IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 October 2023, accepted 25 October 2023, date of publication 27 October 2023, date of current version 2 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3328208

RESEARCH ARTICLE

Track and Noise Separation Based on the Universal Codebook and Enhanced Speech Recognition Using Hybrid Deep Learning Method

S. V. ASWIN KUMER[1], LAKSHMI BHARATH GOGU[1], E. MOHAN[2], SUMAN MALOJI[1], BALAJI NATARAJAN[3], G. SAMBASIVAM[4], (Member, IEEE), AND VAIBHAV BHUSHAN TYAGI[5]

[1]Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522302, India
[2]Department of ECE, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu 602105, India
[3]Department of Computer Science and Engineering, Sri Venkateshwaraa College of Engineering and Technology, Ariyur, Puducherry 605102, India
[4]School of Computing and Data Science, Xiamen University Malaysia, Sepang, Selangor 43900, Malaysia
[5]Faculty of Engineering, ISBAT University, Kampala, Uganda

Corresponding author: Vaibhav Bhushan Tyagi (tyagi.fict@isbatuniversity.com)

ABSTRACT The concept of Deep learning is a part of machine learning which is very useful nowadays to achieve accurate voice and speech recognition based on the training data by creating robust algorithms. It is also possible to separate the noise from original speech as well as the separation of tracks in particular audio signal with the help of machine learning algorithms. In this paper, the implementation is applicable for voice assistant to separate the tracks and the noises from the multiple original audio which reproduces simultaneously using the speech enhancement and universal code book. For that, the Hybrid Deep Learning Algorithm has been developed and the training data sets are also created and achieve the accuracy in the speech recognition for the variety of voice assistants. Most of the time, the voice assistant recognizes the voice with noises and musical audio which results in the malfunction of devices which can be controlled by the same voice assistant. The Generative adversarial networks from Deep learning and the blind source separation method from multi-channel model are combined to form this proposed hybrid deep learning model.

INDEX TERMS Blind source separation (BSS) method, deep learning method, generative adversarial networks (GAN), multi-channel method, noise separation, speech recognition, speech enhancement, track separation, voice assistant.

I. INTRODUCTION

The usage of voice assistant in recent trends and technology has become severe nowadays to control the appliances of all types like home appliances, industrial appliances and machines, automated vehicles, smart phones, and other related applications. In that, the major input for that voice assistant is the human voice to control and operate all the terminal nodes. If the input collapses or adds noises, then the voice assistant starts malfunctioning, based on the corrupted inputs. To reduce this effect in the input side, the Hybrid Deep Learning Algorithm has been proposed and the training data sets are also created and tested to achieve accuracy in the speech recognition for the variety of voice assistants. The track separation process is not an easy task if the unwanted information which is also called as interfering source which is like the actual original payload information which is also called as target source. Here, the interfering sources are more and there is only one target source. First, the unwanted tracks should be found out and it must be separated by using the Hybrid Deep Learning Algorithm. Speech separation is a fundamental problem in audio processing, with applications ranging from improving audio quality in communication systems to enhancing speech recognition in noisy environments.

The associate editor coordinating the review of this manuscript and approving it for publication was Yilun Shang.

Traditional methods, such as the Universal Codebook, have been used to separate speech signals from background noise. However, these methods might have limitations in dealing with complex real-world scenarios and varying noise conditions.

The advent of deep learning has revolutionized various fields, including audio processing. Deep neural networks (DNNs) have shown remarkable success in speech enhancement and separation tasks, leveraging their ability to learn complex patterns and features directly from the data. Despite their success, DNNs might require a significant amount of labeled training data and computational resources.

The hybrid approach proposed in this research leverages the strengths of both traditional signal processing methods and deep learning techniques. By combining the Universal Codebook's ability to capture structured speech components and the DNN's capacity to handle intricate patterns, the goal is to achieve improved track and noise separation. Wood et al. [1] developed a framework for speech enhancement using the universal code book to highlight the features of original voice and the noise based on the atomic speech presence probability. Subramanian et al. [2] minimized the error in the automatic speech recognition and optimized the speech enhancement based on the word error rate using multi-channel end-to-end system. He et al. [3] implements the wiener filter to separate the noise and the original information signal with the help of code book which estimates the Auto regressive parameters. Baby et al. [4] making the weighted sum of noisy signals by using decomposition of original signals and enhancing the noise to distinguish the original signal.

II. RELATED WORK

Xiang et al. [5] separates the speech harmonics which consists of residual noise which can be removed by the same driven code book by speech enhancement. Pfeifenberger et al. [6] uses Eiggenet architecture for gain mask estimation from the signal received from the different input sources and all the signals are enhanced to achieve Phase Aware Normalization (PAN), Generalized Eigenvalue (GEV) and Minimum Variance Distortion Less Response (MVDR). Hassani et al. [7] demonstrated the speech processing from more than one source using the noise reduction algorithm which involves the multi-channel wiener filter. Huang et al. [8] uses the multi-band excitation model for speech enhancement along with the Deep neural network to achieve the log power spectra of the signal which are having the noises. Gaich et al. [9] improving the performance and reliability of speech enhancement signal using the phase aware methods which helps to achieve good results in noise reduction. Xiang et al. [10] implements the deep neural network and trains the same network for speech enhancement and noise separation based on multi objective learning. Zhang et al. [11] also uses wiener filtering for noise enhancement using the driven code book which estimates the Signal to Noise Ratio and improving the same for the betterment reproduction of original speech signal. Pirolt et al. [12] uses the phase invariance property to find the harmonics with the help of harmonic phase estimator which separates the signal phase and noise phase. Hussain et al. [13] combines the Adaptive noise cancellation technique and the degenerate unmixing estimation technique to find the correlation between the signals for separating the same. The CEGM approach of Chai et al. [16] can also be used to enhance the speech if the speech is from a single source, if the input is from multiple sources, the frequency and signal strength may vary as well as the sampling rate also changed which results the variation in the DC output power. The hybrid voice activity detection method of Wang et al. [17] can also overcome this effect hence, the output noise power and distortion power dominate which results in the suppression of primary input signal. The multi-channel speech enhancement implementation from Lee et al. and Zang et al. [18] and the improvement in SNR from Nian et al. [20] and Zhang et al. [30] in multi- channel environment are challenging tasks, hence, the bandwidth handling by Zhu et al. [22] of the signal from multiple sources are distinguishable which helps to separate the tracks and noises from the original information. Einizade et al. [34] introduces U-GraphJADE-GL, a method for blind separation of graph signals. It addresses the limitation of assuming known underlying graphs in graph-based methods, proposing a unified objective function optimized using Block Coordinate Descent. U-GraphJADE-GL is compared with other methods in blind source separation tasks and applied successfully to denoise epileptic EEG signals and audio speech separation.

Wang et al. [35] introduces a novel multichannel blind source separation (BSS) method using a convolutive transfer function (CTF) for overdetermined scenarios. It employs a frequency-wise convolutive mixture model, estimating demixing matrix via iterative projection and NMF parameters using multiplicative update. The method, advantageous for representing long impulse responses with short windows, outperforms ILRMA and FastMNMF in separating sources in reverberant environments.

Du et al. [36] presents a computationally efficient algorithm for BSS in overdetermined mixtures. It introduces a modified iterative source steering (ISS) algorithm for overdetermined independent vector analysis (OverIVA) and independent low-rank matrix analysis (OverILRMA). Experimental results demonstrate comparable or superior speech separation performance with lower computational cost compared to conventional iterative projection-based methods.

Brendel et al. [37] addresses convolutive BSS in audio processing, focusing on Independent Component Analysis (ICA) methods. It clarifies the relationships between Frequency Domain ICA (FD-ICA), Independent Vector Analysis (IVA), and TRIple-N Independent component analysis for CONvolutive mixtures (TRINICON), establishing a common framework for these algorithms.

Munakata et al. [38] introduces an unsupervised multichannel method for separating moving sound sources using amortized variational inference (AVI). It enhances the neural full-rank spatial covariance analysis (FCA) method by incorporating time-varying spatial information, improving performance in separating and localizing moving sources compared to existing methods.

Muñoz-Montoro et al. [39] introduces a harmonic constrained Multichannel Non-Negative Matrix Factorization (MNMF) method for BSS. It encodes spatial information using magnitude and phase differences, models source variances with harmonic constrained NMF, and uses the constant-Q transform for the spatial covariance matrix. The method, initialized with steered response power (SRP) with the phase transform (PHAT), exhibits reliable results in music source separation tasks.

Hasuike et al. [40] addresses frequency-domain blind source separation for audio signals. It introduces a deep neural network-based permutation solver to tackle the long-standing permutation problem in estimating source components. Experimental results validate the effectiveness and robustness of the proposed approach across different datasets.

