Speech Recognition and Enhancement for Hearing Aids

Sathwik Chowdary Merla
Northeastern University
Boston, MA
merla.s@northeastern.edu

Rucha Bhandari
Northeastern University
Boston, MA
bhandari.ru@northeastern.edu

Dharm Mehta
Northeastern University
Boston, MA
mehta.dhar@northeastern.edu

***Abstract*— This paper proposes a solution to address the communication challenges faced by people using hearing aids. We propose a hybrid speech recognition and enhancement processing pipeline. A combination of traditional signal pre-processing in MATLAB and a CNN-LSTM architecture implemented in Python is explored through our project. MATLAB is initially used for pre-processing the audio data using multiple methods, i.e., filtering, spectral subtraction, and power spectral density visualization, to know the power of the signal in a certain range of frequencies. The machine learning model is trained on a diverse dataset of various hearing loss profiles. The model showed consistent improvement and was able to enhance speech intelligibility. The aim of this project is to implement an scalable, cost-effective solution that is helpful for people with hearing impairments.**

Keywords: noise reduction, speech enhancement, MATLAB pre-processing, CNN, Bi-LSTM, MBSTOI

I. INTRODUCTION

Hearing loss affects millions of individuals worldwide and has significant effects on communication, social interaction, and quality of life. In contrast to traditional hearing aids, which amplify the sounds to assist the user in hearing better, in challenging acoustic environments where there is background noise and multiple speakers, they simply break down.

Emerging advances in digital signal processing (DSP) and machine learning have made it possible to design intelligent hearing systems to detect and enhance speech signals. The innovations equip hearing aids to transcend rudimentary amplification by using speech recognition, noise cancellation, and context-dependent audio processing to provide a more natural and comprehensible listening experience.

This project, "Speech Recognition and Enhancement for Hearing Aids", seeks to circumvent such limitations by employing state-of-the-art DSP techniques combined with machine learning algorithms. The proposed system is a two-stage processing chain: signal preprocessed using MATLAB, while primary speech recognition and enhancement are processed in Python. MATLAB is chosen because of its robust and efficient signal processing ability, which is used to perform operations such as noise reduction, filtering, and feature extraction. These preprocessed signals are then sent to the ML model, where speech processing and machine learning algorithms are used to identify and enhance the speech components of the audio.

The ultimate objective of this project is to produce a practical and scalable solution that maximizes speech intelligibility for hearing aid users, particularly in noisy or dynamic environments.

With ongoing improvements, a system like this can be implemented for current hearing aid devices and improve them, making them smarter, more efficient, and simpler to use. The project demonstrates how interdisciplinary approaches can be utilized to create impactful results in human-centered assistive technology and design.

II. System Overview and Workflow

The proposed system for Speech Enhancement and recognition for hearing aids is organised as a two-stage processing unit. This integrates conventional signal processing with advanced machine learning and deep learning models. First, the audio signal is pre-processed using MATLAB and its operation such as noise reduction, filtering, normalization and resampling are applied to make a cleaner version of the input audio. Subsequently the enhanced audio is processes by the ml model, here we have applied a hybrid CNN and Bi-LSTM model in Python using the Jupyter notebook environment. This helped us to further improve speech intelligibility.

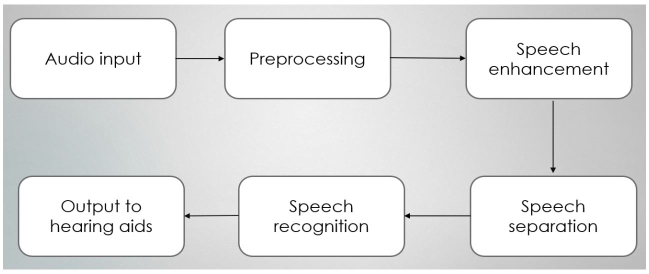


Fig. 1. Comparison with other models

A step-by-step workflow diagram is used to display the entire system from raw audio input to output speech enhancement. The modular nature allows better speech recognition accuracy and significant enhancement of speech intelligibility, and thus it is suited for hearing aid applications. Overview of the overall system parts and how they collaborate to achieve real-time speech recognition and enhancement is given in this section.

