Soft Computing (2022) 26:13037–13047
https://doi.org/10.1007/s00500-021-06291-2

FOCUS

Enhancement of single channel speech quality and intelligibility in multiple noise conditions using wiener filter and deep CNN

D. Hepsiba[1,2] · Judith Justin[1]

Accepted: 15 September 2021 / Published online: 6 October 2021

Abstract
Nowadays, deep neural network has become the prime approach for enhancing speech signals as it yields good results compared to the traditional methods. This paper describes the transformation in the enhanced speech signal by applying the deep convolutional neural network (Deep CNN), which can model nonlinear relationships and compare it with the Wiener filtering method, which is the best technique for speech enhancement among the traditional methods. Denoising is performed in the frequency domain and converted back to the time domain to analyze performance metrics such as speech quality and speech intelligibility. The speech quality is analyzed based on the signal to noise ratio (SNR) and perceptual evaluation of speech quality (PESQ). Speech intelligibility is analyzed by short-time objective intelligibility (STOI). Both the methods evaluated the denoised speech, and the analysis made on the results shows that the SNR of the conventional Wiener filtering method is much improved when compared with Deep CNN. However, the PESQ and STOI of Deep CNN-based enhanced speech outperform the Wiener filtering method. The performance metrics indicate that Deep CNN achieves better results than the conventional technique.

Keywords Deep convolutional neural network · Noisy speech · Speech enhancement · Speech quality · Intelligibility

1 Introduction

Communication through speech is one of the vibrant methodologies to express one person's internal thoughts to another and from the human to machine and vice versa. The original quality of the speech signal becomes distorted as it is delivered into the outside world. Therefore, the speech signal mixed with noise needs to be enhanced. Consequently, speech enrichment needs to enhance the quality and legibility (Wang et al. 2021) of noisy speech signals. The need of the hour in our day-to-day life is the extraction of the clear speech signal from the distorted noisy speech which are prone to background noise and reverberations.

Speech signal enhancement is a tedious process compared to other signals because of its characteristic that changes intensely with time. The algorithms used for this process need to give a spontaneous action for different practical applications. The most common speech processing techniques for denoising that are used for enhancing the speech signal are minimum mean square error method (Schwerin and Paliwal 2014) that is performed by short-time spectral magnitude estimation between the clean speech signal and enhanced speech signal, spectral subtraction method (Paliwal et al. 2010) that deals with the clean speech spectrum estimation by subtraction of noise spectrum from noisy speech spectrum. Various filtering techniques like Wiener filter (Grais and Erdogan 2013) that acts as a linear estimator for reducing the mean squared error (MSE) between the clean speech and enhanced speech signal, and Kalman filter (Dionelis and Brookes 2018) estimates the model from observing a set of the noisy speech signal. These statistical-based (Hu and Loizou 2008; Loizou 2013) unsupervised models are imperfect in

Communicated by Joy Iong-Zong Chen.

✉ D. Hepsiba
hepsiba@karunya.edu

Judith Justin
hod_bmie@avinuty.ac.in

[1] Department of Biomedical Instrumentation Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India

[2] Department of Biomedical Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India

predicting the variations because of dynamic nature of noisy speech signals. The statistical assumptions need to be made in the unsupervised models and that does not improve the performance of the denoised speech. The supervised models are data driven and it eliminates the statistical assumptions that are made on the clean and noisy speech signals.

Nowadays, the enhancement techniques incorporate the taxonomy of artificial intelligence by which the machine learning (Srinivasan et al. 2006) and deep learning technique (Kolbk et al. 2017; Wang and Chen 2018; Chai et al. 2019) is widely applied to improve the clarity of speech (intelligibility) and it increases the listening capability (based on quality) so that it is perceived. It is imperative as the listeners are interested and focused on listening to the speech signal with excellent quality and intelligibility. Denoising is a fundamental strategy that is implemented for applications that deal with speech signals such as telecommunication (Rix et al. 2001), speaker recognition in biometrics (Jain et al. 2004), hearing aids (Healy et al. 2017), hands-free communication (Thiergart and Taseska 2014) and many more.

The drawbacks of the unsupervised techniques could be overcome by applying the deep neural network that deals with training the network with massive data in multiple noise conditions. The data-driven approach (Zhao et al. 2018) of the deep neural network makes it more efficient and is responsive to untrained conditions and unseen noises. In the recent past, the commonly used techniques for supervised speech enhancement (Nossier et al. 2021) technique include the mapping in the frequency domain or time–frequency masking. The speech signal is converted from the frequency domain to the time domain. These methodologies enable the reconstruction of the speech signal from frequency domain to time domain with the phase of the noisy signal (Li et al. 2019).

The order of the content of this research paper is as follows: the recent work carried out in speech enhancement is discussed in the 2nd Section. A clear explanation of the proposed Deep CNN system and a comparison with the Wiener filter is given in the 3rd Section. Section 4 discusses the dataset used, features extracted, algorithm and its description. The description of the results obtained and the conclusion are mentioned in the 5th and 6th Section, respectively.

2 Related works

Similar works carried out in the speech enhancement area helps in removing the background noise that affects the speech signal are the weighted noise encoder for enhancing the speech signal by considering the power spectrum of clean speech and the SNR to build the Wiener filter in the frequency domain (Xia and Bao 2014). Modeling of the time and frequency correlation dimensions by applying the improved minima controlled recursive averaging (IMCRA) and also incorporating the long short-term memory (LSTM) of recurrent neural network (RNN) architecture and CNN exhibits good results in terms of the performance metrics (Yuan 2020). Cycle consistent training (Meng et al. 2018) for enhancement optimizes clean to noisy and noisy to clean speech mapping simultaneously.

