressed as,

$$D = \{E_1, E_2, \ldots E_d, \ldots E_f\} \; ;1 \le g \le f \tag{1}$$

where, E represents the noisy pathological speech signal and D is a dataset. Here, the input pathological voice signal E_d is considered as input for the framing step.

3.2 Input signal Framing using Hanning Window

In this step, E_d is considered as an input for framing process, which is carried out using the Hanning window. The vast speech stream is transformed to continuous stream of chunks, called as frames. The major advantage of using Hanning window [29] is to acquire the signals from the immobile circumstances. In addition, the Hanning window executes a process of fine tuning by the frames, and the equation of Hanning window is formulated as,

$$H(w) = \begin{cases} 0.5\left(1 - \cos\left[\dfrac{2\pi t}{E-1}\right]\right) \cong \sin^2 \dfrac{\pi t}{(E-1)} & ;0 \le t \le E-1 \\ 0 & ;\text{Otherwise} \end{cases} \tag{2}$$

where, $H(w)$ signifies Hanning window function with a dimension of $[1 \times 255]$. The segments are zero-padded, and it is presented to the feature extraction module for further processing. The output of the input signal framing is indicated as F_l.

3.3 Initial Improvement using Multi-band Spectral Subtraction

Once the framing process is performed, then the framing output F_i is considered as the input for feature extraction process, which is performed using the multiband spectral subtraction [30]. Let us consider the additive noise to be stationary and correlated with the clean pathological voice signal, and the corrupted pathological speech can be formulated as,

$$b(l) = k(l) + c(l) \tag{3}$$

where, $b(l)$, $k(l)$, $c(l)$represents the corrupted pathological speech signal, clean pathological speech signal, and noise. The corrupted pathological speech with the power spectrum can be expressed as,

$$\left|U(m)\right|^2 \approx \left|R(m)\right|^2 + \left|V(m)\right|^2 \tag{4}$$

where, $R(m)$and $V(m)$ signifies the magnitude spectra of the clean pathological speech and the noise, respectively. Since, the noise spectrum cannot be calculated, an approximation of $\hat{V}(m)$can be calculated with respect to the state of silence.

The approximation of clean pathological speech spectrum is computed as,

$$\left|\hat{R}(m)\right|^2 = \left|U(m)\right|^2 - \theta\left|\hat{V}(m)\right|^2 \tag{5}$$

where, θ signifies the factor of over-subtraction. The multi-band spectral subtraction method considers the non-uniform cause of the colored noise on the pathological speech spectrum. The pathological speech spectrum is partitioned into non-overlapping bands represented as ,and the spectral subtraction is carried out separately in every bands. Hence, the clean pathological speech spectrum in the band is computed by the equation given below as follows,

$$\left|\hat{R}_i(m)\right|^2 = \left|U_i(m)\right|^2 - \theta_i \nu_i \left|\hat{V}_i(m)\right|^2 \quad h_i \le m \le e_i \tag{6}$$

where, h_i and e_i represents the frequency bins indicating the beginning and the ending frequency of i^{th} frequency band, θ_i represents the factor of over-subtraction in the i^{th} band, and ν_i signifies the tweaking factor that can be independently set for every frequency band to modify the noise elimination properties.

$$SNR_i(dB) = 10\log_{10}\left(\frac{\sum_{m=h_i}^{e_i}\left|U_i(m)\right|^2}{\sum_{m=h_i}^{e_i}\left|\hat{V}_i(m)\right|^2}\right) \tag{7}$$

Using the value of SNR_i computed in equation (7), θ_i can be expressed as,

$$\theta_i = \begin{Bmatrix} 5 & SNR_i < -5 \\ 4 - \frac{3}{20}(SNR_i) & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{Bmatrix} \tag{8}$$

The utilization of over-subtraction factor θ_i produces a degree of control over the noise subtraction level in every band, and hence the utilization of multi frequency bands and ν_i weights offer an extra degree of control over every band.

The enhanced spectrum with the negative values in equation (6) was floored to the noisy spectrum as:

$$\left|\hat{R}_i(m)\right|^2 = \begin{Bmatrix} \left|\hat{R}_i(m)\right|^2 & \left|\hat{R}_i(m)\right| \\ \left|V_i(m)\right|^2 & \end{Bmatrix} \tag{9}$$

where, the parameter of the spectral floor was set to $\delta = 0.002$, respectively. Finally, the output generated by the initial improvement process is I_m .

3.4 Pathological Speech Enhancement using Deep Residual Network

In this section, the pathological speech enhancement procedure is explained. Here, the DRN is utilized for the enhancement process where the initial improvement output I_m is fed as the input to the network. The architecture of DRN is depicted in fig.2,and the optimization process employed for training the classifier is described as follows.