The next subsection provides an overview of the process utilized within the workflow of the project, from a high-level view of the operation of the system. It describes the step-by-step process of the system functioning from receiving raw audio input at step one to finally providing enhanced speech output.

The focus of the research effort in this system is to enhance the accuracy of speech recognition devices and hearing assist devices. The first step in pursuing that goal consists of Audio Input, which means recording from a database of pre-recorded speech or using live microphones.

Additionally, the voice signal will be going MATLAB for the preprocessing step. This involves a number of important processing steps, such as noise reduction, audio normalization, and spectral subtraction. The final result of the process is a dataset of processed audio data ready for use in the later steps.

Next, we fed the enhanced audio into a system that incorporates a hybrid Convolutional Neural Network and a Bi-Directional Long Short-Term Memory model. The CNN focuses on extracting spatial patterns from the audio signals and identifies speech elements from noise elements. CNN emphasises and learns frequency bands associated with human speech and suppresses the irrelevant background noise. The LSTM captures long-term dependencies and models the sequential evolution os speech across various frames. This architecture ensures that individual frames are enhanced, but also that continuity and naturalness of speech are preserved.

The model also performs speech separation, this isolates the speakers voice front the background noise, this can also include multiple speakers. This is pretty important within crowded environments like restaurants or public transport.

Lastly the final audio output represents what the hearing aid would hear, providing the user with an improved and more understandable version of original speech.The system is successful at picking up overlapping speech and background noise, allowing only the relevant speech information to pass on to the next level in the process.

Our system offers a comprehensive solution to significantly improve the hearing experiences of individuals with varying degrees of hearing loss. This complex process simplifies and maximizes the processes involving hearing perception and oral interaction.

III. Data Preprocessing

Data preprocessing is necessary to efficiently run our speech enhancement system and is carried out in MATLAB. We begin with the processing of the raw speech input. It is better to ensure it's mono with a single channel and not silent and misplaced. If it's necessary, we convert the signal to the regular 16 kHz frequency to standardize it. Low-frequency noise is eliminated by a high-pass filter and the high-frequency components of the speech are amplified to make it more clear by pre-emphasis.

The volume is then cranked to the maximum level so the sound will be uniform and easier to enhance. The audio was processed in the segmented manner using 40-millisecond framing with a 50% overlap and Hamming window.

The extensive preprocessing provides a basis for the optimization of speech recognition and improvement through variance reduction and the extraction of speech qualities of importance. A configuration file provides the modifications that will handle different hearing needs in hearing aids. Computation of Power Spectral Density (PSD) graphs before and after preprocessing helps to examine the distribution of energy in frequency bands and determine the success of noise reduction processes. The graphs enable the evaluation of noise suppression achieved through the improvement process without losing the vital speech frequency components.

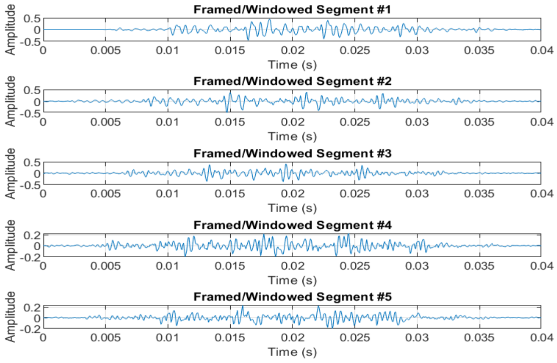


Fig. 2. Frame processing for audio data

aid in examining the energy distribution across various frequency components and assessing how well noise reduction methods work. The PSD comparison also sheds light on how enhancement techniques minimize noise while maintaining speech-dominant bands.