The self-supervised [22] approach is also a solution to find the actual information from the multiple signal environments, if the attentive training [23] is included, then the efficiency can be improved for speech recognition. Involving the Convolutional Neural Network [24] for speech recognition and the trained DNN [25], [26], [27], [29], along with different algorithms can help to achieve efficient results in speech enhancement. The art of separating [28] the track after enhancement produces better results than other approaches, the combined approach followed in the proposed method.

III. PROPOSED METHODOLOGY

The voice assistant receives the input signal from the multiple sources, and it has no knowledge about that, which signal is information and which signal is an unwanted signal. To distinguish that, the speech enhancement process is implemented. The process of speech enhancement having different methodologies, but, in this proposed model, the universal code book is used to enhance the entire input from multiple sources and after that, with the help of code book, the entire input information signal is enhanced and applied to the Hybrid Deep Learning which is shown in Figure 1. The process of discriminating the signals using the data samples in the dynamic system is known as Generative Adversarial Network (GAN). The GAN along with Blind Source Separation (BSS) method represents an algorithm to separate the noise tracks and the original track.

A. UNIVERSAL CODEBOOK FOR TRACK AND NOISE SEPARATION

The Universal Codebook is a traditional method used in speech processing to model speech components and background noise separately. It provides a structured framework for representing different speech elements and has been used effectively in various speech separation tasks.

B. DEEP LEARNING FOR ENHANCED SPEECH RECOGNITION

Deep learning methods, particularly deep neural networks, have gained attention due to their capability to learn complex features from raw data. In the context of speech separation, DNNs can be trained to differentiate between speech and noise patterns, leading to improved quality in separated tracks.

C. HYBRIDIZATION FOR IMPROVED PERFORMANCE

The hybrid approach seeks to capitalize on the complementary nature of traditional methods and deep learning. By integrating the Universal Codebook with a DNN-based model, the research aims to enhance the quality of separated speech tracks, thereby leading to better speech recognition performance in noisy environments.

D. CHALLENGES AND EXPECTATIONS

While deep learning methods have shown promise, they often require a substantial amount of labeled data for training and significant computational resources. The hybrid approach aims to address potential limitations of DNNs by leveraging the structured representations provided by the Universal Codebook, potentially leading to enhanced performance with a reduced need for extensive labeled data.

IV. IMPLEMENTATION METHODOLOGY

The implementation of this proposed method involves the basic steps like speech enhancement, needs reference from the universal code book and the Hybrid Deep Learning Algorithm to separate the noise tracks from the actual information signal.

A. HYBRID DEEP LEARNING ALGORITHM

Pre-processing

- The mixed speech signal is first pre-processed to remove any background noise. This can be done using a variety of methods, such as noise cancellation or spectral subtraction.
- The pre-processed signal is then divided into frames. Each frame is a short segment of the signal, typically 20-30 milliseconds long.

GAN-based separation

- A GAN is trained to learn the distribution of the speech signals in the mixed signal. The GAN consists of two networks: a generator and a discriminator.
- The generator is responsible for creating fake speech signals that are like the real speech signals in the mixed signal. The discriminator is responsible for distinguishing between real and fake speech signals.

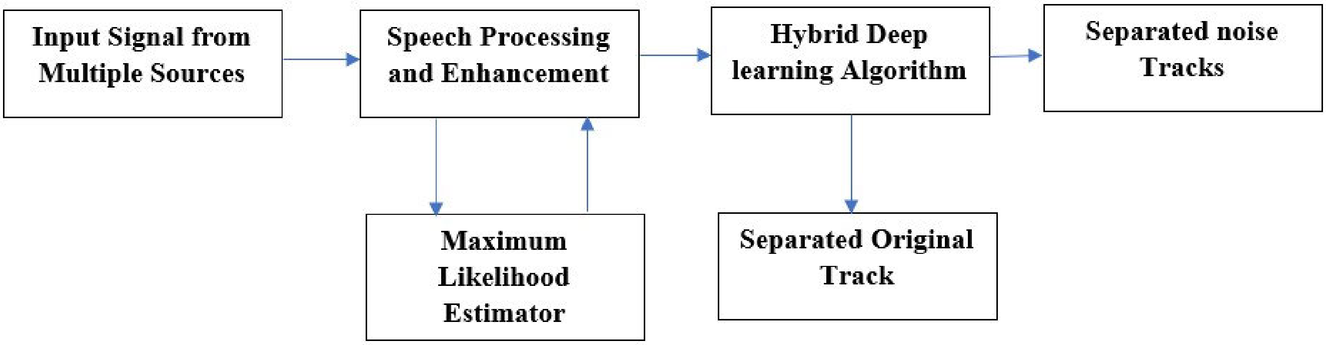

Input Signal from Multiple Sources
Speech Processing and Enhancement
Hybrid Deep learning Algorithm
Separated noise Tracks
Maximum Likelihood Estimator
Separated Original Track

FIGURE 1. Sequence diagram of proposed implementation.

- The GAN is trained by alternating between training the generator to create more realistic speech signals and training the discriminator to become better at distinguishing between real and fake speech signals.

BSS-based separation

- After the GAN has been trained, it can be used to create a preliminary estimate of the individual speech signals in the mixed signal.
- The preliminary estimate is then used as input to a BSS algorithm. The BSS algorithm is responsible for further separating the individual speech signals from the mixed signal.

Post-processing

- The output of the BSS algorithm is then post-processed to improve the quality of the separated speech signals. This can be done by using a variety of methods, such as noise reduction or spectral smoothing.

B. SPEECH ENHANCEMENT AND PROCESSING

The input signal from multiple sources having the various amplitudes which are enhanced for processing the same signals. This implementation is called preprocessing of the input signal for further processing which improves the quality of the speech. Here, the spectral subtraction method is used to increase the perceptual quality and the intelligibility of the noise affected signal. It restores the magnitude spectrum of actual information signal which cancels the noise spectrum in the unbounded input information.

$$I(n) = A(n) + N(n) \tag{1}$$

where, the I(n) indicate the input signal which is equal to the Actual information Signal A(n) and the Noise Information N(n) in time domain. By taking Fourier Transform and magnitude for this signal to find the Actual information Signal in the spectrum,

$$|A\,(w)|^2 = |I\,(w)|^2 - |N\,(w)|^2 \tag{2}$$

The noise information should be subtracted from the input information to obtain the actual information, which is indicated in equation 2, also the representation of the input information spectrum is shown in Figure 3 and the representation of the actual information spectrum is shown in Figure 6. The Speech pauses are combined for averaging the same to estimate the noise spectrum, which is indicated in equation 3, in that the number of pulses is represented as P [14] and the Spectrum of the noise information is shown in Figure 5.

$$|N\,(w)|^2 = \frac{1}{P}\sum\nolimits_{i=0}^{P-1} |I\,(w)|^2 \tag{3}$$

The parameters from Maximum Likelihood Estimator will be processed along with the Actual information Signal and given to the Hybrid Deep Learning Algorithm for further processing.

C. MAXIMUM LIKELIHOOD ESTIMATOR

The code book driven receives the signal spectrum and noise spectrum from the speech enhancement module and the same will be processed to estimate the maximum likelihood parameters as shown in equation 4 [15] which is shown in Figure 7.

$$\left\{k^*, l^*\right\} = \arg\max_{k,l}\max_{\sigma_x^2\sigma_w^2} p_y(Y\,|\,u_x^k, u_w^l; \sigma_x^2, \sigma_w^2) \tag{4}$$

where the variables k and l are represent parameters that maximize the likelihood function, and the functions $arg\max_{k,l}(k, l)$ represents the values of k and l for which the expression inside the parentheses attains its maximum value and $\max(\sigma_x^2\sigma_w^2)$ represents the maximum value with respect to σ_x^2 *and* σ_w^2, and the σ_x^2 *and* σ_w^2 represents the Variance of the random variable x and w, and $p_y(Y|)$ represents the conditional probability density function of $Y|$ given the specified parameters, and u_x^k, u_w^l are parameters related to the random variables x and w, respectively, and k and l index different values of these parameters.

