Radial Basis Function Neural Network Based Speech Enhancement System Using SLANTLET Transform Through Hybrid Vector Wiener Filter

V. R. Balaji, J. Sathiya Priya, J. R. Dinesh Kumar, and S. P. Karthi

Abstract In real environment, humans have the difficulty to recognize the voice signal in noisy situations. Recently, many speech enhancement methods, based on transforms, provide substantial success over the conventional method. In this paper, deep learning along with conventional transforms is introduced for enhancing the speech signal. The deep learning approach is used to train the speech frames for classifying the signal either as voiced or unvoiced, followed by pitch synchronous analysis. A proper windowing technique is used along with the hybrid vector wiener filter. However existing algorithms purely depends on supervised learning which tries to reduce the Mean Square Error. The Minimum mean square error can be reduced by means of comparing the output signal with an appropriate clean speech signal. Processing delay and reliable architecture are the best characteristics of deep learning speech enhancement algorithms. Radial basis function reduces the noise coefficient which in turn increases the speech quality.

Keywords Speech enhancement · Deep learning · RBFNN · Slantlet transform

1 Introduction

In day to day life, the speech signal gets added with the noisy signal, which is inevitable. This unwanted noise is called as background noise which comes from various sound sources like speakers, environment noise, etc. This will lead to the degradation of both speech intelligibility and quality. Understanding noisy speech is tough for normal persons but it will be challenging for hearing-impaired listeners. This leads to loss of effective communication among people. Also, devices which are purely based on speech such as speech recognition devices fail poorly, due to adverse noise conditions. So, speech is one of the attracted research areas in the past

V. R. Balaji (✉) · J. Sathiya Priya · J. R. Dinesh Kumar · S. P. Karthi
Department of Electronics & Communication Engineering, Sri Krishna College of Engineering & Technology, Coimbatore, Tamil Nadu, India
e-mail: balajivr@skcet.ac.in

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 145,
https://doi.org/10.1007/978-981-15-7345-3_61

decade. In recent years, enhancement techniques gradually shifted from conventional methods to newer methods such as deep learning. In Deep learning techniques, the adaptiveness is higher, when compared to signal processing methods. The adaptive nature comes from deep learning by means of training a specified algorithm for certain duration and iterations can be done till it gets converged with the target. [1]. In recent studies, various deep learning algorithms have been used for speech enhancement methods. Promised performance improvements can be achieved by a deep learning approach over the signal processing method. The supervised task involved in deep learning, trains the input signals so that it is easier to process the signals. For enhancing the signal, nonlinear mapping of the clean and noisy signal is performed. The ideal ratio mask concept was also used for speech enhancement for extracting the enhanced speech from noisy speech.

A phase-sensitive mask approach is used to provide good signal to distortion ratio value by finding the phase difference to obtain better performance. In some cases, various ratio mask is used in deep learning for speech enhancement. Many approaches are used to reduce the MSE between the output and training target. If the value of the MSE signal is zero, the iteration will get over and the target will be reached. But the value of MSE may not be zero since the residual is high based on the SNR value. If the SNR value is low, the MSE value will be high. So MSE value cannot be taken as a criterion for enhancing the speech quality. The other possibility to enhance speech is Short Time Objective Intelligibility measure (STOI) which is based on the subject's observation. This metric shows accuracy in terms of measuring intelligibility. Perceptual metrics such as perceptual evaluation of speech quality and perceptual evaluation methods are also used for speech enhancement. These metrics also can be combined with Reinforcement Learning (RL) for enhancing speech. [2]. This research work focuses on slantlet transform combined with deep learning for speech enhancement . The problem in processing speech is differentiated silence sounds and speech signals. Slantlet transform does the processing by taking the speech signals as frames and deep learning is used in depth into the signal for analyzing the coefficients. To further improve efficiency, a distinct filter is used to reduce the noise. In this transform, the speech frames are overlapped between ranges of 50% and 75%. SLT minimizes artifacts even though the speech signals get overlapped. So it is combined with pitch synchronous analysis for speech enhancement [3]. In the deep learning technique, large data sets are used for training with specified activation function for processing and restoration. The pitch synchronous analysis combined with the Maximum Alignment technique also helps in selecting the proper window. The Hanning window filters signal coefficients for further processing.