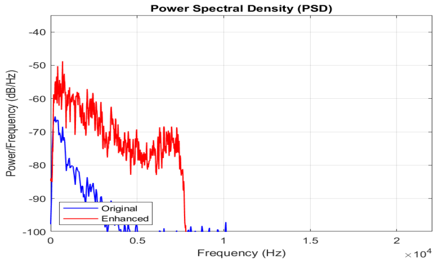
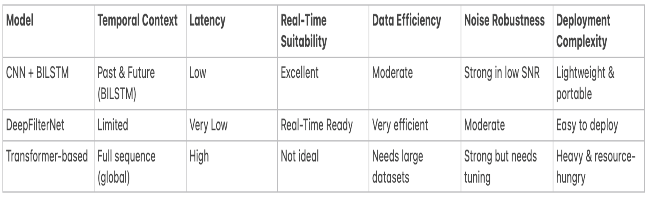


Fig. 3. PSD comparision between orignal vs enhanced audio

IV. BACKGROUND

It is estimated that 466 million people worldwide have hearing loss and, by 2050, that number is expected to increase to 900 million. Modern devices use digital signal processing techniques to improve speech intelligibility in noisy environments, which remains one of the biggest issues for hearing aid users. In complex acoustic scenarios with nonstationary noise, traditional methods based on beamforming, Wiener filtering, and spectral subtraction have shown limited efficacy.

Deep learning has transformed hearing aid speech enhancement in recent years, providing notable advancements over conventional techniques. While Recurrent Neural Networks (RNNs) and their variations, such as Long Short-Term Memory (LSTM) networks, have shown efficacy in modeling sequential data with temporal dependencies, Convolutional Neural Networks (CNNs) have shown remarkable ability in extracting local spectro-temporal patterns from audio spectrograms. For applications involving hearing aids, the combination of these models has demonstrated significant promise than the other models present in the market.



Model	Temporal Context	Latency	Real-Time Suitability	Data Efficiency	Noise Robustness	Deployment Complexity
CNN + BILSTM	Past & Future (BILSTM)	Low	Excellent	Moderate	Strong in low SNR	Lightweight & portable
DeepFilterNet	Limited	Very Low	Real-Time Ready	Very efficient	Moderate	Easy to deploy
Transformer-based	Full sequence (global)	High	Not ideal	Needs large datasets	Strong but needs tuning	Heavy & resource-hungry

Fig. 4. Comparison with other models

Leveraging the added advantages of both, the CNN-BiLSTM hybrid model offers a convincing method for speech enhancement. While CNNs are fairly effective in detecting local patterns and spectral features in time-frequency representations when it comes to separating speech from noise, bidirectional LSTMs capture long-term relationships in both forward and backward time directions. This bidirectional processing is especially helpful for uses involving hearing aids since it preserves both anticipatory and retroactive cues, improving speech intelligibility.

Recent advances like the DeepFilterNet models have gone beyond simple masking approaches to apply complex filtering techniques that better preserve speech characteristics. By applying domain knowledge of speech production and psychoacoustic perception, these models achieve higher performances while maintaining computational efficiency suitable for hearing aid devices.

The Modified Binaural Short-Time Objective Intelligibility (MBSTOI) metric has emerged as the industry standard for evaluating speech enhancement algorithms for hearing aids because it provides an objective score that closely resembles subjective intelligibility. This metric is quite useful for evaluating algorithms designed for binaural hearing aid systems since it considers both better ear listening and binaural unmasking effects.

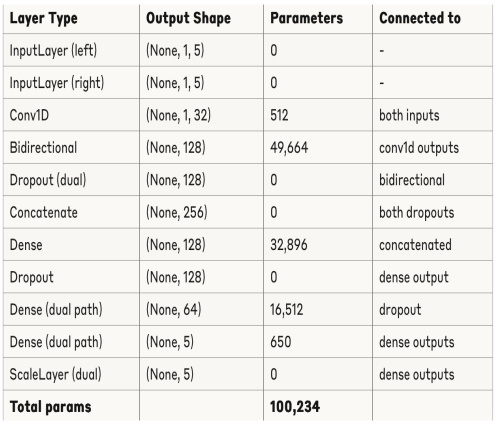
Our research work builds on these bases to build a CNN-BiLSTM architecture optimized especially for hearing aid applications, with particular attention to preserving low computational complexity, minimal latency, and maximum speech intelligibility measured by MBSTOI.

