Acoustic Signal Enhancement Using Deep Neural Networks

Shibani Kar
Department of Electronics and Communication Engineering, Sambalpur University Institute of Information Technology, Sambalpur University, Jyoti Vihar, Burla, Odisha, India
shibanikar@gmail.com (corresponding author)

Vishwajeet Mukherjee
Department of Basic Science and Humanities, Sambalpur University Institute of Information Technology, Sambalpur University, Jyoti Vihar, Burla, Odisha, India
vishwajeet10@gmail.com

Received: 14 February 2025 | Revised: 8 May 2025 | Accepted: 12 May 2025

 DOI: https://doi.org/10.48084/etasr.10571

ABSTRACT

The presence of background noise in acoustic signals, such as speech, audio, and sound signals, degrades listening quality and causes hearing fatigue to the listener. Standard methods offer better signal enhancement under high SNR conditions. Deep neural networks employed in image processing and speech recognition have demonstrated significant performance improvements. This motivates the usage of deep neural networks for denoising speech signals corrupted with multiple noises under low SNR conditions (0 dB). This study applied two different types of deep neural networks, convolutional neural networks and deep generative networks, to remove background noise from speech signals under low SNR conditions. The noise reduction networks were trained to estimate the noise signal present, which was then subtracted to obtain the denoised speech signal. Two convolutional neural network architectures, the UNet and the Convolutional Encoder-Decoder network (CED), and two deep generative networks, Vector Quantized Variational Autoencoders (VQVAE) and Variational Autoencoders (VAE), were trained on STFT magnitude features of noisy signal frames. Four objective quality measures were used to determine the quality of the enhanced speech, namely Perceptual Evaluation of Speech Quality (PESQ), Short Time Objective Intelligibility (STOI), Segmental Signal to Noise Ratio (SSNR), and improvement in SNR. Spectral subtraction and logMMSE methods were used to evaluate the performance of these networks in two datasets. The results of the comparative analysis support the superiority of CED for signal denoising and enhancement of speech signals for multiple noises under low SNR conditions, with a much smaller number of model parameters compared to other methods for both seen and unseen noise conditions.

Keywords-background noise estimation; speech signal; deep neural networks; deep generative networks; speech enhancement

I. INTRODUCTION

Acoustic signals corrupted with background noise are difficult to comprehend for speech-based applications. Background noise consisting of multiple speaker voices reduces the intelligibility of the speech signal and causes listening fatigue [1]. The estimation of this type of background noise is termed babble noise and is very difficult to estimate [1, 2]. Traditional methods such as spectral subtraction [3] and logMMSE [4] estimate the noise present in the signal. The spectral subtraction method estimates the noise signal from the silent regions present in the speech signal. The estimated noise is subtracted from the observed (noisy) signal to obtain a clean signal estimate [3]. In [5], signal enhancement was performed in the modulation domain using Short-Time Fourier Transform (STFT) features of the speech signal. The noise signal estimated in the STFT domain was subtracted from the STFT of the noisy signal to receive the clean STFT signal [5]. In [6], a gain function was applied using a priori SNR and a posteriori SNR. The gain function was applied to the noisy signal to provide a clean estimate of the signal [6-8]. In [9-10], a comparative analysis of traditional speech enhancement methods was performed on a common speech dataset. This analysis inferred that not all traditional methods are suitable for the enhancement of signals for all types of noise and noise conditions. The analysis showed that traditional methods improve the quality of denoised speech in a high SNR environment. However, the improvement in speech intelligibility is much lower in low-SNR environments.

Nowadays, deep neural networks offer better results in processing image signals, denoising image signals, and speech recognition. Deep neural networks are capable of modeling complex signals. Deep neural networks can be trained with

large data, which helps in increasing generalizability. The network is trained with speech data to learn the complexity of the speech signal and applied to signal estimation applications. In [11], deep denoising autoencoders were proposed to enhance speech signals. This method improves the Perceptual Quality (PESQ) of the denoised speech but offers less improvement in the average intelligibility, i.e., the STOI score for the enhanced speech signal. In [12], a fully convolutional deep neural network was trained with noisy speech samples but offered less improvement in signal intelligibility and quality for babble-corrupted speech signals under low SNR conditions. In [13], a Kalman filter was proposed to enhance speech signals, improving perceptual quality and intelligibility but offering less improvement in SSNR scores. In [14], a complex ideal ratio mask was proposed to enhance speech signals, where the target mask was calculated from the real and imaginary parts of clean and noisy speech signals and used as the target function. The network estimated the CIRM from the input signal and compared it with the CIRM mask of the target function. The estimated mask was then applied to the noisy signal to obtain the denoised speech signal.