The different DNN-based speech enhancement methodologies adopted vary based on neural network architecture, training the target and selection of training features. Nowadays, the deep learning models that are becoming popular in the field of speech enhancement are the CNN (Zheng et al. 2020; Li et al. 2020), LSTM (Li et al. 2019), and RNN (Xian et al. 2021), which incorporate the transformation function to convert the spectral features of the noisy speech signal and clean speech signal. As CNN is widely used for image processing and recognition, it would be a good solution for the problems faced with the degradation of speech signals due to background noise. The SNR-aware (Fu et al. 2016) CNN for the enhancement process shows that the CNN suits well for extracting the time–frequency features and moves forward in achieving the goal. Loss functions based (Fu et al. 2018; Li et al. 2020) on the performance metric STOI are used for modeling the utterance as a whole.

CNN implemented to perform end-to-end speech enhancement (Du et al. 2017) task can estimate the phase of clean speech that improves the quality and intelligibility of speech. Some of the speech enhancement methods perform direct enhancement on the raw speech waveforms by mapping (Fu et al. 2017; Pandey and Wang 2019) and are referred to as the waveform-based approaches. The fully convolutional neural network (Park and Lee 2017) is one among them that allows direct mapping and feature selection from the convolutional encoder-decoder model (Lan et al. 2020). Obtaining the mean absolute error loss for the training of CNN is done by taking the magnitude of the enhanced STFT and clean STFT (Pandey and Wang 2019). In some cases, a combination of the CNN and RNN model (Hsieh et al. 2020) works out to be more suitable to capture the local and sequential correlations (Wang et al. 2021). Another approach uses sequence to sequence model (Kameoka et al. 2020) using LSTM RNN to model the encoder by encoding the input sequence and decoder to decode the output sequence for voice conversion.

The mapping function created based on the noisy and clean speech signal by the nonlinear-based regression model (Xu et al. 2013) shows that the ability to handle the unseen noise is diminished. In the ILMSAF-based speech enhancement, the performance of the network is reduced

for the volvo noise (Li et al. 2016; Sungheetha and Rajesh 2021; Kumar 2021). As the task is to enhance the speech signal by removing the noise, the CNN is applied for the speech enhancement as it was observed that it gives improved results compared to multi-layer perceptron (Grais and Plumbley 2017).

The CNN is robust and suits well for speech enhancement. Therefore, in the proposed work, the Deep CNN is designed to give outperforming results. Deep CNN takes the noisy speech signal as the input and converts it into the frequency domain to train the network. It is because the noise and the clean speech signal can be discriminated only in the frequency domain. The training is performed until the mean squared error is minimum between the clean speech signal and the denoised or enhanced speech signal.

3 Speech enhancement system

In today's scenario, the best of all techniques are the Deep algorithms, as they can handle a lot of data and design a model by themselves. In this work, the Deep CNN is designed to perform speech enhancement and a comparative study is done by analyzing its performance with the best conventional technique, i.e., the Wiener filter as shown in Fig. 1. Therefore, the best conventional Wiener filter and Deep CNN are taken for comparison. The comparison results show that each technique is best in its way.

3.1 Model of speech signal

The noisy speech signal is acquired from adding the clean speech signal with the different types of noise as given in Eq. 1. The task is to retrieve the clean speech signal from the noisy speech signal by eliminating the noise.

$c(n)$: Clean Speech Signal
$b(n)$: Noise Signal
$s(n)$: Noisy Speech Signal

$$s(n) = c(n) + b(n) \tag{1}$$

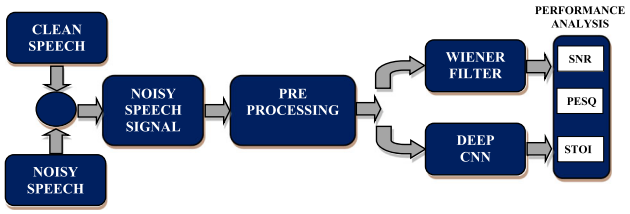


Fig. 1 Speech enhancement system for performance analysis

3.2 Wiener filtering

The presence of noise is unavoidable in real-world scenarios of speech processing. The most fundamental methodology in noise reduction of a speech signal is the optimal Wiener filter. The Wiener filter acts as a linear filter that could be utilized to separate the clean speech signal from the noisy speech signal by reducing the MSE between the estimated signal and the original signal. As the Wiener filter can achieve noise reduction, it also has the disadvantage of losing the speech signal's integrity. Therefore, the speech misrepresentation should be managed in such a way by adequately manipulating the Wiener filter or to have explicit knowledge of the speech signal. In any speech communication system, the speech signal could be distorted by background noise and reverberations. Therefore, noise reduction methodologies and speech enhancing techniques are needed to obtain the desired speech signal from the corrupted ones.

$$R(\omega) = \frac{C(\omega)}{S(\omega)} = \frac{C(\omega)}{C(\omega) + B(\omega)} \tag{2}$$

where $C(\omega)$—Signal Spectrum, $B(\omega)$—Noise Power Spectrum, $S(\omega)$—Noisy Speech Spectrum

$$R_{\text{Wiener}}(\omega) = \frac{C(\omega)}{S(\omega)} = \frac{S(\omega) - B(\omega)}{S(\omega)} \tag{3}$$

E_s—Estimation of enhanced signal

$$\widehat{E_s}(\omega, k) = R_{\text{Wiener}}(\omega) S(\omega, k) \tag{4}$$

$$\left|\hat{d}[n]\right| = \text{IFFT}\left[\sqrt{\widehat{E_s}(\omega, k)}\right] \tag{5}$$

By combining the magnitude of the clear speech spectral data with the phase of the noisy speech, the estimate of the enhanced speech is obtained. It is given as,

$$\left|\hat{d}[n]\right| = \left|\hat{d}[n]\right| \angle s[n] \tag{6}$$

$\hat{d}$—Estimate of Enhanced Speech.

