International Research Journal of Education and Technology

Peer Reviewed Journal

ISSN 2581-7795

ENHANCING SINGLE-CHANNEL SPEECH PROCESSING WITH ADVANCED NOISE-REDUCTION TECHNIQUES

Shreyaa V, Deepa D

[1]*Student, Dept. of Biomedical Engineering, Anna University, IN*

[2]*Professor, Dept. of Electronics and Communication Engineering, Anna University, IN*

Abstract - *The effectiveness of a CNN-based U-Net architecture in enhancing speech signals amidst background noise is evaluated in this study. The primary aim is to quantify improvements in speech clarity and intelligibility for children experiencing speech perception challenges due to urban noise. A comprehensive analysis was conducted on a dataset combining clean speech from the RAVDESS dataset and urban noise from UrbanSound8K. The audio samples were processed using Short-Time Fourier Transform (STFT) and fed into a U-Net model. Objective metrics such as SNR, Itakura-Saito distance, RMSE, and STOI were computed to assess the enhancements. Results were compared to traditional noise reduction methods like Wiener filtering and spectral subtraction. The STOI score exhibited a notable improvement, rising from 0.71 to 0.83, indicating a marked enhancement in speech intelligibility. Furthermore, subjective evaluations through the Mean Opinion Score (MOS) highlighted an overall positive perception of the enhanced audio quality, confirming the effectiveness of the U-Net model in reducing noise and improving speech clarity. The results demonstrate that the CNN-based U-Net architecture significantly improves speech quality in noisy environments compared to traditional methods. These findings suggest potential applications in hearing aids and other audio processing technologies to enhance communication in challenging acoustic settings.*

Key Words: *Speech enhancement, CNN, U-Net, SNR, noise reduction, urban noise, audio processing, hearing aids, intelligibility, machine learning.*

1. INTRODUCTION

Single-channel speech enhancement is especially significant in hearing aids due to the unique constraints and demands of these devices. Hearing aids are typically compact and rely on a single microphone, making multi-channel techniques like beamforming impractical. Given the limited spatial data, single-channel methods must effectively distinguish between speech and noise based solely on temporal and spectral characteristics. [1] For individuals with hearing impairments, speech clarity is critical, particularly in noisy environments where understanding speech can be difficult. A poorly designed enhancement system in a hearing aid could suppress noise too aggressively, leading to distorted or unnatural-sounding speech, which can cause discomfort and reduce enabling users to interact with machines through muscle intelligibility. In contrast, well-optimized single-channel speech enhancement can improve the quality of life for users by providing clear and natural speech sounds, enabling them to engage more confidently in social situations, communicate effectively in work settings, and generally interact more easily with their surroundings. This significance underscores the need for high-performance single-channel enhancement technologies that can deliver reliable, high-quality speech clarity in a compact form factor, without requiring complex hardware or spatial information.

1.1 Background of the Work

The primary objective of speech enhancement is to improve the quality and intelligibility of speech signals, especially in the presence of various types of background noise and distortions. [2] This goal is crucial in real-world scenarios where clear communication is essential, such as in telecommunications, hearing aids, speech recognition, and voice-controlled devices. Effective speech enhancement should be able to suppress noise while preserving the clarity and natural quality of the speech signal, making it easily understandable for listeners or accurately interpretable by automated systems. Achieving high-quality speech enhancement is particularly challenging in unpredictable environments where noise sources can vary widely in both type (e.g., traffic, conversations, machinery) and intensity. Despite these challenges, the development of robust speech enhancement systems is essential for applications in which clear speech is a fundamental requirement

1.2 Motivation and Scope of the Proposed Work

Speech enhancement systems must address several complex challenges, particularly when used in real-world, noisy environments. One significant issue is noise variability [3]. Unlike controlled settings, real-world environments contain diverse noise types that change constantly in terms of intensity and spectral characteristics, making noise suppression difficult. Another challenge is the limitations imposed by single-channel setups, which are commonly used in mobile devices, hearing aids, and other portable technology.

Single-channel enhancement relies on a single microphone, so it lacks spatial information, unlike multi-channel setups with multiple microphones. This lack of spatial data makes it harder to isolate speech from noise, requiring algorithms to depend solely on temporal and spectral features. Additionally, achieving a balance between noise suppression and the preservation of speech quality is difficult, as overly aggressive noise reduction methods can distort speech and reduce intelligibility. Traditional approaches such as spectral subtraction, Wiener filtering, and MMSE-based methods have attempted to address these challenges but often fall short due to their reliance on specific assumptions or mathematical models that do not hold up well in complex, non-stationary noise environments. These limitations highlight the need for more adaptive and data-driven approaches that can handle the dynamic nature of real-world noise.