2 Related Works

2.1 Slantlet Transform

Slantlet Transform (SLT) is similar to DWT. But the advantage is better time localization available in this transform. DWT is constructed as a tree structure by having iterated banks but Slantlet transform will have a parallel structure with parallel branches. The input speech data is applied to filter structures as shown and then it is downsampled using the factor of four to obtain coefficients. The coefficients are the then threshold by a selective factor. To reconstruct the data, the inverse transform has to be applied based on a threshold value. SLT has a higher capacity to secure the speech data and this enhances the value. So SLT based transform is suitable choice for enhancing the speech signals. For reducing the noise further, hybrid vector wiener filtering is used.

2.2 Slantlet Transform Based Speech Enhancement

The proposed method is illustrated in Fig. 1. The degraded signal is segmented into speech frames. Then filtering is done with the help of a suitable filter. With the help of speech frames voiced or unvoiced decision is made for further processing of the signal [4]. Based on this decision, the time shift will be either a fixed one or shifted by one type period. One pitch period will be shifted if the frame is a voiced signal or it will be retained as such. Here the window has to be adapted based on the speech properties. The window should be flexible and must not be fixed [5] (Fig. 2).

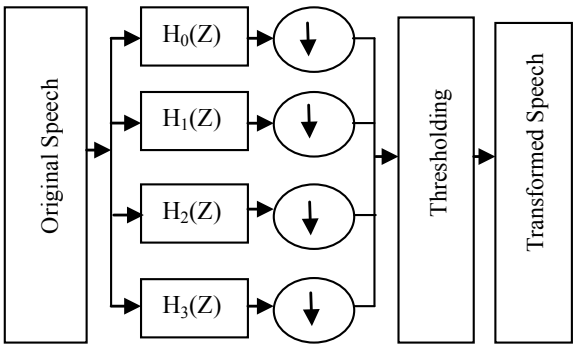


Fig. 1 Two-level SLT based data transformation

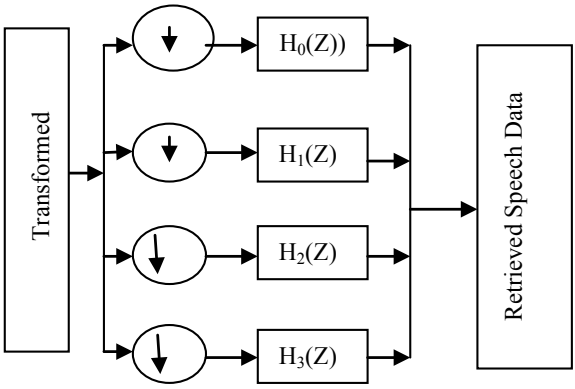


Fig. 2 Two-level SLT based reconstruction scheme

2.3 Pitch Synchronisation

For the implementation of SLT, the extraction of pitch period should be done. But, most of the algorithms can be able to obtain the pitch period only in clean situations. So the noise reduction has to be done initially by filtering Wiener filtered speech $\hat{\xi}_{m,k}$ can be given by:

$$\hat{S}_{m,k} = \frac{\hat{\xi}_{m,k}}{\hat{\xi}_{m,k} + 1} Y_{m,k} \tag{1}$$

where SNR $\hat{\xi}_{m,k}$ is the estimated apriori signal. Once the noise reduction is done, enhanced speech can be calculated by performing an inverse slantlet transform. One of the simple methods is the autocorrelation method due to robustness in a noisy environment. For performing time shifting, the extraction of the pitch period is done followed by clipping. The level of clipping is done by taking the samples from the speech frame. The autocorrelation function of the output signal $\hat{s}(n)$ is:

$$R(n) = \sum_{m=0}^{N-m-1} \hat{s}(m)\hat{s}(n+m) \tag{2}$$

The presence of a voiced signal can be found by the presence of peak magnitude. The absence of peak magnitude implies the absence of a voiced signal [6]. The pitch period determines the length of the window followed by window analysis.

The window length is to be higher for longer pitch period. The weighting function has to be calculated based on the current and the previous frames in real time. The weighting function is to be optimized for getting the enhanced speech signal [7]. The maximum alignment technique is used further to improve the pitch synchronization. Based on the SLT coefficients, the window shift is to be realigned and the signal

can be represented using different frequencies and amplitudes. The coefficients are reordered with the help of basis functions [8].

