Journal of Integrated
SCIENCE & TECHNOLOGY

Acoustic signal enhancement using autoregressive PixelCNN architecture

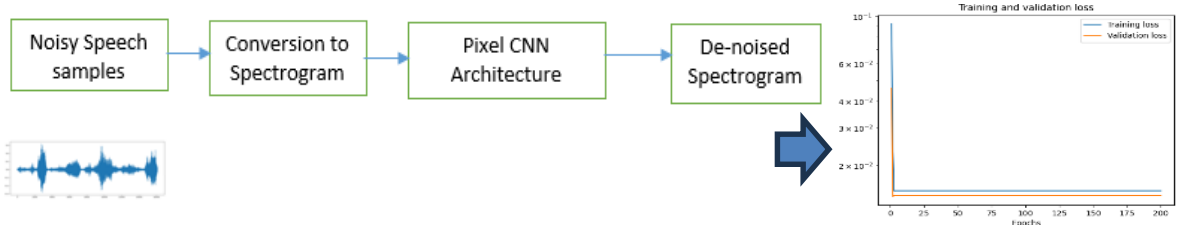
Shibani Kar

Department of Electronics and Communication Engineering, Sambalpur University Institute of Information Technology, Jyoti Vihar, Burla, Odisha. India.

Received on: 17-Sept-2023, Accepted and Published on: 11-Dec-2023

Article

ABSTRACT



Acoustic Signals such as speech and sound are easily degraded by interferences present in our surroundings. The present work explores the usage of the Pixel CNN architecture for the removal of non-stationary noises from the speech signal. The presence of noise in speech signals affects the performances of applications that use speech signal as a medium for communication such as automatic speech recognition systems, hearing aid, mobile phones. Pixel CNN is a deep generative network architecture implemented as an autoregressive model. The dataset "NOIZEUS" is used for noise mixed speech samples and clean speech samples. The architecture learns the feature from the input speech using the spectrogram representation of speech signal. The performance of Pixel CNN architecture is compared with three methods: VAE, UNET, Auto encoders. The performance evaluation metrics used for comparison are "PESQ" and "STOI".

Keywords: Pixel CNN, deep generative model, auto regression, non-stationary noises, speech de-noising

INTRODUCTION

The speech signal is a useful means of communication among human beings. It is an important acoustic signal that is one dimensional in nature and the amplitude varies with respect to time. "In real time environments", the speech signal[1] gets easily corrupted by a variety of "stationary and non-stationary signals"[2]. The signal is degraded more by the "non-stationary" noise such as babble[1] noise[1]. The presence of noise in speech signals affects the performances of applications that use speech signal as a medium for communication such as automatic speech recognition systems, hearing aid, mobile phones.The traditional method introduces musical noises in the signal after denoising[3] the signal. Also, the traditional methods can not remove babble noise completely from the signal.Hence, "deep neural network architectures"[4]are used nowadays for the de-noising of speech signals. Deep neural networks learns the complex speech features for the prediction of the denoised speech. In this paper, the babble noise is used as noisy signal that degrades the signal of interest. The signal degradation is confirmed when the listener faces difficulty in understanding the message communicated by the signal. The popular speech de-noising methods are: spectral subtraction[5], minimum mean square error[6], wiener filter for speech enhancement[7,8], deep neural networks[4,9,10], Convolution neural network architectures[11,12], LSTM networks[13], WaveUnet[14], Unet[15,16], Variational autoencoders[17,18], Convolutional autodencoders[19]. These deep neural network architectures are used for image processing. In this paper, the de-noising of speech signal is performed using a deep generative networks known as PixelCNN architecture. This is an autoregressive type of neural network that predicts the output value based upon the previous values of the signal. Deep generative networks are nowadays popular for generating images of high quality. In this paper, these generative networks are used for the denoising and enhancement of speech signal. The Pixel CNN method is a generative architecture. In the propsed work, the PixelCNN is applied for speech enhancement . The objective of the paper are as follows:

- To apply the PixelCNN architecture for speech de-noising
- To apply the de-noising model for the removal of non-stationary noises such as babble noise.

