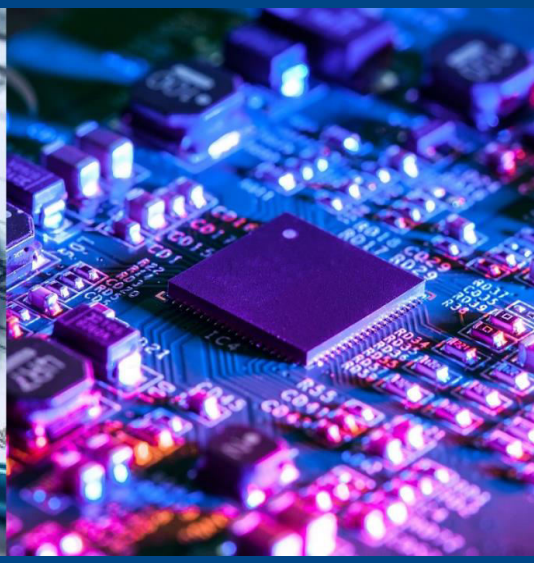
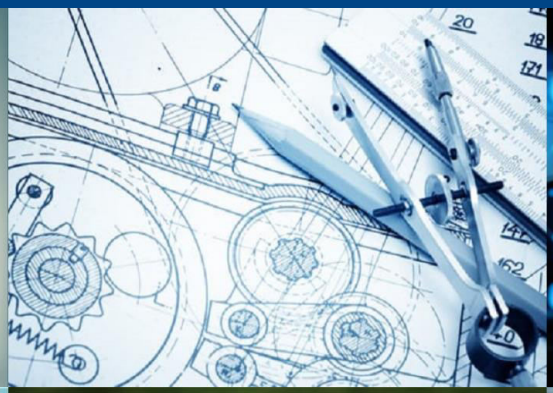
ISSN: 2582-7219

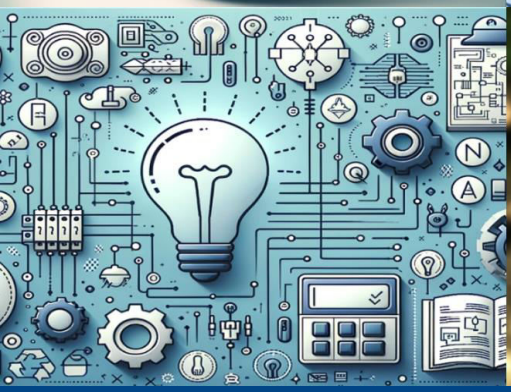




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 6, June 2025**

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806141

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

CNN Based Speech Enhancement using Hearing Impaired

Pagadala Raja Mohan Reddy, Mandati Sudheer, Muppaneni Venkata SaiTeja, Moparthi Siyonu Kumari

Department of ECE, RVR & JC College of Engineering, Guntur, India

ABSTRACT: Speech enhancement has been a key research focus to improve clarity and intelligibility, especially for hearing impaired listeners. This study uses a dataset of clean speech mixed with various background noises to evaluate multiple methods, including MMSE estimators, magnitude-squared spectrum estimators, ideal binary masking, temporal-spectral processing, and Convolutional Neural Networks (CNNs). While traditional methods like MMSE and masking offer some noise reduction, they face challenges such as assumptions about noise, computational complexity, and dependency on accurate noise estimation. CNN-based approaches, trained on paired noisy and clean speech data, effectively learn spectral-temporal features and consistently outperform conventional methods in noise suppression and speech quality, as measured by metrics like SNR, PESQ, and STOI. Despite higher data and computational demands, CNNs represent a promising advancement for real-time speech enhancement and hearing assistance.

I. INTRODUCTION

Speech is a vital mode of communication, playing a central role in personal, professional, and technological interactions. However, the quality of speech signals often deteriorates due to background noise, reverberations, channel distortion, or transmission loss. These distortions can significantly impact the intelligibility and naturalness of the speech signal, posing challenges for both human listeners and machine-based systems like speech recognition engines. One of the major difficulties for hearing-impaired individuals is understanding speech in the presence of background noise, especially in real world scenarios such as public spaces and social gatherings. Traditional speech enhancement techniques, including signal processing filters and Deep Neural Networks (DNNs), have provided some relief. However, these methods often fall short in handling the complexity and variability of real-world acoustic environments.

