Microphone Array Speech Enhancement Via Beamforming Based Deep Learning Network

Original Scientific Paper

Jeyasingh Pathrose

Research Scholar , Department of Electronics and Communication Engineering ,
B.S. Abdur Rahman Crescent Institute of Science and Technology,
Chennai 600048, India
Jeyasingh.p@jasmin-infotech.com

Mohamed Ismail M

Professor and the Dean (Academic Affairs) of B.S. Abdur Rahman Crescent Institute of Science & Technology,
Chennai, 600048 India.
mmismail@crescent.education

Madhan Mohan P

Jasmin Infotech Pvt Ltd,
Chennai, India 600100
madhanmohan.p@jasmin-infotech.com

Abstract *– In general, in-car speech enhancement is an application of the microphone array speech enhancement in particular acoustic environments. Speech enhancement inside the moving cars is always an interesting topic and the researchers work to create some modules to increase the quality of speech and intelligibility of speech in cars. The passenger dialogue inside the car, the sound of other equipment, and a wide range of interference effects are major challenges in the task of speech separation in-car environment. To overcome this issue, a novel Beamforming based Deep learning Network (Bf-DLN) has been proposed for speech enhancement. Initially, the captured microphone array signals are pre-processed using an Adaptive beamforming technique named Least Constrained Minimum Variance (LCMV). Consequently, the proposed method uses a time-frequency representation to transform the pre-processed data into an image. The smoothed pseudo-Wigner-Ville distribution (SPWVD) is used for converting time-domain speech inputs into images. Convolutional deep belief network (CDBN) is used to extract the most pertinent features from these transformed images. Enhanced Elephant Heard Algorithm (EEHA) is used for selecting the desired source by eliminating the interference source. The experimental result demonstrates the effectiveness of the proposed strategy in removing background noise from the original speech signal. The proposed strategy outperforms existing methods in terms of PESQ, STOI, SSNRI, and SNR. The PESQ of the proposed Bf-DLN has a maximum PESQ of 1.98, whereas existing models like Two-stage Bi-LSTM has 1.82, DNN-C has 1.75 and GCN has 1.68 respectively. The PESQ of the proposed method is 1.75%, 3.15%, and 4.22% better than the existing GCN, DNN-C, and Bi-LSTM techniques. The efficacy of the proposed method is then validated by experiments.*

Keywords*: Speech Enhancement, Microphone, Deep Learning, Beamforming, Noise Reduction*

1. INTRODUCTION

Today Speech enhancement (SE) is a pre-processing step in speech recognition that is also required to accommodate the growing demand for higher-quality speech. The speech signal is now used in a variety of systems including speaker identification, speech control, speech-to-text systems, voice over internet protocol (VOIP) accessibility of web applications, and interactive voice response system (IVRS) services. Voice recognition and other speaker activities [1], interaction [2], sound aids [3], and coding of speech all require SE [4]. The SE is a difficult operation when the noisy signal is generated at a lower frequency [5]. The quality of the voice signal should not be sacrificed while designing a speech signal-based system. However, speech signals can be damaged in practice due to a variety of disturbances such as echo, noise in the background, babbling noise babbling sound, and so on. Speech enhancement technology [6] can improve not just the signal-to-noise ratio (SNR) and audio perception of collected speech as well as the resilience of speech improvement and speaker verification systems. As a result, speech improvement in noisy contexts has gotten a lot of attention.

Speech intelligibility when utilizing in-vehicle speech applications has been impacted by engine noise and other noise sources, such as airflow from electric fans or automobile windows. Inside the car, the reflection of speech waves is employed to communicate particularly between the front and back seat passengers. In addition to in-car disturbances, the quality of speech communication is generally poor. The speech signals are picked up by the microphone and the microphones are placed front seat headrest position.

Hands-free car kits and in-car speech recognition systems increasingly use single-channel noise reduction and beamformer arrays to reduce noise. Microphone array processing focuses on speech improvement and localization, particularly in noisy or reverberant situations [7]. A microphone array is used in the car to increase voice communication quality. [8] A microphone array may gather data in the spatial domain as well as the temporal and frequency domains. The passenger dialogue inside the car, the sound of other equipment, and a wide range of interference effects are major challenges in the task of speech enhancement in-car environment. The noise can be split into interference and desired noise depending on how the noise source and interference path differ from one another. Interference noise makes up the noise in a car. The aim of this paper is to improve speech quality under interference noise conditions. The primary goal of the proposed approach is to improve the speech quality in cars. To overcome this issue, a novel Beamforming based Deep learning Network (Bf-DLN) has been proposed for speech enhancement. The major contribution of the proposed method is;

- Initially, a pre-processing method known as the Least Constrained Minimum Variance (LCMV) is used to pre-process the collected microphone array signals. Consequently, the pre-processed signal is converted into an image using a time-frequency representation.
- The Smoothed Pseudo-Wigner-Ville Distribution (SPWVD) transforms time-domain speech signals into images.
- The convolutional deep belief network (CDBN) receives these transformed signals as input to extract the most pertinent characteristics. By removing the interference source, the desired source is selected using the Enhanced Elephant Heard Algorithm (EEHA).
- The experimental results show that the proposed strategy is effective in removing background noise from the original speech signal.

The rest of the work is organized as follows: Section 2 describes the literature survey, and the problem about the array position inside the car is addressed in section 3. The proposed Source Separation for the car is given in section 4, outcomes are presented in section 5, Section 6 encloses with conclusion and future work.

2. LITERATURE SURVEY

This section outlines the various investigations that have been carried out throughout the year to improve speech signaling. An overview of recent developments in the speech signal is given in this study.

Gentet et al. [9] presented a speech enhancement algorithm it increases the signal-to-noise ratio (SNR). The result showed that the technique has low-frequency noise. Speech intelligibility optimization problem with a fixed perceived loudness restriction is a major drawback.