3 Methodology

3.1 Deep Learning

Machine learning is one of the emerging fields and Deep learning is one among them. It purely depends on algorithms that can be trained for certain iterations. The set of neurons is used for this training purpose. The input signal is sent to one set of neurons and the output can be obtained from the other set. The input signal gets modified in the second layer whereas the first layer receives the actual input. Based on the application, the number of layers between input and output gets varied and parallel processing is done to get the exact output. The accuracy of the output depends on the number of iteration. It is used in applications such as speech processing, speech recognition, and various recognition techniques [9]. Various architectures used in deep learning are Generative deep architectures, Discriminative deep architectures and Hybrid deep architectures. Among these complex deep learning architectures, the most commonly used are Deep feed-forward networks, Convolution networks, and Recurrent Networks [10].The main aim of an RBFNN is used for approximation of a function. In a classifier, $y = f*(x)$ maps an input x to a category y. The mapping is done using $y = f\ (x; \theta)$. It updates the parameter value θ which gives the best approximation [11]. For analyzing the speech signal 15 ms frame is used. Based on voiced speech or unvoiced speech, the speech frame can be taken either as long or short frame [12] (Figs. 3 and 4).

3.2 Radial Basis Forward Neural Network

RBFNN uses radial basis functions as an activation function for an artificial neural network. The RBFNN structure has 3 layers that uses feed forward technique.

1. Input Layer: Input signal is given which is linear.
2. Second layer: This layer is Non-linear and uses Gaussian functions.
3. Final layer: Gaussian outputs are combined.

During training, the tap weights are updated in the second layer and third layer. RBFNN is best suited for optimization. It can be done with the help of 5 parameters such as

