An Integrated Deep Learning Model for Concurrent Speech Dereverberation and Denoising

Vijay M. Mane [1,*], Seema S. Arote [1], and Shakil A Shaikh [2]

[1] Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Technology, Pune, India
[2] Department of Electronics and Computer Engineering, Pravara Rural Engineering College, Loni, India
Email: vijay.mane@vit.edu (V.M.M.); seema.arote@gmail.com (S.S.A.); shaikhshakil1968@gmail.com (S.A.S.)
*Corresponding author

***Abstract*—Speech is most likely the simplest and efficient type of human-human communication, as well as the most intuitive and effective way of human-machine interaction. Human voice is often damaged in real-world contexts by both reverberation and noise from the surroundings, which has a detrimental impact on speech intelligibility and quality. In terms of denoising, a model-based approach has been thoroughly researched, and several practical solutions have been created. In comparison, study on dereverberation has been sparse. Significant advances have been achieved in the study of a model-based strategy for dereverberation. The resultant approach may be used to any deep neural network that provides masks in the time-frequency domain with just a few extra variables that can be trained and an overhead of computation that is low for state-of-the-art neural networks. A deep learning-based approach in this article is developed that eliminates early reverberations, late reverberations, and noise from speech signals in order to enhance speech signal quality. The method is tested using data from three simulated rooms—a conference room, a seminar hall, and a room from reference paper number seven—with Reverberation Time (RT_{60}) of 0.3 s and variety of noise like Additive White Gaussian Noise (AWGN), realistic noise such as babble, restaurant and a variety of signal-to-noise ratio values. The proposed technique outperforms baseline multichannel dereverberation and denoising algorithms as well as a cutting-edge multichannel dereverberation and denoising algorithm, resulting in a considerable improvement.**

***Keywords*—deep learning, dereverberation, denoising, Room Impulse Response (RIR)**

I. Introduction

Speech is most certainly the most basic and efficient method of communication between humans, and it is also probably the more natural and effective form of interaction between humans and machines [1]. In hands-free communication, speech is captured using only one microphone or a cluster of microphones set up at various locations around the room.

Multiple microphones capture more than just the intended voice signal; reverberation, background noise, and other interferences are also included in the recorded audio [2]. However, speech in rooms is degraded by acoustic reverberation as well as ambient noise. Room reverberation is one of the two primary sources of speech deterioration (the other being background noise), therefore there is a rising need for voice dereverberation in different speech processing and communication applications. Disturbing sounds, on the other hand, often disrupt genuine speech signals and undermine the efficacy of information transmission in real life [3]. Reverberation occurs when an audio signal travels from its origin to many recording devices through multiple paths. Because of the many reflections, the received sound (e.g., a distant microphone or a listener) lasts even after the originating sound ends. The combination of direct transmitted and reflected sound waves affects speech intelligibility or perception of the received acoustic wave and lowers the performance of many signal processing applications such as automated speech recognition systems, speaker identification systems, and so on. The first sound you hear and the sounds that bounce back right away (called early reverberation), as well as reflections arriving beyond the early reflections (called late reverberation) all contribute to the received signal. This negative perceptual impact often rises as the distance between the source and microphone increases. One of the most difficult issues in the current context is improving the quality of a damaged voice signal. When a voice signal is captured by a remote microphone, reverberation is one of the key elements that degrade its quality. Speech intelligibility suffers as a result of this reverberation [4–6]. We tried to design a “memory efficient” dereverberation technique for Additive White Gaussian Noise (AWGN) noise that produces few artefacts in this research. Deep Dense Neural Network (DDNN) is a novel network architecture that removes redundant representations of a noisy and reverb spectrum during the decomposition step and maps them back to a clean spectrum during the reconstruction stage. This may be thought of as mapping the spectrum to higher

Manuscript received August 10, 2023; revised September 13, 2023; accepted October 7, 2023; published February 24, 2024.

doi: 10.12720/jait.15.2.281-287

dimensions (e.g., the kernel approach) and then projecting the characteristics back down to lower dimensions.

Additional tests are carried out in this work by taking into account three different room sizes (seminar hall, conference room and room size given in reference paper) and acoustic circumstances The suggested method of using a Deep Neural Network with a delay sum beamformer produces improved results. In addition, the source position is changed at six various positions in the room, and results are produced. The goal of this project is to create and implement a dereverberation algorithm that eliminates early reverberations, late reverberations, and noise from speech signals in order to enhance the speech quality.

The following is how the paper is structured. Section II includes a review of the literature. Section III presents deep neural network architectures, including the proposed deep dense neural network. Section IV describes the experimental techniques. The results are described in Section V, and the research is concluded in Section VI.

