Noise reduction and accuracy improver for a Speech Recognition machine learning model

Aman Hogan-Bailey, Divya Tejasbha Shah, Karan Bhavin Shah

CSE, The University of Texas at Arlington

Contents

Introduction

In the realm of audio processing, the clarity of sound is paramount for a multitude of applications ranging from voice-command systems and hearing aids to the broader scopes of telecommunications and automatic speech recognition. As we venture deeper into an era where technology becomes increasingly voice-activated, the demand for pristine audio quality is more pressing than ever. This has spurred the development of advanced noise reduction techniques aimed at distilling clarity in audio data. Our *Audio Noise Reduction Model*, the subject of this report, is a machine learning construct designed to transform noisy audio inputs into clarified outputs. The model was honed utilizing an open-source dataset taken from *Columbia University*, which served as the training ground for various machine learning paradigms explored in this project. The challenge at hand was to engineer a solution adept at diminishing background noise within audio files, thereby achieving the highest fidelity in the resultant output. This undertaking was not without its hurdles; the team grappled with constraints on computation, data storage, and temporal resources. This report delineates the journey of the *Audio Noise Reduction Model* from conception to implementation. It documents the comparative analysis of different machine learning models, the selection criteria for the most effective methodology, and the process of transforming noisy data into clear audio.

Related Works

Noise suppression in audio processing is a well-trodden path with substantial advancements made over decades. Traditional techniques have leveraged *spectral estimation* methods, heavily reliant on noise spectral estimators and voice activity detectors. These systems, while effective, often require extensive manual tuning. A significant shift has been observed with the introduction of deep learning techniques into the realm of noise suppression. Notable among these is the hybrid DSP/deep learning approach to real-time full-band speech enhancement developed by Jean-Marc Valin at *Mozilla Corporation*. Valin's work exemplifies the integration of traditional digital signal processing (DSP) techniques with modern deep learning models, particularly recurrent neural networks (RNNs), to enhance speech quality while maintaining low computational complexity suitable for real-time applications (Valin, Mozilla). This approach is especially relevant to our project as it mirrors our methodological pivot—balancing classic filtering techniques with the computational power of neural networks. Valin's model utilizes a deep RNN to estimate ideal critical band gains, complemented by a traditional pitch filter to attenuate noise between pitch harmonics. This methodology aligns with our project's objective to employ a synergy of machine learning models and traditional. Moreover, many deep learning-based approaches in recent studies target automatic speech recognition (ASR) systems and are not optimized for low-latency applications. Our project extends this concept by not only focusing on low-latency solutions but also ensuring that our models are feasible for operation without the need for high-powered GPU resources. In contrasting our work with Valin's, while both approaches aim for high-quality speech enhancement with low computational costs, our project explores various machine learning techniques including CNNs and RNNs in a comparative framework to determine the most efficient model in terms of both performance and efficiency.

Methodology

Traditionally, methods such as Wiener Filtering and Spectral Subtraction have been employed to 'subtract' unwanted noise from audio signals, particularly in real-time scenarios. However, these techniques come with their own set of limitations, such as the assumption of stationary noise and the potential introduction of speech distortions, not to mention the requirement for dual audio inputs - one clean and one contaminated by noise. In contrast, harnessing the prowess of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) presents a different avenue for exploration. These networks are trained on extensive datasets to unearth and learn the intricate, non-linear correlations between noisy inputs and their clean counterparts, typically used in post-processing applications. The downside is their insatiable appetite for vast amounts of labeled training data and their significant computational expense. Our proposed methodology navigates a middle path between these two extremes. By amalgamating the traditional filtering approach with modern machine learning techniques, we strive for an equilibrium that optimizes both accuracy and speed. (*Figure 1*). This section outlines the systematic approach adopted in the development of our Audio Noise Reduction Model, detailing the procedures from data preparation to post-processing evaluation.

Dataset Description: The project utilized an open-source dataset provided by Columbia University, comprising 30 minutes of varied audio recordings. These recordings include diverse acoustic scenarios and background noises. The dataset was partitioned into separate sets for training and validation purposes, ensuring a representative distribution of noise conditions across both.

Preprocessing Steps: Prior to feeding the audio data into our models, a crucial preprocessing step was padding. This was executed to standardize the length of all audio samples, ensuring consistency when batch processing during the training phase.

Wiener Filtering Technique: As an initial step in our noise reduction pipeline, Wiener filtering was applied to an audio sample. This statistical approach aimed to estimate the desired signal by minimizing the mean square error between the estimated and true signals, thereby reducing noise that is assumed to be additive and stationary.

