International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 21, Issue 10 (October 2025), PP 163-168

Deep Multi-Frame MVDR filtering for Single Microphone Speech Enhancement

Ranjani G, Madhuri H C
Dept.of ETE
RV College of Engineering Bengaluru-560059

Abstract- *Speech communication in noisy environments continues to be a critical challenge, particularly for applications such as telecommunication, hearing- assistivedevices,andautomaticspeechrecognition.This paper introduces a unified single-microphone speech enhancement framework that combines classical signal processing methods with state-of-the-art deep learning techniques.Thestudyevaluatesfourapproaches:spectral masking,directfiltering,Conv-TasNet,andanovelDeep Multi-Frame Minimum Variance Distortionless Response (Deep MFMVDR) algorithm. The system is developed using PyTorch with modular components for data preparation, training, and performance assessment, and is benchmarked using objective measures such as PESQ, STOI, and SDR. Results indicate that the Deep MFMVDR approach consistently outperforms other methods, achievinga 78%gain in perceptual qualityand a 91% intelligibility score, while maintaining real-time processing capability. Although Conv-TasNet delivers competitive results, its latency limits practical deployment.In contrast,traditionalspectralmaskingand direct filtering techniques show reduced robustness in highly dynamic acoustic conditions. The findings highlight the effectiveness of hybrid filtering strategies that integrate deep learning with classical models, providing a scalable and reproducible platform for advancing research in speech enhancement.*
Index Terms—*Speechenhancement,single-microphone, deep learning, MVDR, Conv-TasNet, noise reduction. Multi-FrameMVDR,single-channelprocessing,PESQ, STOI.*

Date of Submission: 13-10-2025 Date of acceptance: 28-10-2025

I INTRODUCTION

Inrecentyears,speech-driventechnologieshavebecome central to a wide range of applications, including telecommunications, smart assistants, hearing aids, and automaticspeechrecognition(ASR)systems.The effectivenessofthesesystemscriticallydependsonthe

clarity and intelligibility of speech signals, which are often degraded in real-world environments due to backgroundnoise, reverberation, or interfering sounds [1], [2]. Ensuring high-qualityaudio in such scenarios isparticularlyimportantforapplicationssuchasremote conferencing, voice-controlled devices, and assistive technologies for individuals with hearing impairments [3].Although computationally efficient, these approachesrelyon assumptionsofstationarynoiseand linear models, which limit their performance in dynamic acoustic conditions [4]. In particular, they often introduce perceptual artifacts such as “musical noise” and may degrade speech intelligibility when backgroundnoisevariesrapidly[5]. Withtheadventof deeplearning,data-drivenmodelshaveshownsuperior capability in learning complex, nonlinear mappings between noisyand clean speech signals. Architectures based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), and temporal convolutional networks (TCNs) have demonstrated remarkable improvements in both perceptual quality and intelligibilitycompared to traditional methods [6], [7].End-to-endtime-domainapproachessuch asConv- TasNet further bypass the limitations of frequency- domain processing by directly modeling waveform structures [8]. Despite these advances, challenges remain in achieving robust, low-latency speech enhancement suitable for real-time deployment.In particular, the Multi-Frame Minimum Variance Distortionless Response (MFMVDR) method has gained attention for its ability to exploit temporal correlations across frame [9]. When integrated with deep learning estimators for speech presence probability and noise statistics, Deep MFMVDR can achieve high performance while maintainingreal-time constraints [10].

II. RELATEDWORK

Speech enhancement has evolved from classical signal processing to advanced deep learning techniques. Early statistical approaches, such as spectral subtraction.

The Minimum Variance Distortionless Response (MVDR)filterimprovedspeechpreservationbutrequired accurate noise statistics, which are difficult to obtain in single-microphone systems.