Alkaher et al. [10] presented the dual microphone speech enhancement for enhancing speech communication in cars. The Pareto optimization decreases the overall speech distortion and relative gain reduction. The result demonstrates the dual-microphone system enhanced howling detection sensitivity. The drawback is that howling sounds may occur even before the speech reinforcement (SR) system reaches instability.

Saleem, N., et al. [11] suggested a Kalman filtering model with an augmented Bidirectional Gated Recurrent Unit (BiGRU) based on residual connections for speech enhancing and recognizing. With the use of the LibriSpeech dataset, the suggested method increased the quality, intelligibility, and word error rates under varied noisy situations by 35.52%, 18.79%, and 19.13%, respectively.

Chuang, S.Y., et al. [12] proposed an improved lite audio-visual speech enhancement (iLAVSE) algorithm for a car-driving scenario. Three stages are involved in the iLAVSE system: data preprocessing, AVSE based on CRNN, and reconstruction. It is also demonstrated that iLAVSE is suitable for real-world scenarios where superior audio-visual sensors might not always be available.

Tao et al. [13] presented the enhanced sound source localization and speech enhancement algorithm which reduce microphone cost and also reduces the complexity. The dual-microphone sound algorithm effectively identifies the sound location, as well as the speech enhancement algorithm, is more resilient and adaptive than the previous method, according to experimental data. The biggest disadvantages are the high cost and high design requirements.

Kothapally, V., et al. [14] proposed a subband spatio-temporal beam former based on DL that can perform speech separation in a care setting with a reduced computation cost and inference time.They showed that the suggested strategy produces improved WER and objective scores using a variety of sub-band configurations.

In Jolad, B and Khanai, R [15], speech signal quality can be improved by using a fractional competitive crowd search algorithm (FCCSA). When the suggested technique is evaluated using the UA speech database, it yields 0.930, 0.933, and 0.934 in terms of accuracy, specificity, and sensitivity.

Qian et al. [16] presented the car speech enhancement system based on a combination of a deep belief network and wiener filtering. The deep belief networks (DBN) parameters are optimized by using the Quantum Particle Swarm Optimization (QPSO) algorithm. The results of the experiment demonstrated that the suggested strategy may successfully reduce the original speech signal's noise signal and improve the speech signal.

Zhou, W., et al. [17], suggested a Meta-reinforcement learning paradigm by concentrating on few-shot learning for improving speech. The experiment's findings demonstrate that in comparison to state-of-the-art DNN-based SE methods under difficult conditions, where the environment noises are varied and the signals are non-stationary, this work achieves at least improvements of 1.3%~12.5% for a single shot and 3.1%~14.3% for a five-shot scenario.

Table 1. Comparison of existing with the proposed method

Method	Advantage	Disadvantage
Speech intelligibility enhancement method for typical in-car [9]	To improve understanding by employing specialized speech transformation methods without altering the initial SNR	High computational complexity
dual microphone speech enhancement [10]	decreases the overall speech distortion and relative gain reduction using Pareto optimization.	howling sounds may occur even before the speech reinforcement (SR) system reaches instability
Kalman filtering model with augmented Bidirectional Gated Recurrent Unit (BiGRU) [11]	for speech enhancing and recognizing.	Do not eliminate echo
improved lite audio-visual speech enhancement (iLAVSE) algorithm [12]	suitable for real-world scenarios	Failure of the sensor to record the visual signal is another source of low-quality visual data.
enhanced sound source localization and speech enhancement algorithm [13]	reduce microphone cost and also reduces the complexity	high cost and high design requirements
DL-based mel-subband spatiotemporal beamformer [14]	decreased computation cost and inference time	unstable speech signal
Fractional Competitive Crow Search Algorithm-based speech Enhancement Generative Adversarial Network [15]	does not investigate in-vehicle speech recognition	high cost and low performance of the microphone
car speech enhancement system based on a combination of a deep belief network and wiener filtering [16]	eliminate the noise signal of the original speech signal and enhance the speech signal	High computational complexity
Meta-reinforcement learning paradigm by concentrating on few-shot learning [17]	decreased computation cost and complexity	higher complexity with an optimum number of layers
Proposed Beamforming-based Deep Learning Network (Bf-DLN)	improve the speech quality in cars, high performance for microphones, less computation and complexity	A limited number of the dataset is used for training and testing

Based on the literature review, a variety of deep learning techniques for improving speech were suggested. However, speech enhancement systems face challenges from unstable voice signals, poor microphone performance, expensive computing, and the problem of echo cancellation in the presence of background noise. To overcome the above challenges this research proposed a novel Beamforming based Deep learning Network (Bf-DLN) and its detailed process is presented in section 4.

3. PROBLEM FORMULATION

Consider a microphone array with N elements that captures D desirable voice signals. The received signal is corrupted by additive noise, which might be made up of white noise from a point source, a diffuse source, or both. The short-time Fourier transform (STFT) domain is used to construct the speech enhancement problem, with a time frame index x and the frequency index f. The microphone signal can be expressed as

$$P(x,f) = [p_1(x,f), p_2(x,f), p_3(x,f) \dots p_n(x,f)]^t \quad (1)$$

Where t represents transpose. The decomposition of the received signal $p(x, f)$ is given as follows:

$$P(x,f) = \sum_{i=1}^{D} U_{i,j}(x,f) + W_i(x,f), i = 1,2, \dots D \quad (2)$$

where $U_{i,j}$ (x,f)stands for the voice signal from the j^{th} speaker as acquired by the i^{th} microphone and W_i (x,f) stands for background and sensor noise.

Assuming that the i^{th} microphone receives anechoic speech signals E_j (x,f) and a time-invariant ATF $H_{i,j}$ (f) (assuming a static scenario), the observed speech signal can be approximated in the STFT domain as a multiplication of the anechoic signals and i^t hmicrophone, i.e.

$$U_{i,j}(x,f) = H_{i,j}(f)E_j(x,f) \quad (3)$$

Designing a beamformer *Z(x,f)* to ensure that the output signal is accurate is the problem.

