Neural Network-Based Approaches for Transient Noise Reduction: An Empirical Evaluation

Muqaddaspreet Singh Bhatia[1], Ali Ayub[2]
Department of Computer Science and Software Engineering,
Concordia University, Montreal, Canada
muqaddaspreetsb@gmail.com[1], ali.ayub@concordia.ca[2]

Abstract

The aim of this work is to investigate the possibility of applying neural networks to suppress transient noise in speech signal which is an important feature in enhancing speech understanding and beneficial to hearing applications such as hearing aids. The CSTR VCTK dataset is used which contains different accents and transient noise portions. The work considers three kinds of neural networks: Multi-layer Feed-Forward Networks (FNNs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) which uses LSTM (Long Short Term Memory) networks. The evaluation of performance is carried out over noise conditions and model configurations making use of Signal-to-Noise Ratio (SNR), Mean Squared Error (MSE), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI) for modeling techniques. Some trade-off between performance and architectural efficiency is also offered in comparative study of the architectures. The study finishes with thorough insight into existing works on this area of research on hearing aids where neural networks have plenty of opportunities. The RNN structure appears to effectively address sequential information therefore, transients noise should be largely eliminated using RNN techniques. This is achieved by having low MSE levels, with a maximum of up to 0.0051, which is promising compared to how well the audio was reconstructed. Since transient in nature are undetermined, they can tropically be expressed through the sequential structure of RNNs thus RNNs provide satisfactory accuracy.

Keywords: Noise Reduction, Transient Noise, Neural Networks, Hearing Aids, Signal Processing

1 Introduction

1.1 Problem Statement

The sudden and rapid sounds called transient noises has become a significant concern in the realm of speech intelligibility and listener comfort. They include noises such as door slams, coughs, clicks, drops, and explosions [5], which can have a distinctive negative impact on spoken material, especially for a hearing aid user [5]. Automatic Gain Control (AGC) and other methods for noise suppression face difficulties in dealing with the unpredictable nature of transient noises, leading to poor noise suppression and even degradation of speech signals.
This problem can be divided into a number of problems:

Real-time Processing: Because a transient noise is so abrupt, there is a need for noise reduction strategies suited for real time ai applications such as hearing aids; in this case, instantaneous noise reduction strategies are used [6].

Preprint. Under review.

Generalizability: There is a need for improvement on the generalization of noise reduction algorithms since they need to be applicable on wide varieties of transient noise and different speech variations including accents and speakers [5] [3].

Preservation of Speech Quality: In the effort of eliminating transient noise, the primary objective such as preserving the quality, naturalness or authenticity of the speech signal has to be adhered to as much as possible without introducing artifacts or distortions [6] [8].

1.2 Motivation

Transient noise remains a significant challenge for the enhancement of voice communication as it is a unique interference which cannot be completely eliminated in a typical environment. It has also been established that deep learning architectures, notably Long Short-Term Memory (LSTM) networks, have been effective in reducing noise from complementary sources such as wind and babble noise. [5] [2].

This study focuses on Multi-layer Feed-Forward Networks (FNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) to decrease the transient noises that distort hearing [5]. It compares neural network architectures to identify the most efficient option for real-time applications, balancing performance and computational cost.

1.2.1 Research Questions and Hypotheses

This work presents a study into neural networks for the reduction of transient speech noise for the assistive technology area of hearing aids. The analysis is driven by the following research questions:

- **RQ1:** How effective are neural networks in reducing transient noise or interference in speech while maintaining speech intelligibility?
- **RQ2:** Which transient noise suppression (FNN, CNN, or RNN LSTM) architecture is most favorable?
- **RQ3:** How do various models of the neural network architectures react to the unseen data and different levels of noise and noise conditions used in the experiments?
- **RQ4:** What are the computational and architectural limits for each neural network architecture and how do these limits affect clinical applications like real time speech enhancement.?

In response to these questions, the overall objective of this study is to give a contextualized answer which will assist the development of the neural networks used for transient noise reduction. Such deep networks will also be evaluated so as to outline particular performance metrics, application potential, and trade-offs for hearing aid technology.