V. Network architecture

The proposed hearing aid enhancement system employs a novel binaural architecture that balances independent channel processing and cross-channel information sharing to maintain spatial cues. The network topology, feature processing pipeline, and component interactions are mentioned in this section below.

A. Model Topology

Our architecture features a dual-input, dual-output structure designed specifically for binaural hearing aid applications. While it provides a comprehensive overview of the model architecture, including layer sizes and parameter counts, Table below lists the key architectural elements.



Layer Type	Output Shape	Parameters	Connected to
InputLayer (left)	(None, 1, 5)	0	-
InputLayer (right)	(None, 1, 5)	0	-
Conv1D	(None, 1, 32)	512	both inputs
Bidirectional	(None, 128)	49,664	conv1d outputs
Dropout (dual)	(None, 128)	0	bidirectional
Concatenate	(None, 256)	0	both dropouts
Dense	(None, 128)	32,896	concatenated
Dropout	(None, 128)	0	dense output
Dense (dual path)	(None, 64)	16,512	dropout
Dense (dual path)	(None, 5)	650	dense outputs
ScaleLayer (dual)	(None, 5)	0	dense outputs
Total params		**100,234**	

Fig. 5. NETWORK ARCHITECTURE SUMMARY

The model accepts two parallel input channels (left_input and right_input) with dimensions (None, 1, 5), where:

Here, none indicates that the batch size of the input layer can vary. The temporal dimension (single time step) is represented by 1. The acoustic feature vector taken from each channel is represented by 5.

This input representation preserves the crucial interaural information needed for spatial audio perception while enabling effective processing of binaural signals.

B. Feature Extraction and Processing

The first step in the feature extraction process is a shared convolutional layer which is a CNN layer (Conv1D), which will processes both left and right inputs and has 32 filters. This layer transforms the 5-dimensional input features into a 32-dimensional representation while preserving the temporal structure. Consistent feature extraction is ensured while lowering the total number of parameters by sharing the convolutional layer across channels. Following feature extraction, a bidirectional recurrent layer processes convolutional features, capturing both forward and backward temporal dependencies. This BiLSTM layer transforms the time-series data into a 128-dimensional vector for each channel, accounting for the majority of the model's parameters (49,664 out of 100,234 total).

C. Channel Fusion and Enhancement

A key innovation in our architecture is the controlled information sharing between channels. When the BiLSTM outputs are subjected to dropout regularisation (dropout_3 and dropout_4), a concatenation layer merges these features into a thorough 256-dimensional representation, allowing for cross-channel information sharing while preserving channel-specific features. After that, dropout regularisation is applied after this combined representation has gone through a shared dense layer with 128 units. After that, the processing path divides into two parallel branches:

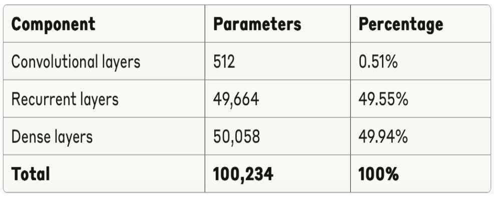
Left channel branch: dense_6 (64 units) → dense_8 (5 units) → left_scaled

Right channel branch: dense_7 (64 units) → dense_9 (5 units) → right_scaled

This design guarantees that shared information can be utilised while channel-specific enhancement can be implemented. The improved 5-dimensional features for every audio channel are provided by the left_scaled and right_scaled ScaleLayer outputs, which are subsequently utilised to reconstruct.