1. Weights in the Second and Final layer.
2. Activation function.
3. Center of activation functions.

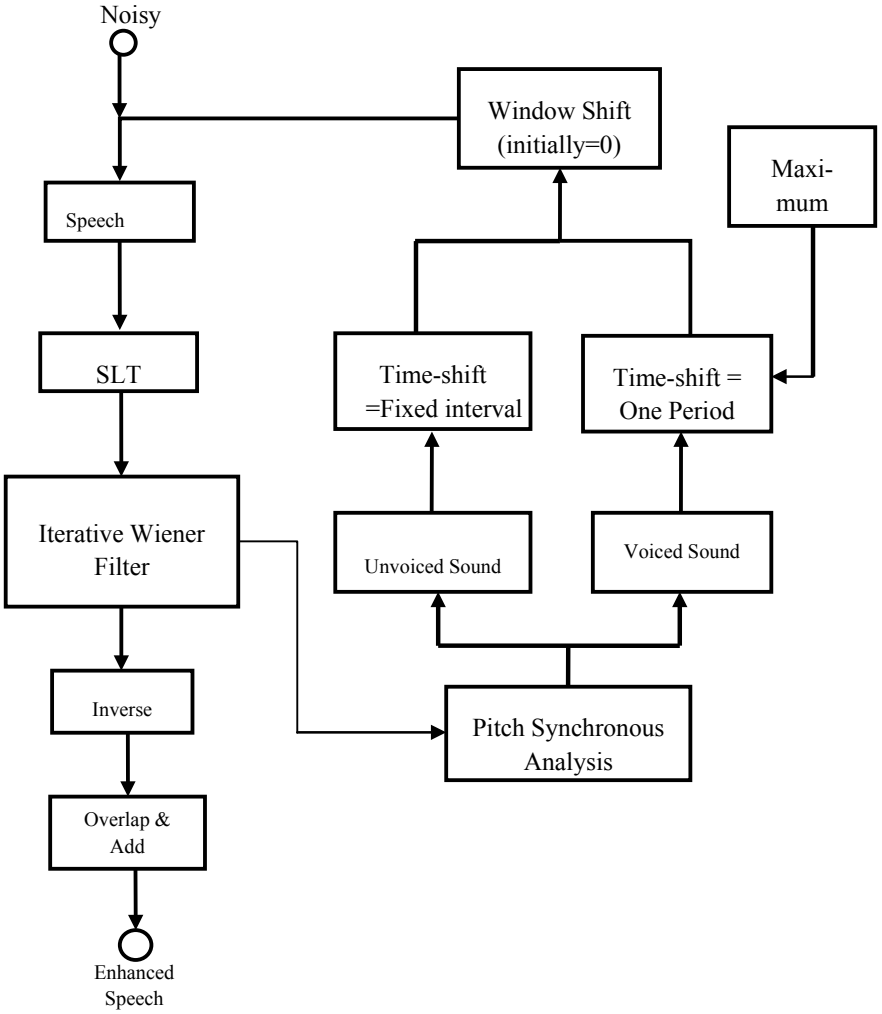


Fig. 3 Proposed block diagram for speech enhancement

4. Distribution of the center of activation functions.
5. The number of hidden neurons.

The network training starts by determining the weights between the second layer and final layer. This algorithm has many advantages when compared to conventional algorithms to overcome several issues. Since the training duration in this method is reduced and has the feature of generalization capacity it is applicable for real-time cases [13]. The activation functions will have similar characteristics based on center and distribution functions. The Gaussian coordinates are determined by means of the K-means clustering algorithm. The approximation of a function can be done with the

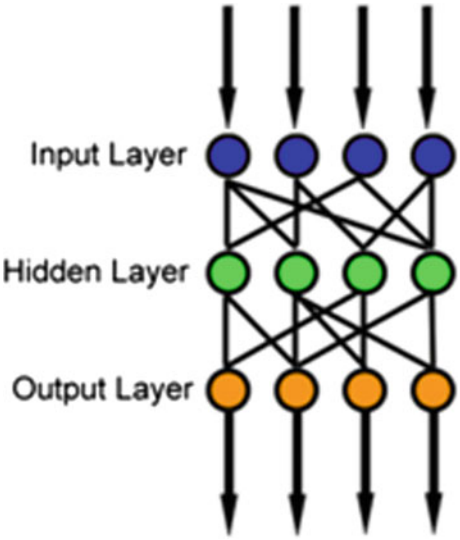


Fig. 4 Feed-forward neural network

help of the hidden layer [14]. The accuracy can be achieved by adding more number of hidden layers. RBFNN is popularly used in signal processing applications even though it has low classification speed. The optimization is done by means of faster training.

3.3 Window Function

For manipulating signals, the window function is needed for analysis of certain duration in a continuous signal. In most signal processing applications, a rectangular window is used for analysis. But the problem in using the rectangular window is the spectral leakage effect. If two signals have the same frequency, the overlapping of spectral leakage will happen. But if there are two signals with varied frequencies, one signal will obstruct the other signal. While selecting the window, side lobes play an important role [5]. A rectangular window with strong side lobes will be selected. Slantlet transform has some disadvantages when using a rectangular window because of having discontinuities at the endpoints. Due to discontinuities, the coefficient's value may get changed. This leads to finding a new windowing technique for speech application. For audio coding applications, a sine window is used to provide better attenuation. Similarly, a rectangular window is used when the signal is having high strength due to the narrow main-lobe. So while selecting the window, there should be a mutual relationship between spectral resolution and leakage effect. Hann window is very popular since it can avoid spectral leakage.

3.4 Hybrid Vector Wiener Filter

An optimal filter is needed to handle the real-time coefficients in Mean square error logic. It is based on a priori SNR for wiener filter implementation. Various ways can be used to calculate the value. Decision directed approach is one way to calculate the values. With respect to multiple noisy signals, observations are done over a continuous signal [15]. There is a need to recover the original speech signal from the continuous speech frames [16]. It may also lead to inaccuracy if the algorithms are not able to separate the discrete signals. In that case, the signal has to be modeled as a continuous time signal. The extracted samples are used for calculating minimum Mean square error [17].This wiener filter helps in noise reduction and also to reduce the Minimum Square Error of the speech signal. The extracted signal which is noise free is processed by Time Domain Pitch Synchronous Overlap-Add method. With the help of pitch synchronization, the processing is done by dividing the speech into voiced or unvoiced based on the pitch period [3].