II. Literature Survey

Gannot *et al.* [7] proposed multimicrophone speech dereverberation and denoising using Minimum Variance Distortionless Response (MVDR) beamformer and wiener post filter. Masuyama *et al.* [8] created a ground-breaking end-to-end architecture for automatic speech recognition that incorporates de-embracement, beamforming, self-supervised data methodology, and neural network-based de-noising. Han *et al.* [9] propose a parallel interpreting structure based on Distributed Beam-Forming and Multiple-Channel Linear Prediction (DB-BFMCLP) consisted of a Generalised Sidelobe Canceller (GSC) and multiple channels linear prediction for concurrently speech dereverberation as well as noise reduction by sharing the same desired response vector. Lemercier *et al.* [10] provide a strategy for converting multiplicative maskers built with deep neural networks into deeper subband filters for time-frequency audio restoration. Lemercier *et al.* [11] evaluate the efficacy of generative diffusion models and discriminatory techniques on different speech restoration tasks using earlier contributions on diffusion-based speech improvement in the complicated time-frequency domain, and then apply this knowledge to the goal of band with extension. Zheng *et al.* [12] explored both single and multi-speaker recordings are recovered in the Deep Learning (DL) based monaural voice augmentation methods. For this challenging speech augmentation challenge, Convolutional Neural Network (CNN) based models are provided in particular since to their parameter effectiveness and state-of-the-art performance. Sheeja *et al.* [13] developed a novel approach to voice separation and dereverberation using Principal Component Analysis (PCA) based on Locally Weighted Projection Regression (LWPR) and Weighted Prediction Error (WPE) based on a Deep Neural Network (DNN), The technique uses Blind Source Separation (BSS) as well as Blind Dereverberation (BD) after the preprocessing of the reverberant signal, resulting in a mixture of sources. BSS and BD are abbreviations for blind source separation and blind dereverberation, respectively. Lemercier's [14] demonstrate of a two-stage online lightweight dereverberation approach focused on hearing aids. Combining a single-channel post-filter with a multi-channel linear filter can result in a better output. Both of these components are dependent on the DNN's estimates of Power Spectral Density (PSD). Routray *et al.* [15] provided Deep Neural Network (DNN) technique for concurrent denoising and dereverberation of speech. The technique that is being proposed may be broken down into two stages: denoising and dereverberation. Denoising is the process of reducing additive noise by developing a phase-sensitive mask with the use of DNN. The process of dereverberation is the next step that is taken in order to obtain noise-free reverberant speech. During the deverberation phase, we dereverberate using a reverberation time-aware DNN-based model. This model takes advantage of superposition attributes and frame-wise temporal correlations for a variety of reverberation circumstances via two parameters that are time-dependent on the reverberation time: frameshift size and acoustic context size. Ai *et al.* [16] demonstrate a hierarchical neural vocoder called DNR-HiNet that is capable of denoising and dereverberation to clean up acoustic data, we make some adjustments to the Magnitude Spectrum Predictor (ASP) of the default HiNet vocoder so that we may build the DNR-HiNet vocoder. With the help of this improved enhanced Denoising and Dereverberation ASP (DNR-ASP), it is possible to anticipate clean log amplitude spectra from distorted input. DNR-ASP is able to accomplish this goal by first using signal processing methods to anticipate the log amplitude spectra of noisy as well as reverberant speech, the log amplitude spectra of additive noise with the room impulse response, and then carrying out initial denoising and dereverberation. Fu *et al.* [17] proposed that voice augmentation and dereverberation can be performed simultaneously with former, an Unet-based dilated complex & real dual-path conformer system in both the complex as well as magnitude domains. In order to represent dimensional data, we use both local and global context, as well as temporal attention and dilated convolution. Li *et al.* [18] presented a work which showed that voice denoising performance may be enhanced by self-supervised learning. The proposed Pre-training Auto Encoder (PAE) needs just a few unpaired and unseen clean speech signals to obtain speech latent representations. Following an analysis of existing multi-microphone speech dereverberation methods, we came to the conclusion that deep neural networks provide enticing outcomes. Performance metrics like Perceptual evaluation of speech quality and Log-spectral distance can be used to gauge how well an algorithm performs, allowing for the creation of a robust, adaptive algorithm capable of handling early and late reverberation, as well as additive noise in the presence of acoustic parameter variations.