3.3 Speech denoising and enhancement using deep convolutional neural network

Deep learning adopts the learning methodologies to create a model based on the data given to it. Neural network is the basic building block of deep learning. Speech enhancement is much required as the speech signal gets easily corrupted due to multiple noise conditions and noise levels. The noises can be stationary or nonstationary with varying acoustic characteristics. As the DNN can possess the model of highly nonlinear parameters, it makes the speech enhancement process simpler. The DNN architecture adopts the multi-layer feedforward network. The Deep

CNN is designed with multiple hidden layers with rectified linear unit (ReLU) activation function for speech enhancement. The input applied to the Deep CNN system is the frames of the noisy speech signal, and the expected output is the denoised speech signal.

The clean and noisy speech signal is converted to the frequency domain using STFT. The magnitude spectrum of the clean speech signal is taken as the target. The noisy speech signal is taken as the predictor and presented to the Deep CNN for denoising the speech as shown in Fig. 2. The regression network uses the magnitude of the noisy speech signal to reduce the mean square error between the denoised speech signal and the clean speech signal. The output from the Deep CNN gives the denoised signal in the frequency domain. The denoised speech signal is converted to the time domain using the output magnitude spectrum from the Deep CNN network and the phase of the noisy speech signal.

4 Algorithm description

Algorithm

- **Adding noises of various levels 0dB, 5dB, 10dB, 15dB to the clean speech**
- **Apply STFT to generate magnitude STFT vectors from clean and noisy speech signals**

$$X(w,n) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-jwm}$$

- **Extract the Magnitude STFT feature of clean and noisy speech**
- **Normalize the feature to zero mean and unity standard deviation**
- **Holdout validation for splitting training and testing data**
- **Apply Deep Convolution Neural Network**
- **Compare performance metrics SNR, PESQ and STOI of denoised speech with noisy speech**

4.1 Dataset

The clean speech signal is taken from the University of Edinburgh, Centre for Speech Technology Research (CSTR) (https://datashare.is.ed.ac.uk/handle/10283/2791). The dataset contains nearly 400 speech sentences. These speech sentences are taken for training and different types of noise are added with different decibels. The dataset is divided into training and testing data by applying holdout validation method. 80% of the dataset is taken as training data and 20% is taken as the testing data. The noisy speech signal is generated for feeding the Deep CNN. The noisy data set is created by mixing the clean speech with the different noise types such as washing machine noise, rainbow noise, jet airplane noise and train whistle noise with different noise levels such as 0 dB, 5 dB, 10 dB and 15 dB.

The dataset contains 400 utterances and it is split into 3:1 for training and testing. Deep CNN is trained with 300 sentences and tested with 100 sentences. The training set is created by mixing the noise with the clean speech signal at different noise levels. From the testing set, the noisy speech signal is randomly chosen to check the denoising ability of the network.

4.2 Feature extraction

The first step is to convert the speech signal from the time domain to the frequency domain using STFT to extract features. The magnitude STFT vectors of the clean speech and the noisy speech are input features to the Deep CNN Model. Therefore, the speech signal is divided into a 10 ms frame with no frameshift. In converting from the time domain to frequency domain using STFT, the hamming window is utilized with a window length of 256 samples and 75% overlap. For training and testing purposes, the speech signal is down sampled to an 8 kHz signal and a 256-point FFT is implemented and the number of frequency bins is 129. The clean speech corpus taken from the open-source dataset was contaminated by the noise signals at different noise levels.

The discrete Fourier transform is applied on the overlapped frames for acquiring the STFT of the signal. Due to the overlap, the successive frames cause the nearby frames to have common samples at the boundary of the overlap.

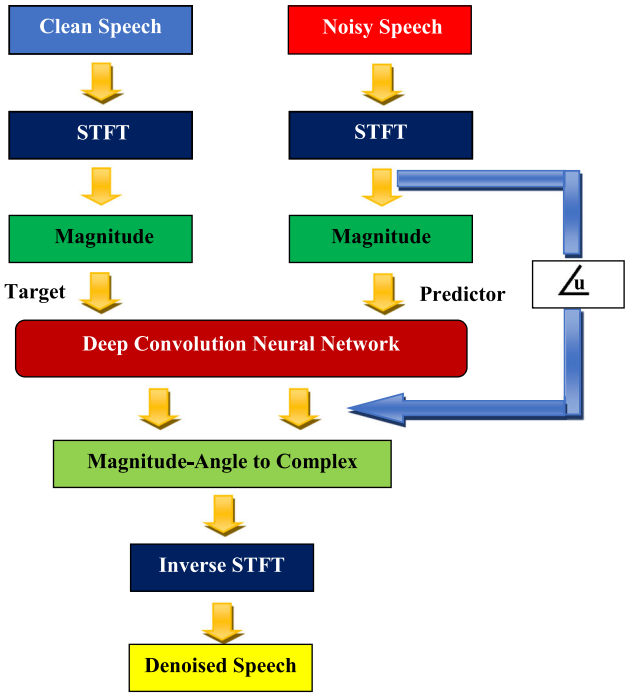


Fig. 2 Proposed deep CNN for speech enhancement

The relationship between STFT magnitude and the STFT phase is due to the correlation between the adjacent frames in the frequency domain. The original speech signal is reconstructed by maintaining a relation between the STFT magnitude and phase.