*Corresponding Author: Shibani Kar
Email: shibanikar@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(3), 770. URN:NBN:sciencein.jist.2024.v12.770*

- To demonstrate the efficiency of PixelCNN architecture for the removal of noise from speech signal using "PESQ" and "STOI" metric.
- To use variational autoencoders , autoencoders, UNET architecture for comparison purpose by training these models with same speech data samples.

The paper is organised in following sections. First the literature review is given. Then, the methodology and simulation setup is explained. In the last section result and conclusion is presented.

LITERATURE REVIEW

The Table 1 represents a brief review of deep learning methods for speech enhancement.

Table 1: Review of Methods for Speech Enhancement

Ref	Objective	Method Used	Evaluation Metrics used
1	De-noising of single channel speech signal	Deep neural Network and Wiener Filter	Better Performance in terms of PESQ[20], SDR, BAK, SIG for the removal of babble noises at low SNRs
13	Single channel speech enhancement	LSTM based de-noising method and convolutional auto encoder for the synthesis of enhanced speech	Better PESQ scores for the removal of non-stationary noises.
14	De-noising of speech signal in time domain	UNET architecture is combined with Attention mechanism	Better PESQ score for the de-noised speech
21	Time domain de-noising of speech.	UNET method combined with attention mechanism and compressed sensing loss function	Better PESQ, STOI metrics
22	Comparative analysis of DNN methods for de-noising speech signals	Multilayer Perceptron(MLP), Convolutional Neural Network(CNN), De-noising Auto-encoders(DAE)	PESQ, STOI
23	Time domain speech enhancement	Generative Adversarial Network Architecture	PESQ, STOI
24	Single channel speech enhancement	Variational Auto-encoders and Non "Negative matrix Factorization" method for modeling of noise	SDR
25	Enhancement of Speech Signal	Time Frequency correlation using LSTM and CNN	PESQ, STOI
11	Enhancement of speech for hearing impaired listeners	Convolutional Neural Network Architecture	PESQ, STOI, Segmental SNR
26	Time domain speech enhancement	Convolutional Neural Network	PESQ, STOI, SI-SDR
27	Singing voice separation	UNET and WaveUnet architecture	SAR, SDR, SIR

The literature suggest that "the state of the art"[4] deep neural networks[4] for "speech enhancement"[1] model uses convolutional neural network architecture, autoencoders, UNET , adversarial scenarios, such as "generative adversarial network(GAN)", generative networks such as "variational autoencoders, non-negative matrix factorization" for noise modeling. The "model" uses different forms of "speech signal" represenation as input to the model[9], such as time domain signal representation, log power spectra, mel frequency cepstral coefficient, time frequency representation as input features representing speech parameters for both noisy and clean speech. These networks are initially applied for image processing, biomedical image processing. The review of literature informs extensive use of convolution neural networks for speech enhancement purpose. The simple autoencoders and variational autoencoders are also tried and tested for speech enhancement. The deep generative network having autoregressive type nature has not been tested for the de-noising purpose. The objective of the paper is to test the PixelCNN architecture for speech de-noising.

METHODOLOGY

The paper uses the PixelCNN architecture proposed by Oord et al.[28] for the removal of non-stationary noises from the speech signal. The PixelCNN architecture is popular for the generation of images and used as a generative architecture. The application of this architecture for the removal of noise from speech uses the masked filter feature of Convolutional layers. This is a neural network that does not apply on the whole image but on a selected pixels. The masking of pixels is implemented by selectively keeping the pixels values as '0' or '1'. These masked feature allows selective selection of pixels and not all the pixels for the feature extraction. The architecture computes the joint probability distribution from the selected pixels by scanning each pixel present in each row and pixel present to the left of the pixel

whose value need to be predicted by using conditional probability. Each pixel is assigned value based upon conditional probability. The conditional probability is given in equation 1[28].