III. METHODOLGY

In Fig. 1, reverberant signals are generated by convolving an anechoic sound with the impulse response of a modelled room prior to adding white Gaussian noise. The reflected signals from each microphone may be represented as

$$y_l[k] = h_l[k] \times s[k] + v_l[k] = x_l[k] + v_l[k] \qquad (1)$$

where $h_l[k]$ Impulse response of acoustic channel from source to microphone l

s$[k]$ is Speech signal

$v_l[k]$ is additive noise component in l^{th} microphone signal

$x_l[k]$ is reverberant speech component

When there are many nodes between the input and output layers, we call it a Deep Neural Network (DNN). No matter how linear or non-linear the connection between input and output, the DNN will determine the appropriate mathematical manipulation to make the transformation. The network iteratively processes through the layers, determining the likelihood of each output as it goes. Data in DNNs is often sent from the input layer to the output layer in a feed forward fashion. At first, the DNN builds a network of hypothetical neurons and gives each link an arbitrary numerical value (the "weights"). An output value between 0 and 1 is calculated by multiplying the weights and inputs.

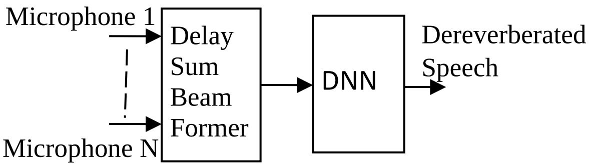


Fig. 1. DSB-DNN system.

Each block in Fig. 1 stands in for a feature in the proposed Deep Dense Network (DDN). The proposed DDN comprises of symmetric decomposition layers and reconstruction layers. Each stage of the decomposition—convolution, batch normalization, max pooling, and ReLU activation—is repeated until the desired result is achieved. Layers of convolution, batch normalization, and up sampling are iterated in order to reconstruct the original data. Normal DDN operation involves compressing features along the decomposition and reconstructing them along the reconstruction. To solve our specific challenge, we replaced the original SoftMax layer in the last layer of the DDN with a thick layer. In the latter phases of a neural network, a layer that is referred to as a dense layer (also known as a totally coupled layer) is used. This layer contributes to the process of adjusting the output dimensionality of the layer that came before it. This makes it possible for the model to offer a more precise description of the relationship that exists among each value of the data that it is processing. In a model, the neurons of the dense layer all receive input from the neurons of the layer above it. Additionally, these neurons do matrix and vector multiplication. Matrix vector multiplication involves multiplying the row vector supplied by the dense layer by the column vector provided by the sparse layer. It is necessary for the row vector to have the identical number of columns as the column vector in a matrix multiplication. Backpropagation is widely used as a training technique for feedforward neural networks. Backpropagation is a common method used in neural network training, and it entails determining the gradient of the loss function with respect to the weights of the network for a single input or output. Mini-batch stochastic gradient descent is used to minimize the error function below,

$$E_{sgd} = \frac{1}{M}\sum_{n=1}^{M}\left\|\widehat{S_n}(T_{n-\tau}^{n+\tau}, W, b) - S_n\right\|_2^2 \qquad (2)$$

where E_{sgd} is the mean squared error

$\widehat{S_n}(T_{n-\tau}^{n+\tau}, W, b)$ is estimated signal frame

S_n is reference normalized frame at index n

M is Minibatch size

(W, b) indicating the trainable parameters of weight and bias.

The revised prediction of W^ι and b^ι in the ι^{th} layer, at a certain learning rate, may be repeatedly calculated as follows:

$$\Delta(W_{n+1}^{\iota},\ b_{n+1}^{\iota}) = -\lambda\frac{\partial E_{sgd}}{\partial(W_n^{\iota}, b_n^{\iota})} - \kappa\lambda(W_n^{\iota},\ b_n^{\iota}) + \omega\Delta(W_n^{\iota}, b_n^{\iota}), 1 \le \iota \le L + 1 \qquad (3)$$

where L shows the number of layers under the surfaces and $L + 1$ show the last layer of output. κ is the weight decay coefficient. And ω is the momentum. If given enough training samples, DDNN may automatically learn the complex connection required to isolate speech from the noisy and reverberant sounds.

The foregoing intuition suggests that the dense layer's output will be a vector with N dimensions. It's clear that it's decreasing the size of the vectors involved. Therefore, a dense layer is used to alter the vectors' dimensions, with each neuron playing a role in the process. DDN compresses the features during the reconstruction phase and encrypts them into higher dimensions during the decomposition phase. Symmetry in the number of filters is maintained by progressively increasing the number of filters during the decomposition and decreasing the number of filters during the reconstruction. Due to its completely convolutional nature, DDN includes a convolution layer as its last layer. As the function that converts noisy speech qualities to clean ones, a DDNN is used. The well-trained DDNN model is used in the improvement phase to analyze the noisy speech characteristics and forecast the clean and anechoic speech features. The DNN's input characteristics were normalized to have a mean of zero and a standard deviation of one.