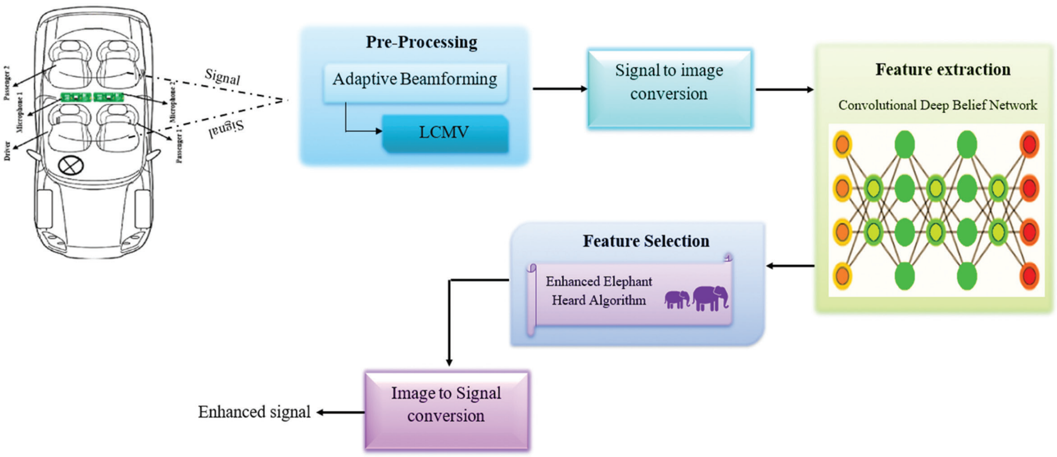
$$L(x,f) = Z^{Ht}(x,f)P(x,f) \qquad (4)$$

Where H_t stands for conjugate-transpose, maximizes SNR for each of the S desired sources while preserving a response that is distortion-free for everyone.

4. PROPOSED METHODOLOGY

There In this section, a novel Beamforming based Deep learning Network (Bf-DLN) has been proposed for speech enhancement. The overall block of the proposed methodology is depicted in Fig.1. Initially, the captured microphone signals are pre-processed using an Adaptive beamforming technique named Least Constrained Minimum Variance (LCMV). Pre-processed signals are converted into images via Smoothed Pseudo-Wigner-Ville distribution (SPWVD). Therefore, the convolutional deep belief network (CDBN) is fed these modified images as input in order to extract the most relevant features. Enhanced Elephant Heard Algorithm (EEHA) is used for selecting the desired source by eliminating the interference source.

Fig. 1. Overall block of proposed Methodology



4.1. PRE-PROCESSING VIA ADAPTIVE BEAMFORMING

Beamforming refers to the process by which signals from the microphone array create a beam pattern. Beamforming is a technique for reducing interference signals that come from various noisy directions. The angle and frequency arriving from different directions cancel out the interference signal. It is utilized to increase the volume of speech signals coming from different directions. When beamforming, it generates a signal that is less noisy than the reference microphone's signal while keeping the speech component intact. The LCMV beamformer is a method frequently employed to solve the issue of many desirable speakers' augmentation in noisy environments. It is defined as follows:

$$Z_{LCMV} = \underset{z}{\operatorname{argmin}}\, Z^{Ht}\Phi_w Z,\;\; B^{Ht}Z = k \qquad (5)$$

Where, $B = \widetilde{Q_1}, \widetilde{Q_2}, \widetilde{Q_3}, \ldots \widetilde{Q_D}$. The required sources are constrained by an N×D matrix, where $\widetilde{Q_1} = \frac{Q_i}{Q_g}$ for i=1,2,…D and $g\epsilon\{1\ldots.N\}$ is the index of the reference microphone, which is often set to be 1 or the microphone with the best input SNR. The associated restrictions are represented by the vector k=[1 1…1]^t, of length D, and the spatial noise coherence matrix is represented by Φ_w. The well-known formula for solving the problem (5) is given by

$$Z_{LCMV}\Phi_w^{-1}B(B^{Ht}\Phi_w^{-1}B)^{-1}k \qquad (6)$$

In response to the necessary speech components, as detected by the microphone, the LCMV beamformer maintains a distortion-free response. When taking into account systems with scattered microphone arrays or arrays with huge apertures, the reference microphone's signal might not be optimal in terms of SNR, or with less speech component power, for all intended speakers. For instance, when one of the desired speakers is placed closest to a microphone $n\epsilon\{1,\ldots,N\}/\{g\}$ in a diffuse noise setting, this is to be expected. SNR for that speaker at the beamformer output is subsequently decreased as a result. The enhanced speech signals are converted into images by using SPWVD, which is briefly discussed in the next section.

4.2. SIGNAL TO IMAGE CONVERSION

Using Smoothed Pseudo-Wigner-Ville distribution, the pre-processed signals are transformed into pictures. At low frequency, Wigner-Ville distributions result in a cross-term and a decrease. SPWVD is used to transform the time-domain filtered speech signals into Time-Frequency Representation (TFR), which overcomes these difficulties. The SPWVD is subjected to independent time and frequency smoothing to attain maximum resolution. The time-frequency resolution of STFT and CWT is problematic, whereas SPWVD has a good resolution. The time-domain signals are transformed into time-frequency representations in order to follow the spectral domain. TFR stands for time, frequency, and amplitude representation in area simultaneously. To enhance frequency resolution while carrying out the quadratic time-frequency transforming, a sliding window is added to the SPWVD signal in both the frequency and time domains. In both the frequency and time domains, it is possible to separately choose the cross-term lessening window's length and sort. SPWVD has improved time-frequency cluster properties as a result. The mathematical formulation of SPWVD is as follows:

$$\varphi(p,q) = \int_{-\infty}^{+\infty} m(p-p')\varphi(p',q)dp' \tag{7}$$

$$\varphi(p',q) = \int_{-\infty}^{+\infty} n(\tau)\, o\left(p' + \frac{\tau}{2}\right) o^{*}\left(p' - \frac{\tau}{2}\right) e^{-j2\pi uv} d\tau \tag{8}$$

Where the cross-terms *m(p)* and *n(p)* decrease the temporal and frequency domain windows, respectively. It is easy to alter the scales for frequency and temporal domain smoothing. The windows' *m(p)* and *n(p)* lengths can be chosen separately.