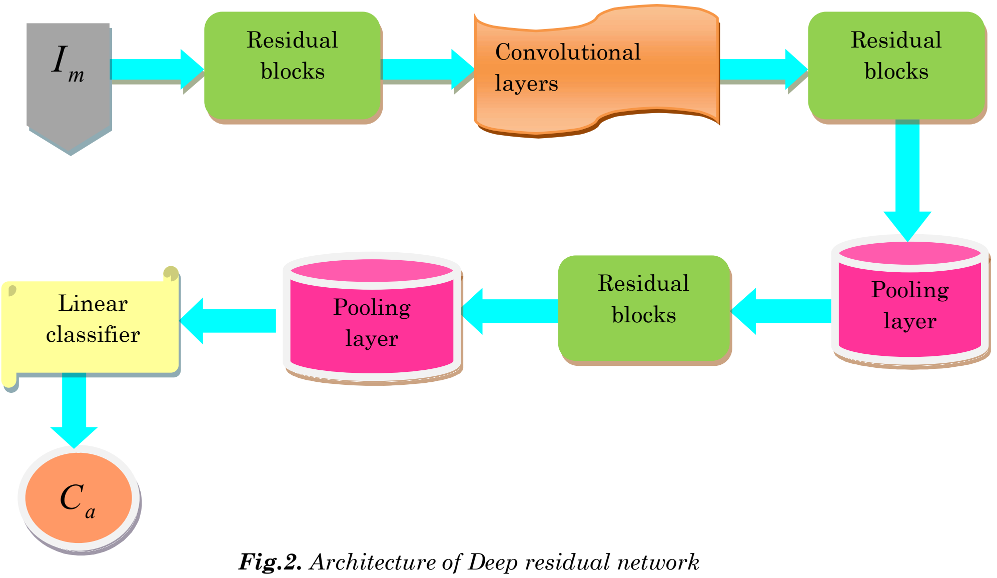


Fig.2. *Architecture of Deep residual network*

Convolutional (conv) layer: It is a two-dimensional layer, which minimizes the training parameters, and it facilitates the reimbursement in order to share the weights. The input image in the conv layer is managed by the filters called kernel. This layer applies the mathematical models to reduce the input of the filter matrix such that the dot product value of the kernel is evaluated, and the computation of the conv layer is formulated by,

$$B2a(T) = \sum_{p=0}^{L-1} \sum_{q=0}^{L-1} M_{p,q} \bullet T(s+p),(u+q) \tag{10}$$

$$B1a(T) = \sum_{J=0}^{C_{in}-1} G_J * T \tag{11}$$

where, T represents the features of CNN from the input image, the terms s and u are utilized for the coordinates recoding, the $L \times L$ kernel matrix called learnable parameter is denoted by G, and the terms p and q signifies the index position of the kernel matrix. Therefore, the size of the kernel for J^{th} input neuron is indicated as M_j, and the operator for the cross correlation is denoted as $*$.

Pooling layer: It is a layer associated among the conv layers, which reduce the spatial size of the feature maps. Hence, the average pooling function is selected to function on all slices and depth of the feature maps

$$p_{out} = \frac{p_{in} - J_p}{\lambda} + 1 \tag{12}$$

$$q_{out} = \frac{q_{in} - J_q}{\lambda} + 1 \tag{13}$$

where, p_{in} indicates the width of input matrix, q_{in} represents the height of the input matrix, p_{out} and q_{out} are the values of output. In addition, the terms J_a and J_s indicates the height and width of kernel size.

Activation function: Activation function understands the non-linear features and the difficult features for enhancing the non-linearity features. Rectified linear unit (ReLU) is a type of non-linear activation function, which is applied for enhancing the speech. The ReLU function is formulated as,

$$ReLU(T) = \begin{cases} 0\,; H < 0 \\ H;\, H \geq 0 \end{cases} \tag{14}$$

Here, H represents the feature.

Batch normalization: In this batch normalization, the training set is partitioned into various small sets called mini batches to train the method. It facilitates an exchange between the computational complexity and convergence. Moreover, the input layers are normalized by scaling and modifying the activations for improving the consistency and training speed.

Residual blocks: Residual blocks indicate the alternative connection between the conv layers. If the size of the input and output are same, then the input can be directly attached to the output. For the varying sizes, the dimension matching factor is employed in order to match the input with the output.

$$R_b = O(T) + T \tag{15}$$

$$R_b = O(T) + \lambda_C T \tag{16}$$

where, T and R_b denotes the input residual blocks of input and output residual blocks, R_b signifies the mapping link, and λ_C indicates the factor with respect to dimension matching.

Linear classifier: The linear classifier enhances the pathological speech from the input voice. Moreover, the linear classifier is the combination of the fully connected layer and the soft max function.

$$R_b = \lambda R_b + \omega \tag{17}$$

Here, the classified output is denoted as C_a.

b) Training procedure of Deep Residual network using FATFA-DRN algorithm

The training process of DRN is done by developing the FATFA optimization algorithm, which is formed by the integration of the Feedback Artificial Tree (FAT) and Firefly Algorithm (FA). FA is the bio-inspired optimization algorithm for solving the complex optimization problems. This algorithm is inspired by the flashing performance of the fireflies. Here, the randomly generated solutions are considered as the fireflies and the brightness is assigned with respect to their performance on the objective function. FAT algorithm is motivated by the transportation of organic matters and the update theories of branches. By integrating the flashing behavior of the fireflies with the FAT algorithm, the enhancement of the pathological voice signal is made more effective. The algorithmic steps involved in the FATFA-based DRN is explained as follows,

Step 1: Initialization: The population of fireflies is randomly initialized, represented as Q, and it is formulated as,

$$Q = \{Q_1, Q_2, \ldots Q_x, \ldots Q_y\} \quad ;1 \le x \le y \tag{18}$$

Step 2: Fitness function estimation: The fitness function is estimated for computing the best solution in order to enhance the pathological voice signal, and it is expressed as,

$$X = \frac{1}{f}\sum_{\beta=1}^{f}\left[W_Z - Y^{\beta}\right] \tag{19}$$

where, Y^{β} represents the target output, W_Z signifies the classified output, X represents the fitness function, and f represents the number of samples.