Machine Learning Model Architecture: Following the Wiener filter pre-processing, two distinct architectures were employed for noise reduction:

1. *Convolutional Neural Network (CNN)*: The CNN architecture was designed with 31D convolution layers, each consisting of a ReLU activation and batch normalization, followed by a max-pooling layer to downsample the signal. The network concluded with a global average pooling layer to reduce the feature maps to a single vector, which was then passed through dense layers culminating in a linear output layer to predict the clean audio signal. Additionally, we used a batch size of six and trained over 100 epochs.
2. *Recurrent Neural Network (RNN)*: The RNN utilized three GRU layers of a hidden size of 128, capable of capturing temporal dependencies within the audio signal. Bidirectional layers were included to process the data in both forward and reverse temporal order. The RNN also featured a linear output layer similar to the CNN, providing the final clean audio prediction. Additionally, we used a batch size of six and trained over 50 epochs.

3. *Non-Neural Nets*: We used Linear regression, KNN regression, and Random Forrest Bagging Regression utilizing various hyperparameters and analyzed them using mse, rmse, and r2 score.

Training Process: The models were trained using the padded audio samples, with the mean squared error (MSE) serving as the loss function. Optimization was performed using the Adam optimizer. Throughout the training, techniques such as dropout were implemented to combat overfitting, thus enhancing the models' generalization capabilities.

Evaluation Criteria: Post-training, the models' performance was evaluated using metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2.

Post-Processing Evaluation: After the noise reduction process, further evaluation was conducted to assess the qualitative improvement in audio clarity. This involved subjective listening tests and comparisons between the original noisy audio, the Wiener filter output, and the machine learning-enhanced output.

Results

Performance Metrics: The evaluation metrics used was measured in Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination, R^2. The comparative analyses between training and testing datasets are presented in an illustrative table (*Figures 2-4*).

KNN, RFB, LRG Evaluations: KNN's performance, with a training MSE of 9 and an RMSE of 3.1, dipped slightly on the testing grounds, indicating commendable generalization. RFB showcased a marginal improvement from training to testing, heralding its robustness. LRG, on the other hand, trailed with higher MSEs and a negative R^2, casting shadows of overfitting or underperformance (*Figures 2-4*).

CNN and RNN Evaluation: CNN unfolded linearly, holding steady around an MSE of 10. RNN's had an oscillating graph between 5 and 20. These fluctuations are captured in a graph (*Figure 5*), illustrating the contrasting stabilities of the two models across 35 epochs.

The KNN Elbow Plot: The optimal number of neighbors was charted on an elbow plot (*Figure 6*), peaking at the triad - three neighbors - beyond which the returns diminished.

Audio Quality Insights: When it came to the qualitative realm of audio clarity, KNN and CNN stole the limelight, outshining their counterparts. LRG lagged, RFB and RNN settled in the median.

Conclusions

We aimed to transform noisy audio inputs into clear, comprehensible outputs. Utilizing a diverse array of machine learning algorithms, including KNN, RFB, LRG, CNN, and RNN, the project undertook a rigorous comparative analysis to determine the most effective approach in mitigating background noise while ensuring audio clarity. Our findings demonstrated that KNN and CNN delivered superior audio quality, indicating their robustness in noise reduction capabilities. However, KNN showed potential signs of overfitting, suggesting a fine balance must be maintained between model complexity and generalization. CNN, with its consistent performance over 35 epochs, emerged as a promising model, adept at capturing the non-linear intricacies of audio signals. Conversely, LRG underperformed in comparison, likely due to its linear nature, which limits its ability to handle the complex patterns present in audio data. The elbow plot for KNN highlighted an optimal performance at three neighbors, underpinning the importance of model tuning in achieving the best results. Meanwhile, RNN's fluctuating performance indicated challenges in model stability, which could be a focal point for further refinement. The RFB algorithm offered a middle-ground performance, though it did not reach the heights of KNN or CNN. In conclusion, our explorations into audio noise reduction have underscored the need for continual model evaluation and enhancement. Future work could delve into more advanced neural architectures, such as attention-based models, and explore the utilization of larger datasets to further improve the performance of the system. The project's outcomes contribute to the ever-evolving landscape of audio processing, providing a foundation upon which more refined and nuanced models can be built.

References

1. J.-M. Valin, A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement, International Workshop on Multimedia Signal Processing, 2018.