4 Results and Discussion

The proposed system is evaluated using objective measures such as STOI (Short Time Objective Intelligibility), Segmental SNR (SegSNR), Perceptual Evaluation of Speech Quality (PESQ) [18]. Ten different segments of speeches (half females and half males), are randomly chosen from the TIMIT database. They are resampled at 8 kHz and corrupted by three additive noise types including white noise, fan noise and car noise. The total speech duration of all these test speech segments is 313.998 s including the silence period. The speech segments may contain voiced and silence period. It is assumed 50% are voiced speech [19].The proposed RBFNN based technique along with SLT transform was evaluated with the above-said parameters for two different windows hanning and rectangular [4]. The three different noise types are evaluated with the two windows and rectangular window produced better results under various SNR than hann window, since it has stronger side lobes. Since SegSNR is similar to the opinion score of subjects, it is used to determine the enhanced speech values. The Segmental SNR is calculated for various noise types and it is trained with RBFNN. The values are compared before and after the training with the network under different SNR values. The value of segmental SNR is based on input SNR values [8]. It will predict the quality of enhanced speech since each speech frame is fixed to a threshold level. An objective measurement tool that is used to measure the quality of speech is PESQ which is described in the ITU-T recommendation. It is purely based on listening tests by subjects. The subjects are made to observe both the signals, processed and noisy speech frames. Based on the observation, they provide the score as high, low or medium [15]. By taking the average of that score the mean opinion score can be calculated. So the two parameters are used to evaluate the speech intelligibility [20]. SegSNR is an accurate evaluation of speech quality,

whereas the PESQ provides accurate values based on speech distortion. PESQ is a better measure when compared to conventional methods since it is highly reliable. Before verifying the SLT based speech enhancement method, the window functions are to be compared. The window length is predefined as 32 ms. SegSNR results are shown in Fig. 5. It clearly shows that the rectangular window is better than Hann window under various SNR.

To illustrate the benefits of the proposed method, the segmental SNR results are compared before and after training with the help of Radial basis function neural network. The results are better after RBFNN training. Based on the noise type, there is a slight variation in the values of Seg SNR. Table 1 lists the comparison of ΔSegSNR results. Various noise types are taken for comparison and the input SNR is 0, 5, 10 and 15 [9]. The results indicate the proposed method gives better Δ SegSNR after training with RBFNN (Figs. 6 and 7).

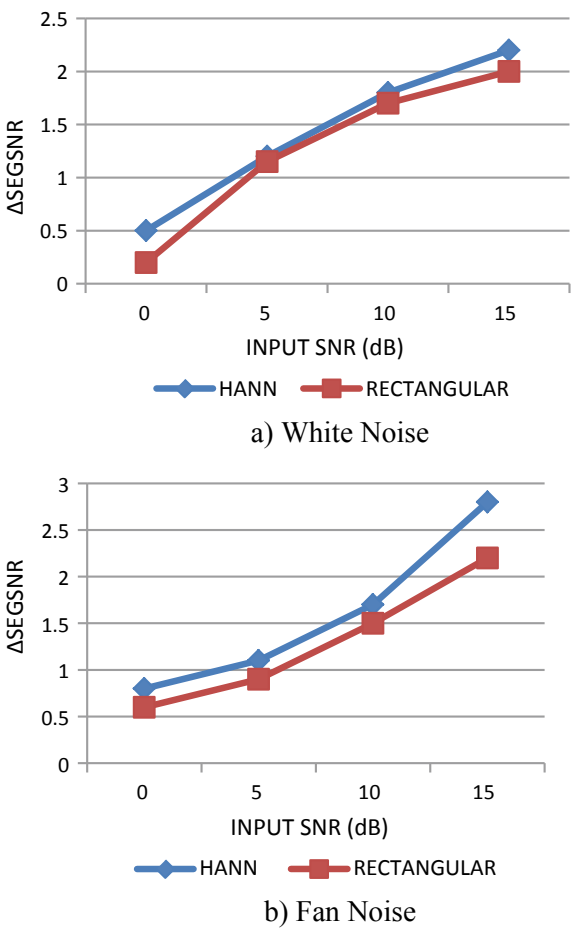


Fig. 5 Comparison of SegSNR results for various noise sources

Table 1 Δ SegSNR results comparison

Noise type	SNR (dB)	SLT using RBFNN—Before training	SLT using RBFNN—after training
White noise	0	6.23	6.01
	5	5.33	5.13
	10	4.92	4.63
	15	3.67	3.28
Fan noise	0	9.67	9.41
	5	9.33	9.13
	10	8.92	8.61
	15	8.22	8.01
Car noise	0	12.66	12.36
	5	12.01	11.99
	10	11.63	11.12
	15	10.22	9.83