$$p(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{n^2} p(x_i|x_1, ..., x_{i-1}, \mathbf{h})$$

Equation (1)

This conditional probability is implemented using neural networks i.e. PixelCNN architecture for the prediction of new pixels. The architecture uses gated pixel CNN method where the "activation function[19]" "ReLU[28]" is replaced by the "activation function[28]" as shown in equation 2[28].:

$$\mathbf{y} = \tanh(W_{k,f} * \mathbf{x} + V_{k,f}^T \mathbf{h}) \odot \sigma(W_{k,g} * \mathbf{x} + V_{k,g}^T \mathbf{h})$$

Equation (2)

DESIGNED MODEL

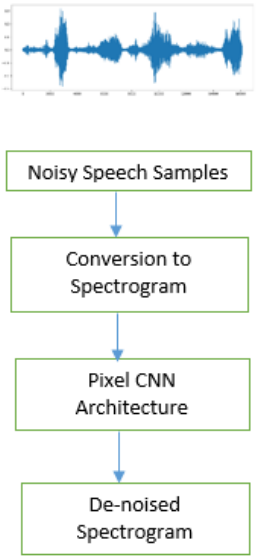


Figure 1: Designed Model

The noisy and clean speech samples cannot be processed at once as speech is in itself a non-stationary signal. Hence, small parts of the speech signal are processed one by one. They are converted to small fragments of speech by methods of framing and windowing. The PixelCNN architecture operates on pixel values,therefore, the speech samples "transformed from time domain" to "time frequency domain"[16] representation. These T-F domain representation of speech is known as spectrograms. These are 2-D representation of speech signal. Pixel CNN architecture processes these samples and provides denoised spectrograms at the output.

SIMULATION SETUP

The "clean speech and noisy speech"[1] samples are obtained from NOIZEUS dataset. These samples are available for free to the research community. The model is trained for babble noise[1] at SNR: 0db, 5db, 10db, 15db. The clean speech file contains recordings from six different speakers(three males and three females). The noisy file consist of thirty speech utterances for each snr level. The "babble[1] speech signal file "[1] at "snr levels 0db[1],5db,10db,15db"[1] contains 120 noisy speech utterances.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 128, 128, 1)]     0

 pixel_conv_layer (PixelConv  (None, 128, 128, 128)    6400
 Layer)

 residual_block (ResidualBlo  (None, 128, 128, 128)    98624
 ck)

 residual_block_1 (ResidualB  (None, 128, 128, 128)    98624
 lock)

 pixel_conv_layer_3 (PixelCo  (None, 128, 128, 128)    16512
 nvLayer)

 pixel_conv_layer_4 (PixelCo  (None, 128, 128, 128)    16512
 nvLayer)