2. *Real-time noise suppression using Deep Learning*. NVIDIA Technical Blog. (2022, August 21). https://developer.nvidia.com/blog/nvidia-real-time-noise-suppression-deep-learning/

3. Ellis, D. (n.d.). Sound examples. https://www.ee.columbia.edu/~dpwe/sounds/

Figures

Figure 1

Project Flow Diagram

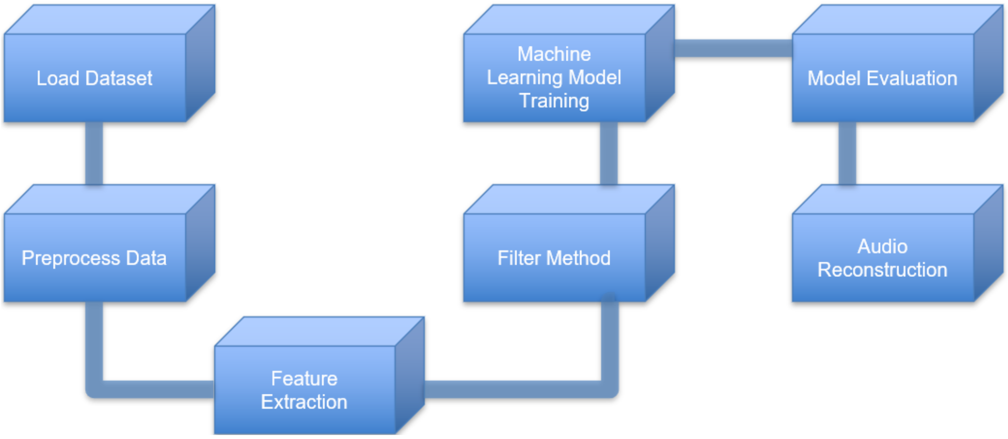


Figure 2

MSE for KNN, RFB, and LRG

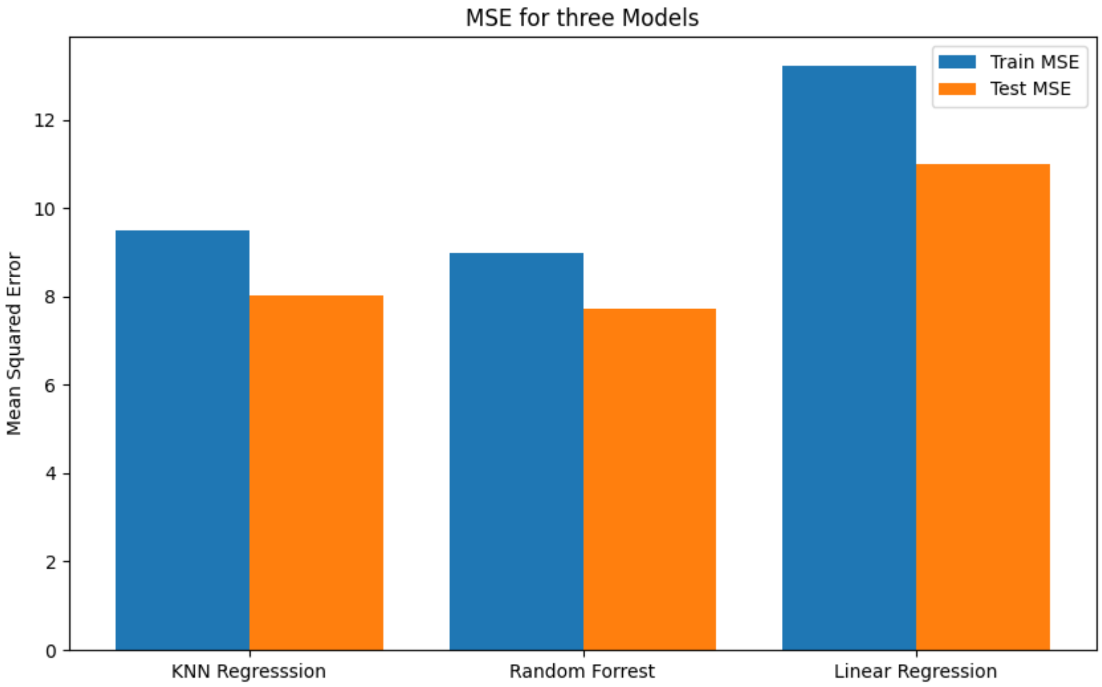


Figure 3

RMSE for KNN, RFB, and LRG

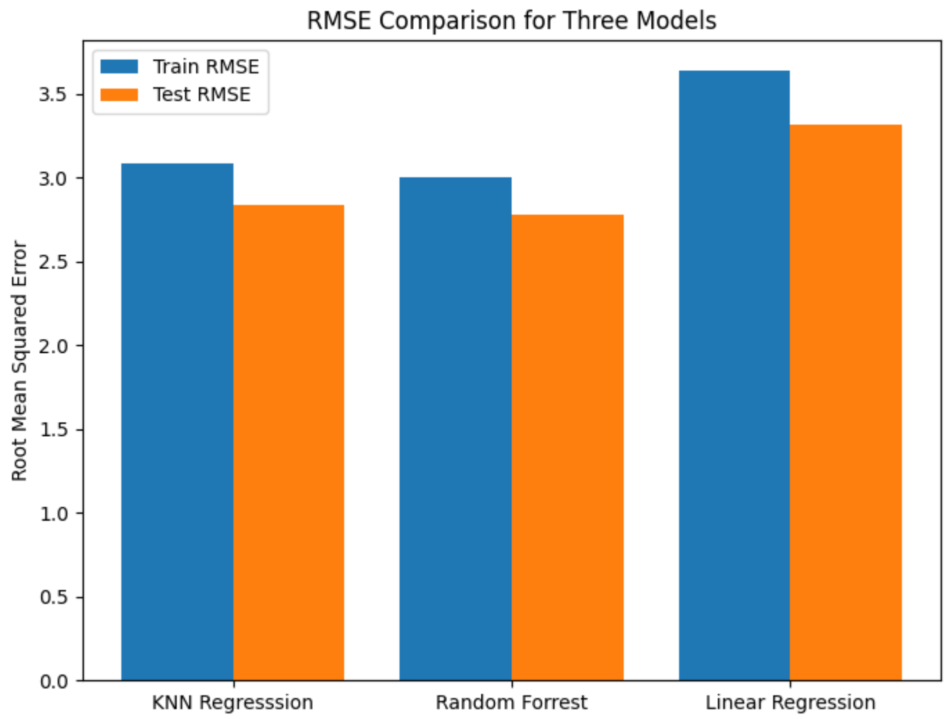


Figure 4

R2 Score for KNN, RFB, and LRG

