Convolutional Deep Neural Network and Full Connectivity for Speech Enhancement

https://doi.org/10.3991/ijoe.v19i04.37577

Ban M. Alameri[1,2(✉)], Inas Jawad Kadhim[3], Suha Qasim Hadi[1], Ali F. Hassoon[1], Mustafa M. Abd[1], Prashan Premaratne[4]
[1]Department of Electrical Engineering, Faculty of Engineering, Mustansiriyah University, Baghdad, Iraq
[2]Department of Telecommunication Engineering, Malaga University, Malaga, Spain
[3]Electrical Engineering Technical College, Middle Technical University, Baghdad, Iraq
[4]School of Electrical and Computer and Telecommunications Engineering, University of Wollongong, North Wollongong, Australia
Ban.alameri@uomustansiriyah.edu.iq

Abstract—The speech signal that is received in real-time has background noise and reverberations, which have an impact on the quality of speech. Therefore, it is crucial to reduce or eliminate the noise and increase the intelligibility and quality of speech signals. In this study, a proposed method that is the most effective and challenging in a low SNR environment for three types of noise are removed, including washing machine, traffic noise, and electric fan noise, and clean speech is recovered. with three samples of noise which are mixed and added to the clean speech signal with a lower level of SNR value fixed at (−5, 0, 5) dBs, that noise source takes equal weights. The enhancement of the corrupted speech signal is done by applying a fully connected and convolutional neural network-based denoising algorithm and comparing their performance. The proposed network shows that a fully connected network (FCN) has less elapsed time than a convolutional network (CNN) while still achieving better performance, demonstrating its applicability for an embedded system. Also, the results obtained show that, overall, the CNN is better than the FCN regarding maximum coloration, PSNR, MES, and STOI.

Keywords—speech enhancement, deep learning, fully connected network, convolutional network, signal-to-noise ratio (SNR)

1 Introduction

Under diverse communication circumstances like speech, speech signals are continuously distorted by numerous noises. The effect of noise on the sound signal quality is an important issue for many communication companies due to the demand for the best quality in voice and video technology. The speech signal is hampered by many types of noise, including white noise, traffic noise, babble noise, additive noise, and channel noise [1]. Noise reduction or speech enhancement are common terms used to describe how to deal with background noise [2]. The two primary categories of

speech enhancement techniques are: traditional techniques and deep learning-based techniques. The traditional techniques include estimating the noise by Wiener filtering, which removes the additive noise [3], and spectral subtraction [4]. In order to improve speech signal, Lim and Oppenheim compressed the bandwidth, which marginally cleaned the noisy speech signal [5]. A class of minimum mean-square error (MMSE) estimators were employed by R. Martin [6] to improve the short-time spectral coefficients of a noisy speech signal. These estimators are based on super Gaussian densities, which result in an improved signal-to-noise ratio. The authors in [7] proposed a method for improving single-channel speech that is semi-supervised using non-negative matrix factorization (NMF). In recent years, Deep Neural Network (DNN) based speech enhancement has also gained popularity and produces significantly better results than traditional techniques [8], [9]. Deep neural networks (DNNs) are used in more recent techniques, such as those described in [10–12], to describe the nonlinear relationship between noisy and clean speech inputs. These techniques have enabled the use of non-stationary audio environments [13], and can be further classified into two types: mapping-based techniques, such as those found in [14–16], which use the log power spectra as the input and output signal of DNN, and masking-based techniques [17–19], which estimate a mask to perform denoising. Pandey and Wang formed complex spectral mapping for improving speech with enhanced cross-corpus generalization [20]. Xie et al. [21] proposed a complex recurrent variational autoencoder for improving speech signals. The computing procedure is substantially more intensive in conventional analogue models than it is in DNN. Therefore, a DNN-based strategy is a much better option for speech enhancement. Studies on single-channel and supervised multi-channel speech enhancement, for example, have been conducted [22–24]. These studies demonstrate how speech enhancement performance improves as the number of channels rises. In [25–27] and [28–29], a fully convolutional neural network (FCN) was used to improve multi-channel speech, whereas multiple recordings were used directly in the time domain. Neural networks have surpassed traditional techniques in several fields, providing enough data and sufficient hardware [30]. Deep neural networks (DNN) have the ability for generalization, which is one of their main advantages. This implies that a trained net could use previously unheard speech or samples to apply its knowledge of speech enhancement. This is crucial for both real-time deployment and training. The deep neural network system for speech enhancement uses the short-time Fourier transform (STFT) [31], [32], due to its simplicity, natural similarity to the auditory processes occurring within the human ear, and the availability of effective windowing techniques for the time-domain synthesis of the modified speech. This is because it has a lower computational complexity thanks to the use of a fast Fourier transform (FFT). Even though powerful DNNs can be used to improve speech, doing so comes at the expense of increasing complexity. In this work, we compare fully connected and convolutional DNNs to demonstrate how we can improve speech quality and intelligibility with simpler architectures and come to a conclusion about which of them is best suited for enhancing voice signals. The speech signal is processed using a denoising approach after random noise has been added for the evaluation. The objective measurements are short-time objective intelligibility (STOI) [33], signal-to-noise ratio (SNR) [34], mean square error (MSE) [6], peak signal-to-noise ratio (PSNR) and

maximum correlation (MC) are taken into consideration as figures of merit [30] for performance evaluation and comparison.