4.3. FEATURE EXTRACTION USING CONVOLUTIONAL DEEP BELIEF NEURAL NETWORK

Convolutional deep belief networks (CDBNs) are developed from deep belief networks in several ways, including how the building blocks are stacked, the way the network is trained, and even the building blocks themselves. In Fig. 2, the overall structure is depicted, and the main difference is how the building blocks are formed. CRBBMs (Convolutional Boltzmann Machines) are created using Boltzmann machines with convolutional restrictions. The purpose of CRBM is to provide realistically sized images with scaling approaches.

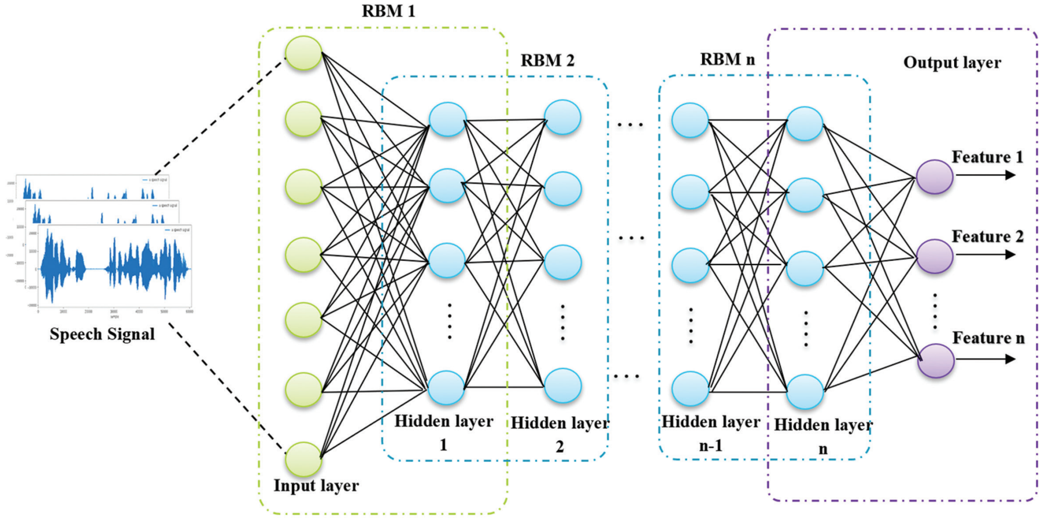


Fig. 2. Architecture of Deep Belief Network

The "convolutional" RBM is an extension of the "regular" RBM in which the weights between the hidden units and the input units are distributed over all locations in the hidden layer. A hidden layer H_l and an input layer I_l make up the two layers of the CRBM. The input units are real-valued or binary-valued, whereas the hidden units are binary-valued.

Assume that the input layer's size is n_{I_l} and the hidden layer's size is n_{H_l}. There are M filters (weights), W_M each of which is convolutional with the input layer, and there are biases b_M for each of the weights as well as c for the input layer. As defined below, the energy function with binary input is

$$E(I_l,H_l) = -\sum_{M=1}^{M}\sum_{y=1}^{n_{H_l}}\sum_{s=1}^{n_w} h_y^s w_s^M i_{y+s-1} - \sum_{M=1}^{M} b_M \sum_{y=1}^{n_{H_l}} h_y^s - c\sum_{x=1}^{n_{I_l}} i_x \tag{9}$$

Accordingly, a CRBM's energy function can be expressed as follows:

$$E(I_l,H_l) = \frac{1}{2}\sum_{x}^{n_{I_l}} i_x^2 - \sum_{M=1}^{M}\sum_{y=1}^{n_{H_l}}\sum_{s=1}^{n_w} h_y^s w_s^M i_{y+s-1} - \sum_{M=1}^{M} b_M \sum_{y=1}^{n_{H_l}} h_y^s - c\sum_{x=1}^{n_{I_l}} i_x \tag{10}$$

Following is a definition of the joint and conditional probability distributions:

$$p(I_l,H_l) = \frac{1}{z}\exp\left(-E(I_l,H_l)\right) \tag{11}$$

$$p(h_{xy}^s = 1|\, n_{I_l}) = sigmoid((\widetilde{W_M} *_v I)_y + b_M) \tag{12}$$

$$p(i_x = 1|\, n_{H_l}) = sigmoid(\textstyle\sum_M (W_M \ *_f h^M)_y + b_M) \tag{13}$$

Here, $*_v$ denotes valid convolution and $*_f$ denotes full convolution. where, $\widetilde{W_M} = \Delta W_{nw-y+1}^{M}$. Similar to that of RBM, CRBM is trained using block Gibbs Sampling as an extension of Gibbs Sampling in order to maximize the similarity of distribution between the constructed input layer and the hidden layer and, in that case, obtain the equilibrium state. Convolutional deep belief networks (CRBMs) use probabilistic max-pooling as a fundamental building component.