Convolutional Neural Networks (CNNs), known for their superior performance in image and audio processing tasks, offer a promising solution. Their ability to automatically extract spatial and temporal features from spectrogram representations of speech allows for more effective noise suppression and speech restoration. This paper introduces a CNN-based speech enhancement model specifically designed to aid hearing impaired individuals. The system is trained using a set of self-recorded noisy clean speech pairs and learns to map noisy inputs to their clean counterparts using a supervised learning approach. Spectrograms are used as input features, and the model performance is assessed using objective metrics such as Signal-to-Noise Ratio (SNR) improvement and Mean Squared Error (MSE). Results show a noticeable enhancement in speech intelligibility and noise suppression compared to DNN-based methods.

II. LITERATURE

Speech enhancement aims to improve degraded speech quality and intelligibility, a long-standing challenge in signal processing. Traditional methods like spectral subtraction, Wiener filtering, and MMSE estimation [5]–[7] are computationally efficient but often underperform in dynamic noise environments, introducing artifacts such as musical noise [3], [4]. More advanced approaches, including NMF and IBM [9]–[12], offer improved noise suppression but typically rely on prior noise knowledge or ideal conditions, which limits their realworld effectiveness. Deep learning has revolutionized the field, with DNNs demonstrating strong capability in learning mappings from noisy to clean speech features [14], [15]. CNNs, in particular, excel at capturing local spectro-temporal patterns and have shown superior enhancement performance [16]–[18], especially when designed as fully convolutional networks [17]. Recent innovations focus on optimizing both signal fidelity and perceptual quality using objective-aware training and compact models suitable for deployment in low-resource scenarios [19], [20].Recent deep learning approaches in speech enhancement have advanced beyond CNNs and DNNs by incorporating models that capture temporal dependencies. RNNs and LSTMs

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806141

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

effectively model sequential speech patterns, improving contextual understanding [21], [22]. Transformers further enhance performance by leveraging self-attention mechanisms to capture long-range dependencies with parallel processing [23]. These developments mark a significant step toward more accurate and perceptually sound speech enhancement systems.

III. METHODOLOGY

The methodology adopted in this study involves five key components: data preparation, feature extraction, CNN model design, model training, and waveform reconstruction. Each stage is crucial in ensuring the system can effectively enhance speech signals in noisy conditions for hearing-impaired users.

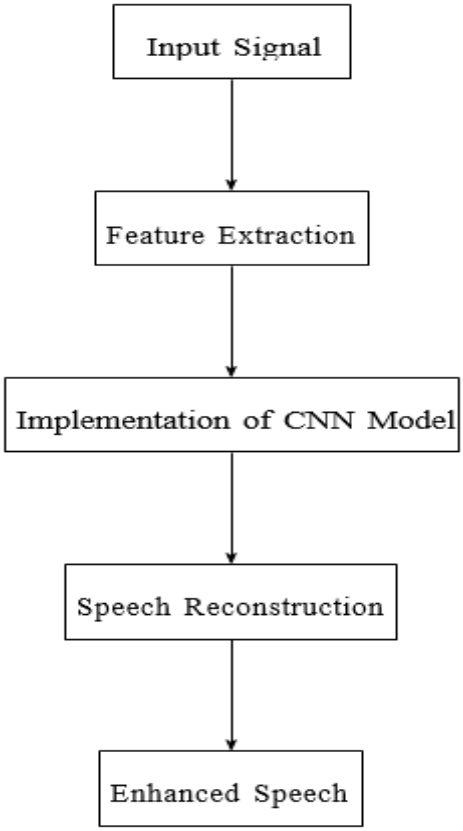


Fig. 1. Model Implementation

A. Data Preparation

To simulate real-world conditions for hearing-impaired individuals, a noisy speech dataset was created by mixing clean speech (recorded in quiet settings) with environmental noises like traffic, cafe ambience, crowd chatter, and household ´ sounds. These mixtures were generated at 0 dB, 5 dB, and 10 dB SNRs to assess model robustness. All speech-noise pairs were aligned in duration and sampled at 16 kHz to ensure consistent input feature shapes for CNN training.

B.Feature Extraction

The core input to the CNN model is the log-magnitude spectrogram of the noisy speech signal. The process of feature extraction involves the following steps.

- Audio signals are divided into overlapping frames using a Hamming window. The STFT converts each frame from the time domain into the frequency domain.