D. GAN AND BSS BASED SEPARATION

A novel algorithm has been developed to separate the noise tracks and the actual information track with the help of Hybrid Deep Learning algorithm, which combines generative Adversarial Network approach and Blind Source Separation approach for speech enhancement. This algorithm has been shown to be effective in separating multiple speech signals

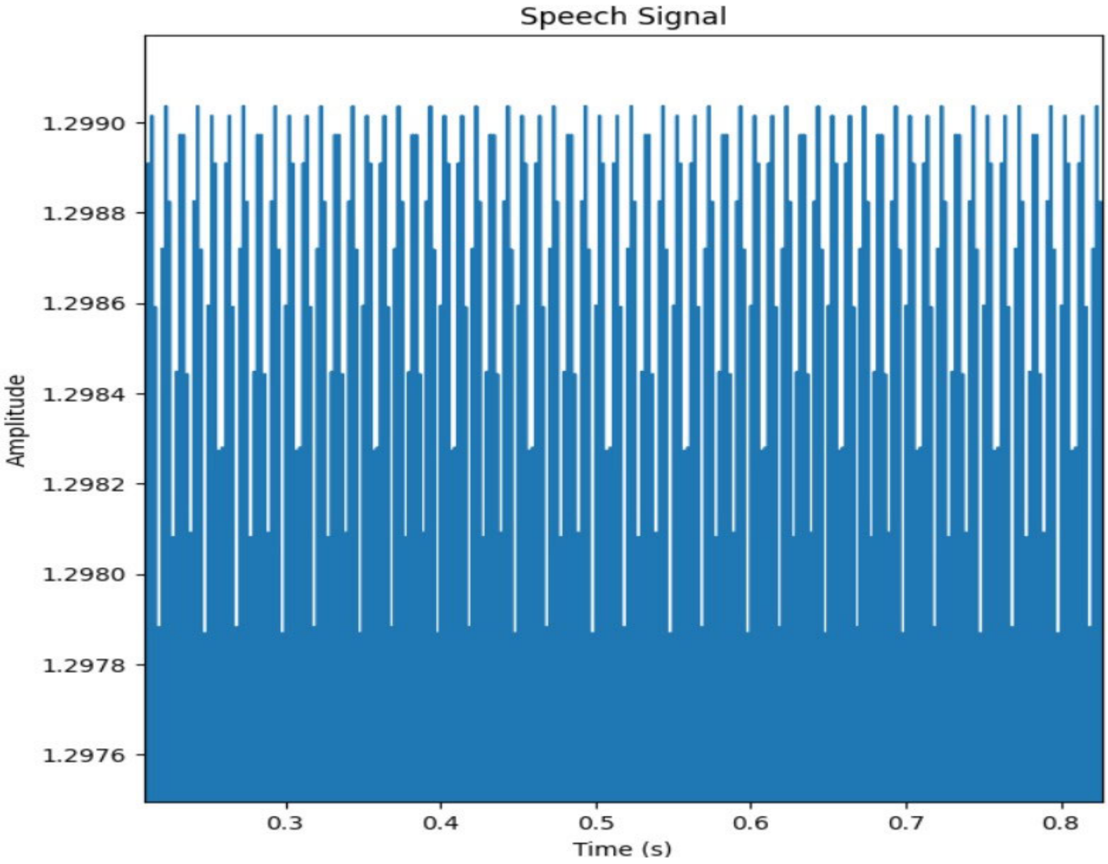


FIGURE 2. Input information.

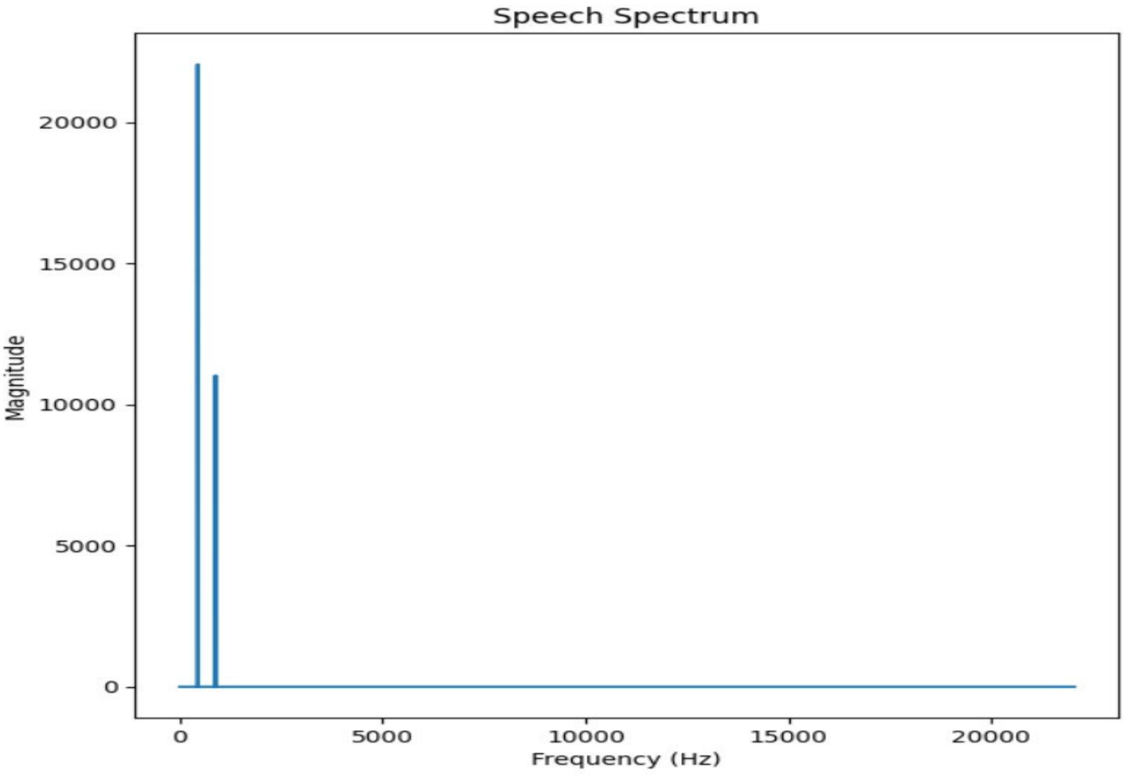


FIGURE 3. Spectrum of input information.

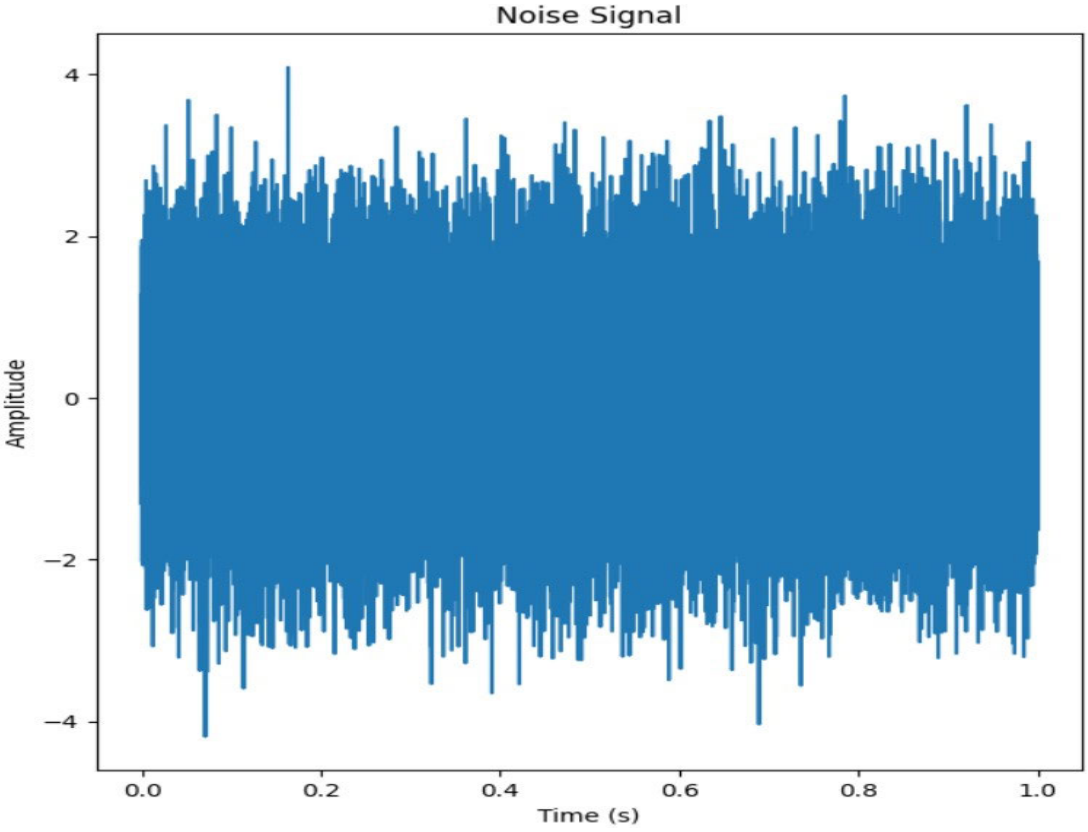


FIGURE 4. Noise information.