 conv2d_9 (Conv2D)           (None, 128, 128, 1)       129

=================================================================
Total params: 236,801
Trainable params: 236,801
Non-trainable params: 0
_________________________________________________________________
```

Table 2: Simulation Model of Pixel CNN based Speech Denoising

The model is trained with the 120 clean speech samples and 120 noisy speech samples corrupted with babble[1] noise at 0db,5db,10db and 15db. The number of epochs selected as 200. The batch size selected for model simulation is 32. The "ADAM" optimizer is selected for optimization of learned weights during training. For the simulation of model "Huber loss function" is used. The model is simulated using Python, keras and tensorflow framework. The simulation model of PixelCNN based speech de-noising is given in Table 2.

RESULTS

The simulation results for the Pixel CNN architecture for the removal of babble noise is provided in Table 3,4,5 6 and 7. The table presents the comparative analysis of the results. The designed method PixelCNN performance is compared with three methods: UNET, Autoencoder and Variational autoencoder. Five performance comparison metrics are used for the analysis of the methods. The First metric is Perceptual Evaluation Of Speech Quality[20](PESQ). PESQ measures the perceived quality of speech. The normal score for this metric is between 0.5 and 5. The higher the score, the better is the quality. The second metric is short time objective intelligibility score(STOI). The STOI[2,29] score access the percentage of speech that a listener can understand. The higher the score, the better is the performance w.r.t intelligibility content of speech. The third metric is segmental signal to noise ratio(Seg SNR)[30]. This is a time domain objective metric that measures the distortion in speech waveform at the output of the speech enhancement model. The Segmental

SNR is computed on short frames or segments of speech having a duration of 15ms-20ms. The higher the value, the better is the performance of the method. The fourth metric is the signal to noise ratio. This is another time domain metric that operates upon a waveform to measure the distortion in speech post denoising. The higher the value, the better is the performance. Fifth performance measuring metric is Signal to Distortion Ratio. This measures the quality of signal as compared to its noisy counterpart post denoising. This measure represents the loudness of the signal as compared to its