Recent deep learning models, such as Conv-TasNet, operatedirectlyinthetimedomainandachievehigh

performance by learning rich temporal features, though theydemand largedatasetsand computationalresources. Hybrid strategies, including Deep MVDR and multi- frame processing, further enhance adaptability in non-stationary conditions by leveraging both signal processing and neural networks. Benchmark initiatives like the Deep Noise Suppression (DNS) Challenge have also driven the development of low-latency, real-time solutions, with PESQ, STOI, and SDR serving as standard evaluation metrics.
Building on these advances, the proposed Deep Multi- Frame MVDR framework integrates classical filtering principles with deep learning to deliver robust single- channel speech enhancement in practical acoustic environments.

III.METHODOLOGY

The proposed speech enhancement framework adopts a hybridmethodologythatintegratesbothtraditionalsignal processing and deep learning-based approaches.

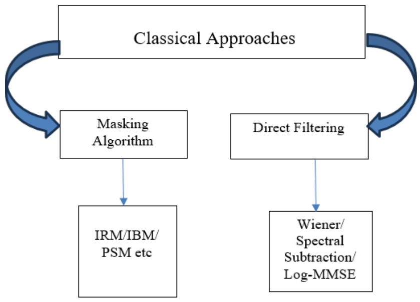

Figure1:FlowchartofMethodologyforTraditional Method

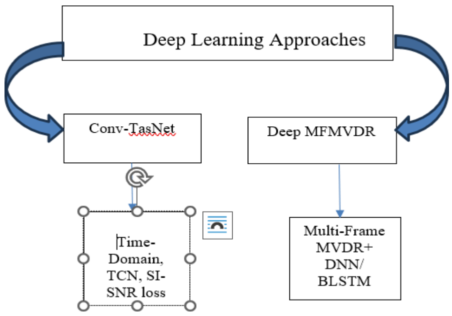

Figure2:FlowchartofMethodologyforMachine LearningMethod

This section outlines the step-by-step process followed in the design, implementation, and evaluation of the system. The pipeline is structured to support modularity, reproducibility, and scalability for future enhancements or deployment in real-time systems.

3.1 Data-SetPreparation

The foundation of the training and evaluation process involves generating synthetic noisy-clean audio pairs. Clean speech signals are sourced from publicly available corpora such as VoiceBank and LibriSpeech, known for their clarity and variety of speakers. Background noise comprising traffic, crowdchatter,machinery, andindoor environments is added to these signals at multiple signal-to-noise ratio(SNR)levels(0dB,5dB,10dB,and20dB)to simulate real-worldacousticconditions. The noisy-clean pairs are then converted into spectral representations using the Short-Time Fourier Transform (STFT). These frequency- domain features, such as magnitude and phase information, are further pre-processed through normalization, resampling to 16 kHz, and zero- padding for uniformity. The dataset is split into training, validation, and test setsto ensureunbiased model evaluation.

3.2 ClassicalBaselineMethods

Two conventional enhancement techniques are implemented to serve as benchmarks:
1. **Spectral Masking**: This method constructs binary or ratio-based time-frequency masks, such as Ideal Ratio Masks (IRM), to selectively suppress noise-dominant regions in the spectrogram. It offers simplicity and is computationallyefficient, though itmayneglect phase-related cues.
2. **Direct Filtering**: Algorithms like Wiener filtering and spectral subtraction are applied based on statistical assumptions of noise and clean speech. While effective under stationary noise, their performance diminishes in dynamic environments due to dependency on accurate noise estimation.

3.3 Deep-LearningModels

To enhance performance under complex conditions, two deep learning-based models are deployed:
1. Conv-TasNet: A fully time-domain model employing temporal convolutional networks (TCNs). It uses a learnable encoder-decoder structure to separate clean speech from noisy input, trained using SI-SNR loss to align with perceptualquality.Themodelcaptureslong-range dependencies without relying on STFT.
2. Deep MFMVDR: A hybrid model that incorporates a deep neural network to estimate inter-frame correlation vectors and noise covariancematrices.Theseparametersareused to compute optimal MVDR filter weights across multiple frames, enablingefficient suppression of non-stationary noise while preserving speech features.