4.3 Denoising using convolutional layers

The denoising algorithm utilizes convolutional layers in which each neuron is connected to all the activations in the previous layer. Deep CNN is used to learn the spectral mapping from the noisy speech signal to the clean speech signal. The Deep CNN in this work is designed with 2-D convolutional layer and applies the sliding filter to the input as shown in Fig. 3.

The inputs to the convolutional layer are the features taken from the magnitude vector of STFT and the number of segments of the noisy speech signal. The convolution layer convolves by moving the filter on the input vertically and horizontally. The dot product is determined by the weights and the input and it is added to the bias.

The convolutional layers are defined as a group of layers, i.e., Convolutional Layer, Batch Normalization Layer and ReLu Layer and repeated 6 times, with the filter width of 9, 5 and 9 and the number of filters are 18, 30 and 8. The final convolutional layer is given a filter width of 129 along with 1 filter. The mean and standard deviation of outputs are normalized using the Batch Normalization Layers. The maximum epoch is set to 15; therefore, the network makes 15 passes through the training data. The shuffle is made for the training sequence at the starting of every epoch. During the training phase, the Adam Optimizer is used for optimizing the parameters and the MSE is taken as the loss function.

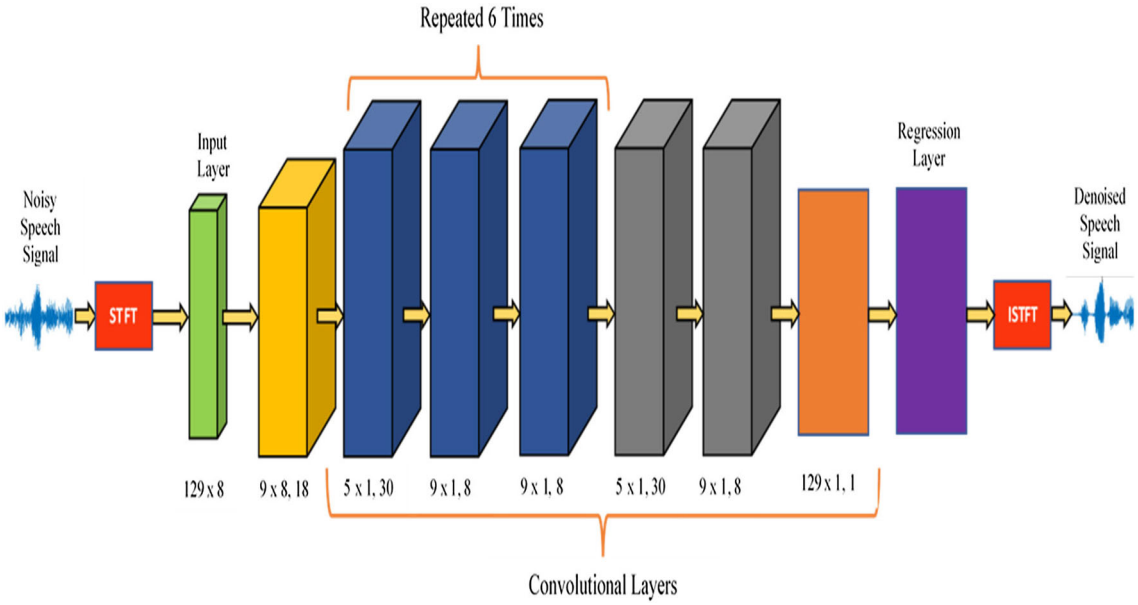


Fig. 3 Deep CNN architecture for denoising speech signal

Table 1 PESQ description

PESQ Score	Description
4–5	Excellent
3–4	Good
2–3	Fair
1–2	Poor
0–1	Bad

5 Results and discussions

The clean speech signal is added with different noise types such as washing machine noise, rainbow noise, train whistle noise and jet airplane noise with different noise levels such as 0 dB, 5 dB, 10 dB and 15 dB. The noisy speech signal generated by adding washing machine noise is given as input to the Wiener filter and DNN-based speech enhancement system. The SNR of the denoised signal is improved compared to the SNR of the noisy signal.

Fig. 4 Performance improvement comparison for noise levels 0 dB, 5 dB, 10 dB, 15 dB **a** washing machine noise, **b** rainbow noise, **c** train whistle noise, **d** jet airplane noise

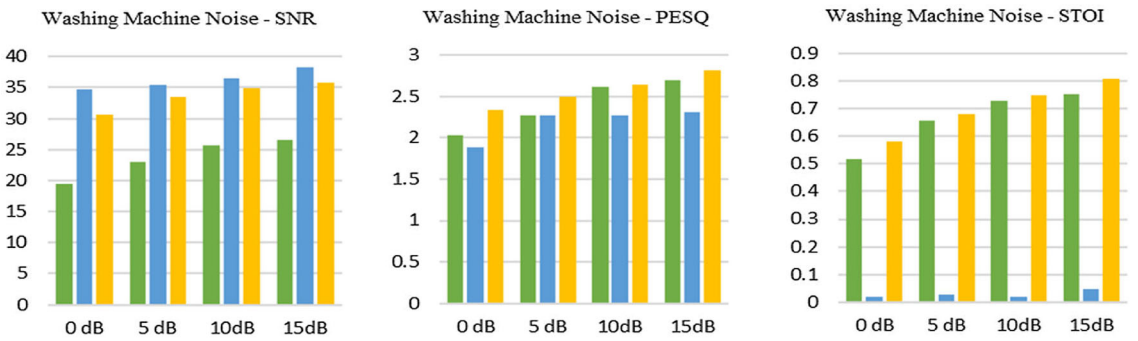
The noisy signals are taken for different noise levels such as 0 dB, 5 dB, 10 dB and 15 dB for the different noise types and were added with the clean speech signal to form the noisy speech signal. For analyzing the enhanced speech signal, the performance metrics considered are SNR, PESQ and STOI.