distorted or noisy version. The model simulation results of Pixel CNN, UNET , autoencoder and Variational autoencoder are presented in fig.2, fig 3, fig 4 and fig 5 respectively. The models are trained for 200 epochs with 120 noisy speech samples at SNRs 0dB , 5dB, 10dB and 15dB. The methods performance was compared based on the five metrics stated above. The findings of the analysis of result state that the Pixel CNN method provides an average score of 1.31 at low SNR condition(0dB). The UNET and autoencoder provide an average score of 1.31 and 1.37. Hence, the Pesq score of the designed method is similar to UNET and convolutional autoencoder at low SNR(0dB). The comparative summary of PESQ[29] score is given in Table 3. The Pixel CNN architecture provides an average STOI score of 64% at low SNR(0dB). The intelligibility score is higher at 0dB as compared with UNET and autoencoder. At 15dB the average score of Pixel CNN is 0.91, representing a high intelligibility score. The comparative summary of STOI scores is given in Table 4. The segmental SNR score of Pixel CNN is very low at 0dB and increases the SegSNR value at 15dB. The comparative summary of Segmental SNR score is given in Table 5. The average SNR score for Pixel CNN is 2.09 at 0dB SNR. The performance of Pixel CNN SNR score decreases at high SNR but the decrease in performance score is lower than other competing methods.The comparative summary of SNR score is given in Table 6. The SDR ratio of PixelCNN increases from low to high SNR. The comparative summary of SDR score is given in Table 7.

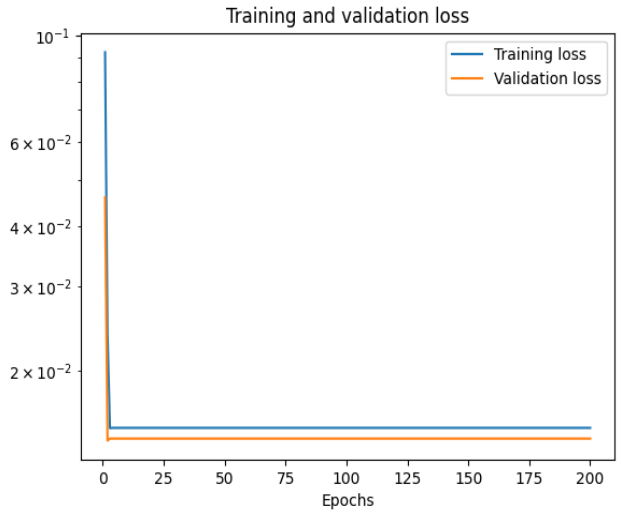


Figure 2: Model Training Graph for PixelCNN architecture

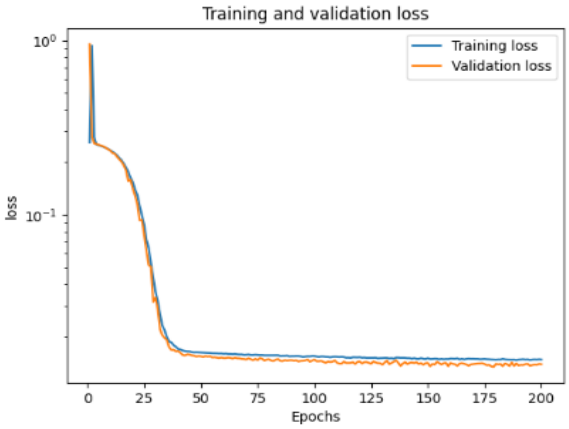


Figure 3: Model Training Graph for VAE model

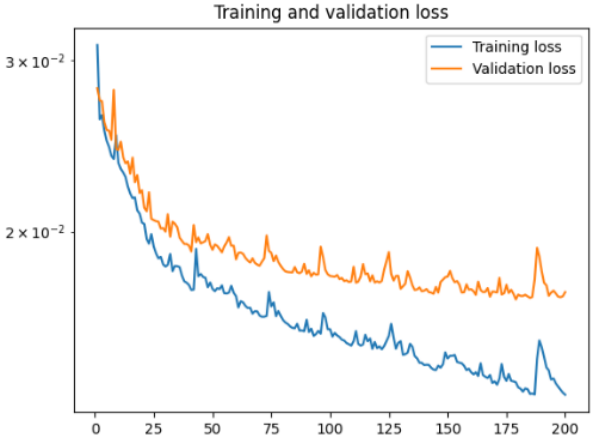


Figure 4: Model Training Graph for Autoencoder

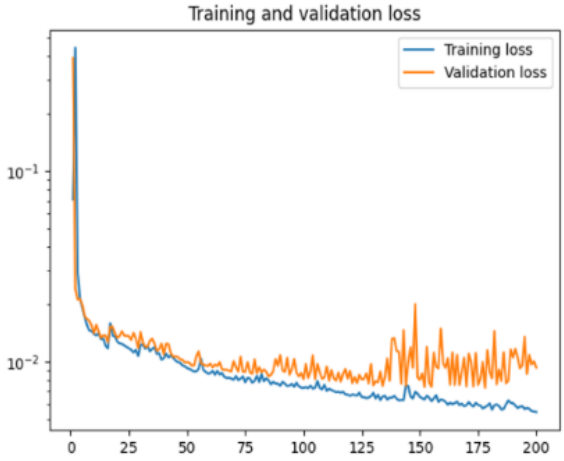


Figure 5: Model Training Graph for UNET

Table 3: Comparative Performance for PESQ[29] for Babble[1] Noise

PESQ Measure				
	PixelCNN	**UNET**	**Autoencoder**	**VAE**
0dB	1.31	1.31	1.37	1.28
5dB	1.37	1.45	1.49	1.35
10dB	1.52	1.58	1.70	1.45
15dB	1.58	1.77	1.81	1.61

Table 4: Comparative Performance for STOI[31] for Babble Noise

STOI Measure				
	PixelCNN	**UNET**	**Autoencoder**	**VAE**
0dB	0.64	0.62	0.62	0.57
5dB	0.74	0.72	0.76	0.71
10dB	0.84	0.82	0.85	0.80
15dB	0.91	0.88	0.88	0.86

Table 5: Comparative Performance for Segmental SNR for Babble Noise

Segmental SNR Measure				
	PixelCNN	**UNET**	**Autoencoder**	**VAE**