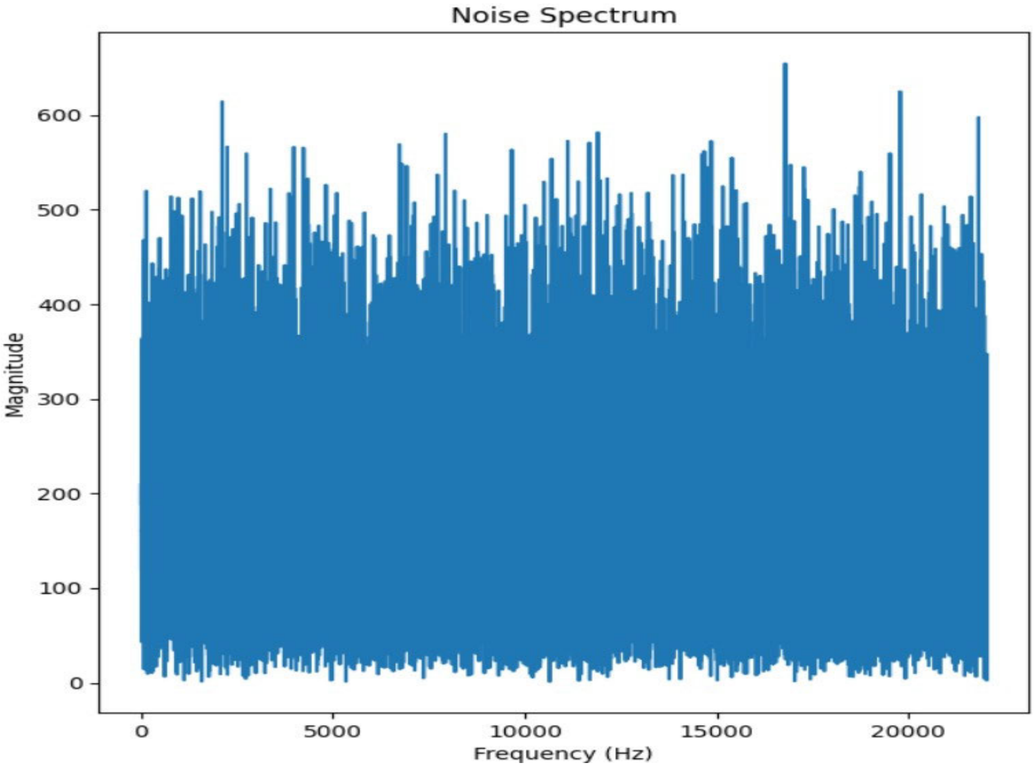


FIGURE 5. Spectrum of noise information.

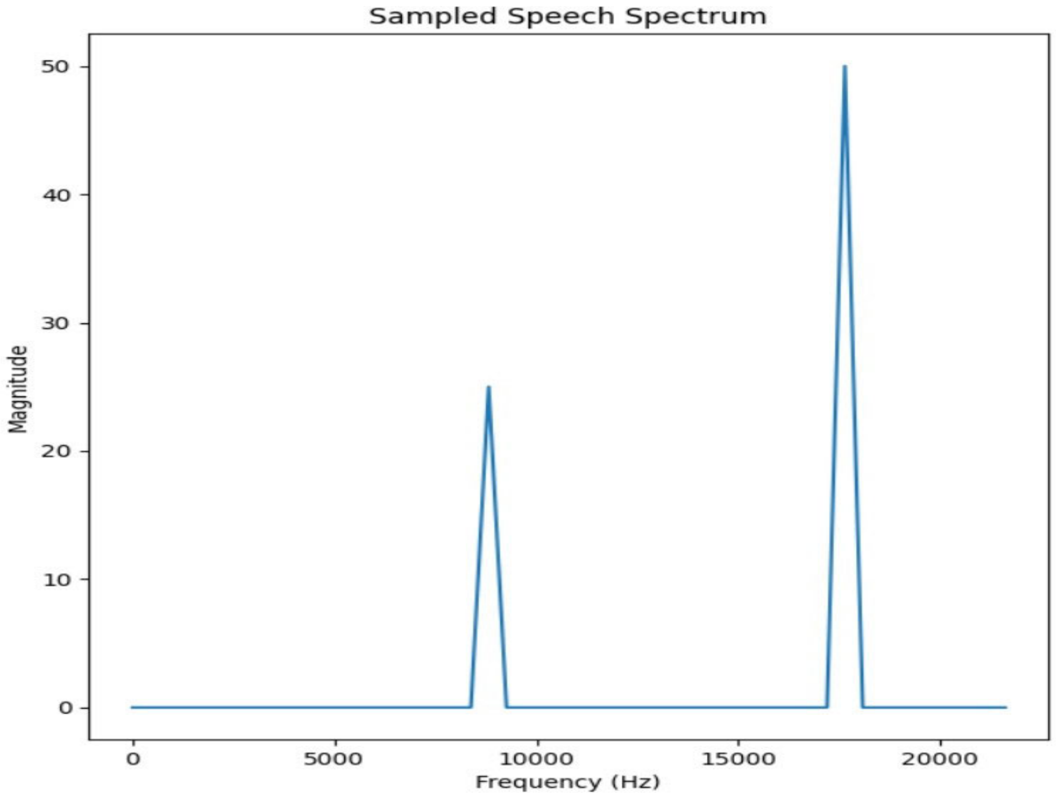
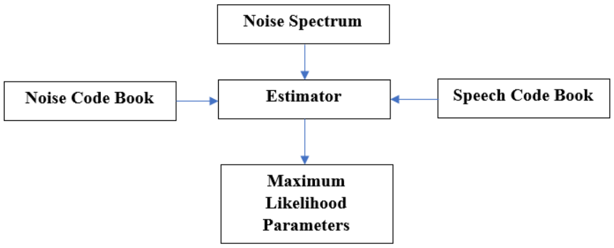


FIGURE 6. Spectrum of sampled information with 44100 sampling rate.



Noise Spectrum

Noise Code Book

Estimator

Speech Code Book

Maximum Likelihood Parameters

FIGURE 7. Block diagram of maximum likelihood estimator.

from a mixed signal. It can achieve good separation quality even in the presence of background noise. Usually, the signal parameters have a similar range of values, so, the separation of tracks becomes complicated. So, the algorithm uses the frequency of the signal to separate the tracks with the constant time slots. Each frequency received from the sources will be put up in the frequency Bins and based on the value of frequency, the bin has been chosen. The rest of the parameters like harmonics, Bin power and total power may change based on the instants and not by the frequency.

The actual sampling rate for hi fidelity audio 44100 Hz is taken as consideration, with that threshold, the sampling rate is decreased to 43200 Hz for our proposed method. The fundamental frequency taken for the sample is 1687.500000 Hz and the Sampling Frequency is 43200 Hz. The Spurious Free Dynamic Range (SFDR) and the Signal to Noise and Interference Ratio (SINR) and the Signal to Noise Ratio (SNR) are also listed in the following tables with the different output powers of the processing speech signal.

$$\mathrm{SFDR} = \mathrm{Ps} - \mathrm{Pn} \tag{5}$$

$$\mathrm{SINR} = 10 * \log10(\mathrm{Ps}/(\mathrm{Pn} + \mathrm{Pi})) \tag{6}$$

$$\mathrm{SNR} = 10 * \log10(\mathrm{Ps}/\mathrm{Pn}) \tag{7}$$

The Ps represents the signal power, Pn represents the noise power, and the Pi represents the interference power.

V. RESULTS AND DISCUSSIONS

By observing the values tabulated in Table 1 and Table 2, the sampling rate of 43200 provides better results than other sampling rates. The same approach is applicable for multiple speech signals received to identify the actual informative signal. The complex power spectra of multiple speech signals are shown in figure 9.

The frequency of instants 11, 20 and 27 are the low frequencies which are put up in the low value bins which can be separated using the Hybrid Deep Learning which are shown in Table 3, 4 and 5. The frequency of the instants 38, 39, 40, 42, 46, 48 and 49 are also the low frequencies which are put up in the low value bins which can be separated using the Hybrid Deep Learning which are shown in Table 6 and 7.

To measure the performance of the proposed method, the parameters like Signal to Noise Ratio (SNR) (Equation 8),

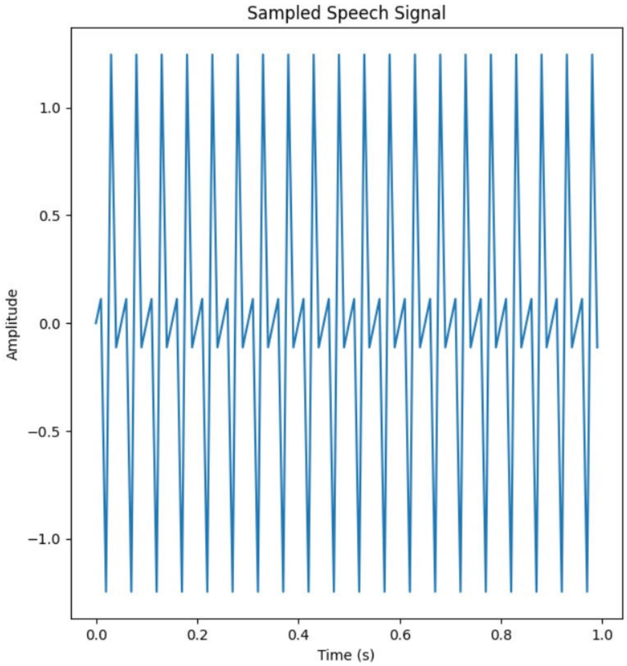


FIGURE 8. Sample signal plot with 44100 sampling rate.