Calculating the precise gradient for the log-likelihood term is difficult while training convolutional RBMs. Contrastive divergence, however, is an efficient method for approximating the gradient. A sparsity penalty term is added to the log-likelihood goal since a typical CRBM is significantly overcomplete. The training goal might be stated more precisely as

$$maxi_{W,b,c} = \mathcal{L}_{likelihood}(W,b,c) + \mathcal{L}_{sparsity}(W,b,c) \quad (14)$$

where $L_{sparsity}$ is a penalty term that requires the hidden units to have sparse average activations and $L_{likelihood}$ assesses how well the CRBM approximates the distribution of the input data. The "capacity" of the network can be understood as being limited by this sparsity regularization, which frequently produces feature representations that are simpler to understand. We stack the CRBMs to create a convolutional deep belief network once the parameters for each layer have been trained. Each training involves training the RBM of the lowest layer through each subsequent layer until the top layer is reached. In order to achieve its standard convolutional layer, CBRN uses a 3×3 convolutional kernel with 64 channels. Table 2 displays the CBRN Net hyper parameter.

Table 2. Hyper parameter setting

Parameter	Value
No. of Neurons	512
Learning rate	0.02
Activation function	ReLu
No. of epochs	50
Batch size	100 -250
Dropout	0.03

There are several hyperparameters to tune, including the number of neurons, activation function, optimizer, learning rate, batch size, and epochs. Typically, the connection weights of a neural network serve as its parameters. During the training phase, these traits are revealed in this circumstance. Finally, CDBN feature learning is unsupervised, allowing for extensive use of unlabeled images.

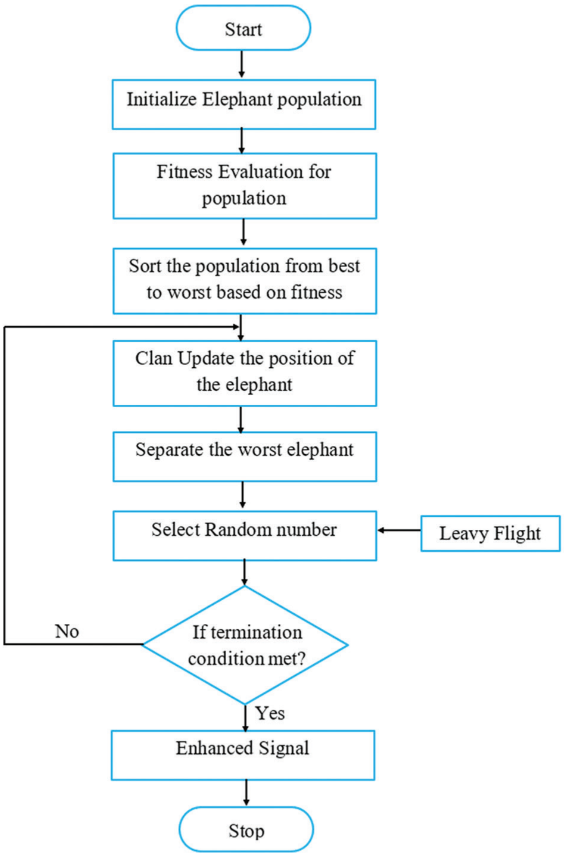
4.4. FEATURE SELECTION VIA ENHANCED ELEPHANT HEAD ALGORITHM

The Elephant Herd Algorithm (EHA) algorithm is modeled by the behavior and way of life of elephants. EHA is a heuristic intelligence system based on elephants' nomadic lifestyles. Elephants exhibit social behavior and have a complicated structure of females and calves. The EHA algorithm selects the most pertinent features from the extracted speech image features.

The number of elephants in this algorithm represents the features that were taken from the input layer; the most pertinent features are the best female elephant of the clan after the matriarch has passed away; and the irrelevant features represent the male elephants with the lowest fitness value. The suggested Enhanced Elephant Herd Algorithm's flowchart is shown in Figure 3. Figure 3 shows the flowchart for the proposed Enhanced Elephant Herd Algorithm.

A group of elephants is made up of several clans, each of which is led by a matriarch who may also have calves or other related females under her care.

Following are some of the algorithm's recommended rules: Elephants live in clans, and each tribe has a set number of elephants. Each clan also has a matriarch, who is the clan's chief (the fittest elephant of the clan). A predetermined number of elephants (worst candidates) must depart the clan each generation, and all the elephants of a clan live together under the authority of the matriarch. Clan update and separation operators make up the two stages of the Elephant Herding Optimization algorithm.



Start
Initialize Elephant population
Fitness Evaluation for population
Sort the population from best to worst based on fitness
Clan Update the position of the elephant
Separate the worst elephant
Select Random number
Leavy Flight
No
If termination condition met?
Yes
Enhanced Signal
Stop

Fig. 3. Flow Diagram of the proposed Enhance EHO

The number of elephants in this algorithm represents the features that were taken from the input layer; the most pertinent features are the best female elephant of the clan after the matriarch has passed away; and the irrelevant features represent the male elephants with the lowest fitness value. The suggested Enhanced Elephant Herd Algorithm's flowchart is shown in Figure 3. Figure 3 shows the flowchart for the proposed Enhanced Elephant Herd Algorithm.

A group of elephants is made up of several clans, each of which is led by a matriarch who may also have calves or other related females under her care. Following are some of the algorithm's recommended rules: Elephants live in clans, and each tribe has a set number of elephants. Each clan also has a matriarch, who is the clan's chief (the fittest elephant of the clan). A predetermined number of elephants (worst candidates) must depart the clan each generation, and all the elephants of a clan live together under the authority of the matriarch. Clan update and

separation operators make up the two stages of the Elephant Herding Optimization algorithm.

