Speech Pre Emphasis as a Simple and Inexpensive Method for Speech Enhancement

Ojaswini kapu
Department of Electronics and Communication Engineering,
Institute of Aeronautical Engineering, Hyderabad
Ojaswinik16@gmail.com

Nivedhitha Thota
Department of Electronics and Communication Engineering,
Institute of Aeronautical Engineering, Hyderabad
nivedhithathota28@gmail.com

Ramu Appala
Department of Electronics and Communication Engineering,
Institute of Aeronautical Engineering, Hyderabad
ramuappala2003@gmail.com

Dr. S. China Venkateswarlu
Department of Electronics and Communication Engineering,
Institute of Aeronautical Engineering, Hyderabad
c.venkateswarlu@iare.ac.in

Abstract **-** This project focuses on improving objective metric values of Speech Enhancement using the LibriSpeech ASR corpus dataset, implemented in a Google Colab environment. The LibriSpeech ASR corpous dataset provides a robust foundation for this task, containing 3600 noise signals with different background noise at different noise level. In this study, we preprocess audio data through normalization and feature extraction. Using the Libros library, we compute Mel-frequency cepstral coefficients (MFCCs) and other acoustic features to create a rich dataset suitable for training deep learning models. Our model architecture combines deep neural networks (DNN) and pre-emphasis to effectively capture both spectral and temporal characteristics of the audio signals, allowing for calculating objective metric values for validating speech enhancement Training is conducted on Google Colab using Keras and TensorFlow, where categorical cross-entropy serves as the loss function, and metrics such as accuracy and confusion matrix analysis are used to evaluate model performance. Techniques like dropout are applied to mitigate overfitting. The model achieves substantial accuracy, demonstrating its capacity to enhance speech effectively and calculate the objective metric of dataset This research highlights the difference between the ML model efficiency in speech enhancement with and without pre-emphasis as a preprocessing technique. The results provide basis for future innovations in speech enhancement technologies.

Keywords: Speech Enhancement, LibriSpeech ASR corpus, Google Colab, Mel- frequency cepstral coefficients, deep learning, deep neural networks, pre emphasis, objective metric values

I.INTRODUCTION

Speech signals are naturally characterized by a spectral roll-off, or tilt, due to the glottal excitation stemming from vocal fold vibration. This characteristic concentrates more energy at lower frequencies, which can lead to the inadvertent pre-emphasis of higher- frequency components by speech processing systems. High-frequency elements, however, are crucial for distinguishing certain phonetic details like fricatives, affricates, and plosives, which hold perceptually significant information. As a result, addressing the spectral tilt is essential for achieving more accurate and detailed speech enhancement. Pre-emphasis filtering is a conventional approach that aims to balance the speech spectrum by amplifying high-frequency components, thereby flattening the spectral tilt. This simple yet effective technique has long been integrated into traditional automatic speech recognition (ASR) systems and speech coding applications, as it enables these systems to capture

more detail across the full range of frequencies. However, despite its recognized benefits in earlier speech processing domains, pre-emphasis filtering has seen limited use in modern deep learning-based speech enhancement models, warranting further exploration. Recent studies, such as those by López-Espejo et al. and the developers of SEAMNET, have considered pre-emphasis in speech enhancement but have found limited evidence for its impact when incorporated into certain loss functions. López-Espejo et al. concluded that pre-emphasis had no marked advantage when integrated with the scale-invariant signal-to distortion ratio (SI-SDR) loss function. Likewise, SEAMNET introduced pre-emphasis into the mean squared error (MSE) loss function, but its benefits were left unexplored in comparison to a non-pre-

emphasized equivalent. These findings suggest that while pre emphasis holds potential, its impact in deep learning- based frameworks remains uncertain and is a promising area for investigation. In this project, we explore the effectiveness of pre-emphasis filtering in a deep neural network (DNN)-based speech enhancement framework trained on the LibriSpeech ASR corpus. Using the Adam optimizer and the MSE loss function, we evaluate the improvements in key objective metrics for enhanced speech quality, with a focus on comparing results with and without pre-emphasis integration. This approach allows us to determine whether pre-emphasis contributes to measurable gains in speech enhancement quality, especially for speech corrupted by various noise types. Our study demonstrates that the inclusion of pre-emphasis filtering within the MSE loss function can yield improvements in objective metrics. Such findings are valuable for applications in fields like telecommunication, assistive technology, and voice-controlled systems, where noise-robust and intelligible speech is critical.