2 Speech enhancement based a deep neural network methodology

A Deep Neural Network (DNN) is a machine learning model, consisting of computing units called artificial neurons. DNNs are used in many different applications, including computer vision and speech processing. DNNs can make use of both the temporal and frequency information contained in an input signal, and produce an output that is more enhanced than the original signal. Figure 1 represents a block diagram and the main stages in the deep neural network. In suggested methodology, two different network types, fully connected and convolutional are used for the same task. Figure 2 illustrates how supervised deep learning-based voice denoising works. A network is taught to cut down on noise by using clear speech signals as its output or target signals. A speech + noisy speech signal n(t) can be expressed as

$$N(t) = C(t) + Ns(t) \tag{1}$$

where clean speech is represented by C(t) and noise is represented by Ns(t).

Since the target signal C(t) and the noise signal Ns(t) are typically not correlated, It is obvious that the clean speech can be recovered when the noise has been removed. The clean speech signal is the desired signal.

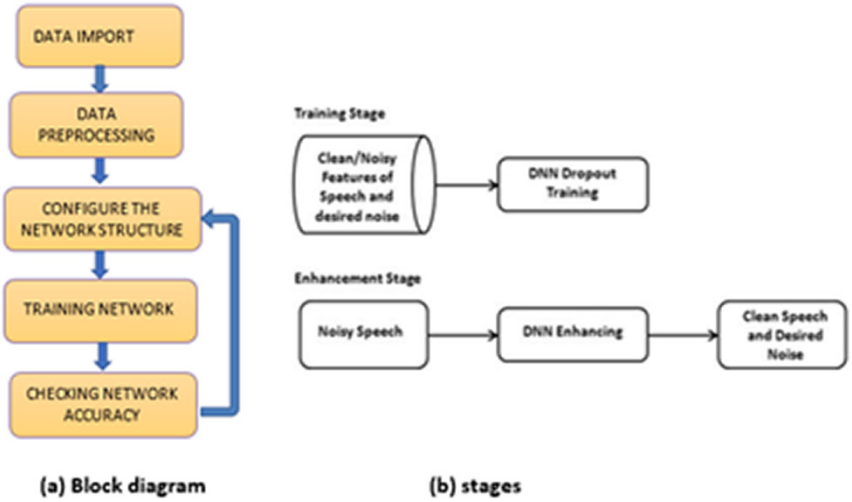

Fig. 1. The deep neural network

A block diagram of the system for improving speech is shown in Figure 2. The deep neural network maps the noisy spectral features to the clean spectral features (DNN).