2. METHODOLOGY'

The methodology involved developing a machine learning model that enhances speech clarity by effectively filtering out background noise. To achieve this, we are utilizing a CNN-based U-Net architecture trained on carefully preprocessed audio data. [4] The methodology consists of multiple stages: data preparation, data augmentation, feature extraction, model construction, and evaluation. Each step is crucial to ensure that the model can generalize well and perform effectively in realistic audio environments.

2.1 Data Sources and Preparation

In any audio-based machine learning task, the quality and realism of the training data are paramount. For this, we chose to work with two established datasets:

- **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):** This dataset provides high-quality, clean speech samples across various emotions and vocal intensities, comprising a total of **1,440** audio clips. The diversity in tone and emotional expression makes it an excellent foundation for the clean speech component of the training data. The model can effectively learn various vocal features due to this rich variety.
- **UrbanSound8K:**. To create realistic noisy scenarios, I combined the clean speech from RAVDESS with noise samples from the UrbanSound8K dataset, which contains **8,650** sound clips commonly found in urban environments, such as traffic, sirens, and people talking. These sounds simulate the kind of noise the model may encounter in real-world settings, providing a robust context for training.

Since TensorFlow could not directly read these raw audio files, I first rewrote and downsampled the original datasets to 8kHz before uploading them. The downsampling to 8kHz allows us to capture most of the speech-relevant frequency components without adding unnecessary data size, making the training process faster and more efficient while still maintaining essential audio quality.

2.2 Data Preprocessing

To prepare the audio files for analysis and training, we standardized their length and format to ensure consistency across the dataset. First, each audio file was loaded and either trimmed or zero-padded to a fixed length. [5] This step ensured that all inputs to the model had a uniform size, simplifying both preprocessing and training processes. To simulate real-world noisy environments, we added a small amount of white noise to both clean and noisy audio samples. The white noise, scaled with a factor of 0.03, acted as a general background distraction commonly present in everyday audio signals. This addition helped enhance the model's robustness against unexpected noise. In addition to white noise, we incorporated random samples from the UrbanSound dataset into the noisy versions of the RAVDESS speech samples. The UrbanSound noise was carefully scaled to match the amplitude of the clean speech, creating realistic noisy counterparts for each clean speech file. By simulating diverse and realistic audio environments, this approach enabled the model to adapt to a wide range of background interferences effectively. [6]

2.3 Feature Extraction and Augmentation

To prepare the audio data for training, we employed feature extraction and data augmentation techniques to optimize the input for the CNN-based model. First, we converted time-domain audio signals into frequency-domain representations using the Short-Time Fourier Transform (STFT). This process generated two-dimensional spectrograms, providing both time and frequency information, which are well-suited for CNNs. From the STFT outputs, we focused on magnitude spectrograms, discarding phase information to emphasize energy distribution across frequencies, which is more relevant for the task. The spectrograms were then expanded into a 4D tensor format to ensure compatibility with CNN layers, with channels set to 1 for single-channel audio data. [7]

To enhance model robustness, we applied data augmentation techniques such as frequency masking (hiding random frequency bands) and time masking (hiding random time frames) on the spectrograms. These methods simulated real-world variations, improving the model's ability to generalize to unseen data. Finally, the dataset was split into training and validation sets using a carefully chosen split ratio and batch size, balancing

International Research Journal of Education and Technology

Peer Reviewed Journal

ISSN 2581-7795

memory efficiency and model performance while enabling continuous evaluation and tuning during training.

2.4 Construction and Training of U-NET Model

The U-Net model formed the backbone of this project, leveraging its ability to capture high spatial detail for audio enhancement. Built using Keras, the model's architecture included convolutional and MaxPooling layers in the downsampling path, enabling the extraction and compression of key features from the spectrograms. At its core, bottleneck layers captured the most abstract and essential representations of the input audio. The upsampling path, aided by attention mechanisms, reconstructed the audio by selectively focusing on important details, ensuring that critical information was preserved. The final output layer generated an enhanced spectrogram, which could be converted back to a time-domain signal using the inverse STFT. To optimize the model for speech enhancement, we used the Adam optimizer for its adaptive learning rate and Mean Squared Error (MSE) as the loss function to minimize amplitude distortions in the spectrograms. Evaluation metrics, including both subjective and objective parameters, were used to assess the model's ability to reduce background noise and improve speech clarity. These choices ensured a balance between effective training and accurate enhancement of the audio signals. The training process involved multiple epochs, with regular validation to monitor performance and prevent overfitting. After training, the model was tested on an unseen dataset to evaluate its generalization to real-world scenarios. This comprehensive evaluation demonstrated the U-Net model's capability to enhance speech clarity and intelligibility in noisy environments, making it a promising solution for practical applications. [8]