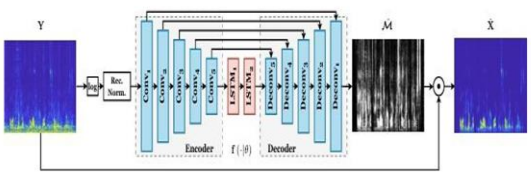

Fig 1.Block diagram of existing model

II.DATASET

For this project, I utilized the LibriSpeech ASR corpus, a widely used and reputable dataset for automatic speech recognition and audio processing tasks. The LibriSpeech corpus offers high-quality, clean speech data, making it ideal for training and evaluating speech enhancement models. Clean Signals: I selected 30 clean speech signals from the LibriSpeech ASR corpus. These signals were carefully curated to provide a range of vocal characteristics and content variety, ensuring robustness in the model's training and evaluation processes. Noise Signals: To simulate realistic noisy environments, I incorporated four different types of background noise:

1. Car noise
2. Babble noise
3. Station noise
4. Airport noise

For each noise type, I generated 30 samples across multiple Signal-to-Noise Ratios (SNRs) to simulate varying levels of noise intensity. Specifically, I created datasets with SNRs of 0 dB, 5 dB, 10 dB, and 15 dB. The inclusion of these different SNR levels enables the model to handle a range of noise conditions, from intense background noise (0 dB) to relatively mild interference (15 dB).

Storage and Organization: To facilitate seamless access and integration with my model, I organized the dataset in Google Drive with a clear and logical folder structure with subfolders categorized by noise type and SNR level. This systematic organization in Google Drive allows for efficient loading and processing of data in Google Colab, streamlining the model development and testing workflow. The structure also makes it easy to reference specific subsets of data, such as a particular noise type or SNR level, improving flexibility in experimental setups and model evaluation. By combining clean signals from LibriSpeech with various noise types at multiple SNR levels, I created a comprehensive dataset that effectively represents the challenges of real world noise, enabling robust training and accurate performance assessment of the speech enhancement model.

III.METHODOLOGY

3.1 Collection of Dataset

For this project, I utilized the LibriSpeech ASR corpus, a widely used and reputable dataset for automatic speech recognition and audio processing tasks. The LibriSpeech corpus offers high-quality, clean speech data, making it ideal for training and evaluating speech enhancement models.

3.2 Reshaping Dataset

To prepare the dataset for input into the deep neural network (DNN) model, the raw audio signals were reshaped to match the model's input specifications. The audio files, sampled at a rate of 16,000 Hz, were processed to ensure uniformity across all samples. Each audio file was reshaped into a fixed-size sequence, with each segment containing 320 samples, representing a temporal frame, and a sample rate of 16,000 Hz. Given that the sampling rate (SR) of 16,000 Hz is consistent across the dataset, the reshaped data reflects a temporal window of 320 samples per input frame. Furthermore, the reshaped data contained a total of 22,050-time steps, allowing for sufficient resolution to capture the speech signal's characteristics over time. The reshaped dataset was organized into a 3D array of shape (Num_samples, 320, 22050), where Num_samples denotes the total number of audio samples, 320 corresponds to the number of frames per sample, and 22,050 represents the number of time steps in each frame. This reshaping ensured the data was suitable for the DNN model's input, facilitating efficient training and inference.

Data Normalization: The mixed (noisy) and clean signals were normalized to ensure that both the input (mixed signals) and output (clean signals) data fell within the range of -1 to 1. This was achieved by dividing each signal by its maximum absolute value, standardizing the data for the model.