TABLE 1. Different output powers of the processing speech signal.

Output Powers	Values at 44100 Sampling rate	Values at 43200 Sampling rate
DC	-30.16 dB c	-31.91 dB c
Noise	-16.93 dB c	-17.49 dB c
Distortion	-34.76 dB c	-34.99 dB c
Fundamental	0.00 dB c	0.00 dB c
Total	0.11 dB c	0.10 dB c

TABLE 2. Different ratios of the processing speech signal.

Ratios	Values at 44100 Sampling rate	Values at 43200 Sampling rate
SFDR	-79.19 dB c	-79.91 dB c
SINR	-16.33 dB c	-16.49 dB c
SNR	-33.79 dB c	-34.09 dB c

Short Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ) and Mean Opinion Score (MOS) are used and the comparison with other methods with different datasets like CHIME-4 [31], NOISEX-92 [32], and WSJO-2 mix [33] are also represented in the table 8.

$$\mathrm{STOI} = 1 - \mathrm{d/D} \tag{8}$$

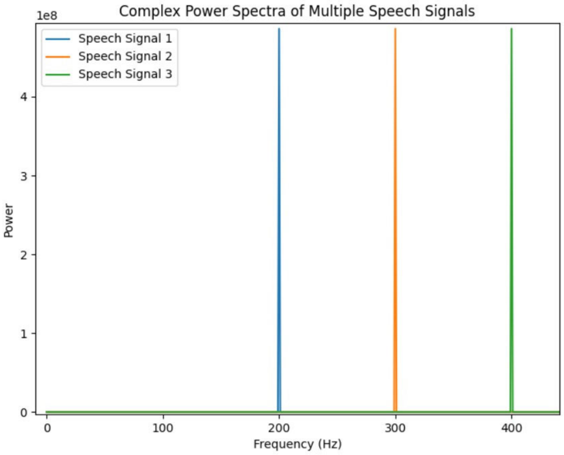


FIGURE 9. Power spectrum of multiple input signal plot.

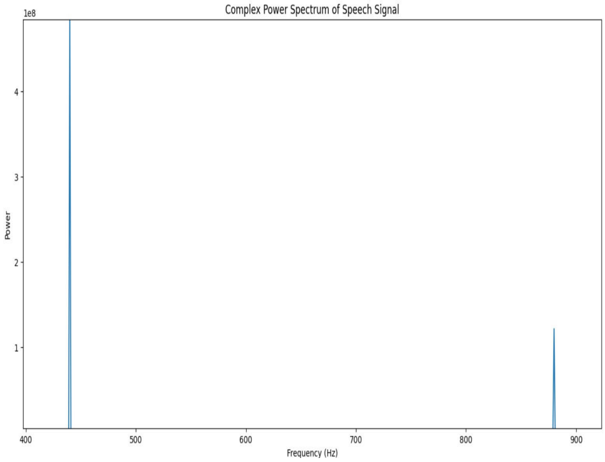


FIGURE 10. Power spectrum of input signal plot with 44100 sampling rate.

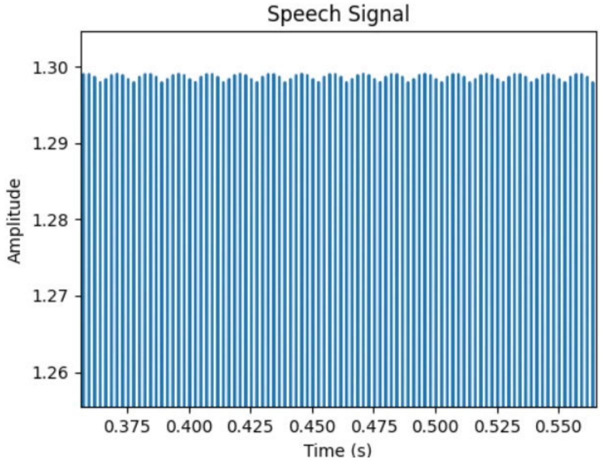


FIGURE 11. Output information.

In STOI measure, d is the normalized difference between the temporal envelopes of the clean and noisy/processed speech and D is the maximum possible normalized difference.

$$\mathrm{PESQ} = 4.5 - 0.1^{*}\mathrm{d}_{\mathrm{SYM}} - 0.0309^{*}\mathrm{d}_{\mathrm{ASYM}} \tag{9}$$

In PESQ measure, $\mathrm{d}_{\mathrm{SYM}}$ is the symmetrical difference between the degraded and reference loudness spectra and $\mathrm{d}_{\mathrm{ASYM}}$ is the asymmetrical difference between the degraded

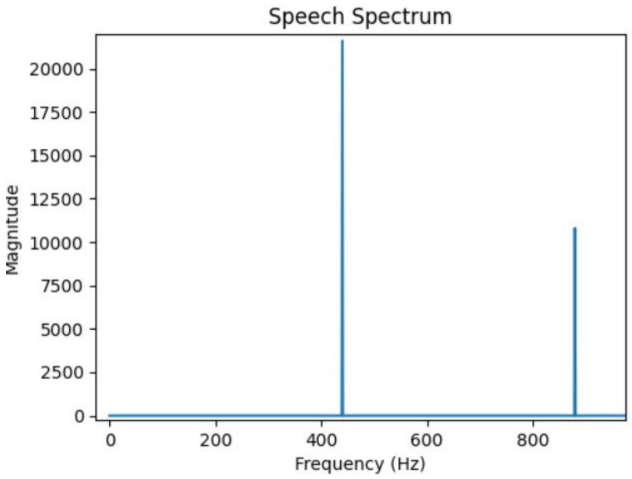


FIGURE 12. Spectrum of output information.

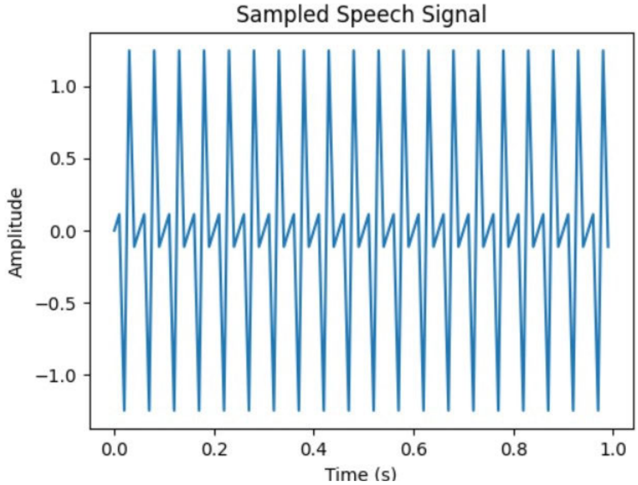


FIGURE 13. Sampled output information with 43200 sampling rate.

TABLE 3. Signal parameters of time slot 1.

Instant	Frequency	Bin	Harmonics	Bin Power in dB	dB below peak in dB	Total Smaller Power in dB
1	1687.500	80	1	0.00	0.00	-16.62
2	1666.406	79	0	-21.53	21.53	-18.31
3	1708.594	81	0	-23.43	23.43	-19.91
4	1645.313	78	0	-27.96	27.96	-20.65
5	1729.688	82	0	-29.95	29.95	-21.19
6	1624.219	77	0	-31.17	31.17	-21.65
7	1750.781	83	0	-33.09	33.09	-21.97
8	1603.125	76	0	-34.20	34.20	-22.24
9	3375.000	160	2	-34.50	34.50	-22.50
10	1582.031	75	0	-34.82	34.82	-22.77

and reference loudness spectra.

$$\mathrm{MOS} = 1 - 5^{*}(Q_{\mathrm{B}} - Q_{\mathrm{W}})/4 \tag{10}$$

In MOS measure, $\mathrm{Q_B}$ is the rating of the best quality speech sample and $\mathrm{Q_W}$ is the rating of the worst quality speech sample.

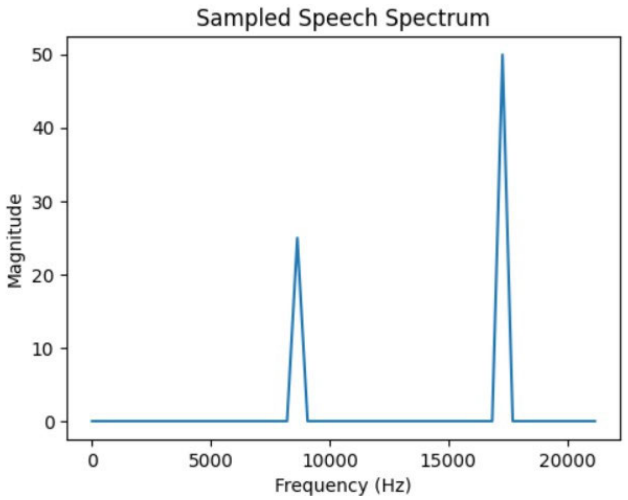


FIGURE 14. Spectrum of output information with 43200 sampling rate.