In [1], a fully convolutional network was implemented in an encoder-decoder framework to denoise speech signals. This study suggested that the network size and parameters are important for the selection of a suitable network and its hardware implementation. In [15], mapping-based training was proposed for deep neural networks, where the clean speech Time-Frequency (T-F) distribution was used as the training target. The hidden layers of the network capture the contextual information of the T-F distribution of the input speech signal to predict the estimate of the clean T-F distribution. This network offered better SSNR and PESQ scores for different noise types and SNR conditions. The main objective was to improve the signal quality, but this work did not provide information regarding the intelligibility of the enhanced signal. In [16], deep neural networks for mapping and masking-based speech separation methods were reviewed. In [17], a Generative Adversarial Network (GAN) was presented to enhance speech signals in the waveform domain. In [18], deep neural networks used to remove babble and factory noise under low SNR conditions were reviewed. The networks were trained to estimate different training targets consisting of different masks, such as IBM and IRM, clean speech estimation, and the combination of masks and clean speech samples as the estimated training target functions. The latter are known as hybrid methods, as they combine different masks and clean samples as training objectives. Conventional methods, such as the Wiener filter and estimation of IBM and IRM masks by DNN, offer fewer improvements in signal quality and intelligibility under low SNR for babble-corrupted speech signals, whereas hybrid methods using masking and mapping-based methods perform better than conventional ones.

The improvement in signal parameters at low SNRs is a very challenging task. Not all methods can improve speech quality and intelligibility under low SNR conditions. The results indicate that improving signal parameters under low SNR conditions is a challenging task. Therefore, a comparative analysis of speech denoising methods under low SNR conditions is required. This study employed two types of deep neural networks, convolutional and deep generative networks, to denoise speech signals corrupted with multiple noises under different SNR conditions.

II. DESCRIPTION OF THE PROPOSED NETWORK

This study implemented a speech enhancement system using deep neural networks in an encoder-decoder framework. As the STFT of the signals captures important contextual information, the STFT magnitudes are used to train the network. During training, the STFT magnitude of the noisy signal is applied as input, and the STFT magnitude of the noise signal is used as the target function. In this way, the network learns to estimate the noise signal. During testing, the predicted noise signal estimate is subtracted from the noisy signal to obtain the denoised signal. The phase of the noisy signal is merged with the STFT of the denoised signal during the reconstruction of the time domain signal. The inverse STFT is applied to the denoised signal to convert it to the time domain. Figure 1 shows the data flow model.

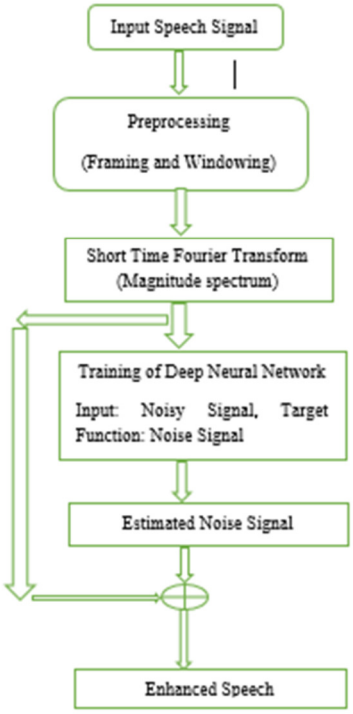

Fig. 1. Dataflow model for speech enhancement.