3.3 Training DNN model

For training the deep neural network (DNN) model, 80% of the dataset was used for training, and 20% was reserved for validation. The training process was carried out using the following steps:

1.Training Process: The model was trained for 100 epochs, meaning the entire dataset was passed through the network 100 times. This allowed the model to learn from the data over an extended period, with each epoch consisting of several iterations based on the batch size. The batch size was set to 32, meaning that during each iteration, the model used 32 samples at once to update the weights. This batch size strikes a balance between computational efficiency and model performance. **Validation split**: During training, 20% of the data was set aside for validation, ensuring that the model's performance could be evaluated on unseen data and preventing overfitting. 17 This means that for each epoch, 80% of the data was used for training, while 20% was used to evaluate the model's performance on data it hadn't seen before.

2.Model Evaluation: During training, the model's loss was monitored at the end of each epoch. The validation loss was also tracked to ensure the model was not overfitting to the training data and generalizing well to unseen data. The training process continued for 100 epochs, with the training and validation loss being printed at the end of each epoch to observe the learning progress.

3.Model saving: After the training process was completed, the model was saved using Keras built-in save function, allowing it to be used for future predictions or fine-tuning without the need to retrain from scratch. The model was saved as an H5 file, which can be easily loaded for inference on new, unseen noisy data.

3.4 Model Compilation

After defining the architecture of the deep neural network (DNN), the next step was to compile the model by specifying the optimizer, loss function, and metrics used during training. The compilation process is critical as it determines how the model's parameters will be updated during the training phase.

1.Optimizer: The Adam optimizer was used to optimize the model during training. Adam is an adaptive optimizer that computes individual learning rates for each parameter by considering both the first-order momentum (gradient) and second-order acceleration (variance) of the gradients. This makes

Adam highly effective for training deep learning models, especially for tasks such as speech enhancement, where the data is noisy and requires stable convergence. Adam has been widely adopted due to its efficiency and ability to handle sparse gradients. For this model, the default learning rate of Adam was utilized, which typically offers good results without the need for manual tuning.

2.Loss Function: The Mean Squared Error (MSE) loss function was selected as it is a well-suited metric for regression tasks. In the context of speech enhancement, the goal is to minimize the difference between the predicted clean signal and the actual clean signal. MSE computes the average of the squared differences between the predicted values and the ground truth values. The objective is to reduce this error, leading to better signal reconstruction.

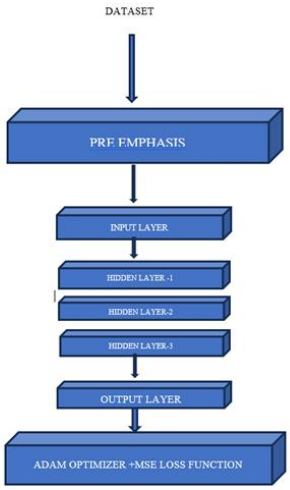

Fig 3.4. Block diagram of proposed model

3.5 Testing the model for Speech Enhancement

Following the training phase, the model was tested using a separate dataset to evaluate its performance in enhancing noisy speech. The testing was conducted in two scenarios: with pre-emphasis and without pre-emphasis applied to the input signals. Additionally, performance was measured at various Signal-to-Noise Ratio (SNR) levels (0, 5, 10, and 15 dB) to observe how the model performed across different noise conditions.

1. Testing Without Pre-Emphasis: In this test, the noisy speech signals were fed directly into the trained DNN without any pre-emphasis filtering. The model was tasked with reconstructing the clean signal from the noisy input. For each test case, the noisy signals were normalized to match the input range expected by the model (as done during training). Quality (PSEQ) and Short-Time Objective Intelligibility (STOI) scores were computed. These metrics provided insight into both the subjective quality and intelligibility of the enhanced speech at different SNR levels (0, 5, 10, and 15 dB). o Additionally, the Mean Squared Error (MSE) and Signal-to-Noise Ratio (SNR) were calculated to further assess the accuracy and enhancement quality of the speech signals.