D. Efficiency Considerations

With approximately 100,000 parameters (Table below), our model achieves a favorable balance between enhancement performance and computational efficiency. The model's ability to learn channel-specific enhancement patterns has been preserved while redundancy has been minimised through the careful optimisation of the parameter distribution across layers, including the strategic use of shared weights in the convolutional and early dense layers.



Component	**Parameters**	**Percentage**
Convolutional layers	512	0.51%
Recurrent layers	49,664	49.55%
Dense layers	50,058	49.94%
Total	**100,234**	**100%**

Fig. 6. PARAMETER DISTRIBUTION

With the right quantisation methods, this architecture can be implemented on contemporary digital hearing aid hardware due to its fewer parameters required. Assuming 5-dimensional input features, the model only needs 400,936 multiply-accumulate operations per frame during inference, allowing for real-time processing on devices with limited power.

VI. Data Processing and Hearing Aid Simulation

We wanted to train our model on audio inputs that would mimic actual hearing loss in humans, thus, we transformed clean speech signals into a form that was suitable for it. Each of the audio clips from our database was split into 30 ms frames with a 15ms overlap. In addition, energy characteristics were extracted across the five standard frequency bands. To simulate human-like hearing loss, we tried lowering the volume in certain frequency ranges, like softening high-pitched sounds, for people with this type of hearing loss. Likewise, we used 7 different patterns to match the common types of hearing loss problems.

To train our model on realistic conditions, we introduced reverberations, environmental noise such as pink, white, babble, and speech-shaped noise. We also added small delays and amplitude between the left and right channel to mimic what the left ear and right ear hear, This helps the model to train on how sounds come from different directions in real-life examples. This pre-processing pipeline produced 423,800 binaural examples, each converted into 5-dimensional feature vectors. 339,040 usable frame pairs were generated per ear for model training.

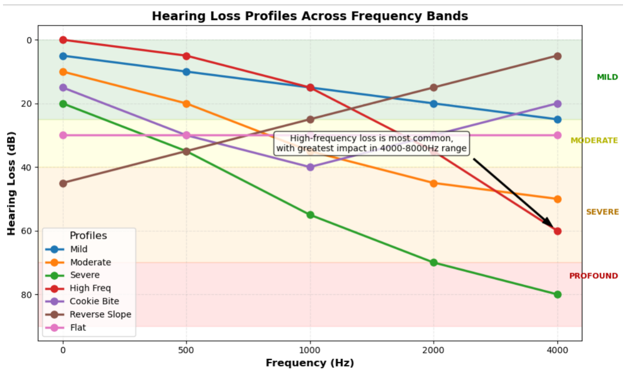


Fig. 7. Hearing loss across various frequencies

VII. Model Training

The CNN-LSTM model was trained for up to 200 epochs. We implemented early stopping to automatically stop training when the model's performance on validation data stopped improving on 10 continuous epochs.. This helped to prevent overfitting of the model. An Adam optimizer, with an initial learning rate of 0.001, is employed with learning rate decay. Thus, as the training progressed, we gradually reduced how much the model adjusted its weights. This helped to fine-tune the model's performance. Our model uses a batch size of 32, and Mean Squared Error(MSE) is used as the primary loss function. We kept track of both training and validation loss, by comparing the 2, we could check if the model was learning useful patterns in the data or just memorizing.

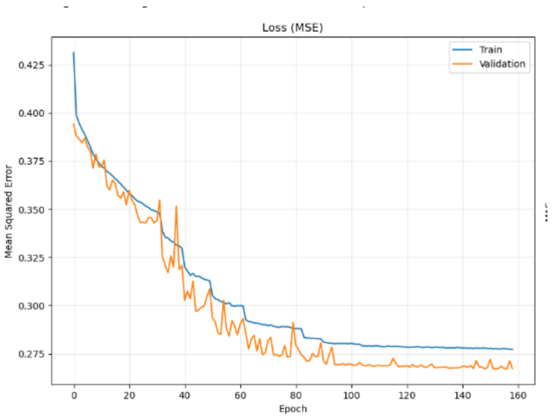


Fig. 8. MSE graph

It also helped us to make sure that the model stayed reliable across various hearing loss profiles, such as mild, moderate, severe, and various noise conditions.