The output of DNN $\widehat{S_n}(v)$ should be transformed back as follows:

$$\hat{S}'(v) = \hat{S}(v) \times D(v) + K(d) \qquad (4)$$

where $D(v)$ and $K(d)$ are the component of the input noisy speech characteristics mean and variance respectively. The equalization factor might then be employed as a post processing step to reduce the variance of the reconstructed signal:

$$\hat{S}''(v) = \hat{S}(v) \times \eta \times D(v) + K(d) \tag{5}$$

Since the DNN output $\hat{S}(v)$ was in the logarithm of the power spectrum, thus the exponential function was used to multiply by the multiplicative factor. Furthermore, this exponential factor has the potential to both reduce residual noise and increase the clarity of the recovered speech's formant peaks.

IV. Experimentation

Three different rooms are taken into consideration during the experimentation. Reference paper's room measures [6.1×5.3×2.7] m, the conference room measures [9.7×5.9×3.5] m, and the seminar room measures [17.7×9.6×3.5] m. The array of four microphone is used, and the space between each microphone is [3, 4, 3] cm. By maintaining the source and receiver microphone positions constant and a 2 m distance between them, RT60 of 0.3 s the RIR is produced using the image source method [19]. Reverberant signal is generated by convolving anechoic speech signal with room impulse response. The reverberant signal is combined with Additive White Gaussian Noise (AWGN), realistic noise, such as babble noise and restaurant noise, at Signal to Noise (SNR) ratios of −10 dB, 0 dB, 10 dB, 20 dB, and 30 dB to produce reverberant and noisy(unprocessed) signal. The signals are processed frame by frame where each frame of 32 ms with 8 ms overlaps between each frame and an 8 kHz sampling rate. Performance of the proposed algorithm is assessed using objective metrics like Log Spectral Distance (LSD) and Perceptual Evaluation of Speech Quality (PESQ) [20]. PESQ has a range of −0.5 to 4.5. In order to get a dereverberated signal, a reverberant and noisy signal is given to a delay sum beamformer, which combines the signals from four microphones into a single channel and then it is passed to DNN. A source-to-microphone distance of 2 m and RT60 of 0.3 s are tested for each room's performance. When the performance of all three rooms is compared to the PESQ and LSD values of an unprocessed signal, it shows that speech quality has improved. Additionally, experiments are run with various source positions in the reference paper room, and performance is assessed using PESQ and LSD with the same types of noise and SNRs. The IEEE database contains phonetically-balanced 720 sentences with relatively low word-context predictability. Out of that 30 IEEE sentences (produced by three male and three female speakers) are used for experimentation [21].

Experimentation for room impulse response generation of simulated room is carried out for three rooms. Figs. 2–4 shows simulated room experimental setup and Tables I–III are used to model the room acoustic setting in which test is conducted for reference paper room, conference room and seminar hall, respectively.

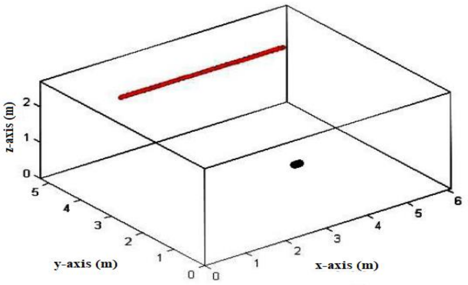


Fig. 2. Experimental set up for room given in reference paper.

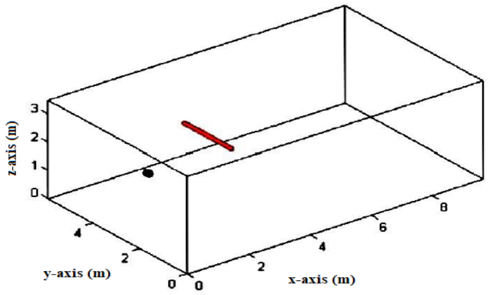


Fig. 3. Experimental set up for conference room.

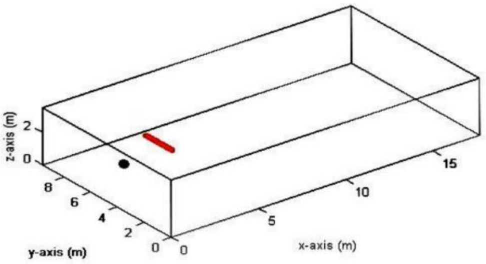


Fig. 4. Experimental set up for seminar hall.