$$X(t,f) = S(t,f) + N(t,f) \quad (1)$$

$$N(t,f) = X(t,f) - S(t,f) \quad (2)$$

The $X(t,f)$, $S(t,f)$, and $N(t,f)$ are the noisy speech, clean, and noise samples in the T-F domain, respectively. The noise signal frames $N(t,f)$ are obtained from subtracting the T-F distribution frames of clean speech samples from the noisy speech signal T-F frames. $N(t,f)$ is used as the target function during network training. The trained network, when applied with test noisy samples as input, predicts the estimated noisy T-F frames, i.e. $N'(t,f)$. This estimated noisy frame is subtracted

from the noisy signal to obtain the denoised signal $S'(t,f)$ as given in:

$$S'(t,f) = X(t,f) - N'(t,f) \quad (3)$$

This study employed two deep learning network types, namely, convolutional neural networks and deep generative networks, for speech enhancement. The networks were trained to minimize the loss between the estimated signal and the target function to update their weights. The encoder network converts the signal into useful representations, and the decoder reconstructs the desired signal at the output. The ability of the convolutional neural network to extract the desired signal features and share them with successive layers makes it a suitable choice to implement in the encoder-decoder network framework, whereas the generative networks determine the probabilistic representation of the input to generate the desired signal from the learned distribution.

III. EXPERIMENTAL SET UP

A. Datasets

The networks were trained and tested with speech samples taken from two datasets: NOIZEUS [9, 10] and noisy speech database for training speech enhancement algorithms and TTS models [19]. The training and testing sets include five types of noise signals, namely Restaurant, Car, Street, Babble, and Airport noises, at three SNR levels of 0 dB, 5 dB, and 10 dB. The networks were trained with 450 samples from the NOIZEUS dataset, and tested on 30 samples from [19].

B. Preprocessing and Training

The speech signals are sampled at 8000 Hz and framed for 512 samples. The Hamming window is applied to smooth the framed signal. Previous studies for speech enhancement used the STFT magnitudes of the signals as features to train the network. Therefore, to standardize the comparison with the literature, STFT features of the signals were used for training and testing of the networks. The network was trained with 128 frames of STFT magnitude, having a dimension of (128, 1) selected for the input noisy signal and the target function. The networks were trained for 200 epochs using the Adam optimizer and mean square error as the loss function.

C. Performance Metrics

The performance of the networks was analyzed using four objective metrics: PESQ, STOI, SSNR, and improvement in SNR [1, 20-22]. PESQ measures the improvement in signal quality, with higher values indicating a better signal quality. STOI offers an intelligibility score, with higher values indicating better intelligibility. SSNR measures the average improvement in SNR scores per frame in the time domain, with a range of -10 to 30 dB. The improvement in the SNR score is measured by computing the SNR between clean and noisy speech samples. The SNR measured between clean and denoised audio samples is represented as the output SNR. The difference between the output and input SNRs offers a measure of the improvement in SNR, with higher values indicating better performance.

D. Models

The UNet architecture consists of a stack of four encoder-decoder networks, where each encoder layer is connected to a corresponding decoder layer using skip connections to restore the spatial information of the signal. Each encoder layer consists of two convolution layers, a ReLU layer, and a max pool layer. The decoder network consists of two convolutional layers, an upsample layer, and a ReLU layer. The last layer uses tanh as the activation function [23].

The CED architecture consists of three encoder and decoder networks. Each encoder network consists of three convolutional layers, a ReLU layer, and a max pooling layer. The decoder consists of three convolutional and upsample layers to reconstruct the signal. No skip connections are used in the CED architecture.

The VAE encoder network consists of two convolutional layers, a ReLU activation layer, a flatten layer, and two dense layers to sample the joint distribution of the input signal to obtain the mean and variance parameters. The decoder network consists of two convolutional layers and a ReLU activation layer to reconstruct the signal at the output. The last output is passed through a convolution layer to make a fully convolutional decoder network. KL divergence and mean square error loss are used during network training [24].

The encoder of the VQVAE architecture consists of three convolutional layers, a max pool layer, and a ReLU activation layer to create a codebook consisting of 64 vectors having a dimension of 16. The decoder consists of two convolution layers, a ReLU layer, and an upsample layer to reconstruct the output [25].

IV. RESULT

The results of the four methods were compared with two baseline methods, spectral subtraction and logMMSE. Tables I, II, and III compare the performance of the methods for babble-corrupted speech signals. The results indicate that at low SNR (0 dB), CED and UNet improve the intelligibility (STOI) and SSNR scores for babble-corrupted speech signals compared to the baseline methods (spectral subtraction and logMMSE). The baseline methods improve the PESQ score under low SNR conditions. The proposed CED and UNet models improve the intelligibility score both at low and high SNRs. The CED and UNet models offer higher STOI and SSNR scores for babble-corrupted speech signals compared to the method proposed in [12] under low SNR conditions.