While training, we observed that early stopping was triggered at epoch 159, restoring weights from the best-performing model at epoch 149. While validation performance is monitored at each epoch, we also evaluated the model's performance using MBSTOI(Modified Binural Short-Time Objective Intelligibility) at every 5 epochs. This metric helped us estimate how understandable speech is when heard through both ears. This helped us ensure the model was learning in a way that mattered in real listening scenarios.

VIII. Results

To evaluate how effective the proposed CNN-LSTM model is specifically for binaural speech enhancement, we calculated both how mathematically accurate the model's predictions were and how much those predictions actually helped listeners understand speech better.

The final model achieved a validation loss of 0.2672 and a mean absolute error(MAE) of 0.2236. This indicates stable and consistent learning. We can also observe a consistent frame-level prediction accuracy, meaning the enhancements are applied smoothly for each of the small frames, which is especially important for real-time use of a hearing aid.

To calculate the perceptual improvements, we used the MBSTOI (Modified Binaural Short-Time Objective) metric. The average intelligibility score for noisy audio was 0.389 and we achieved an enhanced output of 0.4071, thus an average improvement of +0.232 or 2.3%. The CNN-LSTM model achieved a 100% improvement rate across all the test samples of 1000 audio.

The model was not trained to improve only 1 type of hearing problem, it performed well and consistently across various types of hearing loss. This shows it can work accurately in different real-world situations. When we measured how much the model improved speech intelligibility using MBSTOI, we found that individuals with high-frequency hearing loss had a +0.0199 improvement while individuals with a moderate hearing loss got a +0.253 improvement.

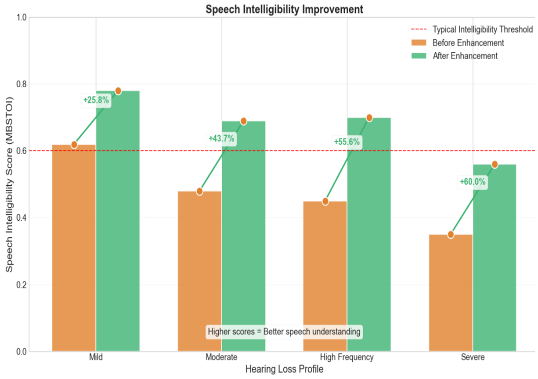


Fig. 9. MBSTOI scores for hearing loss profiles

We also tested the model across various environmental conditions. For clean conditions, we saw the highest improvement of +0.0407, and for noisy conditions, +0.0271, still improving speech clarity. Although our model didn't perform well for reverberant conditions like large halls or tiled rooms where sound echoes. Here we got an improvement of 0.0040, thus leaving scope for model improvement.

Enhancing speech that is already understandable leaves very little room to improve its understandability or intelligibility. Thus, even small increases can reflect meaningful improvements in real-world listening conditions. Although the MBSTOI-based numerical improvements appeared modest, it is a very sensitive metric, and even a small improvement can reflect improvements in how easy speech is to understand for people with hearing loss.

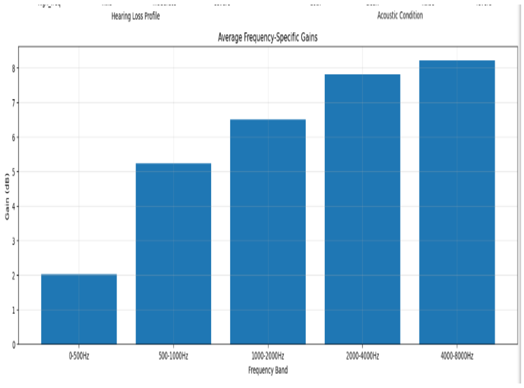


Fig. 10. Speech intelligibility across various frequencies

When we listened to the enhanced audio compared to its noisy version, the difference was clear. The enhanced audio had speech that noticeably sounded clearer and was easier to follow, especially in noisy environments. Also, some of the original files were already fairly understandable, leaving less room for improvement.