2.Testing With Pre-Emphasis: In the second evaluation, pre-emphasis filtering was applied to the noisy speech signals before passing them into the trained DNN. Pre-emphasis is a signal enhancement technique that boosts the high-frequency components of the speech signal, which can aid in recovering fine details that might be obscured by no A first-order high-pass filter with a pre-emphasis coefficient of 0.97 was applied to the noisy signals to emphasize the higher frequencies. The filtered signals were then normalized and passed through the trained model for enhancement. As with the first test, the output was compared to the clean target signals.

3.Comparation of Result: The results from both testing scenarios (with and without pre-emphasis) were compared to determine the effect of pre-emphasis filtering on the performance of the speech enhancement model. The PSEQ values helped assess the perceptual quality of the enhanced speech, while the STOI scores provided insight into the intelligibility of the enhanced speech at each SNR level. A higher STOI score indicates better intelligibility. The SNR and MSE metrics were used to evaluate the quantitative performance in terms of noise reduction and signal fidelity. By testing the model at varying SNR levels (0, 5, 10, and 15 dB), we could observe how robust the enhancement process was to different levels of noise and how pre-emphasis influenced the model's ability to recover clean speech.

IV.RESULTS

4.1 Model loss: The plot below shows the training and validation losses over 100 epochs.

- Training Loss: The training loss consistently decreases throughout the training process, indicating that the model is effectively learning to enhance speech signals. The sharp decline in the initial epochs suggests that the model quickly adapts to the training data, and the loss begins to stabilize as the model approaches convergence, signalling diminishing improvements with more epochs.
- Validation Loss: The validation loss decreases initially but starts to increase after a certain point, which points to a phenomenon commonly referred to as overfitting. Although the model performs well on the training data, it begins to struggle to generalize to unseen data, as evidenced by the rising validation loss. This pattern is typical of deep learning models, where overfitting can occur as the model starts memorizing the training data rather than learning generalizable patterns. The increase in validation loss indicates the need for regularization or other techniques to improve the model's generalization ability

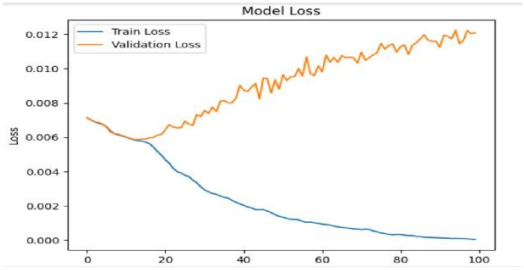

Fig 4.1 Graph of model loss during training

4.2 Model performance: Throughout the 100 epochs of training, the model demonstrated significant improvements in its ability to minimize the training loss. Starting with a relatively higher loss, the training loss decreased steadily and consistently, reaching a very low value by the end of the training process. This decline in training loss indicates that the model successfully learned from the data and was able to improve its performance across the epochs. In contrast, the validation loss showed a more stable pattern. While the training loss continued to decrease, the validation loss fluctuated within a narrow range, showing some minimal changes towards the later epochs. This suggests that the model's performance on the validation set did not improve as dramatically as on the training set, indicating that the model might have reached its generalization limit, or there may be some degree of overfitting. Overall, while the model's ability to fit the training data improved substantially, the validation performance remained relatively unchanged, signifying a potential plateau in generalization after a certain point in training.

```
Epoch 1/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 7s 574ms/step - loss: 0.0071 - val_loss: 0.0071
Epoch 2/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 3s 340ms/step - loss: 0.0069 - val_loss: 0.0070
Epoch 3/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 3s 332ms/step - loss: 0.0069 - val_loss: 0.0069
Epoch 4/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 3s 329ms/step - loss: 0.0068 - val_loss: 0.0069
Epoch 5/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 6s 458ms/step - loss: 0.0068 - val_loss: 0.0068
Epoch 6/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 4s 358ms/step - loss: 0.0067 - val_loss: 0.0066
Epoch 7/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 5s 311ms/step - loss: 0.0064 - val_loss: 0.0063
Epoch 8/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 3s 370ms/step - loss: 0.0063 - val_loss: 0.0062
Epoch 9/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 5s 326ms/step - loss: 0.0061 - val_loss: 0.0062
Epoch 10/100
8/8 ━━━━━━━━━━━━━━━━━━━━ 5s 327ms/step - loss: 0.0061 - val_loss: 0.0061
```