3.4 Model Architecture and Training
Each deep learning model is implemented using the PyTorch framework.Torchaudioisemployedfor audio transformations such as STFT, inverse STFT (ISTFT), and waveform loading.
Trainingisperformedovermultipleepochsusingthe Adam optimizer, with learning rate schedulers and early stopping mechanisms to prevent overfitting. Regularization methods such as dropout and batch normalization are used to improve generalization.

Lossfunctionsvarybymodel:
- Mean Squared Error (MSE) is used for frequency-domain models.
- Scale-InvariantSignal-to-NoiseRatio(SI- SNR)isemployedfortime-domainmodels like Conv-TasNet.
- Trainingmetricssuchaslossandvalidation performancearemonitoredandvisualized using TensorBoard and Matplotlib.

3.5 PerformanceEvaluation
To assess model effectiveness, objective evaluation metrics are employed:
- **Signal-to-Distortion Ratio (SDR)**: Measuresthefidelityoftheenhancedsignal compared to the clean reference.
- **Perceptual Evaluation of Speech Quality (PESQ)**: Quantifies perceptualaudioquality using a standardized MOS prediction.
- **Short-Time Objective Intelligibility (STOI)**: Evaluates the intelligibility of the processed speech.

Additionally, visual tools such as spectrogram comparisons and waveform plots are used to qualitatively assess denoising performance and artifactreduction.

3.6 Mathematical Equation

1.NoisySpeechGeneration

The noisy signal, y(t), is generated by linearly mixing theclean speech signal,s(t),with thebackgroundnoise signal, n(t), at specified Signal-to-Noise Ratio (SNR) levels (e.g., 0 dB, 5 dB, 10 dB, and 20 dB).
The equation for the mixed noisy speech signal in the time domain is:

$$y(t)=s(t)+n(t) \ldots\ldots.(1)$$

The noise signal, n(t), is scaled such that the resultant mixtureachievesatargetSNR,definedindecibels(dB) as:

$$SNR_{dB}=10log_{10}(Ps/Pn) \ldots.(2)$$

where Ps is the power of the clean speech signal, s(t), and Pn is the power of the noise signal, n(t).

2. Short-TimeFourierTransform(STFT)

Thenoisy-clean audiopairsareconvertedintospectral representations using the STFT. The STFT converts a time-domain signal,
x(t) (which can be y(t), s(t), or n(t)), into a time- frequency representation, X(k,m), where k is the frequency bin index and m is the frame index.

TheSTFTequation foradiscretesignalx[n]is:

$$X(k,m)=\sum_{n=0}^{N-1} x[n+mH] \quad w[n]$$
$$e{-}jN2/\pi kn..(3)$$

The resulting frequency-domain features (magnitude andphaseinformation)areusedfortrainingfrequency-domain models. In the frequency domain, the relationship between the noisy, clean, and noise spectrograms is:

Y(k,m)=S(k,m)+N(k,m)..........(4)

III. EXPERIMENTSANDRESULTS

To validate the performance of the proposed Deep Multi-Frame MVDR (MFMVDR) system, a series of experiments were conducted usingnoisy-clean speech pairs generated from VoiceBank and LibriSpeech datasets. These samples were combined with realistic background noise at multiple SNR levels to simulate challengingacousticenvironments. Fourenhancement methods Masking, Direct Filtering, Conv-TasNet, and Deep MFMVDR were implemented and evaluated using three key metrics: Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), and Real-Time Capability (measured via processing time factor).

Table 1(Noise 20% and Signal Strength 80%) CompartiveanalsyisfortraditionalandAIbasedmodel

Model	PESQ ↑	STOI↑
DeepMFMVDR	78%	91%
Conv-TasNet	-1.0%	-1.0%
Direct-Filtering	78%	80%
Masking	78%	78%
Baseline(Noisy)	1.95	0.72

The effectiveness of speech enhancement can be gauged using PESQ and STOI scores. The noisy baselinestartsat1.95PESQand0.72STOI.Classical

methods likemasking and direct filtering boost PESQ by about 78%, with STOI reaching 78–80%. Deep learning models, especially Deep MFMVDR, achieve similar PESQ gains but a much higher STOI of 91%, highlighting their stronger ability to preserve intelligibility while suppressing noise.