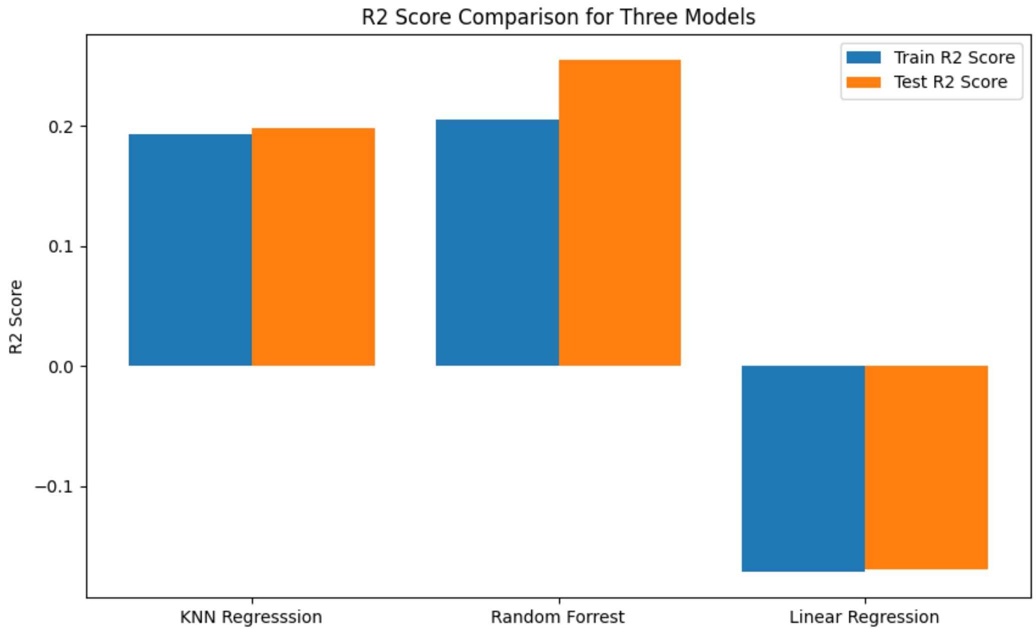


Figure 5

MSE Loss for CNN and KNN over 35 Epochs

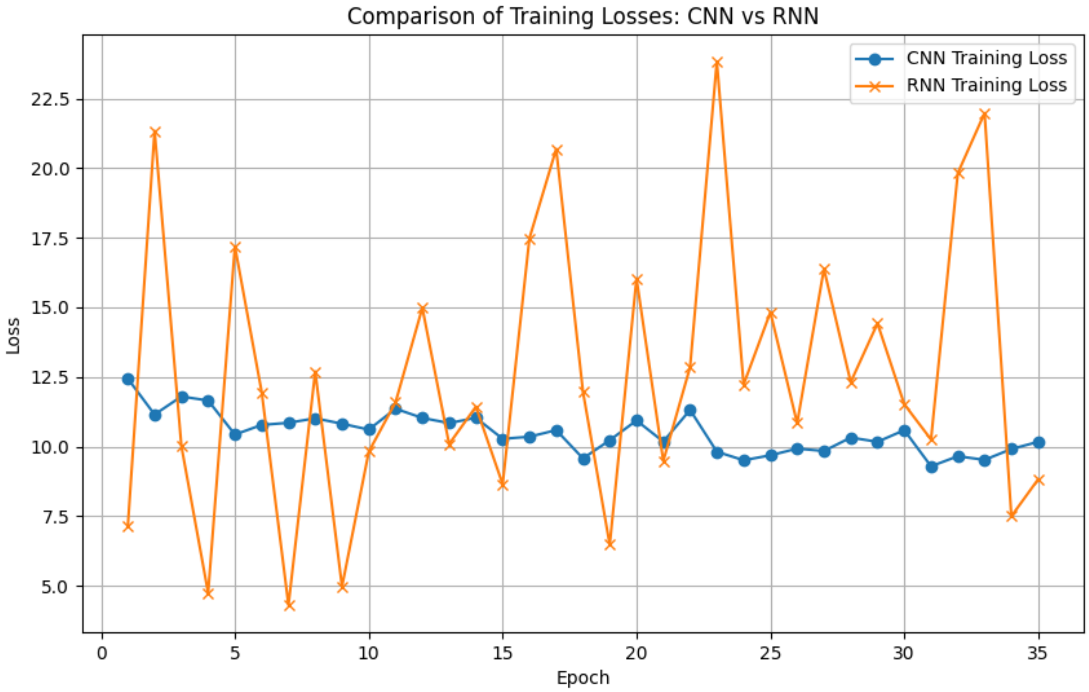
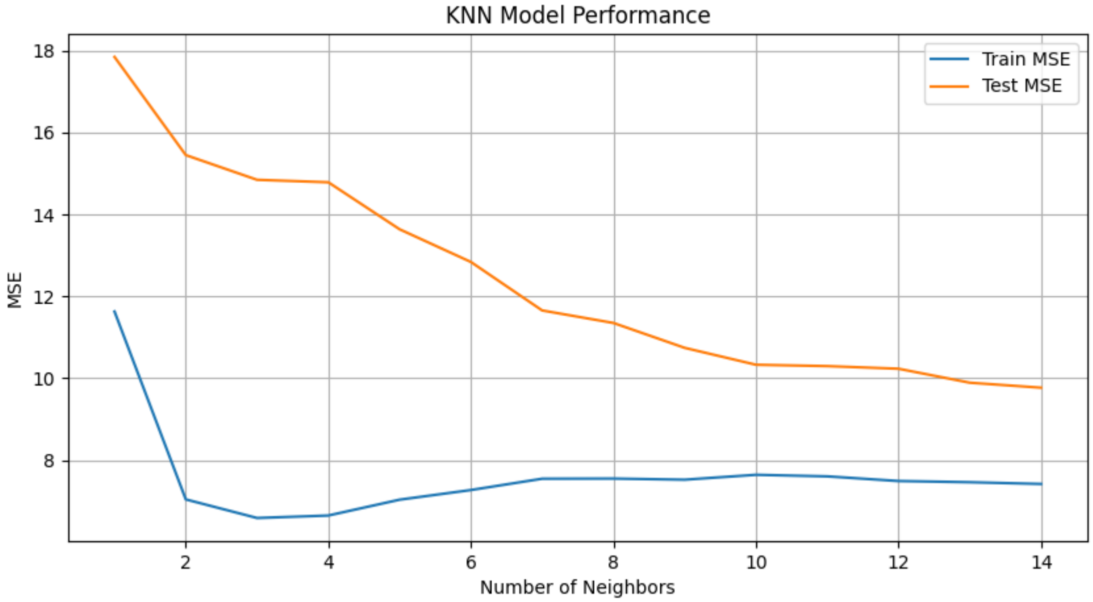


Figure 6

Elbow Plot for KNN


KNN Model Performance
Train MSE
Test MSE
MSE
Number of Neighbors