0dB	-4.49	0.06	-1.73	-4.16
5dB	-2.89	1.38	0.22	-2.37
10dB	-1.26	2.40	1.64	-1.02
15dB	0.26	3.11	2.61	0.22

Table 6: Comparative Performance for improvement in SNR for Babble Noise

SNR Measure				
	PixelCNN	**UNET**	**Autoencoder**	**VAE**
0dB	2.09	3.56	3.05	2.24
5dB	-0.77	-0.95	-1.33	-2.03
10dB	-4.43	-6.40	-6.21	-6.87
15dB	-8.77	-12.84	-12.03	-11.77

Table 7: Comparative Performance for SDR for Babble Noise

SDR Measure				
	PixelCNN	**UNET**	**Autoencoder**	**VAE**
0dB	1.18	-5.29	-0.33	0.64
5dB	1.79	4.32	0.51	1.48
10dB	2.43	-6.40	1.87	2.22
15dB	4.04	-1.31	2.15	2.69

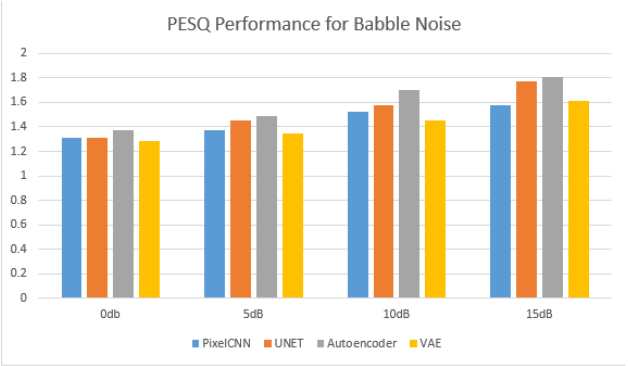


Figure 6: Comparative Chart for PESQ for Babble Noise

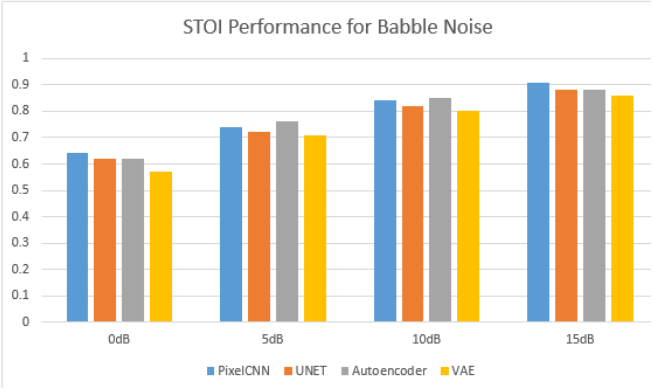


Figure 7: Comparative Chart for STOI for Babble Noise

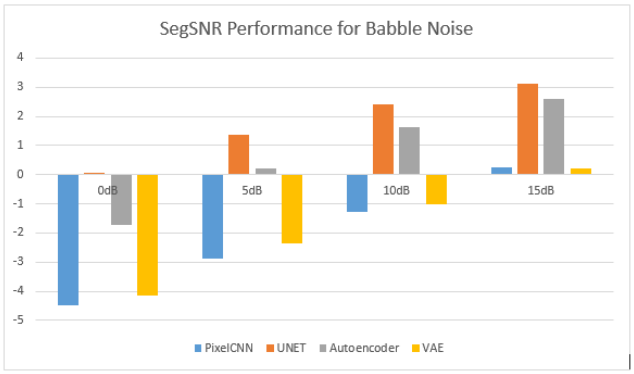


Figure 8: Comparative Chart for Segmental SNR for Babble Noise

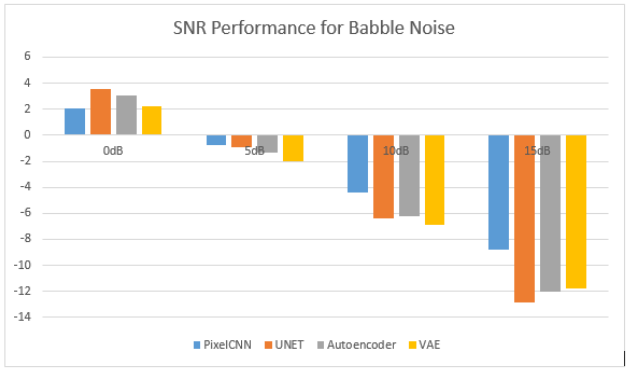


Figure 9: Comparative Chart for SNR for Babble Noise

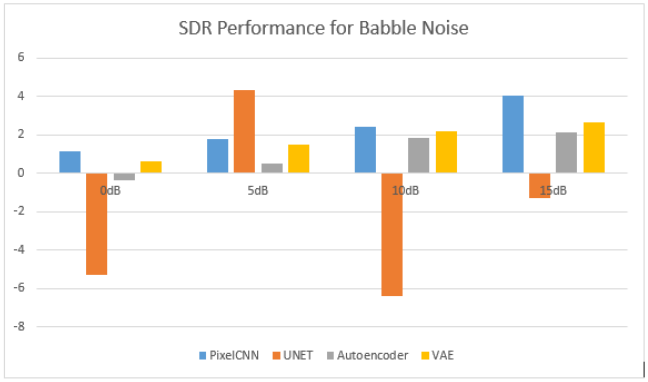


Figure 10: Comparative Chart for SDR for Babble Noise

CONCLUSION AND FUTURE SCOPE

The paper focuses on the de-noising of speech signals corrupted with babble noise[1,2] at low SNRs. The method Pixel CNN effectively de-noise the signal using the Time Frequency representation i.e. spectrogram form of the signal. The Pixel CNN removes babble noise from the noisy spectrograms and provides de-noised spectrogram at the output. This method is a mapping based method and trained end to end. The performance of the Pixel CNN architecture for noise removal from speech signals is compared with convolutional autoencoders, UNET and variational autoencoders. The parameters used for comparison[1] of methods are "PESQ and STOI". The other two time domain objective metrics that are used for speech enhancement are segmental SNR and SNR score. The Pixel CNN method performs

at par with these methods in terms of PESQ[31] and STOI[29] which measures signal "quality and intelligibility". The STOI score is 0.90 for PixelCNN method. The segmental SNR score is low for Pixel CNN as the method works with a time frequency domain feature and the segmental SNR score is useful for waveform based operations. The SNR score that measures speech quality for Pixel CNN is 2.09 at low SNR. The Pixel CNN method gives superior performance in enhancing the intelligibility of speech. This paper is initial work on the application of the PixelCNN architecture for noise removal from speech signal. The method demonstrates satisfactory performance for the removal of non-stationary noises. The method can be trained with different noise types at different SNRs, datasets and feature inputs as a future scope of the method.

