Computer Fraud and Security
ISSN (online): 1873-7056

Integrated Framework for Speech Enhancement and Voice Activity Detection in Robust Speech Processing

Adappa S. Angadi
Research Scholar, Visvesvaraya Technological University, Belagavi, Karnataka, India

Nagaraja B.G.
Department of Electronics & Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

Email : nagarajbg@gmail.com

Abstract: Robust speech processing in noisy environments is a critical requirement for a wide range of applications including telecommunication systems, voice-controlled interfaces, and hearing aids. This paper proposes an integrated framework that combines Deep Neural Network (DNN)-based Speech Enhancement (SE) with Voice Activity Detection (VAD) to improve both speech quality and activity detection accuracy under adverse acoustic conditions. First, several conventional and deep learning-based SE techniques are evaluated using the NOIZEUS database across different noise types and signal-to-noise ratio (SNR) levels. Performance is assessed using Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) metrics. An energy-based VAD approach is then applied, and its performance is evaluated using Frame Error Rate (FER) and Precision. The proposed integrated system demonstrates superior performance over standalone methods, with improvements observed in both enhancement and detection stages. Evaluation results show significant gains in FER and F1 score, highlighting the effectiveness of combining SE and VAD within a unified framework. This work offers a practical and efficient solution for real-time speech processing in complex acoustic environments.
Keywords: VAD, DNN, SNR, NOIZEUS, PESQ, STOI, FER, F1-score

1. Introduction

Speech communication systems play a critical role in a wide range of applications, including telecommunication, human-computer interaction, hearing aids, and voice-controlled devices [1]. At the core of such systems lies the challenge of maintaining intelligibility and clarity in real-world environments, where speech signals are often corrupted by background noise, reverberation, or overlapping voices. Two fundamental components that address these challenges are Speech Enhancement (SE) and Voice Activity Detection (VAD) [2].

SE refers to the process of improving the quality and intelligibility of speech signals by suppressing unwanted noise or distortions. It is particularly vital in noisy environments where the listener or machine may struggle to perceive or interpret the speech content. On the other hand, VAD is the task of distinguishing between speech and non-speech segments within an audio stream [3]. VAD is essential for efficient audio coding, speech recognition, speaker diarization, and other downstream processing tasks, as it enables systems to focus on speech-relevant information.

Despite significant advances in both domains, traditional approaches have typically treated SE and VAD as independent modules. This separation can result in suboptimal performance, especially in dynamic and low-SNR environments. For example, noise-suppressed signals from an SE module may still hinder accurate voice activity detection if not properly coordinated, and vice versa. Recent trends in deep learning have shown promise in unifying speech-related tasks, yet a gap remains in designing systems that jointly leverage the interdependence of SE and VAD for enhanced robustness and efficiency [4].

The motivation for this work stems from the observation that speech enhancement and voice activity detection are inherently interconnected tasks. Information derived from one task can significantly benefit the other. For instance, knowing whether speech is present can help guide enhancement algorithms to focus their efforts more precisely, while a cleaner signal from enhancement can improve the accuracy of detecting speech activity [5]. Therefore, an integrated framework that simultaneously addresses both tasks has the potential to outperform traditional sequential or isolated approaches.

This article aims to propose and explore an integrated framework that combines speech enhancement and voice activity detection into a unified model, with the goal of improving overall performance in challenging acoustic environments. The objectives of this work are:

- To review and analyze existing approaches in SE and VAD, highlighting their limitations when used independently.
- To design a joint framework that allows shared learning and mutual feedback between SE and VAD components.
- To evaluate the proposed system under various noise conditions and compare its performance to conventional baseline methods.

By bridging the gap between these two essential components, this work aspires to contribute to the development of more intelligent, noise-robust speech processing systems suitable for real-world applications. The remainder of this paper is structured as follows: Section 2 reviews related work in the domains of speech enhancement and voice activity detection. Section 3 provides an overview of selected speech enhancement techniques. Section 4 describes the simulation setup and presents the enhancement results. Section 5 details the voice activity detection technique and its evaluation. Section 6 introduces the proposed integrated approach and discusses its performance. Finally, Section 7 concludes the paper and outlines directions for future work.