TABLE I. Experimental Setup Structure for Room Given in Reference Paper

Fs	8000 (Hz)
Room size	[6.1, 5.3, 2.7] in m
Number of microphones	4
RT60	0.3 s
c (speed of acoustic wave)	343 m/s

TABLE II. Experimental Setup Structure for Conference Room

Fs	8000 (Hz)
Room size	[9.7, 5.9, 3.5] in m
Number of microphones	4
RT60	0.3 s
c (speed of acoustic wave)	343 m/s

TABLE III. Experimental Setup Structure for Seminar Hall

Fs	8000 (Hz)
Room size	[17.7, 9.6, 3.5] in m
Number of microphones	4
RT60	0.3 s
c (speed of acoustic wave)	343 m/s

V. Results

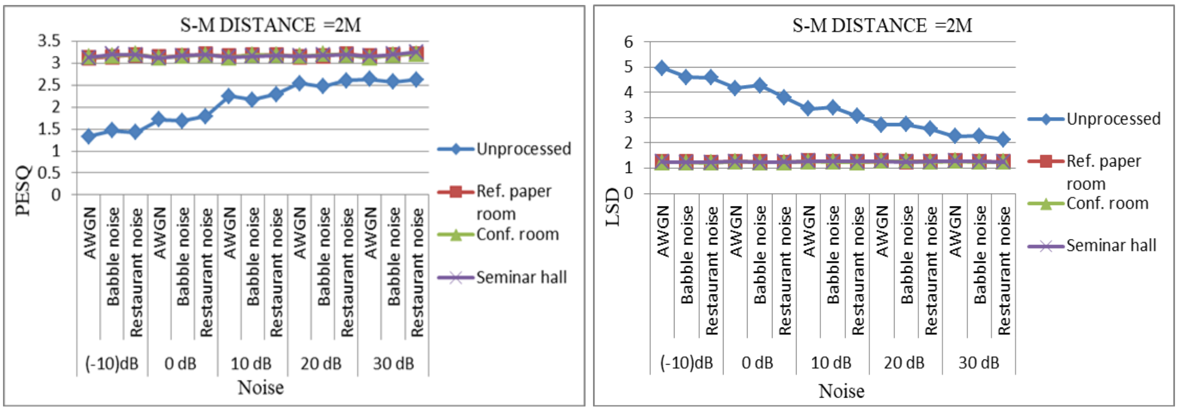
This section is divided in two subsections. In first section of result two parameters are kept constant RT60 of 0.3 s and source to microphone distance is of 2 m. The proposed algorithm is tested for different noise types namely AWGN, babble, and restaurant noise with different values of SNRs −10 dB, 0 dB, 10 dB, 20 dB, and 30 dB for three different rooms, including a reference paper room, conference room, and seminar hall. For this the performance of proposed algorithm is evaluated by using two metrics, Perceptual Evaluation of Speech Quality (PESQ) and Log Spectral Distortion (LSD) for speech quality assessment.

The obtained PESQ and LSD values are summarized in Table IV and plotted in Fig. 5. This shows that increase in PESQ and decrease in the LSD values, which signifies an improvement in the speech quality of the processed signal for different types of noise and several SNR levels in comparisons with the unprocessed signal. It also shows that there is increase in value of PESQ for restaurant signal with decreasing SNR. The variation in room size shows slightly better performance for seminar hall demonstrating that even though room size is increased, algorithm gives better performance.

In second section of result room size is fixed, RT60 is 0.3s and various types of noise are added with speech signal namely AWGN, babble, and restaurant noise with different values of SNRs −10 dB, 0 dB, 10 dB, 20 dB, and 30 dB for different source positions in room as shown in Fig. 6. Performance for this is evaluated by using PESQ and LSD metrics.