Fig 4.2 First 10/100 epoch of DNN model training

4.3 Spectrogram of Enhanced signals: The spectrogram presented below compare the frequency-time representation of a noisy input signal (left) and its enhanced version (right). In the noisy spectrogram, high-intensity bands are observed across various frequency ranges, indicating the presence of significant background noise. This interference can obscure the speech content, particularly in lower frequency bands, which are typically associated with intelligible speech components. In contrast, the enhanced signal's spectrogram shows a marked reduction in noise across the frequency spectrum. The enhanced spectrogram highlights more distinct speech-related patterns with reduced background interference, particularly noticeable in the lower and mid-frequency ranges. This improvement suggests that the enhancement process successfully mitigates noise, making the speech components more prominent and clearer.Overall, the visual comparison indicates an effective reduction in noise levels, supporting the enhancement model's ability to

improve speech intelligibility while preserving essential speech characteristics.

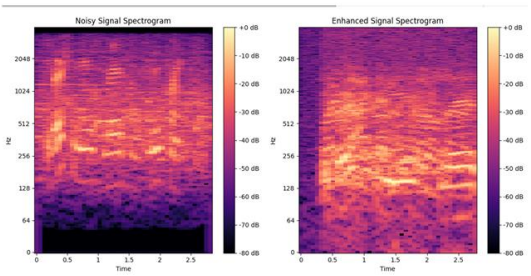

Fig 4.3 Spectrogram of noised and enhanced signal

4.4 Objective metric values :

s.no	Noise Level	Pre-emphasis as preprocessing step	PSEQ	STOI
1	0 dB	✓	1.234	0.253
		✗	1.165	0.253
2	10 dB	✓	1.276	0.248
		✗	1.200	0.248
3	15 dB	✓	1.333	0.248
		✗	1.223	0.251
4	20 dB	✓	1.440	0.246
		✗	1.195	0.251

Fig 4.4 Table comparison of objective metric values with and without pre emphasis

Noise Level 0 dB:
PSEQ: With pre-emphasis, a score of 1.234 was achieved, whereas without pre-emphasis, the score slightly decreased to 1.165. This indicates that pre-emphasis marginally improves perceptual quality at this noise level.
STOI: Both cases yield the same STOI value of 0.253, showing that intelligibility is unaffected by the pre-emphasis step at this noise level.
Noise Level 10 dB:
PSEQ: Pre-emphasis resulted in a higher PSEQ of 1.276 compared to 1.200 without it, suggesting that pre-emphasis enhances perceptual quality under moderate noise.
STOI: The intelligibility score remains constant at 0.248 with or without pre-emphasis, implying that pre-emphasis has a limited impact on intelligibility for this noise level.
Noise Level 15 dB:
PSEQ: The PSEQ score increased to 1.333 with pre-emphasis, compared to 1.223 without it, again demonstrating that pre-emphasis improves the perceptual quality of the enhanced speech.
STOI: A small increase in STOI score is noted without pre-emphasis (0.251) compared to with pre-emphasis (0.248), suggesting a negligible impact on intelligibility.
Noise Level 20 dB:
PSEQ: A noticeable improvement in PSEQ is observed with pre-emphasis (1.440)compared to without pre-emphasis (1.195). This indicates that pre-emphasis is beneficial for perceptual quality at higher noise levels.
STOI: Without pre-emphasis, the STOI score is slightly higher at 0.251 than with pre emphasis at 0.246, showing a minimal impact on intelligibility.

4.5 Graphical Representation of metric values

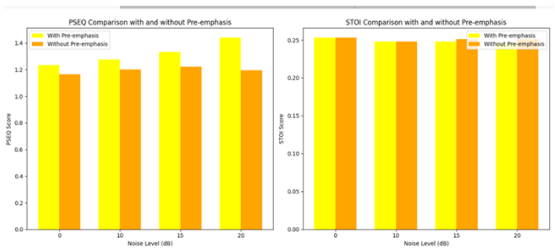

Fig 4.5 Graphical representation of pseq, stoi values

From the above graphs, it is evident that incorporating pre-emphasis as a preprocessing step leads to noticeable improvements in the Perceptual Evaluation of Speech Quality (PSEQ) values across different noise levels. The PSEQ scores with pre-emphasis consistently outperform those without, indicating that pre-emphasis effectively enhances the quality of speech as perceived objectively. However, in terms of the Short-Time Objective Intelligibility (STOI) metric, the results do not show significant differences between the scores with and without pre-emphasis. This suggests that while pre-emphasis improves the perceptual quality of speech, it does not significantly impact intelligibility.