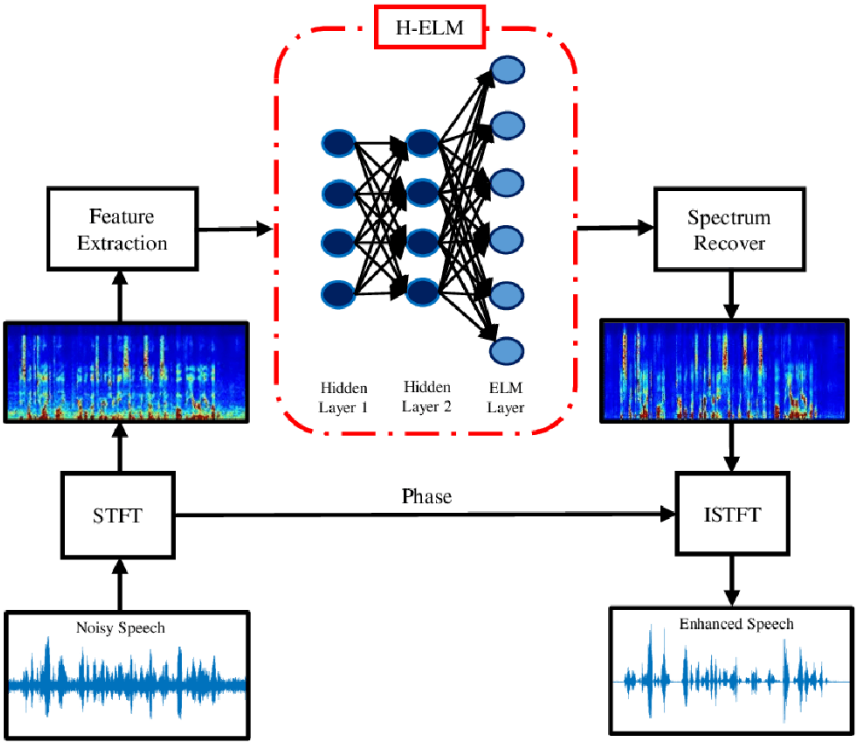

Fig. 2. A deep neural network based sound enhancement methodology

With window lengths of 256 samples, window overlaps of 75%, no zero-padding, and a Hamming window for analysis, the DNN is trained on 8 kHz samples of nosiy speech. The noisy speech is transformed into the STFT domain. During the training stage, labels are used to quantify prediction loss based on features that were extracted from the clean speech signals using STFT. The neural network relies on the features derived from the noisy speech inputs as predictors, together with the targets, throughout its training phase. The magnitude spectra of noisy and clean speech sounds, respectively, represent the predictor and target network signals. The augmented signal's magnitude spectrum, which is the network's output, is scaled using the mean and standard deviation from the training phase. Based on what the predictor gives, as a result of the regression network, the mean square error between the input target and the output is decreased. Following the joining of the modified output magnitude spectra and the previously acquired noisy signal phase, the improved speech is converted back to the time domain. Due to the predictor input consisting of 8 consecutives noisy STFT vectors, each STFT output estimate is created using the current noisy STFT and the 7 prior noisy STFT vectors.

3 Experimental and datasets

For the purpose of evaluating the efficacy of the suggested strategy, five regularly employed objective performance criteria were taken into account. These measures include mean square error (MSE), peak signal-to-noise ratio (PSNR), short-time objective intelligibility (STOI), signal-to-noise ratio (SNR), and maximum correlation (MC). By contrasting the improved speech with the matching clean speech, evaluation metrics are obtained, which measures the subjective quality of the speech. For the STOI metric, the ideal values can be found in the range [0, 1], where the upper border of the range. Better performance is indicated by higher CM. A small value of MSE means the mean square error between clean and denoise speech which is greater than 0. A higher PSNR means better speech quality. The dataset used in the experiment is a subset of the Mozilla Common Voice dataset [35–38]. Clean speech utterances are taken from the common voice. It consists of 2806 recordings. Short sentences are equally divided into three sounds: a male and two female voices. In our experiment, 2000 recordings were chosen as the target speech for training. Another 403 recordings are selected for testing, and another 403 recordings are selected for the validation set. Three different noises (N1-Wishing machine, N2-Traffic noise, N3-Electric fan) are used as noise signals for the speech samples. At first, it began with a small speech signal, like someone speaking a little phrase from the MCV dataset. A clean speech set with three samples of noise, which are mixed and added to the clean speech signal with lower levels of SNR value fixed at (−5, 0, 5) dBs, where the noise source takes equal weights. Table 1. provides additional specific information about the network's parameters.

Table 1. Implemented DNN model configuration

Parameter	Range
Hidden layers in FCN	• 2 Layers (1024 neurons + batch Normalization Layer + RELU Activation Function in each Layer). • 1 Layer (129) with regression Layer. • Number of Weights is 2237440.
Hidden layers in CNN	• 6 layers (([9 8] × 18, [5 1] × 30, [9 1] × 8, [9 1] × 18, [5 1] × 30, [9 1] × 8) with Stride [1 100]) + batch Normalization Layer + RELU Activation Function in each Layer).
Validation data	1%
Epochs	3
Mini batch size	128
Learning Option	Adam
Shuffle	never
Initial Learn Rate	1×10^{-5}
Learn Rate Drop Factor	0.9
Max Iteration	27549
Iteration per Epoch	9183
Input Nodes	129 × 8
Output Nodes	1

All speech signal proceeds are sampled at (8 kHz) with an input frequency of (48 kHz) and a weighted segment of eight. The training data takes about a half hour to process. The sentences are resampled to 48 kHz and extracted with the STFT features with the frame length set to 5 seconds. A model configuration process by using two methods of deep learning, fully connected neural (FCN) and convolutional neural network (CNN), and a comparison of the results obtained by the two methods to get the best enhancement of speech signal. After the training process, the network can model three classes of noise suitable for denoising the speech signal.

4 Results and discussion

The training is done based on various parameters to filter out the undesired noise while keeping the quality of the speech signal. The most important parameters are short-time objective intelligibility (STOI), signal-to-noise ratio (SNR), mean square error (MSE), peak signal-to-noise ratio (PSNR), and maximum correlation (MC). The results of speech signal enhancement (MSE, PSNR, STOI and maximum colouration) based on CFN and CNN with SNR (−5, 0, 5) dBs shown in Tables 2–4 below.