TABLE IV. SIMULATED RESULTS OF PESQ & LSD AT THE OUTPUT OF BF+DNN FOR RT60 = 0.3 S, S-M DISTANCE = 2 M

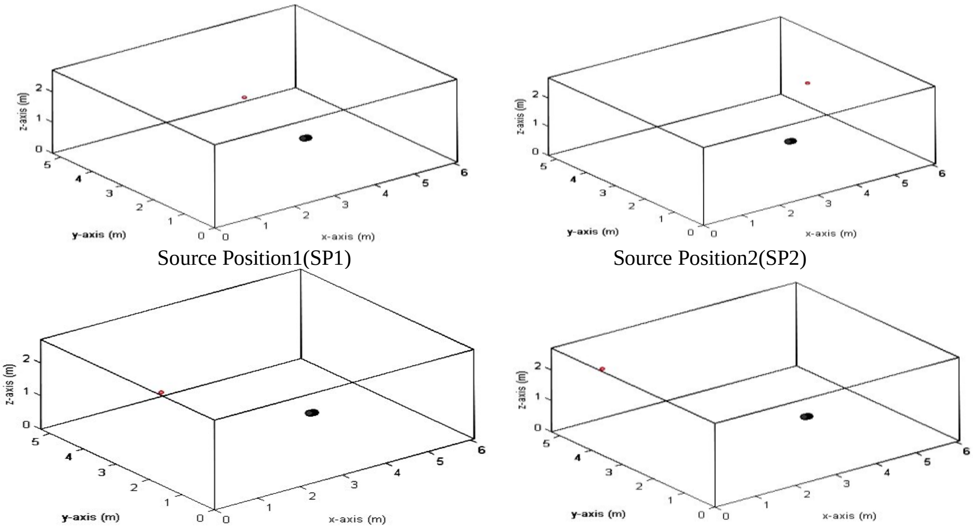
SNR	**(−10) dB**			**0 dB**			**10 dB**			**20 dB**			**30 dB**		
								PESQ							
Noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise
Unprocessed	1.33	1.47	1.42	1.72	1.69	1.79	2.25	2.17	2.29	2.54	2.47	2.61	2.64	2.58	2.63
Ref. paper room [7]	3.12	3.15	3.19	3.14	3.17	3.2	3.16	3.18	3.19	3.15	3.17	3.21	3.16	3.18	3.21
Conf. room	3.14	3.17	3.2	3.13	3.16	3.19	3.13	3.19	3.18	3.16	3.2	3.19	3.12	3.19	3.2
Seminar hall	3.13	3.2	3.19	3.12	3.17	3.19	3.14	3.16	3.17	3.16	3.18	3.19	3.16	3.2	3.25
								LSD							
Unprocessed	4.96	4.6	4.59	4.16	4.27	3.8	3.36	3.4	3.06	2.71	2.73	2.56	2.26	2.28	2.14
Ref. paper room [7]	1.24	1.25	1.23	1.26	1.25	1.23	1.29	1.27	1.25	1.28	1.26	1.27	1.28	1.27	1.25
Conf. room	1.23	1.22	1.21	1.26	1.23	1.22	1.27	1.25	1.22	1.29	1.28	1.25	1.28	1.25	1.24
Seminar hall	1.24	1.22	1.23	1.27	1.22	1.25	1.29	1.27	1.26	1.28	1.25	1.26	1.29	1.27	1.25



(a) Plot of PESQ

(b) Plot of LSD

Fig. 5. Plot of PESQ and LSD at the output of BF+DNN for RT60 = 0.3 s, S-M distance = 2 m.

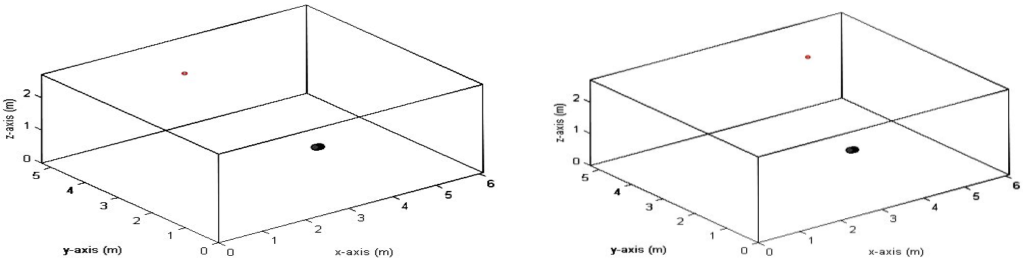


Source Position1(SP1)

Source Position2(SP2)

Source Position3(SP3)

Source Position4(SP4)



Source Position5(SP5) Source Position6(SP6)

Fig. 6. Various position of source in room.