V.CONCLUSION

In this project, we developed and evaluated a deep neural network (DNN) model for speech enhancement, focusing on improving perceptual quality and intelligibility in noisy environments. By employing the LibriSpeech ASR corpus, a comprehensive and diverse dataset, we ensured that the model was trained on a wide range of speech patterns and noise scenarios. The Adam optimizer was used to facilitate efficient training, and the Mean Squared Error (MSE) loss function served as a robust criterion for minimizing reconstruction errors in the enhanced speech output. Our findings demonstrate that pre-emphasis as a preprocessing step significantly improved the Perceptual Evaluation of Speech Quality (PSEQ) values across various noise levels, confirming its effectiveness in enhancing the subjective quality of speech signals. However, this approach had limited impact on the Short-Time Objective Intelligibility (STOI) scores, indicating that while the DNN model with pre-emphasis excels at improving audio quality, its contributions to intelligibility remain modest. Overall, this project highlights the potential of deep learning-based speech enhancement, particularly when augmented with preprocessing techniques like pre-emphasis. Future work may explore advanced architectures, alternative preprocessing steps, or other training objectives to further improve both speech quality and intelligibility. By continuing to refine these methods, we can move closer to achieving clearer and more understandable speech signals in challenging acoustic conditions. Beyond research, the findings from this project have practical implications for real-world applications, such as hearing aids, communication devices, and automated speech recognition systems in noisy environments. The insights gained here suggest that carefully designed preprocessing steps, combined with optimized DNN architectures, could significantly enhance user experience and accessibility in challenging auditory conditions.

VI.REFERENCE

[1] Emma Jokinen and Paavo Alku, “Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network,” The Journal of the Acoustical Society of America, vol. 141, 2017. [2] Tom B¨ackstr¨ om, J´er´ emie Lecomte, Guillaume Fuchs, Sascha Disch, and Christian Uhle, Speech coding: with code-excited linear prediction, Springer, 2017.

[3] Raymond D. Kent and Charles Read, The Acoustic Analysis of Speech, Singular/Thomson Learning, 2002.

[4] Brian B. Monson, Andrew J. Lotto, and Brad H. Story, “Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives,” The Journal of the Acoustical Society of America, vol. 132, pp. 1754–1764, 2012. [5] Marco K¨uhne, Roberto Togneri, and Sven Nordholm, “A new evidence model for missing data speech recognition with applications in reverber ant multi-source environments,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, pp. 372–384, 2011.

[6] Iv´ an L´opez-Espejo, Amin Edraki, Wai-Yip Chan, Zheng-Hua Tan, and Jesper Jensen, “On the deficiency of intelligibility metrics as proxies for subjective intelligibility,” Speech Communication, vol. 150, pp. 9–22, 2023.

[7] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, “SDR- Half baked or well done?,” in Proceedings of ICASSP 2019 44th IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, Brighton, UK, 2019, pp. 626–630.

[8] Bengt J. Borgstr¨om and Michael S. Brandstein, “Speech Enhancement via Attention Masking Network (SEAMNET): An End-to-End System for Joint Suppression of Noise and Reverberation,” IEEE/ACM Transac tions on Audio, Speech, and Language Processing, vol. 29, pp. 515–526, 2020.

[9] Hynek Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” Journal of the Acoustical Society of America, vol. 87, pp. 1738–1752, 1990.

[10] Juan Manuel Mart´ ın Do˜ nas, Online multichannel speech enhancement combining statistical signal processing and deep neural networks, Ph.D. thesis, University of Granada, 2020.