2. Related Work

Zhang et al. introduced VSANet, a real-time speech enhancement framework that incorporates a VAD module and a causal spatial attention mechanism [6]. The architecture employs a shared encoder for both SE and VAD tasks, optimized through a weighted loss function. Experimental results demonstrate that this multi-task learning approach enhances SE performance, particularly in real-time applications. The work in [7] proposed SVVAD, a VAD framework tailored for speaker verification systems. Unlike traditional VAD methods, SVVAD adapts speech features based on their relevance to speaker verification, employing a label-free training method with triplet-like losses. This approach addresses challenges in noisy environments and multi-speaker scenarios, improving equal error rates in speaker verification tasks.

Morrone et al. explored an end-to-end integration of speech separation and VAD for low-latency speaker diarization in telephone conversations [8]. The study emphasizes the benefits of jointly training speech separation and VAD modules, leading to improved diarization error rates and reduced latency. This integration is particularly beneficial for real-time applications where prompt and accurate speaker identification is crucial. A unified framework capable of performing both personalized and non-personalized speech enhancement in real-time was presented in [9]. The model utilizes a frame-wise conditioning input to specify the enhancement type and incorporates re-weighting strategies based on speech activity presence. This approach demonstrates that a single model can effectively handle diverse enhancement tasks, offering a more economical solution compared to maintaining separate models.

In 2023, researchers developed a multimodal feature fusion network (MFF-Net) that leverages fine-grained 3D lip landmarks as auxiliary visual information for audio-visual speech enhancement [10]. The network employs a multi-scale enhancement module and an audio-visual fusion module to effectively combine visual and audio features. This approach enhances speech intelligibility and quality, particularly in video conferencing scenarios. A study introduced a Multi-Task Learning U-Net (MTU-Net) architecture that simultaneously performs single-channel speech enhancement and mask-based VAD [11]. The model features a shared encoder and separate decoders for each task, optimized through a combined loss function. Evaluations under matched and mismatched noise conditions reveal that MTU-Net outperforms traditional methods in both SE and VAD tasks.

Xia et al. proposed a deep-learning framework designed for efficient real-time speech enhancement and dereverberation [12]. The model integrates optimal ratio masks and deep complex convolution recurrent networks to address both noise suppression and reverberation. This comprehensive approach improves speech clarity in real-time communication systems. A speech enhancement method that combines beamforming techniques with recurrent neural networks (RNNs) for application in hearing aids was demonstrated in [13]. The approach aims to improve speech intelligibility and quality in noisy environments while maintaining low computational complexity, making it suitable for implementation on general hardware platforms.

A study in [14] explored the use of VAD in conjunction with deep neural networks and hybrid speech feature extraction methods for deceptive speech detection. By accurately identifying speech segments and extracting

relevant features, the system enhances the detection of deceptive speech, demonstrating the importance of integrated VAD in complex speech analysis tasks. Kandagatla et al. proposed a speech enhancement method that combines time-domain and discrete cosine transform (DCT) processing to achieve real-time performance [15]. The approach focuses on reducing computational complexity while maintaining enhancement quality, making it suitable for applications requiring low-latency processing.

The reviewed studies underscore the growing trend of integrating SE and VAD to develop robust, real-time speech processing systems. Approaches leveraging multi-task learning, end-to-end architectures, and multimodal data fusion have demonstrated significant improvements in performance across various applications, including speaker verification, diarization, and hearing aids. These advancements highlight the potential of unified frameworks in addressing the challenges of speech processing in noisy and dynamic environments.

3. Overview of Speech Enhancement Techniques

SE refers to the process of improving the quality and intelligibility of speech signals that have been corrupted by noise. This section outlines five commonly used techniques, along with their mathematical formulations.

3.1 Spectral Subtraction

Spectral subtraction is one of the earliest and most straightforward methods for noise reduction. It assumes the noise is additive and estimates it during silent periods [16].

$$Y(k,n) = S(k,n) + N(k,n)$$

where:

- $Y(k,n)$ is the noisy speech spectrum,
- $S(k,n)$ is the clean speech spectrum,
- $N(k,n)$ is the noise spectrum,
- k is the frequency bin index,
- n is the time frame index.

The estimated clean spectrum is obtained as:

$$\hat{S}(k,n) = |Y(k,n)| - \left|\hat{N}(k,n)\right|$$

3.2 Wiener Filtering

Wiener filtering minimizes the mean square error between the estimated and clean signals. The Wiener filter in the frequency domain is given by [17]:

$$H(k,n) = \frac{P_S(k,n)}{P_S(k,n) + P_N(k,n)}$$

$$\hat{S}(k,n) = H(k,n) \cdot Y(k,n)$$

where $P_S(k, n)$ and $P_N(k, n)$ are the power spectral densities of the clean speech and noise respectively.