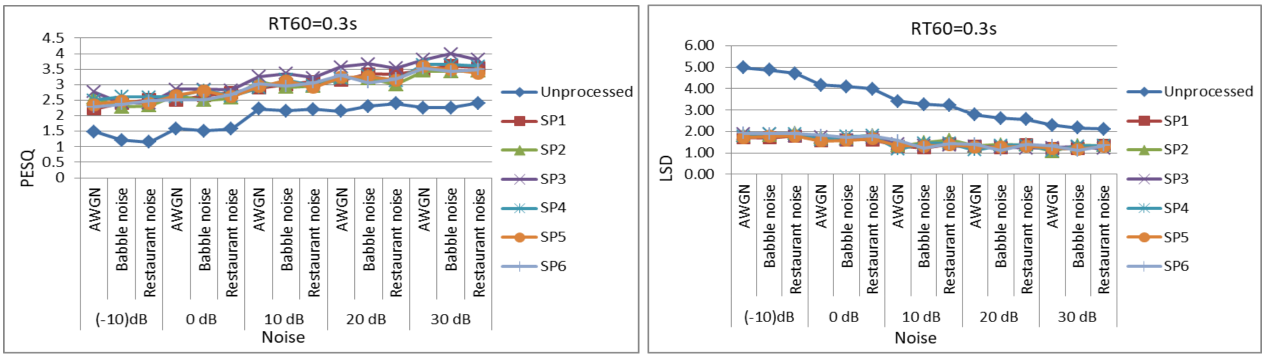
Tables V and VI summarizes the results for various source positions in room and plotted in Fig. 7. Fig. 7(a) is plot of PESQ metric, and Fig. 7(b) is LSD metric. These plots demonstrate that good speech quality is achieved even though the source position is changed at various locations in the room. The comparison of various source position results shows the promising results at source position 3.

TABLE V. SIMULATED RESULTS OF PESQ AT THE OUTPUT OF BF+DNN FOR RT60 = 0.3S WHEN SOURCE POSITION IS VARIED IN ROOM AT DIFFERENT LOCATIONS

PESQ															
SNR	**(−10) dB**			**0 dB**			**10 dB**			**20 dB**			**30 dB**		
Noise/ Position	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise
Unprocessed	1.49	1.21	1.15	1.58	1.51	1.57	2.22	2.15	2.21	2.14	2.31	2.39	2.26	2.25	2.41
SP1	2.22	2.40	2.52	2.50	2.64	2.75	2.89	2.99	3.09	3.14	3.35	3.34	3.48	3.61	3.54
SP2	2.56	2.28	2.31	2.67	2.49	2.57	3.10	2.92	2.95	3.23	3.20	3.00	3.43	3.43	3.44
SP3	2.76	2.41	2.40	2.85	2.85	2.83	3.27	3.35	3.23	3.57	3.67	3.54	3.81	4.00	3.81
SP4	2.50	2.61	2.60	2.61	2.83	2.63	2.89	3.16	2.97	3.18	3.29	3.12	3.64	3.65	3.60
SP5	2.36	2.48	2.41	2.66	2.80	2.62	2.94	3.13	2.92	3.24	3.26	3.15	3.59	3.48	3.37
SP6	2.28	2.37	2.48	2.50	2.50	2.67	3.03	2.95	3.06	3.31	3.07	3.19	3.53	3.42	3.49

TABLE VI. SIMULATED RESULTS OF LSD AT THE OUTPUT OF BF+DNN FOR RT60 = 0.3S WHEN SOURCE POSITION IS VARIED IN ROOM AT DIFFERENT LOCATIONS

LSD															
SNR	**(−10) dB**			**0 dB**			**10 dB**			**20 dB**			**30 dB**		
Noise/ Position	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise	AWGN	Babble noise	Restaurant noise
Unprocessed	4.97	4.86	4.71	4.16	4.10	3.98	3.40	3.27	3.21	2.77	2.62	2.57	2.29	2.17	2.11
SP1	1.72	1.70	1.78	1.57	1.60	1.60	1.32	1.23	1.41	1.31	1.23	1.40	1.19	1.20	1.35
SP2	1.81	1.83	1.92	1.72	1.68	1.82	1.33	1.49	1.61	1.32	1.39	1.38	1.07	1.35	1.33
SP3	1.88	1.87	1.79	1.73	1.61	1.71	1.40	1.34	1.33	1.30	1.29	1.20	1.25	1.28	1.19
SP4	1.78	1.89	1.90	1.61	1.76	1.81	1.16	1.45	1.45	1.13	1.36	1.35	1.10	1.33	1.33
SP5	1.71	1.77	1.79	1.55	1.59	1.67	1.32	1.32	1.38	1.28	1.26	1.32	1.21	1.18	1.31
SP6	1.91	1.9	1.92	1.82	1.71	1.80	1.58	1.23	1.42	1.41	1.12	1.38	1.34	1.11	1.33



(a) Plot of PESQ (b) Plot of LSD

Fig. 7. Plot of PESQ and LSD with BF+DNN for various source positions.