2 Literature Review

2.1 The Properties of Transient Noise

Speech quality and intelligibility are greatly impacted by the presence of these noise signals like the transient noises. Coughs, door bangs, and clicks are examples of transient noises, which are abrupt or brief sounds. [5] [4]. Important speech components may be hid by these noises, making it difficult to listen under difficult listening situation. For those who have hearing loss, this is particularly troublesome because it can significantly damage their capacity to follow conversations.[3] [5].

2.2 Conventional Transient Noise Reduction Approaches and its limitations

With a lot of regards, the traditional methods for the purpose of transient noise reduction lack the effectiveness of adequate addresses the nature of those sounds. The AGC (Automatic gain control) is a feature which is pretty common in hearing aids, that deals with gain of the entire signal in relation to the received input. But AGC systems have been known to have delays when it comes to the sudden fit of transients leading to an aggressive control of the suppression transients [5]. Furthermore, AGC like other systems damages the speech signals at times, particularly at transitions between voiced and unvoiced segments [6].

More sophisticated algorithms, such as multi-channel transient noise reduction (MCTR), are being used to boost transient noise reductions. This method employs many techniques to identify and reduce transient disturbances while examining the signal over several frequency bands [5]. However, time lag, processing constraint, and the need for precise noise estimates can still restrict these approaches.

2.3 The Rise of Neural Networks in Noise Elimination

Neural networks are being used to reduce noise, especially transient noise, as a result of recent developments in deep learning. [2]. Research indicates that deep neural networks can improve hearing aid users' speech intelligibility and sound quality in noisy settings. [3]. These networks can be trained on noisy speech datasets to produce cleaner outputs. However, research specifically on using neural networks for transient noise reduction in hearing aids remains limited [5].

2.4 Gaps in Existing Knowledge

Noise Conditions and Model Configurations: The architectures' performances of models like RNN,CNN or FNN under varying conditions such as noise and model configurations have a need for further research (e.g., depth, width, training data).

Generalization to Unseen Data: Assessed to some extent; therefore a thorough study of the extent of variation that trained networks can generalize to different noise types and speech characteristics must be performed in order to state that these networks are suitable for real world applications.

Computational Efficiency: The different architectures come with different computational requirements, thus the feasibility of the architectures for use in devices with limited resources such as hearing aids needs to be determined.

2.5 Addressing gaps : The focus of the study

This work fills in the gaps in the area by investigating neural networks for transient noise reduction. Architectures, noise levels, and configurations are assessed to improve comprehension of their design for practical applications. When implementing these algorithms in next hearing aids, the emphasis on generalization and resource economy is essential.

3 Methodology and Experimentation

The method includes data augmentation and three different models (FNN, RNN, and CNN) whose purpose is to improve speech signals affected by a transient noise. The study of additional noise types as well as the estimation of computation cost are all parts of the comprehensive work on the performance and the real-life usages of the models studied.

3.1 Data

3.1.1 Data source and sample population

In this research, the CSTR Voice Cloning Toolkit (CSTR-VCTK) corpus was used. It has a collection of speech data from various speakers with different accents and voices which can be beneficial for the training and assessment of the performance of noise-reducing models. There are voices of 109 native speakers of English with different accents and each speaker was recorded while reading about four hundred sentences. In addition, all of the audio files are sampled at 48, 000 Hz.

Only speakers with a complete set of identifying parameters (ID, age, sex, accent, geographical area) were taken. Lines with such a column number that did not comply with the compliance solution were eliminated. Thus after filtration, there were 102 speakers participating in the research.

Noise Samples The main focus was the usage of ten noise files which are transient in nature. Examples of transient noises used include:

- A loud clap

- A car crash sound
- A police siren
- The sound of falling glass objects
- A sub static sound
- A hard snare hit

3.1.2 Data collection

This involves loading speaker information together with all the voice metadata (e.g., ID, Age, Gender, Accents, and Region) through the VCTK Corpus followed by data cleaning procedures such as whitespace removal, data type casts, and confirmation of gender. After this, the list of speakers is combined and shuffled. For reproducibility, 80 % of the examples with 81 speakers went into the training set, 10% with 10 speakers into validation, and the rest of the 10% with 11 speakers was used in testing [5].

Audio Sample Allocation: The number of audio samples in the training set was balanced by ensuring each speaker contributed a specific number of samples. The training set included audio from the training speakers, while validation samples came from surrogate speakers to fine-tune the system's hyper-parameters and enable early stopping where applicable. Final model evaluation was conducted on the testing set, which consisted of audio from separate testing speakers.