Step 3: Update solution: Once the fitness function is computed, the update solution is obtained using the proposed FATFA-based DRN. Here, firefly j gets attracted to brighter firefly i and the association between the fireflies is computed in order to find the best solution, and hence the update solution of the fireflies is formulated as,

$$Q_{j+1} = Q_j + \lambda_0 e^{-\beta q_{ji}^2}\left(Q_i - Q_j\right) + \alpha\varepsilon_j \tag{20}$$

$$Q_{j+1} = Q_j + \lambda_0 e^{-\beta q_{ji}^2} Q_i - \lambda_0 e^{-\beta q_{ji}^2} Q_j + \alpha\varepsilon_j \tag{21}$$

$$Q_{j+1} = Q_j\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right) + \lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j \tag{22}$$

The issues in the FA algorithm can be reduced by combining the FA algorithm with FAT. Hence, the standard equation of FAT is illustrated below using the self evaluation operator as follows,

$$Q_{j+1} = Q_j + \mathrm{rand}(0,1)\left(Q_{best} - Q_j\right) \tag{23}$$

$$Q_{j+1} = Q_j + \mathrm{rand}(0,1)Q_{best} - \mathrm{rand}(0,1)Q_j \tag{24}$$

$$Q_{j+1} = Q_j\left(1 - \mathrm{rand}(0,1)\right) + \mathrm{rand}(0,1)Q_{best} \tag{25}$$

$$Q_j\left(1 - \mathrm{rand}(0,1)\right) = Q_{j+1} - \mathrm{rand}(0,1)Q_{best} \tag{26}$$

$$Q_j = \frac{Q_{j+1} - \text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)} \tag{27}$$

Substituting equation (27) in equation (24)

$$Q_{j+1} = \frac{Q_{j+1} - \text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right) + \lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j \tag{28}$$

$$Q_{j+1} = \frac{Q_{j+1}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right)}{1 - \text{rand}(0,1)} - \frac{\text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right) + \lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j \tag{29}$$

$$Q_{j+1} - \frac{Q_{j+1}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right)}{1 - \text{rand}(0,1)} = \lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j - \frac{\text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right) \tag{30}$$

$$Q_{j+1}\left(1 - \frac{1 - \lambda_0 e^{-\beta q_{ji}^2}}{1 - \text{rand}(0,1)}\right) = \lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j - \frac{\text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right) \tag{31}$$

$$Q_{j+1}\left(\frac{1 - \text{rand}(0,1) - 1 + \lambda_0 e^{-\beta q_{ji}^2}}{1 - \text{rand}(0,1)}\right) = \lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j - \frac{\text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right) \tag{32}$$

$$Q_{j+1}\left(\frac{\lambda_0 e^{-\beta q_{ji}^2} - \text{rand}(0,1)}{1 - \text{rand}(0,1)}\right) = \lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j - \frac{\text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right) \tag{33}$$

$$Q_{j+1} = \frac{1 - \text{rand}(0,1)}{\lambda_0 e^{-\beta q_{ji}^2} - \text{rand}(0,1)}\left[\lambda_0 e^{-\beta q_{ji}^2} Q_i + \alpha\varepsilon_j - \frac{\text{rand}(0,1)Q_{best}}{1 - \text{rand}(0,1)}\left(1 - \lambda_0 e^{-\beta q_{ji}^2}\right)\right] \tag{34}$$

where, ε_j signifies the random numbers taken from Gaussian distribution, and α controls the size of the step and it lies within the range $[0,1]$.

Step 4: Evaluate the feasibility of the solution: In step 4, the feasibility of the solution is computed using the objective function. If the solution obtained is best than the previous solution, then it is replaced by the newly obtained optimal solution.

Step 5: Termination: All the above steps are repeated until the best optimal solution is achieved. Algorithm 1 illustrates the pseudo code of developed FATFA-based DRN.

Algorithm1. Pseudo code of developed FATFA-based DRNalgorithm

Sl. No	Pseudo code of proposed FATFA-based DRN
1	**Input:** Firefly population
2	**Output:** Q_{j+1}
3	Randomly initialize the population with fireflies
4	Compute fitness measure of each firefly
5	while $(y < \text{max generation})$
6	for $j = 1$ to M, all n fireflies do
7	for $i = 1$ to M, all n fireflies do
8	If $W_i > W_j$ then
9	Move firefly j toward i
10	End if
11	End for i
12	End for j
13	Evaluate new solutions
14	Rank the fireflies and identify the current best solution
15	End while

4. Results and Discussion

The analysis of the proposed FATFA-based DRN approach with several classical techniques using Saarbruecken Voice Database [28] is illustrated with respect to the metrics, such as RMSE and PESQ.

4.1 Experimental Setup

The implementation of developed FATFA-based DRN technique is carried out in the Matlab tool with PC having Windows 10 OS, 2GB RAM, and Intel i3 core processor.

4.2 Dataset Description

The experimentation of developed FATFA-based DRN is carried out with Saarbruecken Voice Database.

4.2.1 Saarbruecken Voice Database

This voice database consists of a cluster of recordings related to noise with more than 2000 persons. The ECG signals and the speech signals are stored in different files. Any comments about the recordings are stored in the related text-file. Based on the quality of the recording, there are some sessions in recording where not every vowel in all versions is present.

4.3. Performance Evaluation Measures

The efficiency of the developed FATFA-based deep residual network techniques can be analyzed by considering the RMSE and PESQ metrics.