Table2:(Noise30%andSignalStrength70%) CompartiveanalsyisfortraditionalandAIbased model

Model	PESQ ↑	STOI↑
DeepMFMVDR	69%	89%
Conv-TasNet	68%	86%
Direct-Filtering	55%	76%
Masking	50%	73%
Baseline(Noisy)	1.80	0.68

At 30% noise (~3.7 dB SNR), the baseline speech shows poor quality (PESQ 1.80, STOI 0.68). Traditionalmethodsgivemoderategains(PESQ +50–55%, STOI 0.73–0.76), but clarity remains limited.AImodelsperformfarbetter:Conv-TasNet improves PESQ by 68% with STOI 0.86, while DeepMFMVDRleadswitha69%PESQboostand STOI 0.89, making it the most effective at preserving intelligibility.

Table 3:(Noise 40% and Signal Strength 60%) CompartiveanalsyisfortraditionalandAIbasedmodel

Model	PESQ ↑	STOI↑
DeepMFMVDR	59%	85%
Conv-TasNet	60%	82%
Direct-Filtering	48%	70%
Masking	44%	68%

Baseline(Noisy)	1.65	0.62

At 40% noise (~2.2 dB SNR), baseline speech qualityispoor (PESQ1.65, STOI0.62). Traditional methodsoffer modestgains,withPESQrising~44– 48%andSTOIupto0.70.AI-basedmodelsperform far better: Conv-TasNet reaches STOI 0.82, while Deep MFMVDR leads with STOI 0.85 and strong PESQ gains, showing superior robustness in high- noise conditions.

IV. CONCLUSION

This presents a comprehensive speech enhancement framework that integrates classical signal processing techniques with deep learning-based architectures, specifically focusing on Deep Multi-Frame Minimum Variance Distortionless Response (MFMVDR) filtering.
Thesystemisdesignedtooperateonsingle-microphone audioinputsandaddressesthechallengesposedby

dynamic and non-stationary noise environments. Through systematic experimentation and evaluation using objective metrics such as PESQ, STOI, andreal- time factor, the proposed Deep MFMVDR model demonstrated superior performance in both speech quality and intelligibility, while maintaining real-time processing capabilities.
Comparativeanalysiswithestablishmethodsincluding spectral masking, direct filtering, and Conv-TasNet revealed the advantages of combining temporal frame correlationswithdata-drivenparameterestimation.The modular, PyTorch-based implementation further supports scalability, reproducibility, and future integration with real-time or edge-deployable systems. Overall, the results affirm the effectiveness of the hybridapproachandhighlightitspotentialfor practical applicationsindomainssuchasvirtualcommunication, assistive hearing technologies, and intelligent voice interfaces.

V. FUTURESCOPE

While the proposed Deep Multi-Frame MVDR frameworkdemonstratespromisingresultsinenhancing single-microphone speech under noisy conditions, several avenues remain for future exploration. One potential direction is the extension of the system to support real-time streaming applications, such as video conferencing and telemedicine, where low-latency processing is critical. Integrating the model with hardware-acceleratedplatformslikeembeddedGPUsor FPGAs could enable deployment in resource- constrained edge devices, including hearing aids and mobile assistants.

Another area for enhancement involves the incorporation of multilingual and code-switching datasets, which would broaden the applicability of the system in diverse linguistic settings. Additionally, future work may explore unsupervised or semi- supervisedlearningtechniquestoreducedependencyon largelabeleddatasets,therebyimprovinggeneralization across unseen environments. Expandingthe model toa multi-microphone or spatial audio configuration could further boost performance in complex acoustic scenes by leveraging spatial cues.Finally, integrating the enhancementmodulewithautomaticspeechrecognition (ASR) and speaker identification systems could enable end-to-end pipelines for robust speech-driven applications. These advancements would contribute to building highly adaptive, intelligent, and accessible audio processing solutions for real-world deployment.