TABLE I. PERFORMANCE ANALYSIS FOR BABBLE NOISE AT 0 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.48	0.62	-2.76	4.06
UNet	1.31	0.62	0.06	3.56
CED	1.37	0.62	-1.73	3.05
VQVAE	1.36	0.61	-2.23	3.47
VAE	1.28	0.57	-4.16	2.24
LogMMSE	1.56	0.60	-2.28	4.76

TABLE II. PERFORMANCE ANALYSIS FOR BABBLE NOISE AT 5 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.73	0.76	0.013	3.42
UNet	1.45	0.72	1.38	-0.95
CED	1.49	0.76	0.22	-1.33
VQVAE	1.45	0.73	-0.74	-0.49
VAE	1.35	0.71	-2.37	-2.03
LogMMSE	1.85	0.73	0.15	3.14

TABLE III. PERFORMANCE ANALYSIS FOR BABBLE NOISE AT 10 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	2.1	0.87	2.96	2.87
UNet	1.58	0.82	2.40	-6.4
CED	1.70	0.85	1.64	-6.21
VQVAE	1.57	0.80	0.52	-5.36
VAE	1.45	0.80	-1.02	-6.87
LogMMSE	2.27	0.84	3.18	2.37

Tables IV, V, and VI show the performance of the methods for car noise-corrupted speech signals. CED improved the intelligibility of the speech signal better than other methods under low and high SNR conditions. At 0 dB SNR, CED improved all four metrics compared to the spectral subtraction method. At 5 and 10 dB, CED offers similar PESQ and SSNR scores compared to baseline methods.

TABLE IV. PERFORMANCE ANALYSIS FOR CAR NOISE AT 0 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.48	0.64	-2.19	5.85
UNet	1.29	0.60	-3.86	2.78
CED	1.43	0.71	-0.47	4.22
VQVAE	1.30	0.62	-1.45	3.99
VAE	1.25	0.59	-3.81	2.65
LogMMSE	1.59	0.64	0.64	6.02

TABLE V. PERFORMANCE ANALYSIS FOR CAR NOISE AT 5 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.73	0.76	0.45	4.76
UNet	1.36	0.70	-2.51	-0.64
CED	1.58	0.81	0.87	-0.85
VQVAE	1.45	0.72	-0.13	-1.17
VAE	1.32	0.69	-2.51	-1.93
LogMMSE	1.92	0.74	1.26	4.70

TABLE VI. PERFORMANCE ANALYSIS FOR CAR NOISE AT 10 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	2.12	0.87	3.48	3.72
UNet	1.46	0.80	-1.03	-4.88
CED	1.71	0.88	2.10	-6.39
VQVAE	1.53	0.80	0.66	-7.51
VAE	1.42	0.79	-1.12	-6.77
LogMMSE	2.27	0.83	3.71	2.86

Tables VII, VIII, and IX show the performance of the methods on airport noise-corrupted speech signals. CED achieved higher STOI scores under both low and high SNR conditions. CED also achieved higher SSNR scores compared to the baseline methods at 0 dB SNR.

TABLE VII. PERFORMANCE ANALYSIS FOR AIRPORT NOISE AT 0 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.49	0.66	-2.24	4.35
UNet	1.30	0.65	-1.07	2.56
CED	1.40	0.70	-0.98	3.80
VQVAE	1.35	0.65	-2.10	3.46
VAE	1.26	0.62	-3.62	2.52
LogMMSE	1.57	0.63	-1.90	5.04

TABLE VIII. PERFORMANCE ANALYSIS FOR AIRPORT NOISE AT 5 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.77	0.78	0.24	3.51
UNet	1.46	0.74	0.19	-0.99
CED	1.56	0.80	0.32	-0.89
VQVAE	1.45	0.74	-0.76	-0.67
VAE	1.38	0.72	-2.21	-1.83
LogMMSE	1.91	0.75	0.58	3.44