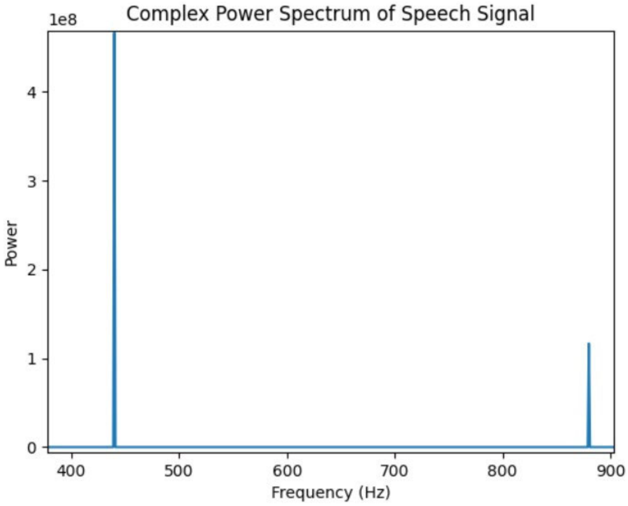


FIGURE 15. Complex power Spectrum of output information with 43200 sampling rate.

TABLE 4. Signal parameters of time slot 2.

Instant	Frequency	Bin	Harmonics	Bin Power in dB	dB below peak in dB	Total Smaller Power in dB
11	105.469	5	0	-36.68	36.68	-22.95
12	1560.938	74	0	-36.95	36.95	-23.12
13	1539.844	73	0	-37.31	37.31	-23.29
14	1771.875	84	0	-37.48	37.48	-23.46
15	1792.969	85	0	-38.33	38.33	-23.60
16	1497.656	71	0	-38.72	38.72	-23.74
17	1518.750	72	0	-39.01	39.01	-23.87
18	1476.563	70	0	-39.99	39.99	-23.98
19	1455.469	69	0	-40.21	40.21	-24.08
20	253.125	12	0	-40.94	40.94	-24.17

The above tabulated values clearly representing the proposed method has providing good results than other approaches, The implementation is applicable for voice assistant to separate the tracks and the noises from the multiple original audio which reproduces simultaneously using the speech enhancement and universal code book. The track sepa-

TABLE 5. Signal parameters of time slot 3.

Instant	Frequency	Bin	Harmonics	Bin Power in dB	dB below peak in dB	Total Smaller Power in dB
21	1413.281	67	0	-41.04	41.04	-24.26
22	1434.375	68	0	-41.12	41.12	-24.36
23	1814.063	86	0	-41.76	41.76	-24.44
24	1392.188	66	0	-41.79	41.79	-24.52
25	1371.094	65	0	-41.92	41.92	-24.60
26	1835.156	87	0	-42.31	42.31	-24.67
27	126.563	6	0	-42.34	42.34	-24.74
28	1328.906	63	0	-42.42	42.42	-24.82
29	1350.000	64	0	-42.51	42.51	-24.89
30	1307.813	62	0	-42.92	42.92	-24.96

TABLE 6. Signal parameters of time slot 4.

Instant	Frequency	Bin	Harmonics	Bin Power in dB	dB below peak in dB	Total Smaller Power in dB
31	210.938	10	0	-42.92	42.92	-25.03
32	1244.531	59	0	-43.03	43.03	-25.10
33	738.281	35	0	-43.26	43.26	-25.17
34	1286.719	61	0	-43.27	43.27	-25.24
35	1265.625	60	0	-43.68	43.68	-25.30
36	1223.438	58	0	-43.84	43.84	-25.36
37	1202.344	57	0	-43.87	43.87	-25.42
38	147.656	7	0	-43.92	43.92	-25.48
39	232.031	11	0	-44.01	44.01	-25.55
40	168.750	8	0	-44.09	44.09	-25.61

TABLE 7. Signal parameters of time slot 5.

Instant	Frequency	Bin	Harmonics	Bin Power in dB	dB below peak in dB	Total Smaller Power in dB
41	1096.875	52	0	-44.09	44.09	-25.67
42	506.250	24	0	-44.19	44.19	-25.73
43	1181.250	56	0	-44.21	44.21	-25.79
44	717.188	34	0	-44.38	44.38	-25.85
45	1139.063	54	0	-44.41	44.41	-25.91
46	316.406	15	0	-44.42	44.42	-25.98
47	1160.156	55	0	-44.42	44.42	-26.04
48	590.625	28	0	-44.44	44.44	-26.10
49	21.094	1	0	-44.55	44.55	-26.16
50	991.406	47	0	-44.56	44.56	-26.23

ration process is not an easy task if the unwanted information which is also called as interfering source which is like the actual original payload information which is also called as target source. For that, the Hybrid Deep Learning Algorithm has been developed and the training data sets are also created and tested to achieve accuracy in the speech recognition for the variety of voice assistants.

TABLE 8. Comparison of proposed method with existing techniques.

Method	Dataset	SNR in dB	STOI	PESQ	MOS
CNN	CHIME-4	10.5	0.95	4.5	4.3
	NOISEX-92	12.0	0.97	4.7	4.5
	WSJO-2 mix	13.5	0.98	4.8	4.7
RNN	CHIME-4	11.0	0.96	4.6	4.4
	NOISEX-92	12.5	0.98	4.8	4.6
	WSJO-2 mix	14.0	0.99	4.9	4.7
GAN	CHIME-4	11.5	0.97	4.7	4.5
	NOISEX-92	13.0	0.99	4.9	4.7
	WSJO-2 mix	14.5	1.00	5.0	4.8
DSBF	CHIME-4	08.5	0.89	4.2	4.0
	NOISEX-92	10.0	0.93	4.4	4.2
	WSJO-2 mix	11.5	0.96	4.6	4.4
MVDRBF	CHIME-4	09.5	0.91	4.3	4.1
	NOISEX-92	11.0	0.94	4.5	4.3
	WSJO-2 mix	12.5	0.97	4.7	4.5
BSS	CHIME-4	10.5	0.93	4.4	4.2
	NOISEX-92	12.0	0.96	4.6	4.4
	WSJO-2 mix	13.5	0.98	4.8	4.6
Hybrid DNN	CHIME-4	12.5	0.98	4.8	4.6
	NOISEX-92	14.0	1.00	5.0	4.8
	WSJO-2 mix	15.5	1.00	5.0	4.9

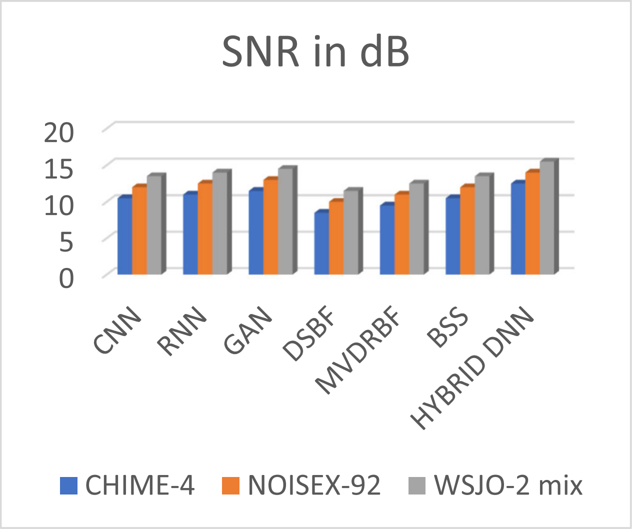


FIGURE 16. Comparison of SNR.

In figure 16, it clearly represents the hybrid DNN approach is having better signal to Noise Ratio (SNR) when compared to the other approaches in all the three different datasets.

In figure 17, it clearly represents the hybrid DNN approach is having better Short Time Objective Intelligibility (STOI) when compared to the other approaches in all the three different datasets.

In figure 18, it clearly represents the hybrid DNN approach is having better Perceptual Evaluation of Speech Quality (PESQ) when compared to the other approaches in all the three different datasets.