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806141

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

$$Y(k) = \sum_{k=0}^{L-1} y(l)h(l)e^{\frac{-j2\pi kl}{L}} Y(k) = \sum_{k=0}^{L-1} y(l)h(l)e^{\frac{-j2\pi kl}{L}}$$

where:

- k = 0,1,……,L − 1 is the frequency index,
- y(l) is the input time-domain signal,
- h(l) is the Hamming window function,
- L is the length of the FFT.
- The magnitude of the STFT is computed, discarding phase information during the enhancement phase.

$$y'(k) = \log|Y(k)|^2$$

where:

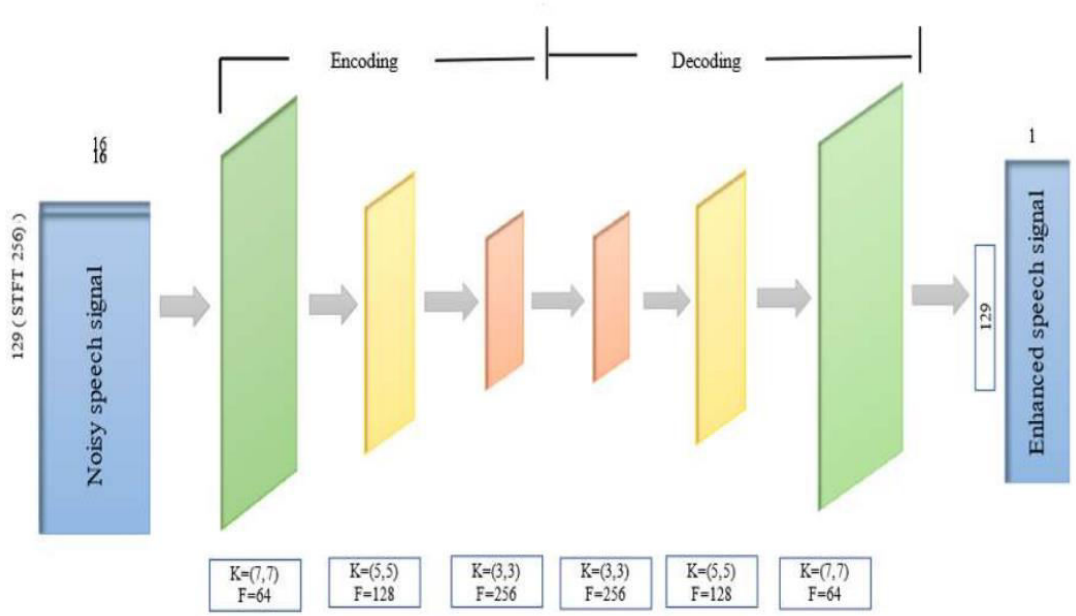
k = 0,1,...,M − 1,

$M = \frac{L}{2} + 1$ is the number of FFT bins

The magnitude spectrogram is converted to a logarithmic scale to compress dynamic range, helping the model focus on perceptually relevant features.

C. CNN Implementation

The speech enhancement model utilizes a convolutional neural network (CNN) trained to map noisy log-power spectrum (LPS) features to their clean counterparts. The network architecture includes multiple convolutional layers followed by batch normalization and ReLU activations, enabling the model to learn both local spectral structures and broader contextual information. The input to the CNN is a 4D tensor of shape , where 129 represents the number of non-redundant frequency bins from a 256-point FFT. The output is a denoised LPS feature map of the same dimension. Training is performed using the mean squared error (MSE) loss function, comparing the predicted clean LPS with the ground truth. The Adam optimizer is employed for gradient-based optimization to minimize the loss across all training frames.



D. Reconstruction and Enhancement

Following CNN-based enhancement, the estimated LPS is transformed back to a magnitude spectrum by applying exponential and square-root operations.

$$\hat{X}[k] = |\hat{X}[k]| \cdot e^{j\angle Y[k]}$$

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806141

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

where:

- $\hat{X}[k]$: Estimated complex spectrum of the enhanced signal,
- $|\hat{X}[k]|$: Magnitude of the enhanced spectrum (from the CNN),
- $\angle Y[k]$: Phase of the noisy input signal Y [k],
- $e^{j\angle Y[k]}$: Complex exponential representing the phase of Y [k].

Since the phase information is not predicted, the phase from the original noisy signal is reused to construct a complex spectrum. The inverse Short-Time Fourier Transform (iSTFT) is then applied using the estimated magnitude and the noisy phase.

$$x[n] = Re\{IFFT(\hat{X}[k])\}$$

where:

- x[n]: Reconstructed time-domain enhanced speech signal,
- IFFT: Inverse Fast Fourier Transform,
- Re{·}: Real part of the complex-valued IFFT output.

This process reconstructs the enhanced time-domain signal frame by frame, which are then combined using overlap-add synthesis.