The entire population of elephants is initially split up into '*y*' clans. Each elephant m_x denotes the new position is influenced by the matriarch m_x. The clan m_x elephant '*y*' can be determined using

$$P_{n,m_{x,y}} = P_{m_{x,y}} + \alpha \times \left(P_{best,m_x} - \lambda_{m_{x,j}}\right) \times L \quad (15)$$

where [0,1] is a scaling factor, $P_{best,\, m_x}$ is the location with the best fitness value inside clan "*x*", and $P_{n,m_{x,y}}$ represent the old and new positions of elephant "*y*" in clan *x*, respectively. With a normal distribution and a value between [0, 1], L is a random number. For each clan, the best elephant is determined using

$$P_{n,m_{x,y}} = \beta \times P_{ct,m_x} \quad (16)$$

where $\beta \in [0,1]$ is a scaling factor that defines how the position of the clan leader $P_{n,m_{x,y}}$ will change for the following iteration depending on the effect of the clan center P_{ct,m_x}. Eq. (12) is evaluated to determine a clan center's value:

$$P_{ct,m_x,k} = \frac{1}{N_{m_x}} \times \sum_{y=1}^{N_{m_x}} P_{ct,m_x,k} \quad where\ 1 \leq k \leq K \quad (17)$$

Where the number of elephants in the clan is signified as N_{m_x}, the k^{th} dimension of an individual elephant. In Eq. (16), the update of the matriarch position is related to the information of all members of the clan.

The worst solution individuals are replaced by randomly initialized individuals during the separation procedure. It expands the population of elephants and enhances their capacity for exploration. The least valuable elephants in each tribe are relocated to the position indicated by

$$P_{w,m_x} = P_{Min} + (P_{Max} - \lambda_{Min} + 1) \times L \quad (18)$$

where P_{w,m_x} is the position with the worst fitness value in clan '*x*'; P_{Min} and P_{Max} are the upper and lower bound of the elephant's position, respectively; L is a random number with a normal distribution in the range [0, 1].

The slower convergence rate that results from using random numbers is due to problems like lack of exploitation and random replacement of the poorest person. To address this problem, the LF mode is used with the EHO. The LF is represented by,

$$LF(L) = \begin{cases} 1 & L < 1 \\ (L)^{-F} & L \geq 1 \end{cases} \quad (19)$$

Equations 18 and 19 are combined to progress EHA with LF as follows:

$$P_{w,m_x} = P_{Min} + (P_{Max} - \lambda_{Min} + 1) \times LeF(L) \quad (20)$$

As a result, the EEHA specifies the following as the most important features:

$$RF_x = \{rf_1, rf_2, rf_3, \dots \dots, rf_n\} \quad (21)$$

The features that were randomly chosen are then added together after the inputs given are multiplied by the feature vectors. The mathematical representation of the input layer is,

$$I_x = \sum_{x=1}^{n} RF_x w_x + B_x \quad (22)$$

The input features are shown as RF_x, the weight values are shown as w_x, and the bias value is shown as B_x where the IL is shown as I_x. The enhanced image of a speech signal is converted into signals by using the Inverse SPWVD technique.

5. RESULT AND DISCUSSION

These sections provide details of the experiment and discuss the simulation's results. In the following section, the effectiveness of a speech enhancement system based on the proposed Beamforming-based Deep Learning Network (Bf-DLN) methodology is assessed. The study includes three types of speech: speech with ambient noise, speech with a source of interference, and speech with both sources of interference and ambient noise. The required interference speech is used as an input source, and MATLAB is used to display the findings.

5.1. EXPERIMENTAL SETUP

Two unidirectional microphones in a silent car were used to record recorded speech at an 8kHz sample rate to create the inputs. Then, to create realistic in-vehicle noisy speech signals, actual in-car noise that was recorded using the identical setup with the car driving in a typical motorway condition was mixed with the clean speech. The center of the vehicle is where the two microphones are situated. The desired source data and interference data are provided by two speakers, SP1 and SP2, respectively. Both microphones are mounted on a car's rear unit layout, 0.1 meters apart from the speakers. The center of the back unit holds the microphone array. The intended source is set to speaker 1, and the interfering source is set to speaker 2.

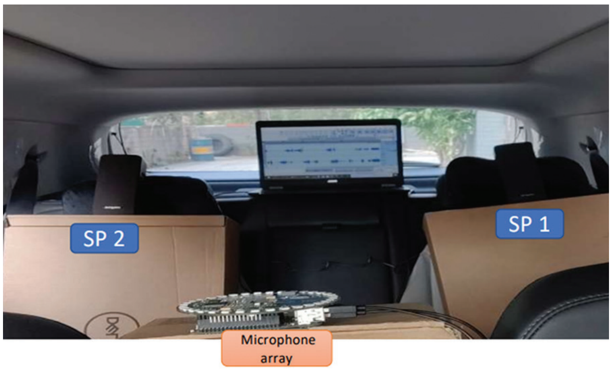


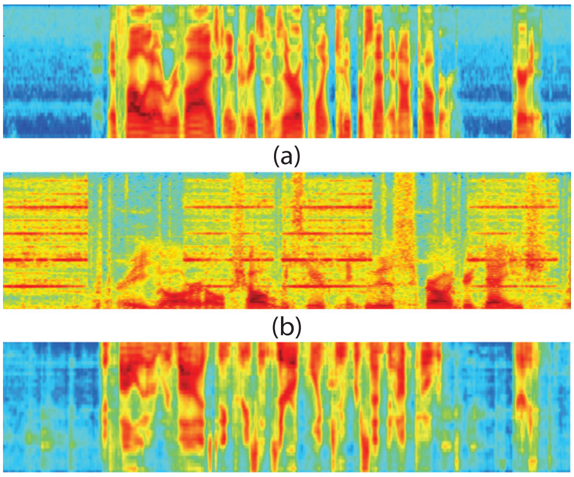
Fig. 4. Realtime data collection setup

With various combinations of input sources at various angles, 30 sets of data in all were obtained.

The parameters used in the suggested method are shown in Table 1. 16 KHz sampling is used for the speech. With segSNR values roughly ranging from -8 dB to -3 dB, noise degrades the signals under various circumstances.