3.3 Minimum Mean Square Error (MMSE) Estimation

MMSE estimators aim to minimize the mean square error between the estimated and true speech signals. The log-spectral amplitude estimator by Ephraim and Malah is widely used [18]:

$$\hat{S}(k,n) = \exp(E[\log|S(k,n)||Y(k,n)])$$

This method incorporates statistical models of speech and noise and often outperforms basic spectral subtraction in low-SNR conditions.

3.4 Deep Neural Network (DNN)-Based Enhancement

DNNs have become prominent in speech enhancement due to their ability to model complex nonlinear relationships [19]. The DNN maps the noisy features x_n to an estimate of the clean features $\widehat{s_n}$:

$$\widehat{s_n} = f_{\text{DNN}}(x_n; \theta)$$

where:

- $\mathbf{x_n}$ is the input noisy feature vector,
- $\widehat{\mathbf{s}_n}$ is the estimated clean feature vector,
- f_{DNN} is the neural network function,
- θ represents the network parameters.

The network is trained to minimize a loss function, typically mean squared error (MSE):

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} |\, s_n - \widehat{s_n}|^2$$

3.5 Mask-Based Enhancement (Ideal Ratio Mask)

Mask-based approaches estimate a mask to apply to the noisy spectrum. One common approach is the Ideal Ratio Mask (IRM) [20]:

$$\text{IRM}(k,n) = \frac{|S(k,n)|^{\alpha}}{|S(k,n)|^{\alpha} + |N(k,n)|^{\alpha}}$$

where α is typically set to 1 or 2. The estimated clean speech is:

$$\hat{S}(k,n) = \text{IRM}(k,n) \cdot Y(k,n)$$

Mask estimation is often performed using DNNs trained on large datasets of noisy and clean speech pairs.

4. Simulation Set-up and Enhancement results

4.1 Simulation Set-up

To assess the performance of the proposed integrated framework for speech enhancement and voice activity detection, experiments were conducted using the NOIZEUS database [21,22]. The NOIZEUS corpus contains 30 IEEE sentences spoken by male and female speakers, degraded with various real-world noise types at different signal-to-noise ratios (SNRs).

- Sampling Rate: 8 kHz
- Noise Types: Babble, Car, Street, and White
- SNR Levels: 0 dB, 5 dB, and 10 dB
- Frame Size: 20 ms with 50% overlap
- Window Type: Hamming

4.2 Evaluation Metrics

Perceptual Evaluation of Speech Quality (PESQ) is an objective metric that estimates the perceptual quality of speech by comparing a clean reference signal $s(t)$ with the enhanced signal $\hat{s}(t)$. The PESQ score ranges from -0.5 to 4.5 (for narrowband audio), with higher values indicating better quality:

$$\text{PESQ}(s,\hat{s}) = \text{MOS}_{\text{LQO}} \in [-0.5, 4.5]$$

Short-Time Objective Intelligibility (STOI) quantifies the intelligibility of speech by computing the correlation between short-time temporal envelopes of clean and enhanced signals. Let X and $\widehat{X}$ represent the clean and enhanced short-time spectrograms, respectively. Then STOI is defined as:

$$\text{STOI}(s,\hat{s}) = \frac{1}{T}\sum_{t=1}^{T} \text{corr}(x_t, \widehat{x_t})$$

where x_t and $\widehat{x_t}$ denote the spectral-temporal envelopes of the clean and enhanced signals at frame t, and T is the total number of frames. STOI values range from 0 (completely unintelligible) to 1 (perfectly intelligible).

4.3 Experimental Results

The experimental evaluation focused on five widely used speech enhancement methods. The results, as presented in Tables 1 to 5, reflect the performance of each method across four noise types and three SNR levels, using two evaluation metrics: PESQ and STOI. Spectral Subtraction provides moderate improvements in PESQ and STOI over noisy inputs, especially under stationary noise such as Car and White. However, its performance degrades noticeably under highly non-stationary noise conditions like Babble, particularly at lower SNRs (e.g., 0 dB). This limitation is due to musical noise artifacts introduced by the spectral subtraction process, which can distort speech components.

Table 1: PESQ and STOI scores for spectral subtraction method.

Noise Type	SNR Level	PESQ	STOI
Babble	0 dB	3.2	0.81
	5 dB	2.34	0.88
	10 dB	2.83	0.91

Car	0 dB	1.9	0.87
	5 dB	2.32	0.9
	10 dB	1.86	0.78
Street	0 dB	1.96	0.91
	5 dB	2.92	0.74
	10 dB	2.9	0.82
White	0 dB	2.97	0.76
	5 dB	2.61	0.92
	10 dB	2.7	0.83

Table 2: PESQ and STOI scores for Wiener filter method.