The performance metrics are calculated as follows:

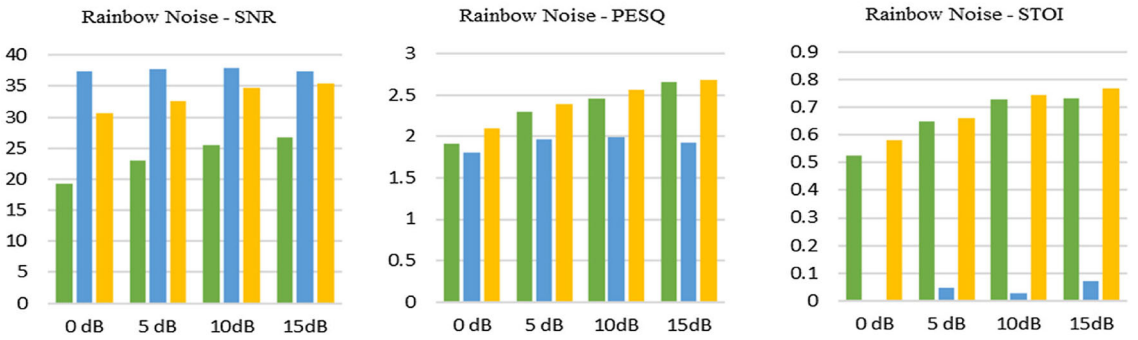
- Signal to Noise Ratio (SNR)

Table 2 Comparison of SNR, PESQ and STOI of noisy signal and denoised signal using Wiener filter and deep CNN

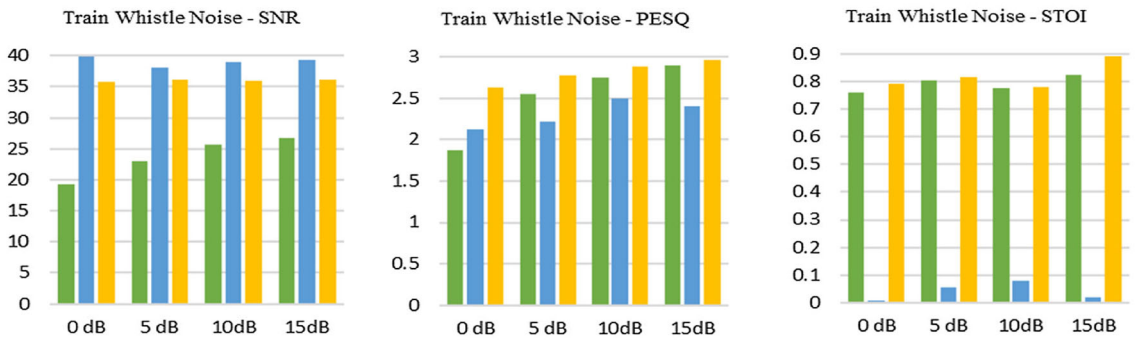
Noise level (dB)	Washing machine noise			Rainbow noise			Train whistle noise			Jet airplane noise		
	SNR			SNR			SNR			SNR		
	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN
0	19.4898	34.7837	30.5801	19.3679	37 .2929	30.5738	19.2925	39.801	35.7682	19.5242	36.4482	33.2925
5	23.0452	35.3452	33.4311	22.9899	37.7086	32.5593	23.0805	38.0964	36.0604	23.1936	38.9859	34.2398
10	25.5831	36.5445	34.8537	25.5454	37.9403	34.6552	25.6373	38.9824	36.0091	25.5068	44.103	35.1301
15	26.6226	38.3302	35.7433	26.7152	37.4407	35.5169	26.6514	39.346	36.1191	27.0348	42.8054	36.1123
Noise level (dB)	Washing machine			Rainbow			Train whistle			Jet airplane		
	PESQ			PESQ			PESQ			PESQ		
	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN
0	2.0374	1.8916	2.3326	1.91	1.812	2.0966	1.8663	2.1188	2.6372	2.2741	1.5465	2.3693
5	2.2703	2.2672	2.4992	2.2918	1.9707	2.3908	2.5498	2.2221	2.7768	2.5966	1.6103	2.6944
10	2.6184	2.2699	2.6496	2.4612	1.995	2.56	2.7473	2.4913	2.8795	2.6331	1.7719	2.8764
15	2.6983	2.3078	2.816	2.6623	1.9265	2.6776	2.9022	2.4074	2.9699	2.6785	1.8318	2.7983
Noise level (dB)	Washing machine			Rainbow			Train whistle			Jet airplane		
	STOI			STOI			STOI			STOI		
	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	wiener filter	Deep CNN
0	0.5166	0.0173	0.5809	0.5252	0.0039	0.5812	0.7598	0.0058	0.7923	0.6334	0.0718	0.6726
5	0.6569	0.0278	0.6814	0.648	0.0459	0.6609	0.8047	0.0535	0.8164	0.6951	0.0724	0.7074
10	0.7284	0.0174	0.7501	0.7291	0.0263	0.744	0.7781	0.0783	0.7814	0.7459	0.0214	0.7685
15	0.7542	0.048	0.8099	0.7331	0.0686	0.7704	0.8256	0.0177	0.8912	0.7463	0.0397	0.7693