Table 2. Results of speech signal enhancement (MSE, PSNR, STOI and maximum colouration) based on CFN and CNN with SNR (5 dB)

SNR = 5 dB									
SNR	**Speech Signal**	**MSE (10^{-4})**		**PSNR**		**STOI**		**Max. Colouration**	
		CNN	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**
N1	S1	12	14	28.11	27.33	0.89	0.87	97.04	97.68
	S2	10	13	26.79	26.02	0.90	0.88	96.72	96.98
	S3	14	17	28.37	27.59	0.91	0.90	96.84	97.58
N2	S1	9.5	11	29.72	29.18	0.88	0.86	97.76	97.53
	S2	8.7	9.3	28.76	25.50	0.91	0.89	97.54	97.01
	S3	10	13	29.77	28.39	0.92	0.91	97.60	98.32
N3	S1	1.67	4.18	37.28	33.29	0.97	0.94	99.18	98.69
	S2	1.87	3.98	35.47	32.18	0.96	0.94	99.15	98.32
	S3	2.74	6.44	35.38	31.68	0.97	0.95	98.94	98.45

From the results in Tables 1–3, it is shown how a fully connected neural (FCN) and convolutional neural network (CNN) can obtain similar results in the performance of parameter intelligibility, while in FCN the value of mean square error (MSE) is less and better than CNN with SNR (−5, 0, 5) dBs for three types of noise chosen. It is also noted that the CNN is better than FCN in term of maximum coloration in the lower level of SNR value fixed at (0, 5) dBs.

Table 3. Results of speech signal enhancement (MSE, PSNR, STOI and maximum colouration) based on CFN and CNN with SNR (0 dB)

SNR = −5 dB									
SNR	**Speech Signal**	**MSE (10^{-4})**		**PSNR**		**STOI**		**Max. Colouration**	
		FCN	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**
N1	S1	305	430	14.66	13.77	0.600	0.599	64.33	65.74
	S2	294	401	13.50	12.16	0.606	0.590	60.64	61.70
	S3	376	501	14.01	12.78	0.635	0.633	61.97	62.11
N2	S1	125	103	18.53	19.37	0.589	0.652	60.67	71.45
	S2	116	94	17.55	18.46	0.629	0.696	64.83	71.18
	S3	153	124	17.92	18.84	0.665	0.704	69.20	72.22
N3	S1	13	18	20.46	26.94	0.874	0.855	95.78	99.52
	S2	12	17	27.32	25.98	0.882	0.852	95.67	94.49
	S3	16	23	27.83	26.22	0.897	0.877	96.60	94.60

Table 4. Results of speech signal enhancement (MSE, PSNR, STOI and maximum colouration) based on CFN and CNN with SNR (−5 dB)

SNR = 0 dB									
SNR	**Speech Signal**	**MSE (10^{-4})**		**PSNR**		**STOI**		**Max. Colouration**	
		CNN	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**
N1	S1	78	77	20.56	20.62	0.790	0.784	92.44	91.65
	S2	72	70	19.63	19.74	0.793	0.790	89.79	88.97
	S3	94	86	20.03	20.45	0.814	0.818	89.77	90.44
N2	S1	37	54	24.77	24.14	0.780	0.783	91.34	91.85
	S2	34	30	22.88	23.43	0.805	0.809	91.25	90.89
	S3	42	39	23.51	23.91	0.832	0.824	91.80	94.45
N3	S1	3.9	6.54	33.51	31.35	0.937	0.912	98.38	97.91
	S2	3.6	6.15	32.56	30.30	0.942	0.920	98.61	97.33
	S3	5.4	9.26	32.40	0.10	0.953	0.934	99.08	97.87

Figures 3–5 give more details about the test results of using the FCN and CNN in speech enhancement by measuring the mean square error (MSE) for three sounds (S1, S2, and S3) for different input SNR at (−5, 0, 5) dBs, with three selected noises (N1-Wishing machine, N2-Traffic noise, and N3-Electric fan).

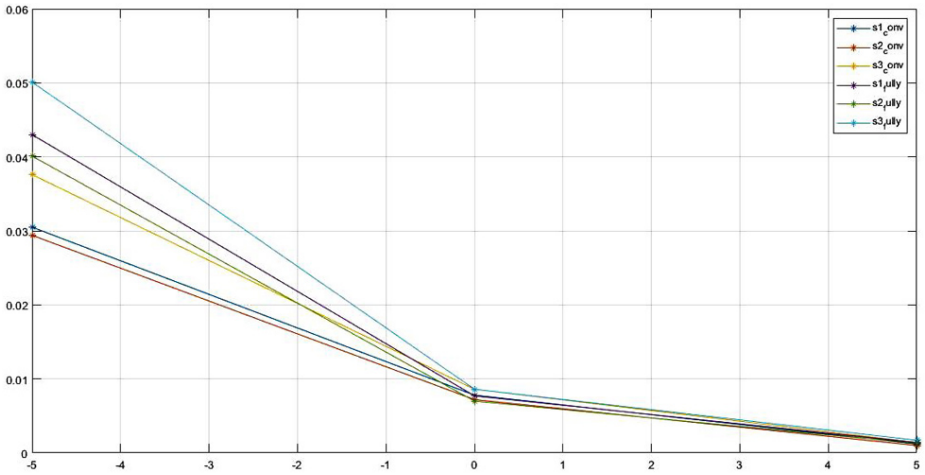

Fig. 3. Mean square error (MSE – Y-axis) for three sounds (S1, S2, and S3) for different input SNR (X-axis) at (−5, 0, 5) dBs for N1-Wishing machine