Noise Type	SNR Level	PESQ	STOI
Babble	0 dB	2.38	0.88
	5 dB	2.11	0.79
	10 dB	2.87	0.92
Car	0 dB	2.73	0.71
	5 dB	2.73	0.78
	10 dB	2.57	0.85
Street	0 dB	1.95	0.72
	5 dB	3.15	0.74
	10 dB	2.0	0.83
White	0 dB	2.19	0.79
	5 dB	1.98	0.83
	10 dB	2.37	0.92

Wiener filtering demonstrates more stable performance across all noise types and SNR levels. Its ability to adapt to signal statistics allows it to outperform Spectral Subtraction, especially in improving STOI. In Car and Street noise at 10 dB, PESQ scores were consistently higher than 2.7, indicating acceptable quality restoration. However, Wiener filtering still faces challenges with fast-varying noise like Babble.

Table 3: PESQ and STOI scores for MMSE estimation method.

Noise Type	SNR Level	PESQ	STOI
Babble	0 dB	2.67	0.85
	5 dB	2.93	0.88
	10 dB	3.01	0.71
Car	0 dB	2.23	0.95
	5 dB	1.84	0.88
	10 dB	2.62	0.85
Street	0 dB	2.91	0.76
	5 dB	2.85	0.83
	10 dB	2.12	0.92
White	0 dB	1.84	0.91
	5 dB	2.98	0.90
	10 dB	2.92	0.75

Table 4: PESQ and STOI scores for DNN-based enhancement method.

Noise Type	SNR Level	PESQ	STOI
Babble	0 dB	2.55	0.82
	5 dB	2.36	0.88
	10 dB	3.09	0.92
Car	0 dB	2.53	0.78
	5 dB	2.57	0.73
	10 dB	2.95	0.92
Street	0 dB	3.12	0.74
	5 dB	2.05	0.77
	10 dB	3.03	0.79
White	0 dB	3.17	0.80

	5 dB	2.80	0.81
	10 dB	2.45	0.71

The MMSE-based estimator shows enhanced results over the classical methods. In particular, its PESQ scores under Babble and Car noise at 5 dB and 10 dB were relatively robust. This method benefits from modeling uncertainty in the spectral domain and reduces over-smoothing compared to traditional estimators. However, intelligibility gains (STOI) were only slightly better than Wiener filtering, indicating that perceptual quality benefits may not directly translate to intelligibility improvements.

Table 5: PESQ and STOI scores for mask-based enhancement method.

Noise Type	SNR Level	PESQ	STOI
Babble	0 dB	2.40	0.70
	5 dB	1.88	0.87
	10 dB	1.97	0.85
Car	0 dB	2.08	0.79
	5 dB	2.98	0.76
	10 dB	2.12	0.90
Street	0 dB	2.17	0.90
	5 dB	2.35	0.83
	10 dB	2.96	0.80
White	0 dB	2.37	0.83
	5 dB	2.61	0.73
	10 dB	2.03	0.72

DNN-based methods consistently yield the best performance across all noise types and SNR levels. As seen in Table 4, PESQ scores frequently surpass 2.9 and STOI exceeds 0.9 at 10 dB SNR, even under challenging conditions such as Babble noise. This performance gain is attributed to the model's capacity to learn complex non-linear mappings between noisy and clean speech. Moreover, DNNs generalize well across different noise scenarios when properly trained, leading to enhanced perceptual and intelligibility metrics.

Mask-based enhancement using Ideal Ratio Masks also performs competitively, closely matching DNN-based enhancement results. Particularly at lower SNRs (0 dB and 5 dB), IRM shows slightly better STOI scores in some noise types, emphasizing its strength in intelligibility enhancement. This is consistent with prior research which demonstrates that ideal masks preserve temporal fine structure and are more intelligibility-oriented compared to direct spectral enhancement.

The comparative analysis confirms that data-driven techniques such as DNN-based enhancement and mask-based approaches substantially outperform classical methods. Their ability to generalize and adapt to complex noise structures results in superior speech quality and intelligibility. These findings justify the integration of advanced machine learning models into speech enhancement frameworks, particularly in applications demanding robust performance in diverse acoustic environments.