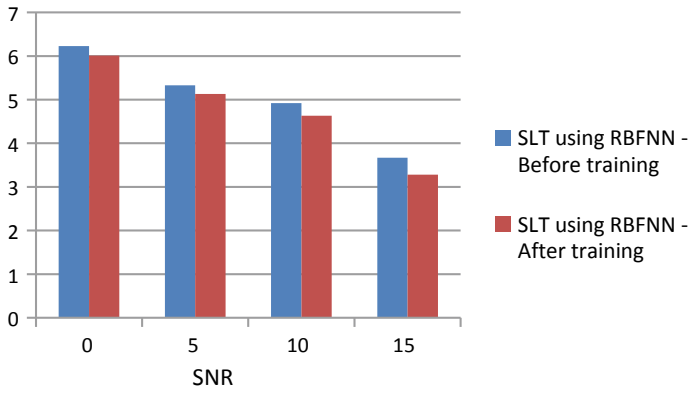


Fig. 6 Comparison of ΔSEGSNR results for white noise

Table 2 shows the comparison of PESQ and STOI measures for white noise and car noise. It also shows that RBFNN provides better results in terms of PESQ and STOI scores (Figs. 8 and 9).

5 Conclusion

This research paper focuses on using RBFNN for improving the speech intelligibility of the signal. Conventional signal processing methods provide speech enhancement with the help of transforms. Here the Slantlet transform is used along with RBFNN

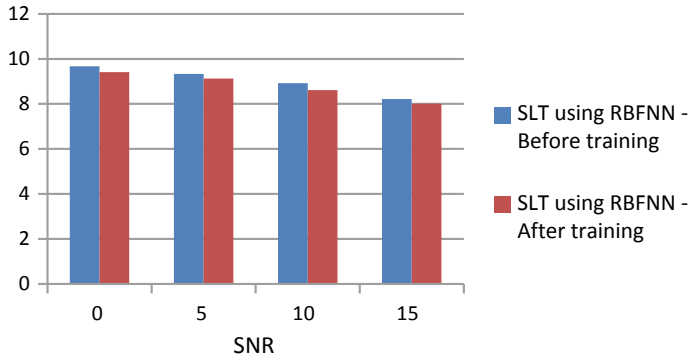
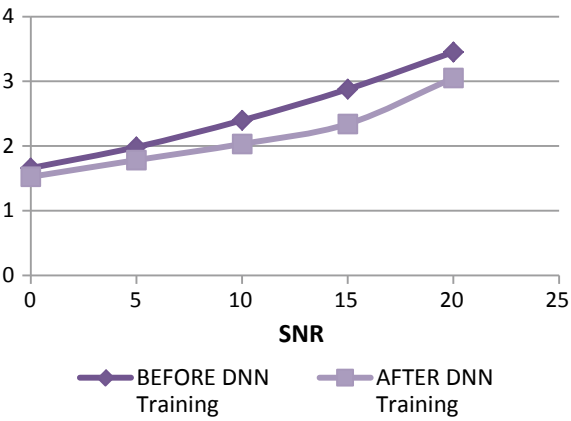


Fig. 7 Comparison of ΔSEGSNR results for fan noise

Table 2 Comparison of PESQ and STOI score for white noise and car noise

SNR	Before DNN training		After DNN training		Before DNN Training		After DNN training	
	White noise				Car noise			
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
20	3.452	0.988	3.052	0.962	4.034	0.963	3.932	0.982
15	2.879	0.921	2.341	0.9	3.634	0.912	3.42	0.943
10	2.397	0.897	2.032	0.812	2.317	0.845	2.124	0.821
5	1.984	0.765	1.782	0.721	1.764	0.723	1.475	0.711
0	1.659	0.526	1.524	0.498	1.329	0.513	1.299	0.467