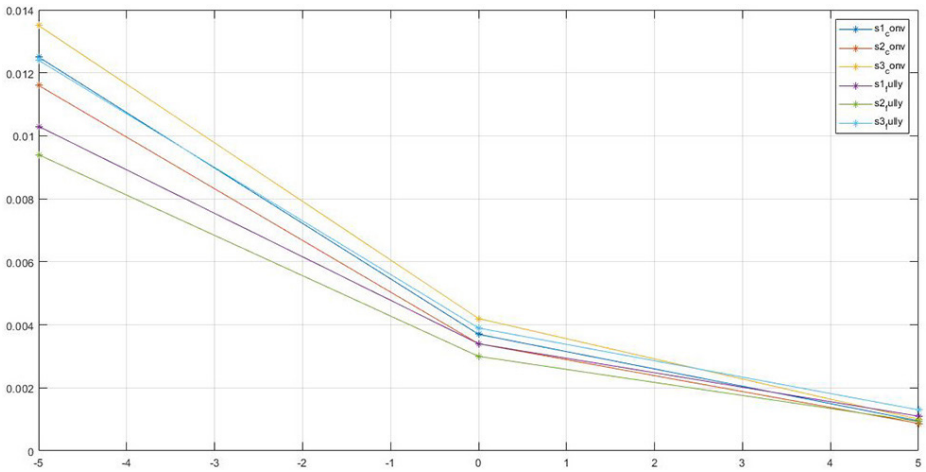

Fig. 4. Mean square error (MSE – Y-axis) for three sounds (S1, S2, and S3) for different input SNR (X-axis) at (−5, 0, 5) dBs for N2-Traffic noise

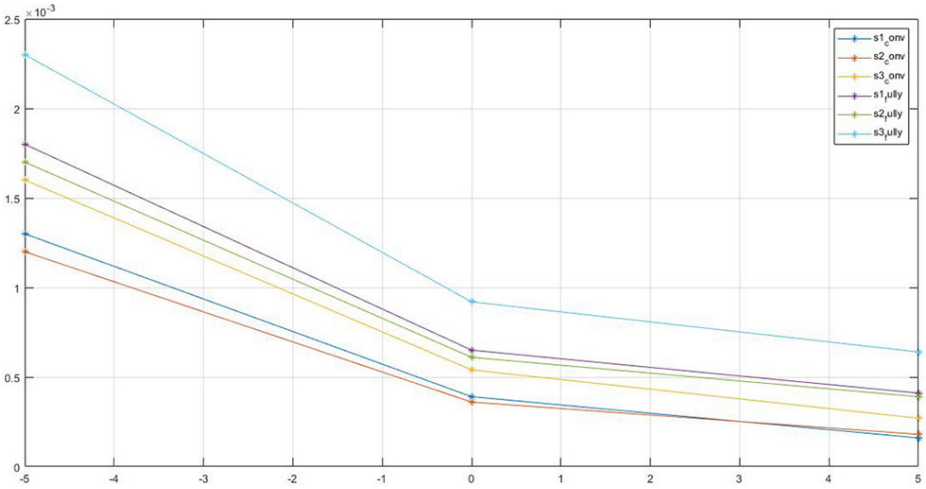

Fig. 5. Mean square error (MSE – Y-axis) for three sounds (S1, S2, and S3) for different input SNR (X-axis) at (−5, 0, 5) dBs for N3-Electric fan

Figures 6–8 are present the speech objective Intelligibility for FCN and CNN in speech enhancement by determining the STOI for three sounds (S1, S2, and S3) for different input SNR at (−5, 0, 5) dBs, with three selected noises (N1-Wishing machine, N2-Traffic noise, and N3-Electric fan).

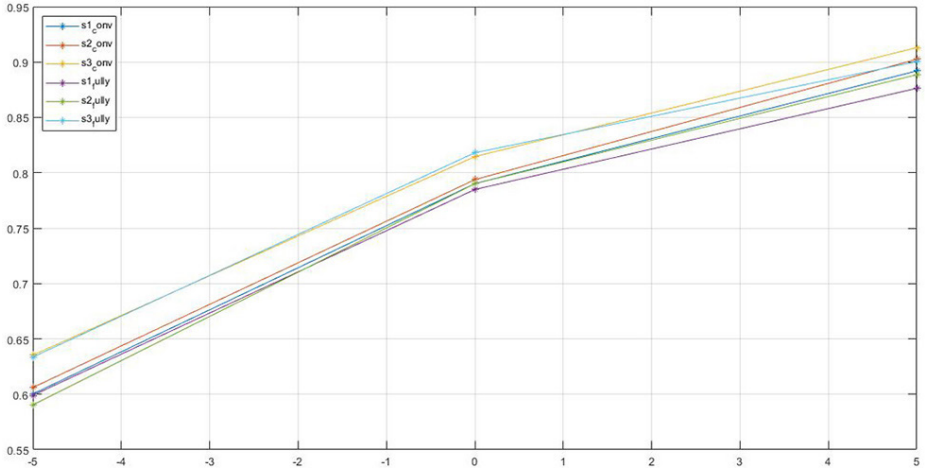

Fig. 6. Intelligibility measure (STOI – Y-axis) for three sounds (S1, S2, and S3) for different input SNR (X-axis) at (−5, 0, 5) dBs for N1-Wishing machine