4.3.1. RMSE: It is a measure which computes the standard deviation of the prediction error, and it is expressed as,

$$RMSE = \sqrt{\frac{(P(r) - K(r))^2}{S}} \tag{35}$$

where, $K(r)$ represents the enhanced speech signal, S denotes the input signal length, and $P(r)$ represents the noisy speech signal.

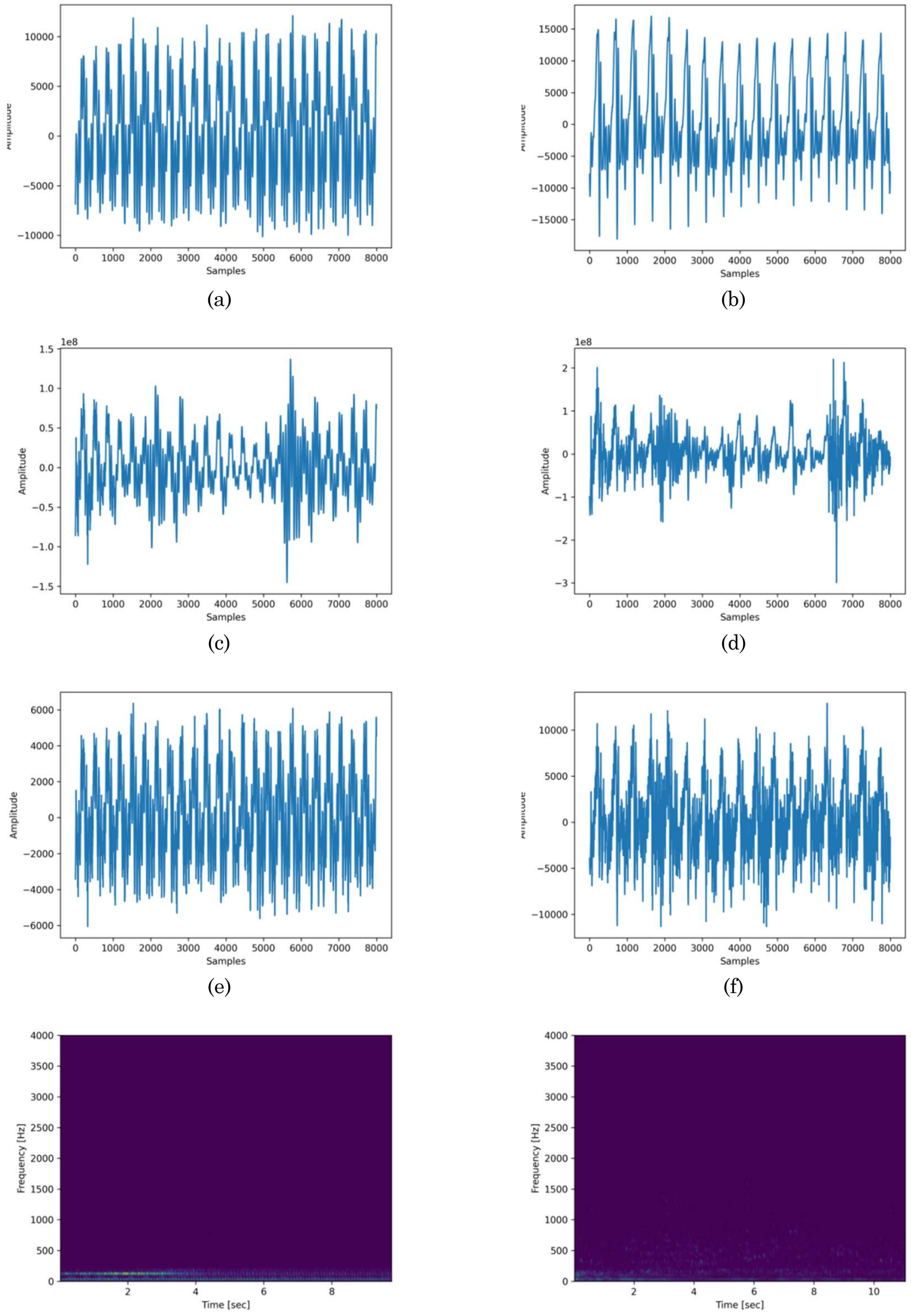
4.3.2. PESQ: PESQ is a measure, which is used for computing the quality of the speech signal.

$$PESQ = N_1 + N_2 A + N_3 B \tag{36}$$

where, B represents the value of the asymmetrical disturbance, and A signifies the value of the average disturbance.

4.4. Experimental Results

The experimental results of developed FATFA-based DRN technique for pathological speech signal enhancement is portrayed in fig. 3. The input signal-1 and input signal-2 is portrayed in fig. 3 a) and 3 b), noise added for the input signal-1 and input signal-2 is depicted in fig. 3 c) and 3d), output signal-1 and signal-2 is illustrated in fig.3 e) and 3 f), and spectrogram for input signal-1 and input signal-2 is depicted in fig. 3 g) and 3 h).