The spectrograms below show the improvement in audio made by our model. For the noisy spectrogram, there is significant energy spread across a wide range of frequencies, which shows a lot of background noise. In contrast, the enhanced audio spectrograms show cleaner and more focused patterns. This indicates that our model was successful not only for noise reduction but also for enhancing the clarity and intelligibility of speech.

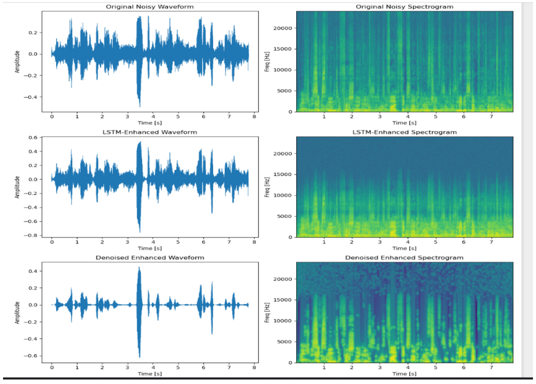


Fig. 11. Spectrogram comparing noisy and enhanced audio

Overall, the CNN-LSTM model has shown consistent and meaningful improvements in not only objective metrics but also in perceptual listening tests. This highlights its potential

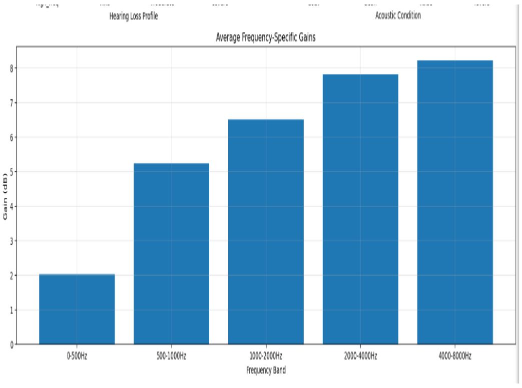


Fig. 12. Speech intelligibility across various frequencies

to enhance speech for people with diverse hearing needs in real-world environments.

IX. FUTURE PROSPECTS

Looking forward, our future work will be focused on optimizing the model for real-time, low power, low latency and power-efficient deployment on edge devices such as hearing aids. This involves exploring possibilities or strategies like model compression, pruning, knowledge distillation and more to significantly reduce memory and computational complexity. While doing so, we will aim not to reduce the perceptual performance of our model.

We would also like to explore expanding our testing beyond a controlled dataset, in real-world environments like public transport, restaurants, movie theaters, outdoor spaces, large reverberant halls, etc.. This is important for validating robustness beyond current test cases. This will be useful in accessing and improving the model's robustness under everyday listening conditions

Further, we also aim to integrate user-specific adaptation by using personalised hearing profiles and user feedback. This dynamic adaptation framework will allow us to fine-tune the model's performance and enhancement strategies based on specific user needs. This will eventually lead to more targeted and effective speech intelligibility improvements. Incorporating a smartphone-based interactive control interface could also provide users with greater customization and a user-centric experience. This type of interactive layer could empower users to personalize their auditory experiences while enabling the system to iteratively improve through adaptive learning. By exploring these future possibilities, we hope to create a scalable, intelligent hearing assistance solution that is capable of delivering personalized, high-quality auditory experiences in complex and dynamic real-world scenarios.

X. CONCLUSION

In this work, we presented a hybrid CNN and Bi-LSTM network for binaural speech enhancement with the application of hearing aid. Preprocessing was done using MATLAB, applying adaptive filtering, spectral subtraction, framing, and windowing to enhance the quality of the signal before training. The CNN layers learned essential spatial features, and Bi-LSTM layers learned long-term temporal dependencies, making the model efficient at dealing with non-stationary noise and dynamic speech patterns.