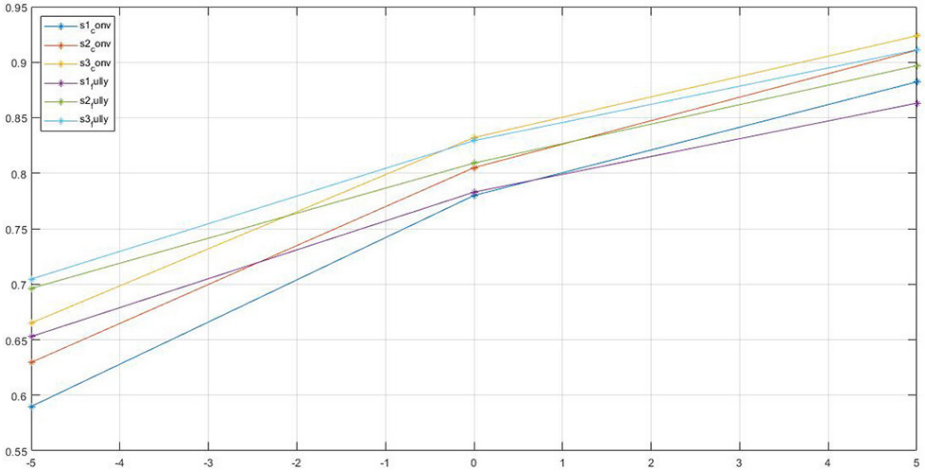

Fig. 7. Intelligibility measure (STOI – Y-axis) for three sounds (S1, S2, and S3) for different input SNR (X-axis) at (−5, 0, 5) dBs for N2-Traffic noise

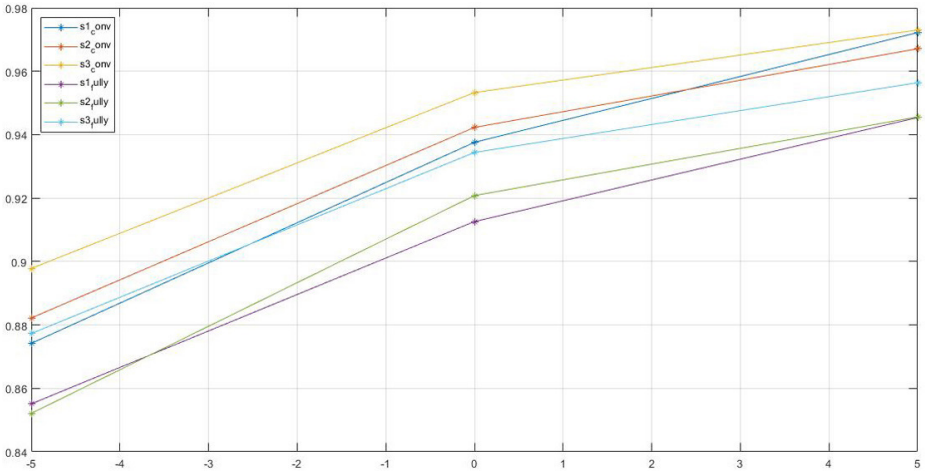

Fig. 8. Intelligibility measure (STOI – Y-axis) for three sounds (S1, S2, and S3) for different input SNR (X-axis) at (−5, 0, 5) dBs for N3-Electric fan

The results are in Table 5. Present the comparison between the run time of a fully connected network (FCN) which has less elapsed time than a convolutional network (CNN), it's clear that FCN is faster by about three times than CNN to process, extract The features, and enhance the speech signal in the system.

Table 5. Variation of validation RSMS and Elapsed time for FCN and CNN with SNR (5, 0, −5) dB

SNR = 5 dB				
Noise	**FCN**		**CNN**	
	Validation RSMS	**Elapsed Time**	**Validation RSMS**	**Elapsed Time**
N1 (Wishing machine)	2.446	52 min, 39 sec	1.998	175 min, 58 sec
N2 (Traffic noise)	2.584	48 min, 54 sec	2.177	185 min, 12 sec
N3 (Electric fan)	2.242	48 min, 28 sec	1.653	177 min, 30 sec
SNR = 0 dB				
N1 (Wishing machine)	3.588	48 min, 48 sec	3.494	184 min, 24 sec
N2 (Traffic noise)	4.060	51 min, 34 sec	4.278	180 min, 52 sec
N3 (Electric fan)	2.679	47 min, 57 sec	2.164	180 min, 56 sec
SNR = −5 dB				
N1 (Wishing machine)	5.978	49 min, 4 sec	6.655	186 min, 25 sec
N2 (Traffic noise)	6.751	49 min, 41 sec	7.730	180 min, 28 sec
N3 (Electric fan)	3.680	44 min, 44 sec	3.160	180 min, 54 sec