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig.3. *Experimental results of developed FATFA-based DRN (a) input signal-1 (b) input signal-2 (c) noise added signal for input signal-1 (d) noise added signal for input signal-2 (e) output signal-1 (f) output signal-2 (g) spectrogram for input signal-1 (h) spectrogram for input signal-2*

4.5. Comparative Techniques

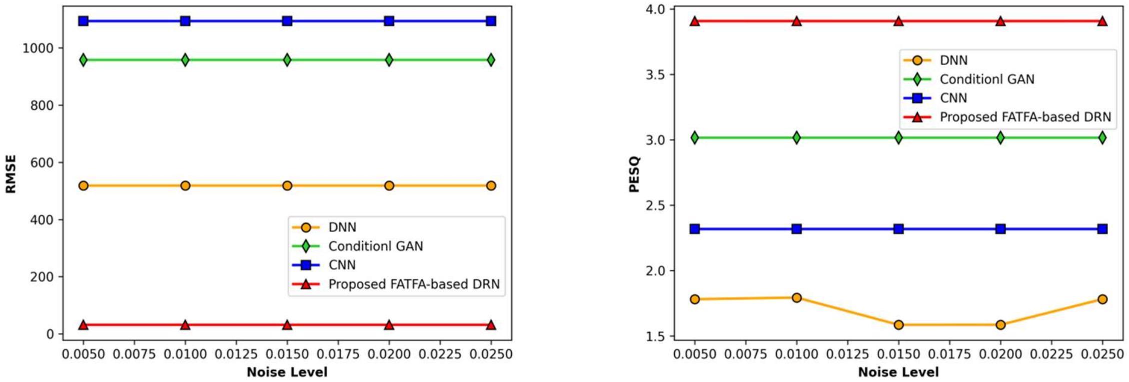
The techniques taken for analysis are DNN [1], Conditional GAN [2], CNN [5], and proposed FATFA-based deep residual network.

4.6 Comparative Analysis

This section illustrates the comparative analysis made by developed pathological speech enhancement method using various noises, such as airport noise, babble noise, car noise, train noise and exhibition noise.

a) Analysis using airport noise

Fig.4 portrays the assessment of RMSE and PESQ by varying the airport noise level. The assessment of RMSE is shown in fig. 4a). When the noise level is 0.0100, the RMSE measured by the proposed FATFA-based DRN is 31.69, whereas the RMSE obtained by the existing DNN, Conditional GAN, and CNN are 518.53, 958.01, 1093.51, respectively.The PESQ metric analysis is shown in fig. 4b). The PESQ value of DNN, Conditional GAN, and CNN method is 1.792, 3.015, 2.316, whereas the developed FATFA-based DRN technique obtained the PESQ value of 3.907 for 0.0050 dB noise level.

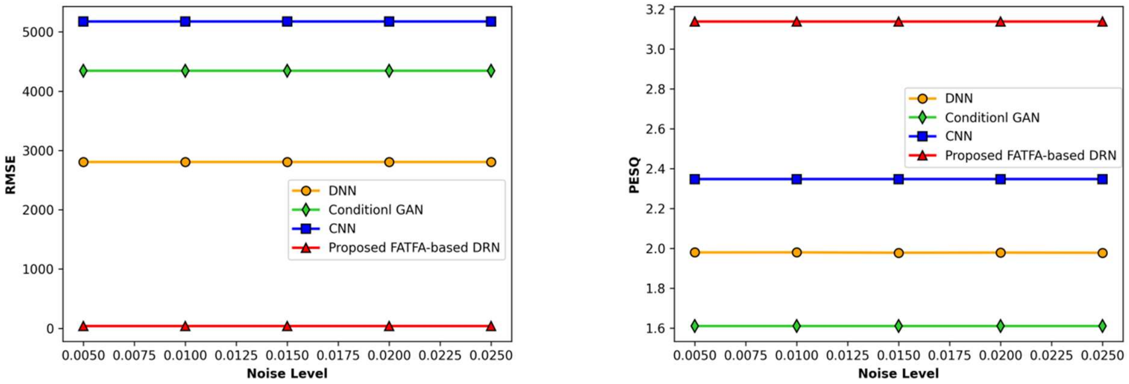


(a) (b)

Fig.4. *Assessment based on airport noise a) RMSE b) PESQ*

b) Analysis using Babble Noise

The analysis of RMSE and PESQ metric based on the babble noise is illustrated in fig.5. The analysis of RMSE by varying the babble noise levels is shown in fig. 5 a). By considering the noise level as 0.0050, the RMSE obtained by the existing DNN is 2806.82, Conditional GAN is 4345.83, CNN is 5174.55, and the proposed FATFA-based DRN is 38.184, respectively. The PESQ analysis is shown in fig. 5 b). When the noise level is 0.0200 dB, the PESQ attained by DNN is 1.978, Conditional GAN 1.610, CNN is 2.347 and developed FATFA-based DRN algorithm is 3.137, respectively.