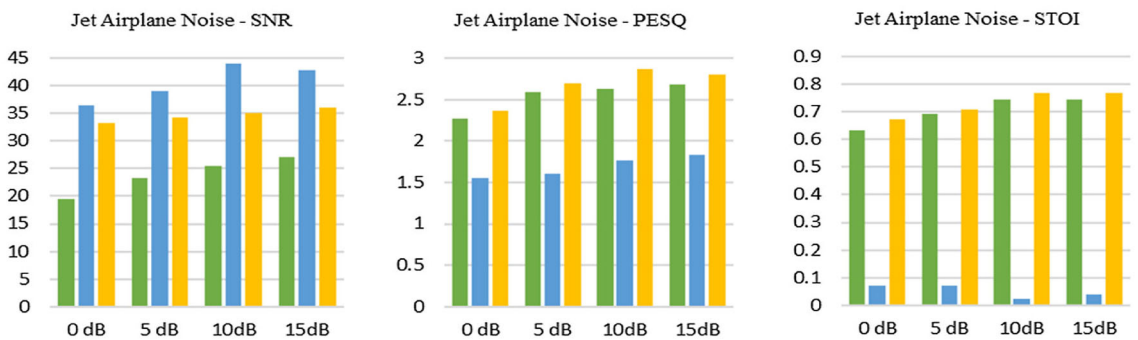
(a) Washing Machine Noise



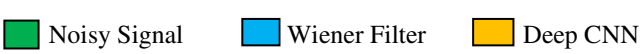
(b) Rainbow Noise



(c) Train Whistle Noise



(d) Jet Airlane Noise p



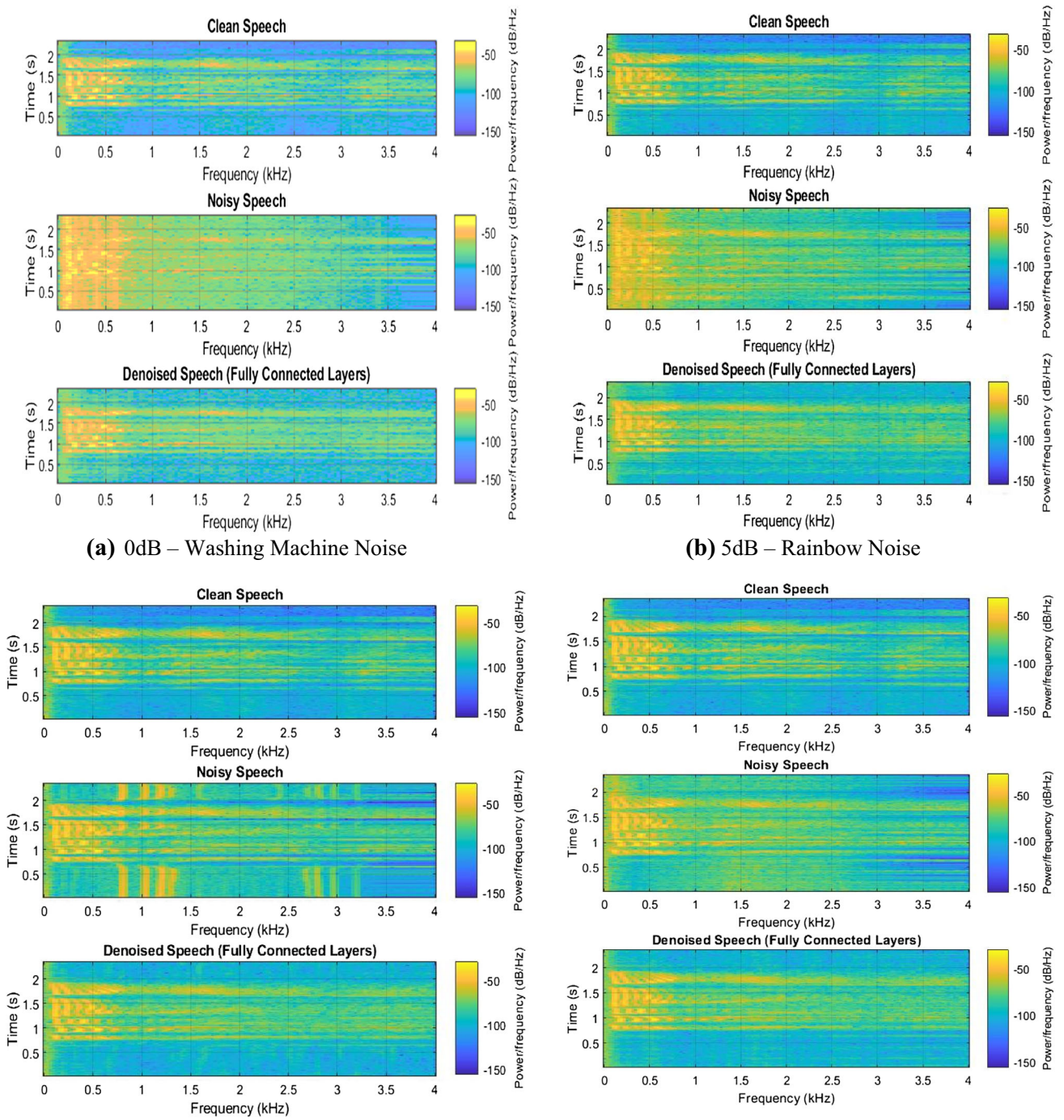


Fig. 5 Spectrogram analysis of clean speech, noisy speech and denoised speech signal **a** 0 dB—washing machine noise, **b** 5 dB—rainbow noise, **c** 10 dB—train whistle noise and **d** 15 dB—jet airplane noise

$$\mathrm{SNR}_{\mathrm{dB}} = 20\log_{10}\frac{S\_\mathrm{rms}}{N\_\mathrm{rms}}$$

where S_rms—root mean square of speech signal, N_rms—root mean square of level of noise.