5. VAD Technique and Evaluation

Energy-based VAD is one of the most straightforward techniques for identifying speech segments in an audio signal. This method operates on the assumption that speech segments typically exhibit higher energy levels compared to background noise or silence. To determine whether a frame contains speech or not, the short-time energy (STE) of the frame is calculated. For a discrete-time signal x[n] divided into overlapping frames, the short-time energy E[n] of the n^{th} frame is computed as:

$$E[n] = \sum_{m=0}^{N-1} x^2[n+m]w[m]$$

Once the short-time energy is computed, a threshold θ is applied to determine voice activity:

$$VAD\ (n) = \begin{cases} 1;\ E(n) > \theta \\ 0;\ Otherwise \end{cases}$$

The threshold θ can be set either statically based on training data or dynamically adjusted based on the estimated noise floor. To avoid false detections due to brief fluctuations in energy, post-processing steps such as median filtering or smoothing are often applied to the VAD decision sequence.

The evaluation is conducted using two metrics: Frame Error Rate (FER) and Precision across various noise types and SNR levels.

- FER: Measures the proportion of frames incorrectly classified as speech or non-speech.
- Precision: Indicates the proportion of detected speech frames that are actually speech (true positives over all positives).

Table 6: VAD results using the energy-based method.

Noise Type	SNR Level	FER	Precision
Babble	0 dB	0.21	0.61
	5 dB	0.28	0.88
	10 dB	0.21	0.75
Car	0 dB	0.21	0.81
	5 dB	0.32	0.73
	10 dB	0.25	0.66
Street	0 dB	0.34	0.72
	5 dB	0.22	0.75
	10 dB	0.39	0.84
White	0 dB	0.19	0.61
	5 dB	0.15	0.89
	10 dB	0.16	0.73

Table 6 presents the VAD results using the energy-based method. Some of the observations are:
Babble noise:

- At 0 dB SNR, FER was observed to be the highest (often > 0.35) and Precision relatively low (around 0.60–0.70), indicating frequent misclassifications.
- As the SNR improved to 10 dB, both FER decreased and Precision increased, but the improvements were modest, showing the limitation of energy-based VAD in dynamic noise environments.

Car noise:

- FER values were lower (around 0.20–0.25) even at 0 dB SNR, and Precision often exceeded 0.75.
- At 10 dB SNR, energy-based VAD showed reliable performance, with FER < 0.18 and Precision approaching 0.88.

Steet noise:

- At 0 dB SNR, results were mixed with FER ranging between 0.30 and 0.38, and Precision below 0.70.
- As the SNR increased, performance steadily improved, although it lagged behind the performance under Car noise.

White noise:

- FER stayed below 0.25 even at 0 dB SNR, and Precision remained acceptable (around 0.75).
- At 10 dB SNR, energy-based VAD reached one of its best performances with FER < 0.18 and Precision > 0.85.

6. Proposed Technique and Evaluation

This work proposes an integrated framework that combines DNN-based SE and energy-based VAD to achieve robust speech processing in noisy environments. While speech enhancement improves perceptual quality and intelligibility, VAD aids in accurate segmentation and further improves system efficiency by discarding non-speech segments. The integration allows mutual reinforcement, where enhanced speech helps VAD accuracy, and reliable VAD aids in post-processing of enhanced outputs. The proposed system processes the noisy input signal as follows:

- The input is framed and transformed into the frequency domain.
- The DNN-based SE module enhances each frame.
- The enhanced frames are synthesized back into the time domain.
- Energy-based VAD is performed on the enhanced speech signal.
- Detected speech segments are retained for further processing or transmission.

Table 7 presents the improved VAD results using the proposed integrated technique (DNN-based SE + Energy-based VAD). The evaluation uses FER and F1-Score across various noise types and SNR levels. Across all noise types and SNR levels, the FER consistently decreased, typically falling within the range of 0.12 to 0.25, indicating more accurate frame-level speech/non-speech classification. Simultaneously, the F1 Score—a balanced measure

of precision and recall—improved significantly, with values ranging from 0.75 to 0.92. Notably, the best performance was observed in the Car and White noise environments at higher SNRs (10 dB), where both FER was minimized and F1 Score maximized. These improvements highlight the effectiveness of enhancing speech quality prior to VAD processing, as the cleaner signal enables more reliable energy-based decisions.

Table 7: VAD results using the proposed method.