VI. CONCLUSION

Reverberation is the process by which a sound travels from its origin to a listener through a number of different paths along its journey. It is a blind issue with an unknown and non-stationary source signal and an unknown and time-varying acoustic channel. As a result, reverberation has effect on speech, making it sound distant and spectrally distorted, as well as less understandable. Human voice is often damaged in real-world contexts by both reverberation and background noise, which has a detrimental impact on speech intelligibility and quality. A noise reduction and dereverberation technique was developed in this research by integrating delay and sum beamformer with deep learning to enhance the speech quality. By using proposed method the two speech quality metrics PESQ and LSD assessment is carried out and the experimental results shows that the quality of speech in seminar hall, conference room and reference paper room gets improved with AWGN and realistic noise such as babble noise, restaurant noise.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Seema S. Arote conducted the research; Vijay M. Mane analyzed the data; Shakil A. Shaikh wrote the paper. All authors had approved the final version.

References

[1] A. R. Jayan, *Speech and Audio Signal Processing*, PHI Learning Pvt. Ltd., 2017, ch. 1, pp. 1–20.

[2] M. Omologo, P. Svaizer, and M. Matassoni, “Environmental conditions and acoustic transduction in hands-free speech recognition,” *Speech Communication*, vol. 25, no. 1–3, pp. 75–95, Aug. 1998.

[3] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. 2013, ch. 1, pp. 1–10.

[4] Y. Takata and A. K. Nabelek, “English consonant recognition in noise and in reverberation by Japanese and American listeners,” *Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 663–666, Aug. 1990.

[5] R. Beutelmann and T. Brand, “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners,” *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[6] A. Warzybok, J. Rennies, T. Brand, S. Doclo, and B. Kollmeier, “Effects of spatial and temporal integration of a single early reflection on speech intelligibility,” *Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, Jan. 2013.

[7] O. Schwartz, S. Gannot, and E. A. P. Habets, “Multi-microphone speech dereverberation and noise reduction using relative early transfer functions,” in *Proc. IEEE/ACM Trans. Audio, Speech and Lang.*, Feb. 2015, vol. 23, no. 2, pp. 240–251.

[8] Y. Masuyama, X. Chang, S. Cornell, S. Watanabe and N. Ono, “End-to-End integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation,” in *Proc.* 2022 *IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, Jan. 2023, pp. 260–265.

[9] Z. Han, Y. Ke, X. Li, and C. Zheng, “Parallel processing of distributed beamforming and multichannel linear prediction for speech denoising and dereverberation in wireless acoustic sensor networks,” *J. Audio Speech Music Proc*., vol. 25, no. 1, pp. 1–17, May 2023.

[10] J. M. Lemercier, J. Tobergte, and T. Gerkmann, “Extending DNN-based multiplicative masking to deep subband filtering for improved dereverberation,” in *Proc. INTERSPEECH Conf.*, 2023, pp. 4024–4028.

[11] J. M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, “Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[12] C. Zheng, Y. Ke, X. Luo, and X. Li, *IoT-enabled Convolutional Neural Networks: Techniques and Applications*, 1st ed. Denmark, River, 2023, ch. 3, pp. 65–95.

[13] J. J. C. Sheeja and B. Sankaragomathi, “Speech dereverberation and source separation using DNN-WPE and LWPR-PCA,” *Neural Comput & Applic*., vol. 35, no. 10, pp. 7339–7356, Apr. 2023.

[14] J. M. Lemercier, J. Thiemann, and R. Koning, “A neural network-supported two-stage algorithm for lightweight dereverberation on hearing devices,” *J. Audio Speech Music Proc*., vol. 18, no. 1, pp. 1–12, May 2023.

[15] S. Routray and Q. Mao, “A context aware-based deep neural network approach for simultaneous speech denoising and dereverberation,” *Neural Comput & Applic*., vol. 34, no. 12, pp. 9831–9845, June 2022.

[16] Y. Ai, Z. H. Ling, W. L. Wu, and A. Li, “Denoising and dereverberation hierarchical neural vocoder for statistical parametric speech synthesis,” in *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, June 2022, vol. 30, pp. 2036–2048.

[17] Y. Fu, “Uformer: A Unet based dilated complex and real dual path conformer network for simultaneous speech enhancement and dereverberation,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 7417–7421.

[18] Y. Li, Y. Sun, and S. M. Naqvi, “Self-supervised learning and multi task pre training based single channel acoustic denoising,” in *Proc. IEEE International Conference on Multi sensor Fusion and Integration for Intelligent Systems (MFI)*, Bedford, United Kingdom, 2022, pp. 1–5.

[19] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer*. vol. 65, no. 4, pp. 943–950, Apr. 1979.

[20] ITU-TRec.P.862. (2001). Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. [Online]. Available: https://www.itu.int/rec/T-REC-P.862

[21] “IEEE recommended practice for speech quality measurements,” in *IEEE Transactions on Audio and Electroacoustic*, vol. 17, no. 3, pp. 225–246, September 1969. doi: 10.1109/TAU.1969.1162058