- Perceptual Evaluation of Speech Quality (PESQ)

PESQ is a subjective quality measurement and it is based on the mean opinion score based on the evaluation given by the listeners and standardized by

International Telecommunications Union (ITU). The PESQ value ranges as per Table 1 given below.

- Short Time Objective Intelligibility (STOI)

STOI is a subjective intelligibility measurement, larger the value better the speech intelligibility. The STOI value ranges between 0 and 1.

The audio of the noisy speech signal was inferior in quality as well as intelligibility. When the signals were fed to the Deep CNN system for speech enhancement, the performance of the denoised speech was well improved in terms of quality which were clearly observed by the values of SNR and PESQ. Also, the intelligibility was improved, which was analyzed from the STOI scores. Table 2 shows the quality (SNR and PESQ) and intelligibility (STOI) of noisy signals and improvement in the denoised signal's performance metrics.

In order to analyze the quality, SNR and PESQ are considered and to evaluate the clarity of speech; the metric STOI is taken. The subjective quality of the spoken speech signal is analyzed by PESQ. The value of PESQ ranges between − 0.5 to 4.5. The higher the value of PESQ on the scale indicates the improvement in quality of the denoised speech. STOI refers to the subjective intelligibility of speech and it ranges between 0 and 1. The improvement in the STOI value is indicated by the higher value.

As per the observations from the performance metrics shown in Table 2, the SNR of the denoised signal through Wiener filtering shows good improvement compared to Deep CNN model for different noise levels as well as different noise types. The PESQ value of the Wiener filter is in the poor range (1–2) for the rainbow and jet airplane noise as per PESQ scores given in Table 1. But the PESQ value of the washing machine noise and train whistle noise of the Wiener filter is in the fair (2–3) range.

For the Deep CNN, the PESQ values for all the noise levels and noise types it falls in the fair (2–3) category of mean opinion score. As the Wiener filter focusses more on the quality of the speech signal, it gives good result in terms of SNR and moderate results for PESQ. But the intelligibility of speech is compromised which reduces the clarity of the speech signal. The STOI scores show that the Wiener filter is not capable of improving the intelligibility. On the other hand, the Deep CNN shows drastic results in the STOI values, which in turn represents the intelligibility of the denoised speech signal.

The consolidated results in Table 2 show the improvement in the performance metrics of Deep CNN compared to the conventional Wiener filtering algorithm for denoising speech signal. The Wiener filtering method shows outstanding results on the SNR and the PESQ. It is clearly observed that the Wiener filter has good capability in improving the quality of the speech signal. When the intelligibility of the speech signal is considered, the performance of the Wiener filter is deficient. However, the DNN shows a drastic increase in terms of the clarity of the speech signal.

The denoised signal shown in Fig. 4 represents that the SNR of the noisy signal is much improved in the Wiener filter compared to the Deep CNN. However, in terms of the other performance metric representing the quality of speech, i.e., the PESQ of the denoised speech signal is much improved in Deep CNN compared to the Wiener filter. When the intelligibility of the denoised speech is analyzed, it is evident that the STOI scores of the Deep CNN give an excellent improvement in the clarity of speech. The spectrograms of the clean speech, noisy speech and denoised speech for the different types of noise and noise levels are shown in Fig. 5.

6 Conclusion

The proposed single channel speech enhancement system estimates the magnitude of the speech signal in the frequency domain. The Deep CNN-based single channel speech enhancement system is compared with the traditional Wiener filtering method. Evaluation is carried out on multiple noise conditions to analyze the denoising capability of the speech enhancement system, and the results indicate that the Deep CNN-based system outperforms in terms of quality and intelligibility compared to the best performing Wiener filtering traditional technique. The quality of the denoised speech signal based on the SNR shows a drastic improvement for the Wiener filtered denoised signal. However, the Deep CNN yields excellent results in terms of quality and intelligibility that are analyzed based on the scores of PESQ and STOI. Thus, it should be recorded that the performance of Deep CNN outperforms the traditional Wiener filter technique.

Funding No funding.

Declarations

Conflict of interest We don't have any conflict of interest.

Human and animal rights statement Humans/animals are not involved in this research work.

Data availability statements The datasets analyzed during the current study are available in the University of Edinburgh, Centre for Speech Technology Research (CSTR). https://datashare.is.ed.ac.uk/handle/10283/2791.

References

Chai L, Du J, Liu Q-F, Lee C-H (2019) Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement. IEEE ACM Trans Audio Speech Lang Process 27(12):1919–1931

Cui X, Chen Z, Yin F (2020) Speech enhancement based on simple recurrent unit network. Appl Acoust 157:107019

De S, Smith SL (2020) Batch normalization biases deep residual networks towards shallow paths. CoRR, vol. abs/2002.10444

Dionelis N, Brookes M (2018) Phase aware single channel speech enhancement with modulation domain Kalman filtering. IEEE ACM Trans Audio Speech Lang Process 26:5

Du et al (2017) Stacked convolutional denoising auto-encoders for feature representation. IEEE Trans Cybern 47(4):1017–1027

Fu S-W, Tsao Y, Lu X (2016) Snr-aware convolutional neural network modeling for speech enhancement. In: Interspeech, pp 3768–3772

Fu S-W, Tsao Y, Lu X, Kawai H (2017) Raw waveform-based speech enhancement by fully convolutional networks. In: Proceedings of the APSIPA ASC, pp 6–12