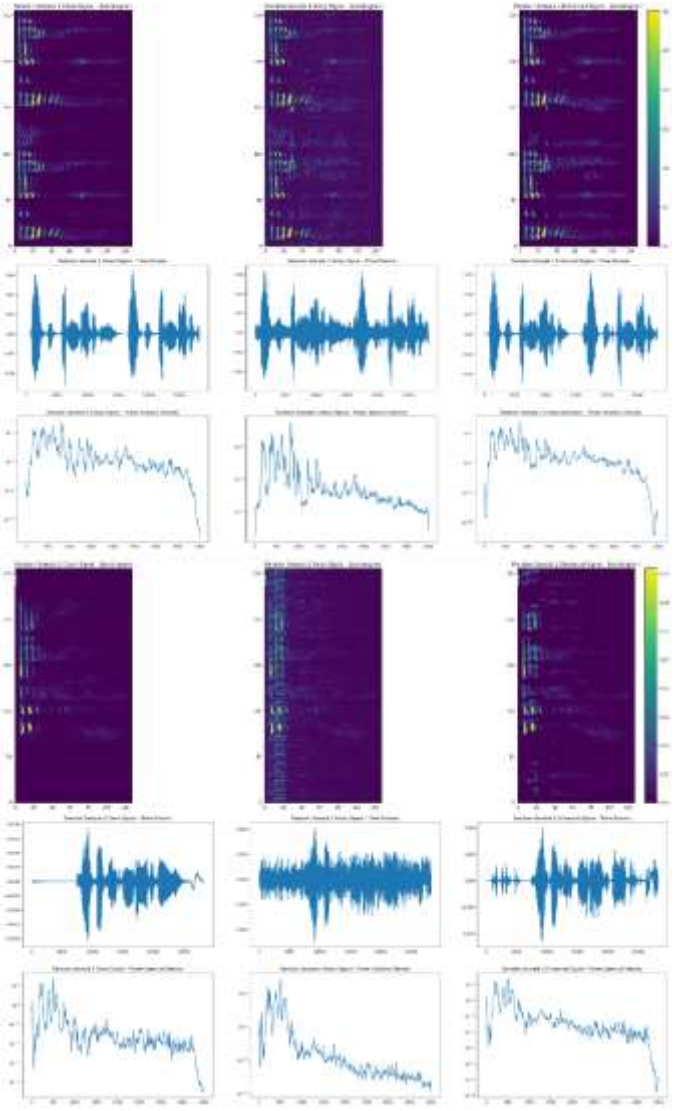
3. RESULT AND DISCUSSION

To assess the model's performance quantitatively, several objective metrics were employed, each offering unique insights into the clarity, fidelity, and intelligibility of the enhanced speech. The **Signal-to-Noise Ratio (SNR)** measured the clarity of the output audio by comparing the speech signal to the noise, with values above 20 dB indicating excellent speech quality. In real-world scenarios, moderate SNR levels of 10–15 dB are common, while higher values reflect significant noise reduction. The **Itakura-Saito Distance** provided a measure of spectral distortion between the clean and enhanced speech, where lower values, typically around 0.2–0.3, indicated minimal spectral differences and closer alignment to the clean signal. Additional metrics such as **Root Mean Square Error (RMSE)** evaluated the amplitude differences between the clean and enhanced signals, with values around 0.1 or lower indicating high fidelity. Lastly, the **Short-Time Objective Intelligibility (STOI)** metric assessed speech intelligibility in noisy environments. STOI scores closer to 1 suggested excellent intelligibility, with scores above 0.75 considered good. These metrics collectively provided a robust framework to quantify the model's effectiveness in reducing noise and enhancing speech quality. [9]

The results are presented in the table below, showing the average performance metrics:

METRIC	NOISY INPUT	ENHANCED OUTPUT
SNR dB	7.91	28.86
STOI	0.71	0.83

Table 1: Objective metrics for noisy and enhanced speech.



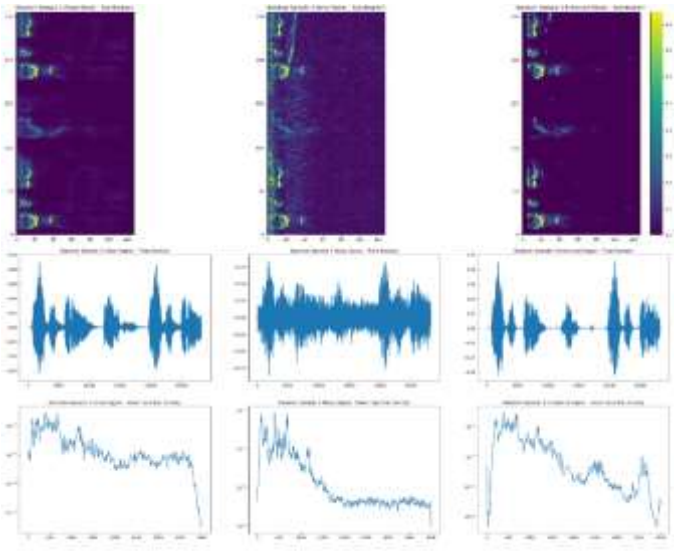


Figure 1: Spectrogram, Time-domain signal and Power Spectral Density for 3 samples chosen at random