CONFLICT OF INTEREST STATEMENT

Author do no have any conflict of interest.

REFERENCES

1. N. Saleem, M. Irfan, X. Chen, M. Ali. Deep Neural Network based Supervised Speech Enhancement in Speech-Babble Noise. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*; IEEE, Singapore, **2018**; pp 871–874.
2. S. Kar, V. Mukherjee. Convolutional Neural Network for Removal of Environmental Noises from Acoustic Signal. In *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*; **2023**; pp 640–645.
3. X. Lu, Y. Tsao, S. Matsuda, C. Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech 2013*; ISCA, **2013**; pp 436–440.
4. D. Hepsiba, J. Justin. Role of Deep Neural Network in Speech Enhancement: A Review. In *Artificial Intelligence*; Hemanth, J., Silva, T., Karunananda, A., Eds.; Communications in Computer and Information Science; Springer Singapore, Singapore, **2019**; Vol. 890, pp 103–112.
5. S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, 27 (2), 113–120.
6. Y. Ephraim, D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, 33 (2), 443–445.
7. Wiener Filtering in Frequency Domain to Enhance Speech Corrupted by Colored Noise. *Int. J. Recent Technol. Eng.* **2019**, 8 (2S11), 1058–1062.
8. D. Ribas, A. Miguel, A. Ortega, E. Lleida. Wiener Filter and Deep Neural Networks: A Well-Balanced Pair for Speech Enhancement. *Appl. Sci.* **2022**, 12 (18), 9000.
9. Y. Xu, J. Du, L.-R. Dai, C.-H. Lee. An Experimental Study on Speech Enhancement Based on Deep Neural Networks. *IEEE Signal Process. Lett.* **2014**, 21 (1), 65–68.
10. Y. Xu, J. Du, L.-R. Dai, C.-H. Lee. A Regression Approach to Speech Enhancement Based on Deep Neural Networks. *IEEEACM Trans. Audio Speech Lang. Process.* **2015**, 23 (1), 7–19.
11. G.S. Bhat, N. Shankar, C.K.A. Reddy, I.M.S. Panahi. A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone. *IEEE Access* **2019**, 7, 78421–78433.
12. A. Pandey, D. Wang. Dense CNN With Self-Attention for Time-Domain Speech Enhancement. *IEEEACM Trans. Audio Speech Lang. Process.* **2021**, 29, 1270–1279.
13. M. Strake, B. Defraene, K. Fluyt, W. Tirry, T. Fingscheidt. Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. *EURASIP J. Adv. Signal Process.* **2020**, 2020 (1), 49.
14. R. Giri, U. Isik, A. Krishnaswamy. Attention Wave-U-Net for Speech Enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*; IEEE, New Paltz, NY, USA, **2019**; pp 249–253.