Table 3. Simulation Parameters

Characteristic	Parameter
No. of sources	2
Source classifications	source of speech
Number of mics	2
Sampling rate	8kHz
Size of the FFT window	512



(a)

(b)

(c)

Fig 5. Spectrograms (a) clean image (b) Noisy image (c) Signal enhanced by the proposed method

Figure 5 displays the spectrograms for (a) clear speech, (b) noisy speech and (c) signal increased by the suggested approach. By comparing (b) and (c), it is clear that the suggested Bf-DLN effectively enhances the noise components by demonstrating the method's efficacy.

5.2. EXPERIMENTAL RESULT

The effectiveness of the proposed technique is assessed in comparison to existing methods using the perceptual assessment of speech quality (PESQ), short-time objective intelligence (STOI), segmental SNR improvement (SSNRI), and signal-to-distortion ratio (SDR). The comparative analysis section evaluates the effectiveness of both the existing and the suggested technique. The effectiveness of the various current approaches is evaluated in this section. In addition, we provide performance under various speech and noise SNRs to provide a more complete evaluation of speech quality. Performance scores such as PESQ, STOI, SSNRI, and low SDR reflect improved performance.

Increased speech signal quality and intelligibility are a benefit of GCN [18], but it predicts lower sound quality when used with non-parallel data. On the other hand, DNN-C [19] has the advantage of improving the corresponding noisy input, and all of the channel-wise enhanced outputs are fed into a DNN fusion model to produce a practically clean signal. However, it performs poorly for improving the signal. Additionally, Bi-LSTM [20] can extract local and global characteristics and achieves competitive results; its main drawback is a high computational cost.

Table 4. Comparison of proposed with an existing method for Microphone 1

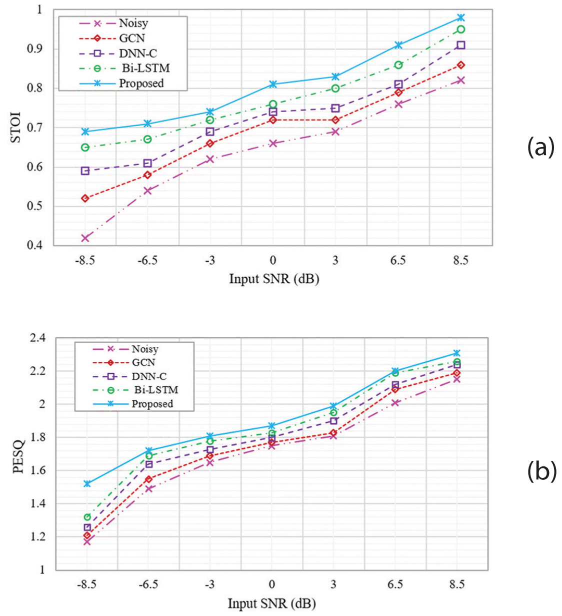
Methods	PESQ	STOI	SSNRI	SDR
Noisy	1.54	0.57	0.00	5.14
GCN [18]	1.68	0.62	7.51	3.84
DNN-C [19]	1.75	0.69	9.52	5.29
Two-stage Bi-LSTM [20]	1.82	0.72	10.45	7.15
Proposed	1.92	0.79	11.5	9.24

Table 5. Comparison of proposed with an existing method for Microphone 2

Methods	PESQ	STOI	SSNRI	SDR
Noisy	1.72	0.668	0.00	3.18
GCN [18]	1.54	0.672	5.12	4.15
DNN-C [19]	1.81	0.752	7.15	6.47
Two-stage Bi-LSTM [20]	1.86	0.842	9.157	7.15
Proposed	1.99	0.954	12.45	9.15

Increased speech signal quality and intelligibility are a benefit of GCN [18], but it predicts lower sound quality when used with non-parallel data. On the other hand, DNN-C [19] has the advantage of improving the corresponding noisy input, and all of the channel-wise enhanced outputs are fed into a DNN fusion model to produce a practically clean signal. However, it performs poorly for improving the signal. Additionally, Bi-LSTM [20] can extract local and global characteristics and achieves competitive results; its main drawback is a high computational cost.

The results show that the proposed method performs better than other GCN [18], DNN-C [19], and Bi-LSTM [20] methods when evaluated at various SNRs, proving the model's efficacy. Tables IV and V show the averaged STOI, PESR, SSNRI, and SDR scores for the input signal and enhanced output signal of the two-microphone used in the proposed method. The proposed models outperform the current methods in terms of STOI, PESQ, SSNRI score, and SDR.



(a)

(b)

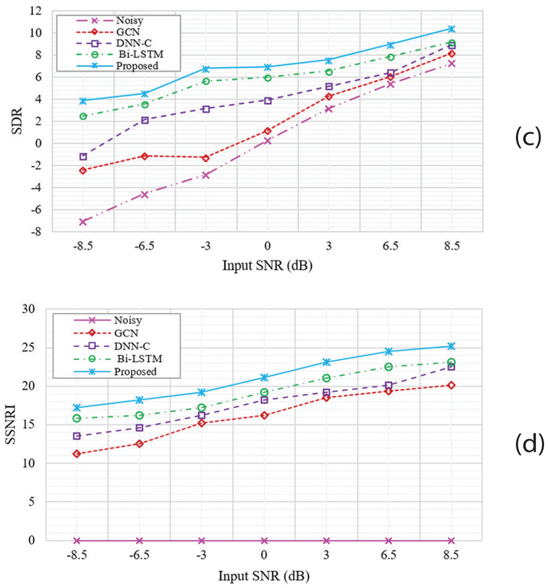


Fig. 6. Performance of (a) STOI, (b) PESQ, (c) SDR, and (d) SSNRI metrics for different SNR levels

In Fig. 6, we analyze the outcomes at various SNR levels. Results for a linear array with two microphones at seven different SNR levels [-8.5, -6.5, -3, 0, 3, 6.5, 8.5] (dB). are displayed. Generally speaking, negative SNR values yield greater performance increases than positive ones. We can see that the proposed Bf-DLN approach improves SDR over the noisy case by more than 8.25 dB for -8.5 dB input SNR. Only 3.48 dB of SDR improvement is seen for the input SNR with the highest values. Additionally, we note that the proposed model outperforms GCN, DNN-C, and Bi-LSTM across the board for SNR values. The proposed model has 0.95, 1.54, and 4.26 improvement on STOI, PESQ, and SDR, respectively. The performance of the SSNIR also improved compared to the existing techniques.