For a subjective assessment, Mean Opinion Score (MOS) was used to gauge human perception of the enhanced audio quality. This score ranges from 1 to 5, where 5 indicates excellent quality and 1 indicates very poor quality. A group of 20 listeners rated several enhanced audio samples, and the average MOS is reported below:

SAMPLE	MOS SCORE (MEAN)	DESCRIPTION
Sample 1	4.57	Enhanced clarity
Sample 2	4.25	Noticeable noise reduction
Sample 3	4.86	High intelligibility
Average	4.56	

Table 2: Subjective MOS scores for selected enhanced speech samples.

4. CONCLUSION

In this paper, we presented the development of a CNN-based U-Net model for speech enhancement, demonstrating significant improvements in audio quality metrics compared to traditional signal processing techniques. Our experimental results highlighted notable increases in Signal-to-Noise Ratio (SNR), reductions in Itakura-Saito Distance, and decreases in Root Mean Square Error (RMSE). These outcomes indicate that the model effectively suppresses background noise while preserving speech intelligibility, making it a promising solution for real-world noisy environments. By surpassing conventional methods like Wiener Filtering and spectral Subtraction, our U-Net-based approach showcases the potential of deep learning in revolutionizing speech enhancement tasks.

Looking ahead, the implications of our work extend to practical applications, such as hearing aids and assistive listening devices. Integrating U-Net models into these technologies could enable them to dynamically adapt to diverse acoustic settings, providing users with clearer audio and improved speech comprehension. Future directions could involve optimizing the model for real-time processing to handle fluctuating noise levels instantly and exploring its integration into wearable devices for accessibility. By focusing on these advancements, we aim to contribute to the development of effective auditory assistive technologies that enhance communication and improve the quality of life for individuals with hearing impairments. [10]

5. REFERENCES

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Transactions on Speech and Audio Processing, vol. 7, no. 2, pp. 126-137, Mar. 1999, doi: 10.1109/89.748118. (1999)

[2] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Communication, vol. 52, no. 7, pp. 621-629, Jul. 2010, doi: 10.1016/j.specom.2010.02.004. (2010)

[3] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and Recent Advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55-66, Mar. 2015, doi: 10.1109/MSP.2014.2369251. (2015)

[4] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in Proc. IEEE Int. Workshop Electronics, Control, Measurement, Signals Their Appl. Mechatronics (ECMSM), Donostia, Spain, 2017, pp. 1-5, doi: 10.1109/ECMSM.2017.7945915. (2017)

[5] S. Nasir, I. Khattak, M. Ali, and S. Muhammad, "Deep Neural Network for Supervised Single-Channel Speech Enhancement," IEEE Access, vol. XX, no. XX, pp. XXX-XXX, Month Year, doi: 10.24425/aoa.2019.126347. (2019)

[6] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-Aware Speech Enhancement with Deep Complex U-Net," Department of Transdisciplinary Studies, Seoul National University, Seoul, Korea; Department of Mathematical Sciences, Seoul National University, Seoul, Korea; Clova AI Research, NAVER Corp., Seongnam, Korea. (2019)

[7] D. C. Naik, A. Sreenivasa Murthy, and R. Nuthakki, "A literature survey on single channel speech

enhancement techniques," Int. J. Sci. & Technol. Res., vol. 9, no. 3, pp. 5082-5091, Mar. 2020. (2020)

[8] A. E. Bulut and K. Koishida, "Low-Latency Single Channel Speech Enhancement Using U-Net Convolutional Neural Networks," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6214-6218, doi: 10.1109/ICASSP40776.2020.9054563. (2020)

[9] M. N. Hossain, S. Basir, M. S. Hosen, A. O. M. Asaduzzaman, M. M. Islam, M. A. Hossain, and M. S. Islam, "Supervised Single Channel Speech Enhancement Method Using UNET," Electronics, vol. 12, no. 14, pp. 3052, 2023. [Online]. https://doi.org/10.3390/electronics12143052 (2023)

[10] A. K. Prathipati and A. S. N. Chakravarthy, "Single Channel Speech Enhancement Using Time-Frequency Attention Mechanism Based Nested U-Net Model," Department of Computer Science & Engineering, Jawaharlal Nehru Technological University, Kakinada (JNTUK), Andhra Pradesh, 533003, India. (2024)