15. A. Jansson, E.J. Humphrey, N. Montecchio, et al. Singing Voice Separation with Deep U-Net Convolutional Networks; **2017**.
16. O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*; Navab, N., Hornegger, J., Wells, W. M., Frangi, A. F., Eds.; Lecture Notes in Computer Science; Springer International Publishing, Cham, **2015**; pp 234–241.
17. S. Leglaive, L. Girin, R. Horaud. Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; **2019**; pp 101–105.
18. H. Fang, G. Carbajal, S. Wermter, T. Gerkmann. Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; **2021**; pp 676–680.
19. S.R. Park, J. Lee. A Fully Convolutional Neural Network for Speech Enhancement. arXiv September 22, 2016.
20. X. Dong, D.S. Williamson. Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks. *J. Acoust. Soc. Am.* **2020**, 148 (5), 3348–3359.
21. Z. Kang, Z. Huang, C. Lu. Speech Enhancement Using U-Net with Compressed Sensing. *Appl. Sci.* **2022**, 12 (9), 4161.
22. S.A. Nossier, J. Wall, M. Moniri, C. Glackin, N. Cannings. An Experimental Analysis of Deep Learning Architectures for Supervised Speech Enhancement. *Electronics* **2020**, 10 (1), 17.
23. S. Pascual, J. Serrà, A. Bonafonte. Time-domain speech enhancement using generative adversarial networks. *Speech Commun.* **2019**, 114, 10–21.
24. Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, T. Kawahara. Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; **2018**; pp 716–720.
25. W. Yuan. A time–frequency smoothing neural network for speech enhancement. *Speech Commun.* **2020**, 124, 75–84.
26. A. Pandey, D. Wang. A New Framework for CNN-Based Speech Enhancement in the Time Domain. *IEEEACM Trans. Audio Speech Lang. Process.* **2019**, 27 (7), 1179–1188.
27. A. Cohen-Hadria, A. Roebel, G. Peeters. Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation. In *2019 27th European Signal Processing Conference (EUSIPCO)*; IEEE, A Coruna, Spain, **2019**; pp 1–5.
28. A. van den Oord, N. Kalchbrenner, O. Vinyals, et al. Conditional Image Generation with PixelCNN Decoders. arXiv June 18, 2016.
29. J. Ma, Y. Hu, P.C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.* **2009**, 125 (5), 3387–3405.
30. X. Dong, D.S. Williamson. Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks. *J. Acoust. Soc. Am.* **2020**, 148 (5), 3348–3359.
31. X. Dong, D.S. Williamson. Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks. *J. Acoust. Soc. Am.* **2020**, 148 (5), 3348–3359.