6. CONCLUSION

In this paper, a novel Beamforming based Deep learning Network (Bf-DLN) has been proposed for speech enhancement. Initially, the captured microphone array signals are pre-processed using an Adaptive beamforming technique named Least Constrained Minimum Variance (LCMV). Consequently, the proposed method uses a time-frequency representation to transform the pre-processed data into an image. Time-domain speech signals are converted into pictures using the smoothed pseudo-Wigner-Ville distribution (SPWVD). These converted images are given as input to the convolutional deep belief network (CDBN) for extracting the most relevant features. Enhanced Elephant Heard Algorithm (EEHA) is used for selecting the desired source by eliminating the interference source. The experimental result demonstrates the effectiveness of the proposed strategy in removing background noise from the original speech signal. The proposed strategy outperforms existing methods in terms of PESQ, STOI, SSNRI, and SNR. Results indicate the superiority of our approach when compared to prior state-of-the-art methods.

7. REFERENCES

[1] J. Benesty, "Fundamentals of speech enhancement", Springer, 2018.

[2] B. K. Khonglah A. Dey, S. R. Prasanna, "Speech enhancement using source information for phoneme recognition of speech with background music", Circuits, Systems, and Signal Processing, Vol. 38 No. 2, 2019, pp. 643-663.

[3] Z. X. Li, L. R Dai, Y. Song, L. McLoughlin, "A conditional generative model for speech enhancement", Circuits, Systems, and Signal Processing, Vol. 37, No. 11, 2018, pp. 5005-5022.

[4] P. Malathi, G. R. Suresh M. Moorthi N. R. Shanker, "Speech Enhancement via Smart Larynx of Variable Frequency for Laryngectomee Patient for Tamil Language Syllables Using RADWT Algorithm", Circuits, Systems, and Signal Processing, Vol. 38, No. 9, 2019, pp. 4202-4228.

[5] T. K. Dash, S. S. Solanki, G. Panda, "Improved phase aware speech enhancement using bio-inspired and ANN techniques", Analog Integrated Circuits and Signal Processing, Vol. 102, No. 3, 2020, pp. 465-477.

[6] J. Yang, B. Xia, Y. Shang, W. Huang, C. Mi, "Improved battery parameter estimation method considering operating scenarios for HEV/EV applications", Energies, Vol. 10, No. 1, 2016, p. 5.

[7] S. Gannot, E. Vincent, S. Markovich-Golan, "A consolidated perspective on multimicrophone speech enhancement and source separation", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 4, 2017, pp. 692-730.

[8] M. Tammen, S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6-11 June 2021, pp. 8443-8447.

[9] E. Gentet, B. David, S. Denjean, G. Richard, V. Roussarie, "Speech intelligibility enhancement by equalization for in-car applications", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4-8 May 2020, pp. 6934-6938.

[10] Y. Alkaher, I. Cohen, "Dual-Microphone Speech Reinforcement System with Howling-Control for In-Car Speech Communication", Frontiers in Signal Processing, Vol. 2, 2022, p. 819113.

[11] N. Saleem, J. Gao, M. I. Khattak, H. T. Rauf, S. Kadry, M. Shafi, "Deepresgru: residual gated recurrent neural network-augmented kalman filtering for speech enhancement and recognition", Knowledge-Based Systems, Vol. 238, 2022, p. 107914.

[12] S. Y. Chuang, H. M. Wang, Y. Tsao, "Improved lite audio-visual speech enhancement", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 30, 2022, pp. 1345-1359.

[13] T. Tao, H. Zheng, J. Yang, Z. Guo, Y. Zhang, J. Ao, Y. Chen, W. Lin, X. Tan, "Sound Localization and Speech Enhancement Algorithm Based on Dual-Microphone", Sensors, Vol. 22, No. 3, 2022, p. 715.

[14] V. Kothapally, Y. Xu, M. Yu, S. X. Zhang, D. Yu., "Deep Neural Mel-Subband Beamformer for in-Car Speech Separation", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4-10 June 2023.

[15] B. Jolad, R. Khanai, "An approach for speech enhancement with dysarthric speech recognition using optimization based machine learning frameworks", International Journal Of Speech Technology, 2023, pp. 1-19.

[16] L. Qian, F. Zheng, X. Guo, Y. Zuo, W. Zhou, "Vehicle Speech Enhancement Algorithm Based on TanhDBN", Proceedings of the IEEE 3rd International Conference of Safe Production and Informatization, Chongqing City, China, 28-30 November 2020, pp. 434-438.

[17] W. Zhou, R. Ji, J. Lai, "MetaRL-SE: a few-shot speech enhancement method based on meta-reinforcement learning", Multimedia Tools and Applications, 2023, pp. 1-20.

[18] S. S. Wang, Y. Y. Liang, J. W. Hung, Y. Tsao, H. M. Wang, S. H. Fang, "Distributed microphone speech enhancement based on deep learning", arXiv:1911.08153, 2019.

[19] P. Tzirakis, A. Kumar, J. Donley, "Multi-channel speech enhancement using graph neural networks", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6-11 June 2021, pp. 3415-3419.

[20] X. Shen, Z. Liang, S. Li, Y. Jiang, "Multichannel Speech Enhancement in Vehicle Environment Based on Interchannel Attention Mechanism", Journal of Advanced Transportation, Vol. 2021, 2021, pp. 1-9.