5 Conclusion

Using deep learning for speech enhancement has recently attracted attention due to its effective and accurate performance. Two techniques of noise removal are done by modelling the signal speech based on the convolution neural network (CNN) and the fully connected network (FCN). A clean speech set with three samples of noise, which are mixed and added to the clean speech signal with a lower level of SNR value fixed at (−5, 0, 5) dBs. The noise source takes an equal weight. To evaluate the viability of the suggested strategy, five performance metrics were taken into account. These metrics consist of short-time objective intelligibility (STOI), signal-to-noise ratio (SNR), mean square error (MSE), peak signal-to-noise ratio (PSNR), and maximum correlation (MC). In order to quantify the subjective speech quality, evaluation metrics are obtained by contrasting the improved speech with the comparable clean speech. Overall, the convolution neural network (CNN) has better performance than the fully connected network (FCN) concerning the mean square error (MSE) and short-time objective intelligibility (STOI), while the FCN is three times faster than CNN, and it has a smaller elapsed time compared with CNN.

6 Acknowledgements

The authors express their gratitude to Mustansiriyah University in Iraq for supporting this study.

7 References

[1] S. A. Shahriyar, M. A. H. Akhand, N. Siddique, and T. Shimamura, "Speech enhancement using convolutional denoising autoencoder," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–5, 2019. https://doi.org/10.1109/ECACE.2019.8679106

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999. https://doi.org/10.1109/89.748118

[3] H. H. Nuha and A. A. Absa, "Noise reduction and speech enhancement using wiener filter," in *2022 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 177–180, 2022. https://doi.org/10.1109/ICoDSA55874.2022.9862912

[4] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Comput Sci*, vol. 54, pp. 574–584, 2015. https://doi.org/10.1016/j.procs.2015.06.066

[5] J. S. Lim and A. v Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979. https://doi.org/10.1109/PROC.1979.11540

[6] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 845–856, 2005. https://doi.org/10.1109/TSA.2005.851927

[7] N. Lyubimov and M. Kotov, "Non-negative matrix factorization with linear constraints for single-channel speech enhancement," *arXiv preprint arXiv:1309.6047*, 2013. https://doi.org/10.21437/Interspeech.2013-132

[8] M. Hasannezhad, H. Yu, W.-P. Zhu, and B. Champagne, "PACDNN: A phase-aware composite deep neural network for speech enhancement," *Speech Commun*, vol. 136, pp. 1–13, 2022. https://doi.org/10.1016/j.specom.2021.10.002

[9] D. Ribas, A. Miguel, A. Ortega, and E. Lleida, "Wiener filter and deep neural networks: A well-balanced pair for speech enhancement," *Applied Sciences*, vol. 12, no. 18, p. 9000, 2022. https://doi.org/10.3390/app12189000

[10] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," *arXiv preprint arXiv:1605.02427*, 2016. https://doi.org/10.21437/Interspeech.2016-88

[11] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017. https://doi.org/10.21437/Interspeech.2017-1428

[12] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by noisy2noisy signal mapping," *Applied Acoustics*, vol. 172, p. 107631, 2021. https://doi.org/10.1016/j.apacoust.2020.107631

[13] J. Singh and K. Kaur, "Speech enhancement for Punjabi language using deep neural network," in *2019 International Conference on Signal Processing and Communication (ICSC)*, pp. 202–204, 2019. https://doi.org/10.1109/ICSC45622.2019.8938309

[14] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," 2015. https://doi.org/10.21437/Interspeech.2015-536

[15] R. E. Zezario, J. W. C. Sigalingging, T. Hussain, J.-C. Wang, and Y. Tsao, "Comparative study of masking and mapping based on hierarchical extreme learning machine for speech enhancement," in *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–2, 2019. https://doi.org/10.1109/ISPACS48206.2019.8986352

[16] C. Zheng, X. Zhang, M. Sun, Y. Xing, and H. Shi, "Throat microphone speech enhancement via progressive learning of spectral mapping based on lstm-rnn," in *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, pp. 1002–1006, 2018. https://doi.org/10.1109/ICCT.2018.8600157

[17] S. Chakrabarty, D. Wang, and E. A. P. Habets, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 476–480, 2018. https://doi.org/10.1109/IWAENC.2018.8521346

[18] S. Chakrabarty and E. A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J Sel Top Signal Process*, vol. 13, no. 4, pp. 787–799, 2019. https://doi.org/10.1109/JSTSP.2019.2911401

[19] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5039–5043, 2018. https://doi.org/10.1109/ICASSP.2018.8462068

[20] A. Pandey and D. Wang, "Learning complex spectral mapping for speech enhancement with improved cross-corpus generalization.," in *INTERSPEECH*, pp. 4511–4515, 2020. https://doi.org/10.21437/Interspeech.2020-2561

[21] Y. Xie, T. Arildsen, and Z.-H. Tan, "Complex Recurrent Variational Autoencoder for Speech Enhancement," *arXiv preprint arXiv:2204.02195*, 2022.