3.1.3 Data augmentation and processing

In terms of the audio, it was then determined in the lower band that the sound would be translated into Mel spectrograms. Mel spectrogram parameters are composed of 64 Mel Frequency bands, the Fast Fourier Transformation (FFT) size of 1024, and a hop length of 512.

A noisy speech dataset was produced by combining speech samples from the CSTR-VCTK corpus with transient sounds at particular signal-to-noise ratios (SNRs: -5 dB, 0 dB, and 5 dB). This made sure that different noise situations were encountered by the neural network. For batch processing, all audio was resampled to 16 kHz, normalized to preserve constant amplitude levels, and padded to equal durations. Accordingly, segments were ready for testing and training.

3.2 Neural Network Architectures

In this work, several neural network architectures were employed in order to determine the best fitting model for the suppression of transient noise.

Multi-layer Feed-Forward Networks (FNNs): Such networks are comprised of one input layer one or more hidden layers and one output layer. Each level performs a linear operation on the level's input and is followed by a nonlinear activation unit. It is employed as a benchmark model in this study.

Convolutional Neural Networks (CNNs): These networks make use of convolutional filters which move over the input data in order to obtain local features which are embedded in the data. They are good in capturing spatial patterns and therefore applicable in image and audio data [4] [6].It is used for locating critical patterns across both the time and frequency domains.

Recurrent Neural Networks (RNNs): RNNs are able to work on time series data due to the presence of an internal memory that is able to store information acquired from earlier time steps. This memory aids the RNNs in dealing with the temporal changes in speech signals. LSTM is a subtype of the RNN architecture that was developed to resolve "the vanishing gradient" problem inherent in standard RNNs so as to enable learning of long-term dependencies in the data [5].

3.3 Model Training and Evaluation

3.3.1 Training procedure

The neural networks that we used implemented backpropagation to coorect their weights and biases by reducing the difference between the model's output and producing clean speech[1, 7].

Important settings like learning rate, the batch size, and number of epochs (between 5-50) were adjusted during training to get the best results. Our implementation worked well, which brought down the Mean Squared Error (MSE) and reducing the gap between the output and the target till the last epoch.

3.3.2 Parameters

Training Parameters: The Adam optimizer, using learning rate of 0.001 was used for its adaptability. Based on Keshavarzi et al. [5], training was limited to 5 epochs with a batch size of 32. LSTM layers used default activation functions ('tanh' for cell state and 'sigmoid' for gates), while the output layer employed a 'sigmoid' activation to get values between 0 and 1 representing ideal ratio mask.

Model parameters: The input shape was determined by the number of time frames and the number of features for which a value of 64 Mel frequency bins was configured in [5]. The number of units in the first LSTM layer is also 128 and the second LSTM layer has 64 units.

Data augmentation parameters Transient noise with Signal-to-Noise Ratio (SNR) levels of -5 dB, 0 dB and 5 dB in conjunction with nine types of transient noise were also applied to enrich the strength of the models against the noise conditions.

3.3.3 Evaluation Metrics

In the end, the model was evaluated with a combination of objective and subjective performance measurements after it has been trained.

Signal-to-Noise Ratio (SNR): SNR is concerned with the ratio of the desired signal in comparison to the background signal at any time. This is crucial since the SNR improvement quantifies how much the model suppressed the noise. A higher SNR improvement signifies an enhancement in noise reduction performance. It is expressed mathematically as follows [5]:

$$\text{SNR (dB)} = 10 \log_{10} \left(\frac{\sum \text{Clean Signal}^2}{\sum (\text{Clean Signal} - \text{Processed Signal})^2} \right) \tag{1}$$

Mean Squared Error (MSE): The average discrepancy between estimated and actual values is evaluated. It functions as a reference to assess how closely the improved signal resembles the clean one, as well as a loss function during training. Better reconstruction and less distortion are indicated by a lower MSE. In terms of mathematics, it is expressed as[5]:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(\text{Clean Signal}_i - \text{Enhanced Signal}_i \right)^2 \tag{2}$$

Perceptual Evaluation of Speech Quality (PESQ): This is an objective measure which predicts the quality of speech as perceived by a listener. Pointers range in core from – 0.5 to 4.5 and higher scores indicate a better perceived quality. It was applied to assess the understanding of the speech quality of the improved signal in relation to the clean signal [5].