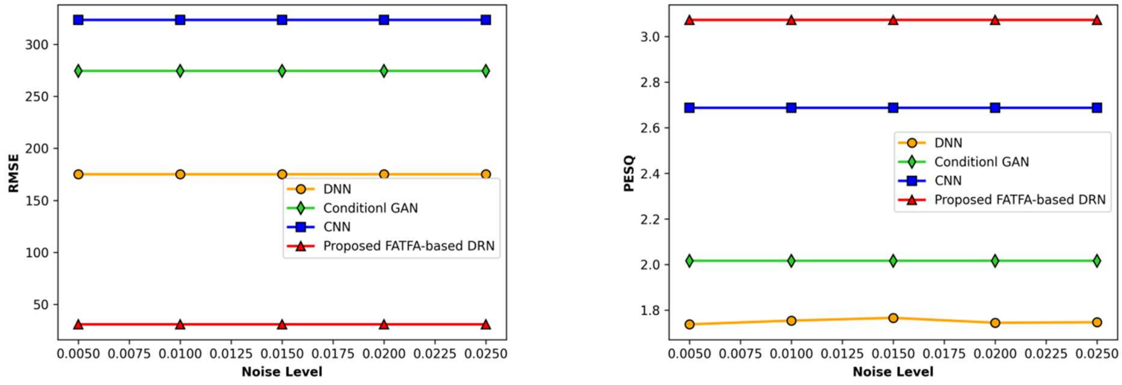


(a) **(b)**

Fig.5. *Analysis with respect to babble noise a) RMSE b) PESQ*

c) Analysis using Car Noise

The assessment of RMSE and PESQ metrics with respect to the car noise level is depicted in fig. 6. The assessment of RMSE metric is portrayed in fig. 6a). The RMSE measured by the proposed FATFA-based DRN is 30.6468. Likewise the RMSE obtained by the existing methods like, DNN, Conditional GAN, CNN are 175.09, 274.55, and 323.59 for noise level 0.0200. The PESQ metric by varying the car noise level is portrayed in fig. 6b). By considering the noise level as 0.0050, the PESQ obtained by the existing DNN is 1.737, Conditional GAN is 2.016, CNN is 2.687, and the proposed FATFA-based DRN is 3.072, respectively.

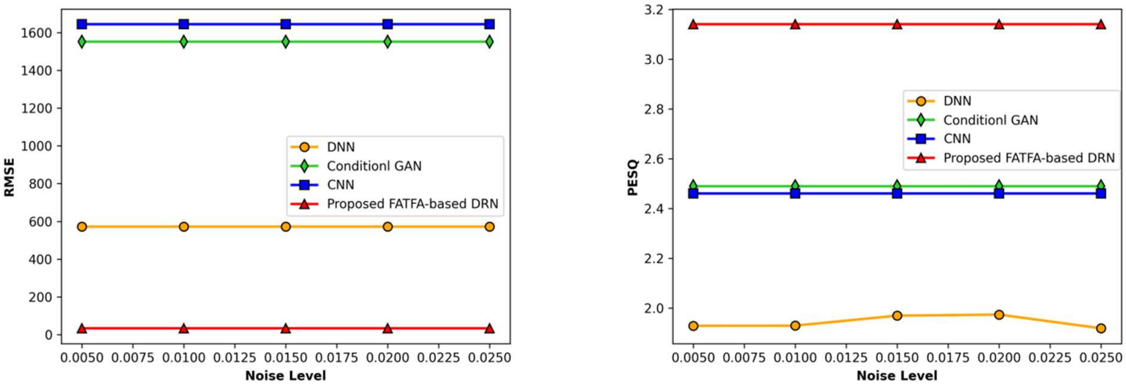


(a) (b)

Fig.6. *Analysis based on car noise a) RMSE b) PESQ*

d) Analysis using Train Noise

The analysis of RMSE and PESQ metric by considering the train noise is illustrated in fig. 7. Fig. 7a) illustrates the analysis of RMSE. The RMSE value of DNN, Conditional GAN, and CNN method is 571.68, 1551.98, 1643.64, whereas the developed FATFA-based DRN technique obtained the RMSE value of 33.387 for 0.0150 dB noise level. Fig. 7b) illustrates the assessment of PESQ. When the noise level is 0.0150, the PESQ measured by the proposed FATFA-based DRN is 3.140, whereas the PESQ obtained by the existing DNN, Conditional GAN, and CNN are 1.969, 2.489, 2.4601, respectively.

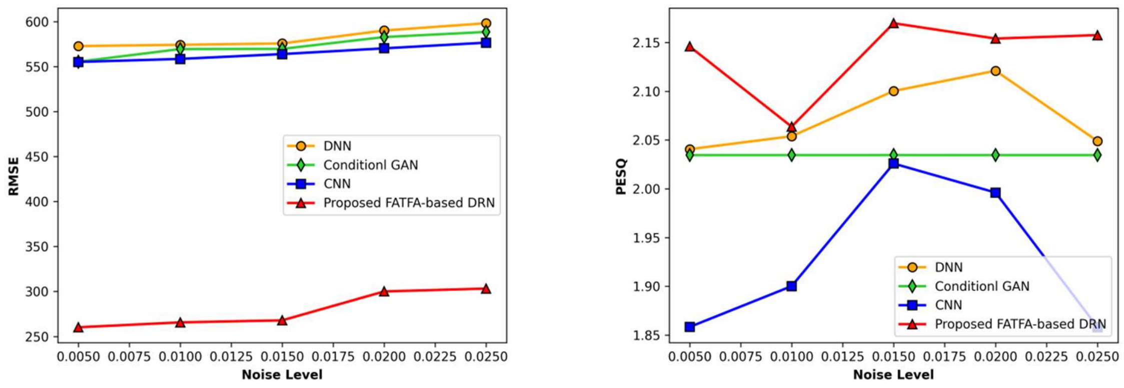


(a) (b)

Fig.7. *Assessment with respect to train noise a) RMSE b) PESQ*

e) Analysis using exhibition noise

Fig. 8 shows the assessment of RMSE and PESQ metric based on the exhibition noise. The analysis of RMSE is depicted in fig. 8a). By considering the noise level as 0.0100, the RMSE obtained by the existing DNN is 574.14, Conditional GAN is 569.42, CNN is 558.45, and the proposed FATFA-based DRN is, 265.67 respectively. The assessment of PESQ is shown in fig. 8b). When the noise level is 0.**0100** dB, the PESQ attained by DNN is 2.053, Conditional GAN is 2.034, CNN is 1.900 and developed FATFA-based DRN algorithm is 2.063, respectively.