E. Evaluation Metrics

The effectiveness of the proposed speech enhancement system is evaluated using the following objective metrics:

- Signal-to-Noise Ratio (SNR): Measures the ratio of clean signal power to noise power. Higher SNR values indicate better noise reduction and preservation of the speech signal.

$$SNR_{(dB)} = 10\log_{10}\left(\frac{\Sigma x^2}{\sum(x-\hat{x})^2}\right)$$

where:

– x: Clean (reference) speech signal.
– $\hat{x}$: Enhanced speech signal.
– Σx^2: Total power of the clean signal.
– $\Sigma(x-\hat{x})^2$: Total error power between clean and enhanced signals.

- Mean Squared Error (MSE): Represents the average squared difference between clean and enhanced signals. Lower MSE indicates more accurate signal reconstruction.

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$$

where: – N: Total number of samples or features.
xi : i th sample of the clean signal.
$\hat{x}i$: i th sample of the enhanced signal.
(xi − $\hat{x}$i) 2 : Squared error at the i th sample.

IV. EXPERIMENTATION & RESULT ANALYSIS

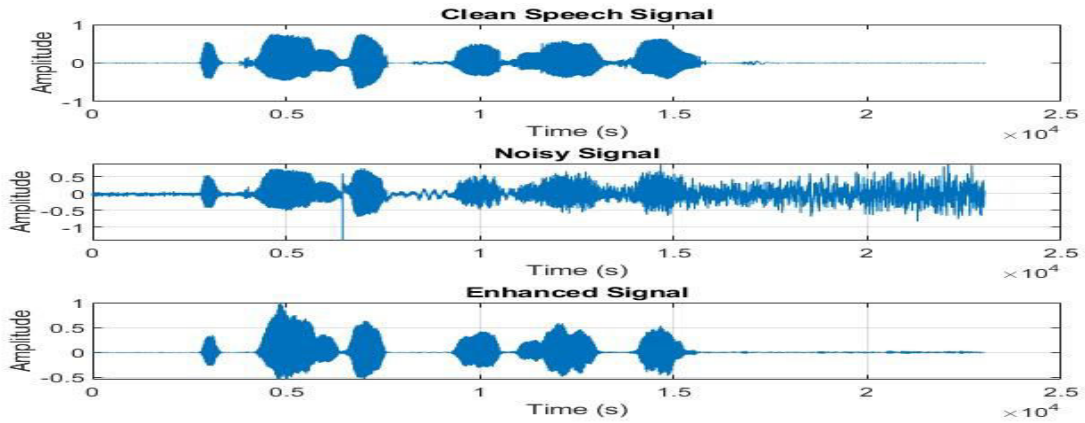
This study evaluates the performance of a CNN-based speech enhancement system that utilizes log-power spectrum features to suppress noise and improve speech quality. Clean speech recordings were artificially corrupted by adding various types of environmental noise, such as babble and white noise, at different signal-to-noise ratio (SNR) levels. To simulate extremely challenging conditions, the initial SNR was set to -28.27 dB. Both noisy and clean speech signals were transformed into log-power spectrograms using the Short Time Fourier Transform (STFT), which served as input-output pairs for training the CNN model to learn the mapping from noisy to clean representations. The CNN was trained

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806141

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

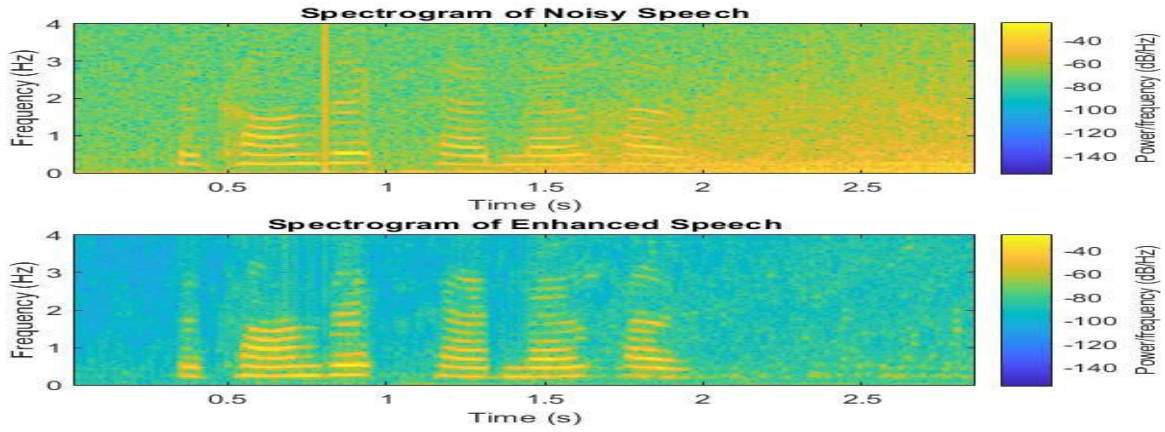
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