Fu S-W, Wang T-W, Tsao Y, Lu X, Kawai H (2018) End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE ACM Trans Audio Speech Lang Process (TASLP) 26(9):1570–1584

Grais EM, Erdogan H (2013) Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation. In: Proc. Inter-speech

Grais EM, Plumbley MD (2017) Single channel audio source separation using convolutional denoising autoencoders. In: Proceedings of the IEEE global conference on signal information processing, pp 1265–1269

Healy EW, Delfarah M, Vasko JL, Carter BL, Wang D (2017) An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker. J Acoust Soc Am 141(6):4230–4239

Hsieh T-A, Wang H-M, Lu X, Tsao Y (2020) WaveCRN: an efficient convolutional recurrent neural network for end-to-end speech enhancement. IEEE Signal Process Lett 27:2149

https://datashare.is.ed.ac.uk/handle/10283/2791

Hu Y, Loizou PC (2008) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 16(1):229–238

ITU, Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs ITU-T Rec. p 862 (2000)

Jain K, Ross A, Prabhakar S (2004) An introduction to biometric recognition. IEEE Trans Circuits Syst Video Technol 14(1):4–20

Kameoka H, Tanaka K, Kwasny D, Kaneko T, Hojo N (2020) ConvS2S-VC: fully convolutional sequence-to-sequence voice conversion. IEEE ACM Trans Audio Speech Lang Process 28:1849–1863

Kolbæk M, Tran Z-H, Jensen SH, Jensen J (2020) On loss functions for supervised monaural time-domain speech enhancement. IEEE ACM Trans Audio Speech Lang Process 28:825–838

Kolbk M, Tan Z, Jensen J (2017) Speech intelligibility potential of general and specialized deep neural network-based speech enhancement systems. IEEE ACM Trans Audio Speech Lang Process 25(1):153–167

Kumar TS (2021) Construction of hybrid deep learning model for predicting children behavior based on their emotional reaction. J Inf Technol 3(01):29–43

Lan T, Lyu Y, Ye W, Hui G, Zenglin Xu, Liu Q (2020) Combining multi-perspective attention mechanism with convolutional networks for monoaural speech enhancement. IEEE Access 8:78979–78991

Li A, Yuan M, Zheng C, Li X (2020) Speech enhancement using progressive learning-based convolutional recurrent neural network. Appl Acoust 166:107347

Li R, Liu Y, Shi Y, Dong L, Cui W (2016) ILMSAF based speech enhancement with DNN and noise classification. Speech Commun 85:53–70

Li J, Zhang H, Zhang X, Li C (2019) Single channel speech enhancement using temporal convolutional recurrent neural networks. In: Proceedings of the APSIPA ASC, pp 896–900

Loizou PC (2013) Speech enhancement: theory and practice, 2nd edn. CRC Press, Boca Raton

Meng Z, Li J, Gong Y, Juang BH (2018) Cycle-consistent speech enhancement. In: Proceedings of the INTERSPEECH, pp 1165–1169

Nossier SA, Wall J, Moniri M, Glackin C, Cannings N (2021) An experimental analysis of deep learning architectures for supervised speech enhancement. Electronics 10(1):17

Paliwal KK, Wojcicki K, Schwerin B (2010) Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun 52(5):450–475

Pandey D, Wang D (2019) TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain. In: Proceedings of the Interspeech, pp 6975–6879

Pandey A, Wang D (2019) A new framework for CNN based speech enhancement in the time domain. IEEE ACM Trans Audio Speech Lang Process 27(7):1179

Park SR, Lee JW (2017) A fully convolutional neural network for speech enhancement. Proc Interspeech 2017:1993–1997

Rix W, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, vol 2, pp 749–752

Schwerin B, Paliwal KK (2014) Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement. Speech Commun 58:49–68

Srinivasan S, Samuelsson J, Kleijn WB (2006) Codebook driven short term predictor parameter estimation for speech enhancement. IEEE Trans Audio Speech Lang Process 14(1):163–176

Sungheetha A, Rajesh Sharma R (2021) Classification of remote sensing image scenes using double feature extraction hybrid deep learning approach. J Inf Technol 3(02):133–149

Tan K, Wang D (2020) Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE ACM Trans Audio Speech Lang Process 28:380–390

Thiergart O, Taseska M, Habets EAP (2014) An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. IEEE ACM Trans Audio Speech Lang Process 22:12

Wang D, Chen J (2018) Supervised speech separation based on deep learning: An overview. IEEE ACM Trans Audio Speech Lang Process 26(10):1702–1726

Wang NY-H, Wang H-LS, Wang F-W, Lu X, Wang H-M, Tsao Y (2021) Improving the intelligibility of speech for simulated electric and acoustic simulation using fully convolutional neural network. IEEE Trans Neural Syst Rehabil Eng 29:184–195

Xia B, Bao C (2014) Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Commun 60:13–29

Xian Y, Sun Y, Wang W, Naqvi SM (2021) Convolutional fusion network for monaural speech enhancement. Neural Netw 143:97–107

Xu Y, Jun Du, Dai L-R, Lee C-H (2013) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 21(1):65–68

Yuan W (2020) A time–frequency smoothing neural network for speech enhancement. Speech Commun 124:75–84

Zhao H, Zarar S, Tashev I, Lee C (2018) Convolutional-recurrent neural networks for speech enhancement. In: International conference on *acoustics,* speech*,* and signal processing, pp 2401–2405

Zheng N, Shi Y, Rong W, Kang Y (2020) Effects of skip connections in CNN-based architectures for speech enhancement. J Signal Process Syst 92:875–884