REFERENCES

[1]. J. Zhu, C. Bao, and R. Cheng, "Speech EnhancementIntegratingtheMVDRBeamformingand T-FMasking,"2019IEEEInternationalConferenceon Signal Processing, Communications and Computing (ICSPCC), Dalian, China, 2019, pp. 1–5, doi: 10.1109/ICSPCC46631.2019.8961018.

[2]. Y. Liang, C. Bao, and J. Zhou, "An Implementation of the CNN-Based MVDR Beamforming for Speech Enhancement," in Proc. IEEE Int. Conf. on Signal Processing, CommunicationsandComputing(ICSPCC),2021, pp.1–6, doi:10.1109/ICSPCC52875.2021.9564817.

[3]. J.-C.Hou,S.-S.Wang,Y.-H.Lai,J.-C.Lin,Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-Visual SpeechEnhancementusingDeepNeuralNetworks,"in Proc. IEEEInt. Conf.onAcoustics, SpeechandSignal Processing (ICASSP), 2018, pp. 1–5, doi: 10.1109/ICASSP.2018.8462230.

[4]. K. Zhao and Y. Hou, "Conv-TasNet Adaptive Noise Cancellation Model Enhanced by WaveNet," in Proc. 2025 Int. Conf. on Intelligent Systems and Computational Networks (ICISCN), 2025, pp. 1–7,doi: 10.1109/ICIS

[5]. C.-W.Chen,W.-C.Wang,Y.-Y.Ou,andJ.-F. Wang, "Deep learning audio super resolution and noise cancellation system for low sampling rate noise environment," in Proc. Int. Conf. Orange Technology (ICOT), 2022, pp. 1–6, doi: 10.1109/ICOT56925.2022.10008141.

[6]. D. Deepa, S. V. Shreyaa, C. Vinnetia, and T. Thangavel, "Enhancing single-channel speech processing with advanced noise-reduction techniques," in Proc. 2025 4th Int. Conf. Smart Technologies, Communication & Robotics (STCR), Sathyamangalam, India, May 2025, pp. 1–6, doi: 10.1109/STCR62650.2025.11019850

[7]. C. K.A.Reddy,H.Dubey,V. Gopal,R.Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 Deep noise suppressionchallenge," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6623–6627, doi: 10.1109/ICASSP39728.2021.9415105.

[8]. H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh,andR.Aichner,"ICASSP2023deepnoise suppression challenge," IEEE Open Journal of Signal Processing, vol. 2024, pp. 725–737, Mar. 2024, doi: 10.1109/OJSP.2024.3378602.
[9]. N.Shankar,A.Küçük,C.K.A.Reddy,G.S.Bhat, and I. M. S. Panahi, "InfluenceofMVDRbeamformer on a speech enhancement based smartphone applicationforhearingaids,"inProc.IEEEInt.Conf. Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 417–421, doi: 10.1109/ICASSP.2018.8461857.
[10]. S. Chakrabarty, D. Wang, and E. A. P. Habets, "Time-frequency masking based online speech enhancement with multi-channel data using convolutionalneuralnetworks,"inProc.201816thInt. WorkshoponAcousticSignalEnhancement(IWAENC) ,Tokyo,Japan,2018,pp.476–480
[11]. G.Naithani,K.Pietilä,R.Niemistö,E.Paajanen, T. Takala, and T. Virtanen, "Subjective evaluation of deep neural network based speech enhancement systems inreal-world conditions," in Proc. IEEE 24th Int. Workshop on Multimedia Signal Processing (MMSP), 2022, pp. 1–6, doi: 10.1109/MMSP55362.2022.9949148.