(a) (b)

Fig.8. *Analysis based on exhibition noise a) RMSE b) PESQ*

4.7 Comparative Discussion

Table 1 illustrates the comparative analysis of the developFATFA-based DRN approach with the existing techniques, like DNN, conditional GAN, and CNN based on PESQ and RMSE metrics by changing the various noises, like airport noise, babble noise, car noise, train noise, and exhibition noise. Here, the PESQ value for DNN is 1.781, Conditional GAN is 3.015, CNN is 2.3169, and developed FATFA-based DRN technique is 3.907 in airport noise for 0.0250 dB noise level. When the noise level is 0.0250 dB in car noise, the RMSE value of DNN is 175.09, Conditional GAN is 274.55, CNN is 323.59 and developed FATFA-based DRN is 30.64. Thus, from the below table it is shown that the developed FATFA-based DRN method obtained high PESQ value of 3.907 using airport noise and less RMSEof 30.64 using the car noise.

Table 1: Comparative discussion

Methods /Noises	Metrics	DNN	Conditional GAN	CNN	Proposed FATFA-based DRN
Airport noise	*RMSE*	518.53	958.01	1093.51	**31.68**
	PESQ	1.781	3.015	2.3169	**3.907**
Babble noise	*RMSE*	2806.82	4345.83	5174.55	**38.18**
	PESQ	1.977	1.610	2.347	**3.13**
Car noise	RMSE	175.09	274.55	323.59	**30.64**
	PESQ	1.746	2.016	2.687	**3.07**
Train noise	RMSE	571.68	1551.98	1643.64	**33.38**
	PESQ	1.91	2.48	2.46	**3.14**
Exhibition noise	RMSE	598.12	588.44	576.52	**303.23**
	PESQ	2.04	2.03	1.85	**2.15**

5. Conclusion

In this research, an efficient strategy is devised using proposed FATFA-based DRN scheme for the pathological speech enhancement. The developed method effectively performed for enhancing the speech signals. At first, the pathological voice signal is considered as an input, and is subjected to the framing phase such that the process is performed using the Hanning window in order to extract the frames from the voice for further processing. Once the frames are extracted, the initial improvement process is carried out in order to improve the quality of the pathological voice signals. Moreover, the improved voice signals are utilized for the pathological speech enhancement process. Finally, the pathological speech enhancement is process is carried out using the DRN classifier, and it is trained by the developed optimization algorithm, called as FATFA. However, the performance of the developed FATFA-based DRN is evaluated using the twoperformance metrics, such as RMSE, and PESQ, and also achieved the higher PESQ of 3.907 and lesser RMSE value of 30.64 for the varying noise levels. The future work would be the concern of designing a novel optimization algorithm in order to improve the performance of the method.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Gelderblom FB, Tronstad TV, Viggen EM., "Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.27, no.3, pp.583-94, November 2018.

[2] Ram RS, Kumar MV, Subramanian B, Bacanin N, Zivkovic M, Strumberger I., "Speech Enhancement through improvised Conditional Generative Adversarial Networks", Microprocessors and Microsystems, pp.103281, September 2020.

[3] Huang Q, Bao C, Wang X, Xiang Y., "Speech enhancement method based on multi-band excitation model", Applied Acoustics, vol.1, no.163, pp.107236, June 2020.

[4] Sekiguchi K, Bando Y, Nugraha AA, Yoshii K, Kawahara T., "Semi-supervised multichannel speech enhancement with a deep speech prior", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.27, no.12, pp.2197-212, October 2019.

[5] Mohammed MA, Abdulkareem KH, Mostafa SA, Ghani MK, Maashi MS, Garcia-Zapirain B, Oleagordia I, Alhakami H, AL-Dhief FT, "Voice Pathology Detection and Classification Using Convolutional Neural Network Model", Applied Sciences, vol.10, no.11, pp.3723, January 2020.

[6] Kadiri SR, Alku P, "Analysis and detection of pathological voice using glottal source features", IEEE Journal of Selected Topics in Signal Processing, vol.14, no.2, pp.367-79, December 2019.

[7] Chen L, Chen J., "Deep Neural Network for Automatic Classification of Pathological Voice Signals", Journal of Voice, July 2020.

[8] Hammami I, Salhi L, Labidi S, "Voice pathologies classification and detection using EMD-DWT analysis based on higher order statistic features", IRBM, January 2020.

[9] Li QQ, He ZC, Li E., "The feedback artificial tree (FAT) algorithm", Soft Computing, vol.14, pp.1-28, February 2020.

[10] Arora, S. and Singh, S., "The firefly optimization algorithm: convergence analysis and parameter selection", International Journal of Computer Applications, vol.69, no.3,2013.

[11] Fang SH, Tsao Y, Hsiao MJ, Chen JY, Lai YH, Lin FC, Wang CT., "Detection of pathological voice using cepstrum vectors: A deep learning approach", Journal of Voice, vol.33, no.5, pp.634-41, September 2019.

[12] Muhammad G, Melhem M., "Pathological voice detection and binary classification using MPEG-7 audio features", Biomedical Signal Processing and Control, vol.11, pp.1-9, May 2014.

[13] Delvaux V, Pillot-Loiseau C., "Perceptual judgment of voice quality in nondysphonic French speakers: effect of task-speaker-and listener-related variables", Journal of Voice, vol.34, no.5, pp.682-93, September 2020.

[14] Hegde S, Shetty S, Rai S, Dodderi T., "A survey on machine learning approaches for automatic detection of voice disorders", Journal of Voice, vol.33, no.6, pp.947-e11, November 2019.