TABLE IX. PERFORMANCE ANALYSIS FOR AIRPORT NOISE AT 10 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	2.14	0.88	3.45	2.76
UNet	1.56	0.82	1.41	-5.25
CED	1.69	0.87	1.75	-6.13
VQVAE	1.55	0.82	0.62	-6.28
VAE	1.49	0.81	-0.71	-6.58
LogMMSE	2.24	0.84	3.56	2.44

Tables X, XI, and XII show the performance of the methods on street noise-corrupted speech signals. CED achieved higher STOI scores under both low and high SNR conditions. CED improved the signal perceptual quality and intelligibility both under low and high SNRs. Tables XIII, XIV, and XVI show the performance of the methods on restaurant noise-corrupted speech signals. CED achieved higher STOI scores under both low and high SNR conditions. CED also achieved higher STOI, SSNR, and SNR scores at 0 dB compared to the baseline methods. Therefore, CED improves perceptual quality and intelligibility both under low and high SNRs.

TABLE X. PERFORMANCE ANALYSIS FOR STREET NOISE AT 0 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.45	0.65	-2.13	4.43
UNet	1.17	0.50	-0.55	1.44
CED	1.41	0.69	-0.78	3.64
VQVAE	1.30	0.61	-1.90	3.29
VAE	1.25	0.60	-3.74	2.65
LogMMSE	1.54	0.62	-1.73	4.72

TABLE XI. PERFORMANCE ANALYSIS FOR STREET NOISE AT 5 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.73	0.76	0.48	3.68
UNet	1.33	0.63	0.36	-2.47
CED	1.53	0.79	0.48	-1.28
VQVAE	1.39	0.71	-0.40	-2.61
VAE	1.33	0.70	-2.47	-1.85
LogMMSE	1.75	0.73	0.54	3.14

TABLE XII. PERFORMANCE ANALYSIS FOR STREET NOISE AT 10 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	2.14	0.87	3.88	2.91
UNet	1.52	0.75	1.45	-7.02
CED	1.72	0.87	1.96	-6.83
VQVAE	1.56	0.79	0.78	-6.28
VAE	1.46	0.79	-1.12	-6.63
LogMMSE	2.19	0.85	3.9	2.27

TABLE XIII. PERFORMANCE ANALYSIS FOR RESTAURANT NOISE AT 0 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.50	0.66	-2.53	3.06
UNet	1.33	0.66	-0.67	3.01
CED	1.40	0.71	-0.95	3.52
VQVAE	1.36	0.66	-2.24	3.41
VAE	1.29	0.63	-3.74	2.61
LogMMSE	1.56	0.61	-2.6	3.10

TABLE XIV. PERFORMANCE ANALYSIS FOR RESTAURANT NOISE AT 5 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	1.72	0.79	0.13	2.46
UNet	1.43	0.79	0.65	-1.2
CED	1.49	0.82	0.43	-0.91
VQVAE	1.42	0.76	-0.83	-0.76
VAE	1.33	0.74	-2.24	-1.90
LogMMSE	1.77	0.76	0.09	2.45

TABLE XV. PERFORMANCE ANALYSIS FOR RESTAURANT NOISE AT 10 DB SNR

Method	PESQ	STOI	SSNR	SNR
Spectral subtraction	2.12	0.89	3.33	2.18
UNet	1.58	0.82	1.92	-7.11
CED	1.68	0.89	1.76	-6.07
VQVAE	1.62	0.82	0.39	-6.10
VAE	1.47	0.82	-0.86	-6.62
LogMMSE	2.18	0.85	3.3	1.70

Table XVI shows the performance of methods on the dataset in [19]. CED improves the perceptual quality and intelligibility of the signal, offering also a higher SSNR score compared to the other methods. Table XVII shows that CED achieves better performance with a smaller number of training parameters. Hence, CED is also a memory-efficient method.