Short-Time Objective Intelligibility (STOI): It aims to assess the intelligibility of recorded speech signals as objectively as possible. It is a metric on a range of 0 to 1 where scores closer to 1 indicate higher intelligibility. It evaluated how well the model enhanced the clarity and context of the speech signal [5].

3.3.4 Computational Analysis

The various neural network structures' computational complexities were studied to determine their practicality in real-time applications such as hearing aids and implantable devices. **Processing Time** computes the Amount of time the network takes to complete audio analysis of one segment.**Memory Requirements:**The memory area needed for the network parameters and operations.

4 Results

4.1 Training Results

The training and validation losses over epochs for the different models are shown in Figures 1, 2 and 3.

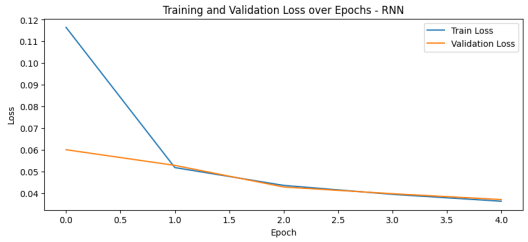

Figure 1: Training and Validation Loss over Epochs for the RNN Model. The loss converges after a few epochs, showing effective learning and minimal overfitting.

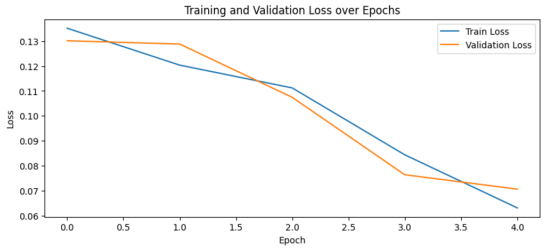

Figure 2: Training and Validation Loss over Epochs for the FNN Model. The model shows stable convergence, with validation loss closely aligning with the training loss throughout the epochs

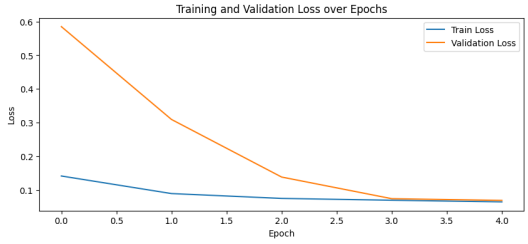

Figure 3: Training and Validation Loss over Epochs for the CNN Model. The model shows stable convergence, with validation loss closely aligning with the training loss throughout the epochs.

4.2 Performance Metrics

4.2.1 Training and Inference Metrics

Table 1 summarizes the training and inference performance metrics for FNN, CNN, and RNN models.

Table 1: Training and Inference Metrics for FNN, CNN, and RNN Models

Metric	FNN	CNN	RNN
Training Time (s)	20.56	113.14	201.65
Inference Time (s)	0.2895	0.0896	0.0813
Memory Usage (MiB)	420.40	301.70	207.43

4.2.2 Test Performance Metrics

Table 2 compares the test performance metrics for FNN, CNN, and RNN models.

Table 2: Test Performance Metrics for FNN, CNN, and RNN Models

Metric	FNN	CNN	RNN
Test Loss (MSE)	0.0626	0.0611	0.0400
SNR of Noisy Audio (dB)	-0.46	-1.96	-0.49
SNR of Enhanced Audio (dB)	-0.09	-1.70	0.09
SNR Improvement (dB)	0.37	0.26	0.58
Mean Squared Error (MSE)	0.0050	0.0205	0.0051
PESQ Score	1.19	1.15	1.14
STOI Score	0.57	0.49	0.58

5 Conclusion

This study provide thorough insight on three neural networks- FNN, CNN, and RNN for transient noise reduction in speech signals. The results focus on the key trade-offs in terms of noise suppression effectiveness, speech intelligibility, and computational efficiency.

5.1 Performance Summary

Training Plots: The training and validation loss curves for the FNN, CNN, and RNN models, shown in Figure 1, 2 and 3 exhibit convergence behaviors. The RNN model Shows stable convergence, with validation loss closely following the training loss, indicating less overfitting. Similarly, the FNN model achieved fast convergence due to its simple architecture. The CNN model, demonstrated a slow convergence compared to RNN and FNN, reflecting its more complex parameter space.