[15] Aichinger P, Pernkopf F, Schoentgen J., "Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices", Biomedical signal processing and control, vol.50, pp.158-67, April 2019.

[16] Al-Nasheri A, Muhammad G, Alsulaiman M, Ali Z, Malki KH, Mesallam TA, Ibrahim MF, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions", IEEE Access, vol.6, pp.6961-74, April 2017.

[17] J. D. Arias-Londo˜no, J. I. Godino-Llorente, N. S´aenz-Lech´on, V. Osma- Ruiz, and G. Castellanos-Dom´ınguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel cepstral coefficients", IEEE Transactions on Biomedical Engineering, vol.58, no.2, pp.370–379, 2011.

[18] J. A. G. Garc´ıa, L. Moro-Vel´azquez, and J. I. Godino-Llorente, "Onthe design of automatic voice condition analysis systems. part I: reviewof concepts and an insight to the state of the art", Biomedical Signal Processing and Control, vol.51, pp.181-19,2019.

[19] Loizou PC., "Speech enhancement: theory and practice", CRC press, February 2013.

[20] Wang ZQ, Wang D,"A joint training framework for robust automatic speech recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.24, no.4, pp.796-806, February 2016.

[21] Kinoshita K, Delcroix M, Gannot S, Habets EA, Haeb-Umbach R, Kellermann W, Leutnant V, Maas R, Nakatani T, Raj B, Sehr A., "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research", EURASIP Journal on Advances in Signal Processing, no.1, pp.7, December 2016.

[22] Xu Y, Du J, Dai LR, Lee CH, "An experimental study on speech enhancement based on deep neural networks", IEEE Signal processing letters, vol.21, no.1, pp.65-8, November 2013.

[23] Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", IEEE Signal processing magazine, vol.29, no.6, pp.82-97, October 2012.

[24]Lo'ai AT, Mehmood R, Benkhlifa E, Song H., "Mobile cloud computing model and big data analysis for healthcare applications", IEEE Access, vol.4, pp.6171-80, September 2016.

[25] Qiao Z, Zhang Z, Pan X, Epel B, Redler G, Xia D, Halpern H., "Optimization-based image reconstruction from sparsely sampled data in electron paramagnetic resonance imaging", Journal of Magnetic Resonance, vol.294, pp.24-34, September 2018.

[26] Zorrilla AM, Zapirain BG, Izquierdo AP, "Computer aided tool for diagnosis of ENT pathologies using digital signal processing of speech and stroboscopic images", SpringerPlus, vol.1, no.1, pp.64, December 2012.

[27] F. Chollet, "Deep Learning with Python", Shelter Island, New York: Manning Publications Co, 2017.

[28] Saarbruecken Voice Database taken from, "http://www.stimmdatenbank.coli.uni-saarland.de./index.php4#target", accessed on October 2020.

[29]Haridas, A.V., Marimuthu, R. and Chakraborty, B., "A novel approach to improve the speech intelligibility using fractional delta-amplitude modulation spectrogram", Cybernetics and Systems, vol.49, no.7-8, pp.421-451, 2018.

[30]Kamath, S. and Loizou, P., "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", In ICASSP, vol.4, pp.44164-44164, May 2002.

[31] Chen, Z., Chen, Y., Wu, L., Cheng, S. and Lin, P., "Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions", Energy Conversion and Management, vol.198, pp.111793, 2019.

[32] Costantini, G., Cesarini, V., Di Leo, P., Amato, F., Suppa, A., Asci, F., Pisani, A., Calculli, A. and Saggio, G., "Artificial Intelligence-Based Voice Assessment of Patients with Parkinson's Disease Off and On Treatment: Machine vs. Deep-Learning Comparison", Sensors, Vol. 23(4), pp.2293, 2023.

[33] Liu, C., Cheng, S., Ding, W. and Arcucci, R., "Spectral Cross-Domain Neural Network with Soft-adaptive Threshold Spectral Enhancement", arXiv preprint arXiv:2301.10171, 2023.

[34] Ksibi, A., Hakami, N.A., Alturki, N., Asiri, M.M., Zakariah, M. and Ayadi, M., "Voice Pathology Detection Using a Two-Level Classifier Based on Combined CNN–RNN Architecture", Sustainability, Vol. 15(4), pp.3204, 2023.

[35] Kim, A.Y., Jang, E.H., Lee, S.H., Choi, K.Y., Park, J.G. and Shin, H.C., "Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach", Journal of Medical Internet Research, Vol. 25, pp. e34474, 2023.

[36] Weise, T., Klumpp, P., Maier, A., Noeth, E., Heismann, B., Schuster, M. and Yang, S.H., "Disentangled Latent Speech Representation for Automatic Pathological Intelligibility Assessment", arXiv preprint arXiv:2204.04016, 2022.

[37] Yang, M., Konan, J., Bick, D., Kumar, A., Watanabe, S. and Raj, B., "Improving Speech Enhancement through Fine-Grained Speech Characteristics", arXiv preprint arXiv:2207.00237, 2022.

[38] Miller, G.F., Vásquez-Correa, J.C., Orozco-Arroyave, J.R. and Nöth, E., "Representation Learning Strategies to Model Pathological Speech: Effect of Multiple Spectral Resolutions", arXiv preprint arXiv:2209.08379, 2022.

[39] Zakariah, M., Ajmi Alothaibi, Y., Guo, Y., Tran-Trung, K. and Elahi, M.M., "An Analytical Study of Speech Pathology Detection Based on MFCC and Deep Neural Networks", Computational and Mathematical Methods in Medicine, 2022.