Fig. 8 Comparison of PESQ score



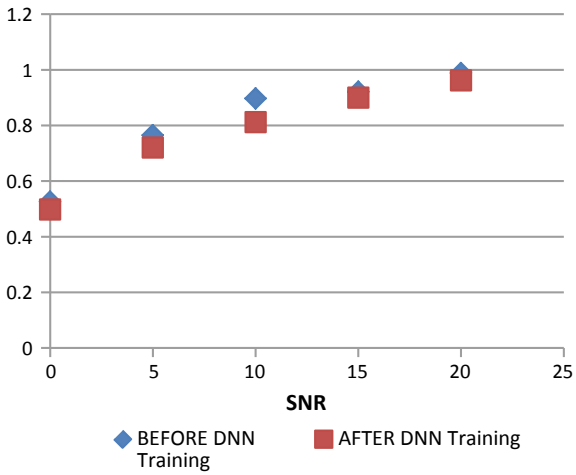


Fig. 9 Comparison of STOI score

which helps to improve the accuracy since it performs training till the convergence is obtained. In this method, the signal taken from the TIMIT database is split into overlapping frames which produces the coefficients. The coefficients are then given to a filter and trained using the algorithm for optimum results. Once the training is done, it is combined with pitch synchronous analysis to detect the silence period. Based on the pitch period, the amount of shift can be decided. The performance measure shows that radial basis function reduces the noise coefficients which in turn increases the speech quality. In the future, the activation function can be adjusted such that the training can be done in less time so that the convergence rate can be faster.

References

1. Le Roux J, Vincent E, Erdogan H (2016) Learning-based approaches to speech enhancement and separation. In: INTERSPEECH tutorials
2. Williamson DS, Wang Y, Wang DL (2016) Complex ratio masking for monaural speech separation. IEEE/ACM Trans Audio Speech Lang Process 24:483–492
3. Balaji VR (2018) A comparison of compression sensing algorithm and DUET algorithm for advanced DCT based speech enhancement system for vehicular noise. Int J Pure Appl Mathematics 119(12):1385–1394
4. Mirsamadi S, Tashev I (2016, May) A causal speech enhancement approach combining data-driven learning and suppression rule estimation. In: Proceedings of inter speech
5. Li Y, Kang S (2017) Deep neural network-based linear predictive parameter estimations for speech enhancement. IET Signal Process 11(4):469–476
6. Zhang X, Wang Z-Q, Wang DL (2017) A speech enhancement Algo Rithm by iterating single- and multi-microphone processing and its application to robust ASR. In: IEEE international conference on acoustics, speech and signal processing, pp 276–280
7. Erdogan H et al (2016) Wide residual blstm network with discriminative speaker adaptation for robust speech recognition. In: CHiME-4 workshop

8. Barker J, Marxer R, Vincent E, Watanabe S (2017) The third "CHiME" speech separation and recognition challenge: analysis and outcomes. Comput Speech Lang 46:605–626
9. Kounovsky T, Malek J (2016) Single channel speech enhancement using convolutional neural network. In: IEEE international workshop of electronics, control, measurement, signals and their application to mechatronics (ECMSM), pp 1–5, 666–676
10. Koizumi Y, Niwa K, Hioka Y, Kobayashi K, Haneda Y (2017) DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements. In IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 81–85
11. Balaji VR, Sathiya Priya J (2018) Enhancement of speech signal modified binary mask based algorithm for vehicular noise. J Adv Res Dyn Control Syst 10(12-Special Issue)
12. Nugraha AA, Liutkus A, Vincent E (2016) Multichannel audio source Separation with deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 24(9):1652–1664
13. Kolbaek M, Tan ZH, Jensen J (2017) Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. In: IEEE/ACM Trans Audio Speech Lang Process 25(1)
14. Balaji VR, Maheswaran S, Babu MR, Kowsigan M, Prabhu Venkatachalam K (2020) Combining statistical models using modified spectral subtraction method for embedded system. Microprocessors and Micro Systems, Elsevier
15. Dinesh Kumar JR, Ganesh Babu C, Balaji VR (2019) Analysis of effectiveness of power on refined numerical models of floating point arithmetic unit for biomedical applications. In: AIP scopus indexed proceedings of international conference on advances in materials processing and characterization. ICAMPC 2019
16. Balaji VR, Sathiya Priya J, Dinesh Kumar JR (2019) FPGA implementation of image acquisition in marine environment. Int J Oceans Oceanography 13(2):293–300. ISSN 0973-2667
17. Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey JR, Schuller B (2015) Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: International conference on latent variable analysis and signal separation, pp 91
18. Barker J, Marxer R, Vincent E, Watanabe S (2015) The third "CHiME" speech separation and recognition challenge: dataset, task and baselines. In:IEEE workshop on automatic speech recognition and understanding, pp 504–511
19. Goehring T, Bolner F, Monaghan J, Dijk B, Zarowski A, Bleeck S (2017) Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. Hear Res 344:183–194
20. Gannot S, Vincent E, Markovich-Golan S, Ozerov A (2017) A consolidated perspective on multi-microphone speech enhancement and source separation. IEEE/ACM Trans Audio Speech Lang Process 25:692–730