The model obtained validation loss of 0.2672 and MAE of 0.2236, which signifies stable and reliable performance. Further, MBSTOI metric demonstrated 2.3% improved speech intelligibility on 1000 varying audio samples. Spectrogram comparisons validated improved speech clarity and noise suppression.

Our findings illustrate the promise of deep learning for enhancing real-time hearing aid technology. By combining traditional signal processing with neural networks, this research moves toward intelligent, user-adaptive hearing aids that more closely satisfy real-world auditory needs.

REFERENCES

[1] A. Patyal, "Review of Subjective Intelligibility Comparison and Evaluation of Speech Enhancement Algorithms," *International Journal of Computing and Corporating Research*, vol. 2, no. 1, Jan. 2012.

[2] M. Karam, H. F. Khazaal, H. Aglan, and C. Cole, "Noise Removal in Speech Processing Using Spectral Subtraction," *Journal of Signal and Information Processing*, vol. 5, pp. 32–41, 2012.

[3] A. Mouchtaris, J. Van der Spiegel, P. Mueller, and P. Tsakalides, "A Spectral Conversion Approach to Single-Channel Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, May 2007.

[4] Y. Yeminy, S. Gannot, and Y. Keller, "Speech Enhancement Using a Multidimensional Mixture-Maximum Model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, Nov. 2007.

[5] Durgesh, A. Garg, and P. Bactor, "Speech Enhancement Algorithms: A Brief Review," *International Journal for Advance Research in Engineering and Technology*, vol. 1, no. 5, June 2012.

[6] M. A. Abd El-Fattah, M. I. Dessouky, S. M. Diab, and F. E. Abd El-Samie, "Adaptive Wiener Filtering Approach for Speech Enhancement," *Ubiquitous Computing and Communication Journal*, vol. 3, no. 2, pp. 23–31.

[7] R. V. Mane and M. T. Kolte, "Implementation of Adaptive Filtering Algorithm for Speech Signal on FPGA," *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. 2, no. 3, Mar. 2014.

[8] Bhat GS, Shankar N, Reddy CKA, Panahi IMS. A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone. IEEE Access. 2019;7:78421-78433. doi: 10.1109/access.2019.2922370. Epub 2019 Jun 12. PMID: 32661495; PMCID: PMC7357966.

[9] Mamun N, Khorram S, Hansen JHL. Convolutional Neural Network-based Speech Enhancement for Cochlear Implant Recipients. Interspeech. 2019 Sep;2019:4265-4269. doi: 10.21437/interspeech.2019-1850. PMID: 34307643; PMCID: PMC8296973.

[10] Strake, M., Defraene, B., Fluyt, K. *et al.* Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. *EURASIP J. Adv. Signal Process.* 2020, 49 (2020). https://doi.org/10.1186/s13634-020-00707-1

[11] Fazal E Wahab, Zhongfu Ye, Nasir Saleem, Rizwan Ullah, Compact deep neural networks for real-time speech enhancement on resource-limited devices, Speech Communication, Volume 156, 2024, 103008, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2023.103008.

[12] DFingerNet: Noise-Adaptive Speech Enhancement for Hearing Aids 17 Jan 2025 · Iosif Tsangko, Andreas Triantafyllopoulos, Michael Müller, Hendrik Schröter, Björn W.

[13] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learn- ing with raw-waveform CLDNNs for voice activity detection," in Proc. Interspeech, San Francisco, USA, 2016.

[14] Pierre Guiraud, Alastair H. Moore, Rebecca R. Vos, Patrick A. Naylor, Mike Brookes, "Using a single-channel reference with the MBSTOI binaural intelligibility metric, Speech Communication", Volume 149, 2023, Pages 74-83, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2023.03.005.