Test Metrics:

- **Test Loss (MSE):** RNN had the lowest test loss (0.0400), outperforming the FNN (0.0626) and CNN with (0.0611), which is better in terms to minimize the prediction error.
- **SNR Improvement:** The RNN achieved the highest SNR improvement (0.58 dB), which indicates the best noise suppression. The FNN achieves SNR improvement of (0.37 dB), while CNN lagged behind with (0.26 dB).
- **Speech Quality and Intelligibility:** FNN has the highest in PESQ Score (1.19) ,RNN closes behind with (1.14) and CNN is slightly lower (1.15). RNN and FNN had achieved comparable STOI scores (0.58 and 0.57, respectively), outperforming CNN (0.49).

Computational Metrics:

- **Training Time:** FNN was the fastest (20.56 s) tto train, followed by CNN (113.14 s), while RNN took the longest (201.65 s) time to train.
- **Inference Time:** The RNN shows the fastest inference time (0.0813 s), followed by CNN (0.0896 s), while FNN shows slower (0.2895 s) times.
- **Memory Usage:** FNN had the highest memory uasge with (420.40 MiB), followed by CNN (301.70 MiB), while RNN performed the best in terms of memory-efficiency, using almost half memory that of FNN (207.43 MiB).

5.2 Insights and Recommendations

- **RNN:** The RNN architecture has the best overall performance, performing well in noise suppression, intelligibility, and inference. Thus, recommended for real-time usage where accuracy and latency are important.
- **FNN:** FNN exhibits good convergence in training and achieved better results in quality and intelligibility of speech. FNN is suitable for applications where speed is prioritized or constrained, but its high memory usage and slow inference time may limits the deployment in resource-constrained environments.
- **CNN:** The CNN shows balanced results, the improvement in SNR was lower and intelligibility scores indicates that it is less suited for tasks involves audio clarity. However, its moderate need of computations make it acceptable for situations where trade-offs between performance and efficiency is possible.

5.3 Future Work

This Work demonstrated the potential of deep neural networks for transient noise reduction in speech for Hearing aids. Future research could explore:

- Increase in General usage with inclusion of more types of noise and speech accents for training.
- further Optimization of the present architectures to reduce computational and memory usage without reduction in performance.
- Testing hybrid architectures that compile the strengths of RNNs, CNNs, and FNNs.

References

[1] Lubna Badri. Development of neural networks for noise reduction. *Faculty of Engineering, Philadelphia University, Jordan*, 2009. Received January 3, 2009; accepted February 25, 2009.

[2] Agudemu Borjigin, Kostas Kokkinakis, Hari M. Bharadwaj, and Joshua S. Stohl. Deep neural network algorithms for noise reduction and their application to cochlear implants. *Weldon School of Biomedical Engineering, Purdue University*, 2022. Unpublished manuscript.

[3] Lars Bramsløw, Gaurav Naithani, Atefeh Hafez, Tom Barker, Niels Henrik Pontoppidan, and Tuomas Virtanen. Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm. *The Journal of the Acoustical Society of America*, 144:172, 2018.

[4] Mahmoud Keshavarzi, Tobias Goehring, Justin Zakis, Richard E. Turner, and Brian C.J. Moore. Use of a deep recurrent neural network to reduce wind noise: Effects on judged speech intelligibility and sound quality. *Trends in Hearing*, 22:1–12, 2018.

[5] Mahmoud Keshavarzi, Tobias Reichenbach, and Brian C.J. Moore. Transient noise reduction using a deep recurrent neural network: Effects on subjective speech intelligibility and listening comfort. *Trends in Hearing*, 25:1–14, 2021.

[6] Hendrik Schröter, Tobias Rosenkranz, Alberto-N. Escalante-B., and Andreas Maier. Low latency speech enhancement for hearing aids using deep filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2716–2728, 2022.

[7] Shin'ichi Tamura. An analysis of a noise reduction neural network. *ATR Interpreting Telephony Research Laboratories*, 1990. Technical report A1a.6.

[8] Luis A. Zavala-Mondragón, Peter H.N. de With, and Fons van der Sommen. A signal processing interpretation of noise-reduction convolutional neural networks. *Journal of IEEE Class Files*, 14(8):1–12, 2021.