TABLE XVI. PERFORMANCE COMPARISON ON VALENTINI DATASET(UNSEEN CONDITIONS)

Method	PESQ	STOI	SSNR (dB)
UNet	2.33	0.84	0.51
CED	2.16	0.82	1.70
VQVAE	2.10	0.81	0.17
VAE	2.16	0.86	-0.60

TABLE XVII. COMPARISON OF MODEL PARAMETERS

Method	Number of parameters
CED	56321
UNet	1941093
VAE	1287317
VQVAE	48913

V. CONCLUSION

This study trained and tested four deep neural networks with T-F representation of speech signals corrupted with five noise types: Babble, Car, Airport, Street, and Restaurant. The noise signal was used as the target function during network training. The networks predict the noise signal, which is then subtracted from the noisy signal to provide a clean signal estimate. The CED network architecture achieved better noise reduction under low SNR conditions compared to other and baseline methods (spectral subtraction and logMMSE). CED achieved higher intelligibility scores under both low and high SNR conditions for each noise type compared to baseline methods. CED improves the intelligibility of speech signals corrupted with babble noise under low SNR conditions. As CED offers better performance with a smaller number of training parameters, it can be identified as a memory-efficient method. The networks were tested with unseen samples from the dataset in [19], and CED achieved higher SSNR scores. All four methods performed well on the dataset in [19]. However, CED performed well in speech enhancement in both matched and unmatched noise conditions.

REFERENCES

[1] S. R. Park and J. W. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," in *Interspeech 2017*, Aug. 2017, pp. 1993–1997, https://doi.org/10.21437/Interspeech.2017-1465.

[2] N. Krishnamurthy and J. H. L. Hansen, "Babble Noise: Modeling, Analysis, and Applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1394–1407, Sep. 2009, https://doi.org/10.1109/TASL.2009.2015084.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979, https://doi.org/10.1109/TASSP.1979.1163209.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985, https://doi.org/10.1109/TASSP.1985.1164550.

[5] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010, https://doi.org/10.1016/j.specom.2010.02.004.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984, https://doi.org/10.1109/TASSP.1984.1164453.

[7] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 497–514, Nov. 1997, https://doi.org/10.1109/89.641296.

[8] 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, Mar. 1999, https://doi.org/10.1109/89.748118.

[9] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, Sep. 2007, https://doi.org/10.1121/1.2766778.

[10] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, Jul. 2007, https://doi.org/10.1016/j.specom.2006.12.006.

[11] A. Azmat, I. Ali, W. Ariyanti, M. G. L. Putra, and T. Nadeem, "Environmental Noise Reduction based on Deep Denoising Autoencoder," *Engineering, Technology & Applied Science Research*, vol. 12, no. 6, pp. 9532–9535, Dec. 2022, https://doi.org/10.48084/etasr.5239.

[12] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by Noisy2Noisy signal mapping," *Applied Acoustics*, vol. 172, Jan. 2021, Art. no. 107631, https://doi.org/10.1016/j.apacoust.2020.107631.

[13] V. Srinivasarao and U. Ghanekar, "Speech enhancement - an enhanced principal component analysis (EPCA) filter approach," *Computers & Electrical Engineering*, vol. 85, Jul. 2020, Art. no. 106657, https://doi.org/10.1016/j.compeleceng.2020.106657.

[14] D. S. Williamson and D. Wang, "Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017, https://doi.org/10.1109/TASLP.2017.2696307.

[15] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014, https://doi.org/10.1109/LSP.2013.2291240.

[16] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018, https://doi.org/10.1109/TASLP.2018.2842159.

[17] S. Pascual, J. Serrà, and A. Bonafonte, "Time-domain speech enhancement using generative adversarial networks," *Speech Communication*, vol. 114, pp. 10–21, Nov. 2019, https://doi.org/10.1016/j.specom.2019.09.001.

[18] A. Azarang and N. Kehtarnavaz, "A review of multi-objective deep learning speech denoising methods," *Speech Communication*, vol. 122, pp. 1–10, Sep. 2020, https://doi.org/10.1016/j.specom.2020.04.002.

[19] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models." University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017, https://doi.org/10.7488/DS/2117.

[20] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008, https://doi.org/10.1109/TASL.2007.911054.

[21] X. Dong and D. S. Williamson, "Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 3348–3359, Nov. 2020, https://doi.org/10.1121/10.0002702.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, Mar. 2010, pp. 4214–4217, https://doi.org/10.1109/ICASSP.2010.5495701.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, pp. 234–241.

[24] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019, https://doi.org/10.1561/2200000056.

[25] A. Van Den Oord, O. Vinyals, and Koray Kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.