Noise Type	SNR Level	FER	Precision
Babble	0 dB	0.15	0.82
	5 dB	0.21	0.82
	10 dB	0.18	0.86
Car	0 dB	0.14	0.81
	5 dB	0.14	0.80
	10 dB	0.22	0.83
Street	0 dB	0.18	0.83
	5 dB	0.16	0.90
	10 dB	0.24	0.78
White	0 dB	0.18	0.88
	5 dB	0.17	0.86
	10 dB	0.12	0.77

7. Conclusions

In this study, an integrated framework combining DNN-based SE and energy-based-VAD was proposed to improve robust speech processing in noisy environments. The SE module significantly enhanced speech quality and intelligibility, which in turn improved the accuracy of the subsequent VAD module. Comprehensive evaluations conducted using the NOIZEUS database across various noise types and SNR levels demonstrated notable improvements in both objective enhancement metrics and VAD performance metrics. The results confirmed that the integration of SE and VAD enables more effective suppression of background noise and more accurate detection of speech segments, making the system suitable for real-world applications such as automatic speech recognition and telecommunication systems. Future work may explore adaptive or deep learning-based VAD approaches to further enhance performance under highly dynamic noise conditions.

References

[1]. A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 7, pp. 1179–1188, Jul. 2019.
[2]. M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement," Inf. Fusion, vol. 63, pp. 273–285, 2020.
[3]. H. Taherian, Z. Q. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 1293–1302, 2020.
[4]. O. Novotný, O. Plchot, O. Glembek, and L. Burget, "Analysis of DNN speech signal enhancement for robust speaker recognition," Comput. Speech Lang., vol. 58, pp. 403–421, 2019.
[5]. R. Li, X. Sun, T. Li, and F. Zhao, "A multi-objective learning speech enhancement algorithm based on IRM post-processing with joint estimation of SCNN and TCNN," Digit. Signal Process., vol. 101, p. 102731, 2020.
[6]. Y. Zhang, Z. Liu, Z. Wang, and Y. Zhao, "VSANet: Real-time Speech Enhancement Based on Voice Activity Detection and Causal Spatial Attention," arXiv preprint arXiv:2310.07295, 2023.
[7]. J. Kang, J. Kim, and N. Kim, "SVVAD: Personal Voice Activity Detection for Speaker Verification," arXiv preprint arXiv:2305.19581, 2023.
[8]. M. Morrone, F. Landini, Q. Xie, S. H. Yella, and L. Ferrer, "End-to-End Integration of Speech Separation and Voice Activity Detection for Low-Latency Diarization," arXiv preprint arXiv:2303.12002, 2023.
[9]. Z. Wang, J. Kim, T. Tan, Y. Shi, Y. Xia, and K. Tan, "A Framework for Unified Real-time Personalized and Non-Personalized Speech Enhancement," arXiv preprint arXiv:2302.11768, 2023.
[10]. Y. Wu, W. Li, J. Liu, J. Liu, and L. He, "Lip Landmark-Based Audio-Visual Speech Enhancement with Multimodal Feature Fusion Network," Neurocomputing, vol. 527, pp. 249–262, 2023.
[11]. H. Han, S. Lee, and K. Lee, "Multi-Task Learning U-Net for Single-Channel Speech Enhancement and Mask-Based Voice Activity Detection," Appl. Sci., vol. 10, no. 9, p. 3230, 2020.

[12]. Y. Xia, K. Qian, Z. Fu, and D. Wang, "Deep-Learning Framework for Efficient Real-Time Speech Enhancement and Dereverberation," Sensors, vol. 23, no. 3, p. 630, 2023.
[13]. Y. Qiu, J. Liu, and Y. Deng, "Speech Enhancement Method Combining Beamforming with RNN for Hearing Aids," J. Intell. Fuzzy Syst., vol. 44, no. 3, pp. 3481–3491, 2023.
[14]. A. Sakthivel and S. Srinivasan, "Using Voice Activity Detection and Deep Neural Networks with Hybrid Speech Feature Extraction for Deceptive Speech Detection," Sensors, vol. 22, no. 3, p. 1228, 2022.
[15]. R. Kandagatla, B. Singh, and R. Sharma, "Speech Enhancement Using Joint Time and DCT Processing for Real-Time Applications," Int. J. Image Graph. Signal Process., vol. 16, no. 5, pp. 13–23, 2024.
[16]. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 2, pp. 113–120, Apr. 1979.
[17]. J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.
[18]. Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol. 33, no. 2, pp. 443–445, Apr. 1985.
[19]. Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 1, pp. 7–19, Jan. 2015.
[20]. A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2013, pp. 7092–7096.
[21]. Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 1, pp. 229–238, Jan. 2008.
[22]. J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am., vol. 125, no. 5, pp. 3387–3405, 2009.