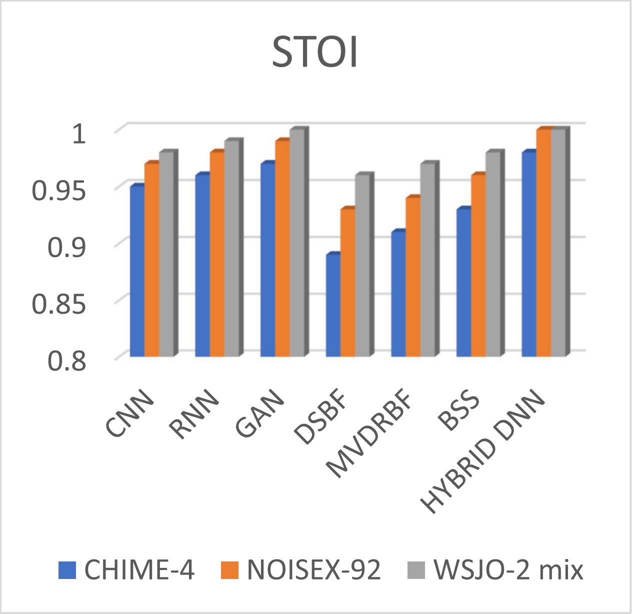


FIGURE 17. Comparison of STOI.

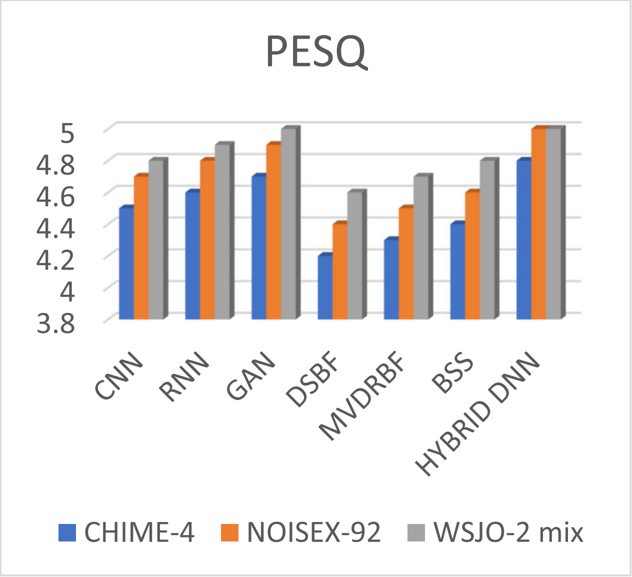


FIGURE 18. Comparison of PESQ.

In figure 19, it clearly represents the hybrid DNN approach is having better Mean Opinion Score (MOS) when compared to the other approaches in all the three different datasets.

VI. CONCLUSION

The combined approach of Generative Adversarial Network (GAN) and Blind source separation (BSS) is a promising technique for track and noise separation from multiple speech signals. GANs can be used to learn the distribution of the speech signals in the mixed signal, while BSS algorithms can be used to further separate the individual speech signals from the mixed signal.

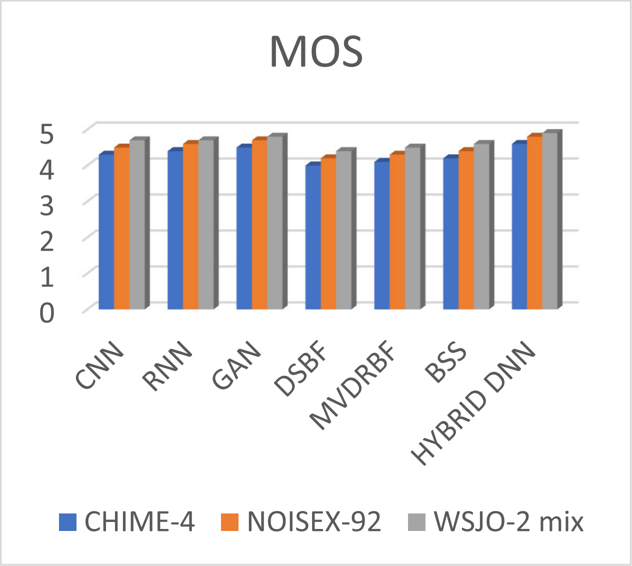


FIGURE 19. Comparison of MOS.

This approach has been shown to be effective in separating multiple speech signals from a mixed signal, even in the presence of background noise. It can achieve good separation quality even when the number of speech signals in the mixed signal is unknown.

However, there are still some challenges that need to be addressed to improve the performance of this approach. One challenge is that GANs can be computationally expensive to train. Another challenge is that BSS algorithms can be sensitive to the choice of parameters.

Despite these challenges, the combined approach of GAN and BSS is a promising technique for tracking and noise separation from multiple speech signals. It is a powerful tool that can be used to improve the quality of speech in a variety of applications.

The work makes a significant contribution by leveraging deep learning techniques for improved speech recognition. One notable aspect is the utilization of the Universal Codebook, a concept likely inspired by the success of adversarial attacks in deep learning models, such as the "Intra-Class Universal Adversarial Attacks on Deep Learning-Based Modulation Classifiers." Moreover, the mention of "Track and Noise Separation" suggests a focus on addressing challenges related to signal processing and background noise in speech recognition systems. The use of a Universal Codebook hints at a method to generalize the separation of various components within the speech signal, contributing to improved accuracy in speech recognition tasks. Overall, this work not only underscores the importance of deep learning but also showcases a novel approach to enhance the robustness and performance of speech recognition systems.

Recent developments in deep learning for speech processing have seen advancements in noise separation and enhanced recognition. The Universal Codebook has been employed for effective speech tracking, offering a comprehensive representation of speech patterns. Additionally, a combined approach of Generative Adversarial Network (GAN) and Blind Source Separation (BSS) has shown promise. GANs contribute by generating realistic noise samples, aiding in training models for robustness against diverse acoustic environments. BSS techniques further improve signal clarity by separating mixed audio sources. This synergistic use of GAN and BSS enhances speech recognition systems, making them more resilient to noise, ultimately improving their performance in real-world scenarios. These innovations address challenges in noisy environments, contributing to the reliability and accuracy of speech-based applications.

The research on "Track and Noise Separation based on the Universal Codebook and enhanced speech recognition using Hybrid Deep Learning Method" seeks to explore the synergy between traditional signal processing methods and modern deep learning techniques. By integrating the strengths of both approaches, the goal is to improve the quality of separated speech tracks and subsequently enhance the accuracy of speech recognition tasks in challenging acoustic environments. The paper's subsequent sections will delve into the methodology, experimental setup, results, and discussions that elucidate the effectiveness of this hybrid approach.

REFERENCES

[1] S. U. N. Wood, J. K. W. Stahl, and P. Mowlaee, "Binaural codebook-based speech enhancement with atomic speech presence probability," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2150–2161, Dec. 2019, doi: 10.1109/TASLP.2019.2937174.

[2] A. S. Subramanian, X. Wang, M. K. Baskar, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "Speech enhancement using end-to-end speech recognition objectives," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2019, pp. 234–238, doi: 10.1109/WASPAA.2019.8937250.

[3] Q. He, C.-c. Bao, and F. Bao, "Multiplicative update of AR gains in codebook-driven speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5230–5234, doi: 10.1109/ICASSP.2016.7472675.

[4] D. Baby, T. Virtanen, J. F. Gemmeke, and H. Van hamme, "Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 1788–1799, Nov. 2015, doi: 10.1109/TASLP.2015.2450491.

[5] Y. Xiang and C. Bao, "A codebook-driven speech enhancement method by exploiting speech harmonicity," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Xiamen, China, Mar. 2017, pp. 1–5, doi: 10.1109/ICSPCC.2017.8242542.

[6] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2162–2172, Dec. 2019, doi: 10.1109/TASLP.2019.2941592.

[7] A. Hassani, A. Bertrand, and M. Moonen, "Real-time distributed speech enhancement with two collaborating microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 6586–6587, doi: 10.1109/ICASSP.2017.8005295.

[8] Q. Huang, C. Bao, X. Wang, and Y. Xiang, "DNN-based speech enhancement using MBE model," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 196–200, doi: 10.1109/IWAENC.2018.8521278.

[9] A. Gaich and P. Mowlaee, "On speech quality estimation of phase-aware single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 216–220, doi: 10.1109/ICASSP.2015.7177963.

[10] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1826–1838, 2020, doi: 10.1109/TASLP.2020.2997118.

[11] D.-m. Zhang, C.-c. Bao, and F. Deng, "Integrating codebook and Wiener filtering for speech enhancement," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Ningbo, China, Sep. 2015, pp. 1–5, doi: 10.1109/ICSPCC.2015.7338795.

[12] M. Pirolt, J. Stahl, P. Mowlaee, V. I. Vorobiov, S. Y. Barysenka, and A. G. Davydov, "Phase estimation in single-channel speech enhancement using phase invariance constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5585–5589, doi: 10.1109/ICASSP.2017.7953225.

[13] A. Hussain, K. Chellappan, and M. S. Zamratol, "Speech enhancement using degenerate unmixing estimation technique and adaptive noise cancellation technique as a post signal processing," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Kuala Lumpur, Malaysia, Dec. 2016, pp. 280–285, doi: 10.1109/IECBES.2016.7843458.