[22] X. Cui, Z. Chen, and F. Yin, "Multi-objective based multi-channel speech enhancement with BiLSTM network," *Applied Acoustics*, vol. 177, p. 107927, 2021. https://doi.org/10.1016/j.apacoust.2021.107927

[23] Z.-Q. Wang and D. Wang, "All-neural multi-channel speech enhancement.," in *Interspeech*, pp. 3234–3238, 2018. https://doi.org/10.21437/Interspeech.2018-1664

[24] H. Chung, E. Plourde, and B. Champagne, "A supervised multi-channel speech enhancement algorithm based on bayesian nmf model," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 221–225, 2018. https://doi.org/10.1109/GlobalSIP.2018.8646634

[25] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 6–12.

[26] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016. https://doi.org/10.21437/Interspeech.2017-1465

[27] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5756–5760, 2019. https://doi.org/10.1109/ICASSP.2019.8683423

[28] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018. https://doi.org/10.1109/SLT.2018.8639585

[29] Y.-J. Li, S.-S. Wang, Y. Tsao, and B. Su, "MIMO speech compression and enhancement based on convolutional denoising autoencoder," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1245–1250, 2021.

[30] H. Reddy, A. Kar, and J. Østergaard, "Performance analysis of low complexity fully connected neural networks for monaural speech enhancement," *Applied Acoustics*, vol. 190, p. 108627, 2022. https://doi.org/10.1016/j.apacoust.2022.108627

[31] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. le Roux, "STFT-Domain neural speech enhancement with very low algorithmic latency," *arXiv preprint arXiv:2204.09911*, 2022. https://doi.org/10.1109/TASLP.2022.3224285

[32] R. Ram and M. N. Mohanty, "Deep neural network based speech enhancement," in *Cognitive Informatics and Soft Computing*, Springer, pp. 281–287, 2019. https://doi.org/10.1007/978-981-13-0617-4_27

[33] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5059–5063, 2018. https://doi.org/10.1109/ICASSP.2018.8462040

[34] D. H. Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, p. 2088, 2006. https://doi.org/10.4249/scholarpedia.2088

[35] H. ALRikabi and H. Tauma, "Secure chaos of 5G wireless communication system based on IOT applications," *International Journal of Online & Biomedical Engineering*, vol. 18, no. 12, 2022. https://doi.org/10.3991/ijoe.v18i12.33817

[36] H. TH. Salim, "Enhanced data security of communication system using combined encryption and steganography," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 16, pp. 144–157, 2021. https://doi.org/10.3991/ijim.v15i16.24557

[37] N. Alseelawi and H. T. Hazim," A novel method of multimodal medical image fusion based on hybrid approach of NSCT and DTCWT," *International Journal of Online & Biomedical Engineering*, vol. 18, no. 3, 2022. https://doi.org/10.3991/ijoe.v18i03.28011

[38] R. Ardila et al., "Common voice: A massively-multilingual speech corpus," arXiv preprint arXiv:1912.06670, 2019.

8 Authors

Ban M. Alameri: A PhD student at department of Telecommunication Engineering, Malaga University, Malaga, Spain. She is also a lecturer and researcher at department of Electrical Engineering, Faculty of Engineering, Mustansiriyah University, Baghdad, Iraq. Her interesting research in communication system, DSP, and interferences issues in the electrical system. (email: ban.alameri@uomustansiriyah.edu.iq, https://orcid.org/0000-0001-8177-0506).

Inas Jawad Kadhim: Dr. lecturer in the department of Electrical Power Engineering in the Electrical Engineering Technical College, Middle Technical University. Her research interests include information hiding, image processing and DSP. (email: inasjk@mtu.edu.iq).

Suha Qasim Hadi: Dr. lecturer, and researcher at department of Electrical Engineering, Faculty of Engineering, Mustansiriyah University, Baghdad, Iraq, from 2009 until now. Her research interests focus on wireless communication, 5G network. (email: drsuhaqasim@uomustansiriyah.edu.iq, https://orcid.org/0000-0002-9458-4865).

Ali F. Hassoon: A researcher in department of Electrical Engineering, Faculty of Engineering, Mustansiriyah University, Baghdad, Iraq. His research interests in Electronic, and communication, DSP. (email: alifattah@uomustansiriyah.edu.iq).

Mustafa M. Abd: A researcher department of Electrical Engineering, Faculty of Engineering, Mustansiriyah University, Baghdad, Iraq. His research interests is Power management systems, Artificial intelligence and DSP. (email: mustafa_albadry@uomustansiriyah.edu.iq).

Prashan Premaratne: Dr. Senior Lecturer Faculty of Engineering and Information Sciences, School of Electrical, Computer and Telecommunications Engineering, Wollongong, Australia from 2020 – present. His research interests in Computer Vision, Image Processing, Radar Signal Processing. (email: prashan@uow.edu.au).

Article submitted 2022-12-20. Resubmitted 2023-01-23. Final acceptance 2023-01-23. Final version published as submitted by the authors.