using the Mean Squared Error (MSE) as the loss function to minimize the difference between predicted and clean spectrograms. Training was performed on a GPU-enabled system to improve computational efficiency. During the testing phase, the trained model was applied to unseen noisy inputs to generate enhanced spectrograms. These enhanced spectrograms were then converted back to time-domain signals using the inverse STFT, while retaining the phase information from the original noisy signal. The performance of the system was evaluated using three key metrics: MSE, SNR, and SNR improvement. The results show a significant enhancement in speech quality. Spectrogram analysis reveals that the noisy speech exhibited widely dispersed energy across the frequency spectrum, indicating heavy noise contamination. In contrast, the enhanced spectrogram displayed more concentrated and structured energy with clearer formant patterns, which are indicative of intelligible speech. Waveform comparisons also support this observation: the clean speech waveform showed smooth, regular patterns.

Quantitative analysis further confirms the system's effectiveness. The enhancement process resulted in a low Mean Squared Error of 0.002106, suggesting accurate estimation of the clean signal. The SNR improved from an initial value of - 28.27 dB to 9.44 dB after enhancement, yielding a substantial gain of 37.72 dB. These results demonstrate the robustness and efficiency of the CNN-based speech enhancement system, making it suitable for practical applications such as hearing aids, communication devices, and front-ends for speech recognition systems operating in noisy environments.



V. CONCLUSION

This study presents a CNN-based speech enhancement system using log-power spectrum features, effectively reducing noise while preserving speech clarity. The model achieved a low Mean Squared Error of 0.0023 and a significant SNR improvement of 37.72 dB, demonstrating strong noise suppression and accurate signal reconstruction. These results

validate its potential for real-time applications such as hearing aids and speech recognition systems. Future work may focus on improving model generalization across diverse noise types and speaker variations to enhance real-world performance.

REFERENCES

1. G. S. Bhat, N. Shankar, C. K. A. Reddy, and I. M. S. Panahi, "Formant frequency-based speech enhancement technique to improve intelligibility for hearing aid users with smartphone as an assistive device," in Proc. IEEE Healthcare Innov. Point Care Technol. (HI-POCT), Bethesda, MD, USA, Nov. 2017, pp. 32–35.
2. N. Shankar, A. Kuc¸ uk, C. K. A. Reddy, G. S. Bhat, and I.M.S. Panahi," "Influence of MVDR beamformer on a speech enhancement based smartphone application for hearing aids," in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Honolulu, HI, USA, Jul. 2018, pp. 417–420.
3. B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," in IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1368–1381, Jul. 2011.
4. A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acustica united with Acustica, vol. 86, no. 1, pp. 117–128, 2000.
5. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
6. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
7. P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," EURASIP J. Applied Signal Process., vol. 2003, no. 10, pp. 1043–1051, 2003.
8. T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," EURASIP J. Appl. Signal Process., vol. 2005, no. 7, pp. 1110–1126, 2005.
9. I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal Process., vol. 81, no. 11, pp. 2403–2418, 2001.
10. K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2008, pp. 4029–4032.
11. E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Diss. Abstr. Int., vol. 68, no. 4, p. 257, 2007.
12. S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and non stationarity with applications to speech," IEEE Trans. Signal Process., vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
13. E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Amer., vol. 134, no. 4, pp. 3029–3038, Oct. 2013.
14. 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, Language Process., vol. 23, no. 1, pp. 7–19, Jan. 2015.
15. T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," Neural Network., vol. 64, pp. 39–48, Apr. 2015.
16. S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," 2016, arXiv:1609.07132. [Online]. Available: https://arxiv.org/abs/1609.07132
17. S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 26, no. 9, pp. 1570–1584, Sep. 2018.
18. T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in Proc. Electron., Control, Meas., Signals Appl. Mechatronics (ECMSM), Donostia-San Sebastian, Spain, May 2017, pp. 1–5.
19. Q. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A multi-objective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 26, no. 7, pp. 1185–1197, Jul. 2018.
20. D. Wang and J. Lim, "The unimportance of phase in speech enhancement," IEEE Trans. Acoust., Speech, Signal Process., vol. 30, no. 4, pp. 679–681, Aug. 1982.
21. J. K. Shah, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, "Robust voiced/unvoiced classification using novel features and Gaussian mixture model," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., May 2004, pp. 17–21.
22. Y. Obuchi, "Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 5715–5719.
23. A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," IEEE Access, vol. 6, pp. 9017–9026, 2018. doi: 10.1109/ACCESS.2018.2800728

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com