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

[15] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006, doi: 10.1109/TSA.2005.854113.

[16] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "A cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 106–117, 2021, doi: 10.1109/TASLP.2020.3036783.

[17] H. Wang, Z. Ye, and J. Chen, "A speech enhancement system for automotive speech recognition with a hybrid voice activity detection method," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 1–9, doi: 10.1109/IWAENC.2018.8521410.

[18] G. Zhang, C. Wang, L. Yu, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for multi-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 9206–9210, doi: 10.1109/ICASSP43922.2022.9746902.

[19] C.-Y. Li and N. T. Vu, "Improving speech recognition on noisy speech via speech enhancement with multi-discriminators CycleGAN," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Cartagena, Colombia, Dec. 2021, pp. 830–836, doi: 10.1109/ASRU51503.2021.9688310.

[20] Z. Nian, J. Du, Y. Ting Yeung, and R. Wang, "A time domain progressive learning approach with SNR constriction for single-channel speech enhancement and recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 6277–6281, doi: 10.1109/ICASSP43922.2022.9746609.

[21] T. Taher, N. Mamun, and Md. A. Hossain, "A joint bandwidth expansion and speech enhancement approach using deep neural network," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Chittagong, Bangladesh, Feb. 2023, pp. 1–4, doi: 10.1109/ECCE57851.2023.10101546.

[22] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, and L.-R. Dai, "A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1927–1939, 2023, doi: 10.1109/TASLP.2023.3275033.

[23] A. Pandey and D. Wang, "Attentive training: A new training framework for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1360–1370, 2023, doi: 10.1109/TASLP.2023.3260711.

[24] R. Soleymanpour, M. Soleymanpour, A. J. Brammer, M. T. Johnson, and I. Kim, "Speech enhancement algorithm based on a convolutional neural network reconstruction of the temporal envelope of speech in noisy environments," *IEEE Access*, vol. 11, pp. 5328–5336, 2023, doi: 10.1109/ACCESS.2023.3236242.

[25] M. Pashaian, S. Seyedin, and S. M. Ahadi, "A novel jointly optimized cooperative DAE-DNN approach based on a new multi-target step-wise learning for speech enhancement," *IEEE Access*, vol. 11, pp. 21669–21685, 2023, doi: 10.1109/ACCESS.2023.3250820.

[26] M. Barhoush, A. Hallawa, A. Peine, L. Martin, and A. Schmeink, "Localization-driven speech enhancement in noisy multi-speaker hospital environments using deep learning and meta learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 670–683, 2023, doi: 10.1109/TASLP.2022.3231700.

[27] H. Fang, D. Becker, S. Wermter, and T. Gerkmann, "Integrating uncertainty into neural network-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1587–1600, 2023, doi: 10.1109/TASLP.2023.3265202.

[28] M. Z. Ozturk, C. Wu, B. Wang, M. Wu, and K. J. R. Liu, "RadioSES: mmWave-based audioradio speech enhancement and separation system," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1333–1347, 2023, doi: 10.1109/TASLP.2023.3250846.

[29] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 54–70, 2023, doi: 10.1109/TASLP.2022.3205757.

[30] Q. Zhang, X. Qian, Z. Ni, A. Nicolson, E. Ambikairajah, and H. Li, "A time-frequency attention module for neural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 462–475, 2023, doi: 10.1109/TASLP.2022.3225649.

[31] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, Nov. 2017.

[32] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993, doi: 10.1016/0167-6393(93)90095-3.

[33] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 31–35, doi: 10.1109/ICASSP.2016.7471631.

[34] A. Einizade and S. H. Sardouie, "A unified approach for simultaneous graph learning and blind separation of graph signal sources," *IEEE Trans. Signal Inf. Process. over Netw.*, vol. 8, pp. 543–555, 2022, doi: 10.1109/TSIPN.2022.3183498.

[35] T. Wang, F. Yang, and J. Yang, "Convolutive transfer function-based multichannel nonnegative matrix factorization for overdetermined blind source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 802–815, 2022, doi: 10.1109/TASLP.2022.3145304.

[36] Y. Du, R. Scheibler, M. Togami, K. Yoshii, and T. Kawahara, "Computationally-efficient overdetermined blind source separation based on iterative source steering," *IEEE Signal Process. Lett.*, vol. 29, pp. 927–931, 2022, doi: 10.1109/LSP.2021.3134939.

[37] A. Brendel, T. Haubner, and W. Kellermann, "A unifying view on blind source separation of convolutive mixtures based on independent component analysis," *IEEE Trans. Signal Process.*, vol. 71, pp. 816–830, 2023, doi: 10.1109/TSP.2023.3255552.

[38] H. Munakata, Y. Bando, R. Takeda, K. Komatani, and M. Onishi, "Joint separation and localization of moving sound sources based on neural full-rank spatial covariance analysis," *IEEE Signal Process. Lett.*, vol. 30, pp. 384–388, 2023, doi: 10.1109/LSP.2023.3264570.

[39] A. J. Muñoz-Montoro, J. J. Carabias-Orti, P. Cabañas-Molero, F. J. Cañadas-Quesada, and N. Ruiz-Reyes, "Multichannel blind music source separation using directivity-aware MNMF with harmonicity constraints," *IEEE Access*, vol. 10, pp. 17781–17795, 2022, doi: 10.1109/ACCESS.2022.3150248.

[40] F. Hasuike, D. Kitamura, and R. Watanabe, "DNN-based frequency-domain permutation solver for multichannel audio source separation," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Chiang Mai, Thailand, Nov. 2022, pp. 871–876, doi: 10.23919/APSIPAASC55919.2022.9979953.

S. V. ASWIN KUMER received the degree in electronics and communication engineering from the Pallavan College of Engineering, Kanchipuram, in April 2008, the master's degree in embedded system technology from SRM University, Kanchipuram, in May 2012, and the Ph.D. degree in the implementation of image fusion using artificial neural network from SCSVMV (Deemed to be University), Enathur, in February 2019. He is an Associate Professor with the Department of Electronics and Communication Engineering, KLEF (Deemed to be University), Guntur. He has more than 15 years of teaching experience. His areas of interests include digital communication and digital signal processing.

LAKSHMI BHARATH GOGU received the B.Tech. degree in electronics and communication engineering and the M.Tech. degree in VLSI system design from JNTUA University, Anantapuram, Andhra Pradesh, India, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in electronics and communication engineering with KLEF (Deemed to be University), Guntur, Andhra Pradesh.

E. MOHAN received the M.E. degree in computer science engineering from Satyabhama University, the Ph.D. degree in computer science and engineering from Vinayaka Missions University, and the M.B.A. degree from Madras University. He has more than two decades experience in academic field. He is currently a Professor with the Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India. Throughout his career, he has good academic records of accomplishment and published many refereed journals in the reputed publications. He titled three books and four scholars successfully completed doctorate under his guidance. His research interests include image processing, WSN, the IoT, ML, and datamining.

SUMAN MALOJI received the B.Tech. degree in electronics and communication engineering from JNTUA, in 2001, the M.Tech. degree in electronics and communication engineering from JNTUH, in 2006, and the Ph.D. degree from K. L. University, in 2014. He has over 20 years of experience in teaching, administration, and research. His areas of interests include speech coding, speech compression, and speech and speaker recognition.

BALAJI NATARAJAN received the Ph.D. degree in computer science and engineering from Pondicherry University, Puducherry, India, in 2017. He is currently a Professor and the Head with the Department of Computer Science and Engineering, Sri Venkateshwaraa College of Engineering and Technology, Ariyur, Puducherry. He has 15 years of teaching, research, and industry experience. He has published more than 50 research papers in various reputed international journals and conferences. His research interests include web services, service oriented architecture, evolutionary algorithms, artificial intelligence, and machine learning.

G. SAMBASIVAM (Member, IEEE) received the Ph.D. degree in computer science and engineering from Pondicherry University, Puducherry, India. He is currently an Assistant Professor with the School of Computing and Data Science, Xiamen University Malaysia, Sepang, Malaysia. Previously, he was the Dean of the School of Information and Communication Technology, ISBAT University, Uganda. His research interests include artificial intelligence, machine learning, deep learning, graph neural networks, web service computing, and soft computing techniques.

VAIBHAV BHUSHAN TYAGI received the B.Tech. degree from UPTU, Lucknow, in 2007, and the M.Tech. and Ph.D. degrees from IIT Roorkee, in 2011 and 2015, respectively. He has more than 13 years of research and teaching experience around the Globe. Currently, he is an Associate Professor (ECE) and the Dean FICT with ISBAT University, Kampala, Uganda. He has worked in several administrative and academic positions in India, Ethiopia, and Uganda. His research interests include sensor applications in signal processing, signal modeling, artificial intelligence, and deep learning.