Neural adaptation perspective on neural networks for speech enhancement and noise cancellation

Romtin Toranji
toranji@usc.edu

Kwuan Wei Chee
kwuanwei@usc.edu

Junxi Liao
junxilia@usc.edu

Hanjin Jiang
hanjinji@usc.edu

DongKyu Kim
dongkyuk@usc.edu

ABSTRACT

Neural adaptation is a phenomenon of decaying neuronal activities in response to repeated short-term stimulation. This can be seen in transient adaptations that occur in the human visual, auditory, and nervous systems. These adaptations do not cause reduction of the stimulus and occur to assist in focusing on the critical features of the stimulus. This project explores the use of neural network based algorithms to mimic the neural adaptation specifically for the auditory data. A convolutional recurrent network (CRN) model was implemented and tested on an audiobook dataset with varying noise levels. The performance of the model was compared with other signal-processing models. Experimental results show that the deep learning method is effective for noise reduction and more robust.

1. INTRODUCTION

Neural adaptation is a biological process (Kwon et al., 2022) that occurs when recurring short-term stimulus attenuates over time, leading to reduced focus on unchanging sensory inputs (Benda, 2021). This process is seen to be a transient non-learning-based adaptation that occurs and is notable in the human visual, auditory, and nervous systems. The following examples are respective sensory effects that correspond to the human systems: the motion after effect, background noise reduction, and pain reduction. These present themselves as short-term and transient effects because they do not cause permanent reduction of the stimulus. It is a biological process that occurs to assist in focusing on processing only dynamical input and a crucial feature that has likely developed in organisms to optimize the processing of sensory input (Rideaux et al., 2023). The scope of our project is centered around exploring the use of neural-network-based machine learning methods to mimic the occurrence in neural adaptation specifically for auditory speech enhancement.

The reduction of background noise without compromising the clarity and integrity of the primary audio signal, particularly in speech, is a complex problem with multiple challenges. Traditional noise cancellation techniques, such as spectral subtraction (Upadhyay & Karmakar, 2015), adaptive filtering (Dixit, 2017), and noise gating have demonstrated limitations. One challenge that these methods face is the requirement of a reference noise signal, which is not always provided. Additional problems are the loss of frequency components in speech and difficulties in managing high-variance background noises. Moreover, these methods assume a linear relationship between the incoming and generated anti-noise, which will likely fail in real-world applications when there present nonlinear distortions (Zhang & Wang, 2021).

On the other hand, while machine learning (ML) approaches have shown promise in surpassing these traditional methods (Zhang & Wang, 2021, Cha et al., 2023), they come with their own set of challenges. These include the need for high-quality training data, consideration of intricate architectural design, substantial computational resources, and processing delays that can hinder real-time application (Nossier et al., 2020).

This paper addresses the critical need to develop a more efficient and adaptive method for noise cancellation in audio signals, drawing inspiration from the biological process of neural adaptation. Human auditory systems naturally exhibit this adaptive trait, diminishing the focus on consistent, repetitive background noise while amplifying important sounds, such as speech. This project aims to replicate this biological efficiency using advanced ML models, particularly focusing on the integration of convolutional and LSTM layers to mimic the neural adaptation process. The goal is to create a system that not only excels in noise reduction and speech enhancement but also adapts dynamically to changing acoustic environments, akin to the human auditory system.

The significance of this problem extends beyond the technical domain into applications such as enhancing safety in noisy environments, improving the quality of virtual and augmented reality experiences, enhancing speech clarity in low data transfer scenarios, and upgrading the audio experience in various entertainment and gaming applications.

2. RELATED WORKS

We reviewed some of the previous studies that are related to signal processing and current machine learning solutions to noise canceling.

2.1 SIGNAL PROCESSING TECHNIQUES

Signal processing has a rich literature with various methods developed. Previously, the spectral subtraction algorithm was proposed for the enhancement of single-channel speech, where the noise spectrum is estimated and subtracted from the noisy speech (Upadhyay & Karmakar, 2015). Alternatively, instead of suppressing the noise spectrum, some algorithms were proposed to enhance the important frequencies. One development is of Wiener filtering techniques where filters are designed to have large gains at frequencies where the signal-to-noise ratio is high(Izquierdo et al., 2002). However, both of these methods are static methods and are less reliable in dynamic settings where the noise statistics may vary as well as the signal-to-noise ratios. There exist some methods that try to remove the noise dynamically, such as adaptive Kalman filtering (Murugendrappa et al., 2020), where the unknown state variables within the speech are estimated using a Gaussian model. This provides a more robust solution in linear systems with additive Gaussian noise, but when the noise is not Gaussian, or there are non-linearities in the system, this method is less effective.

2.2 MACHINE LEARNING TECHNIQUES

Various research studies have been conducted on noise-reduction techniques with different machine-learning methodologies. According to the results from those studies, we see that machine learning techniques are superior in reducing noise and enhancing speech clarity than alternatives. Dogra et al., (2021) presented an efficient algorithm featured on the principle of CNN to detect noise in audio and used a Python module 'noise reducer' to remove similar noise from the audio. Their implementation was broken into three modules: pre-processing, training, and noise removal. According to their results, the model shows an excellent performance with an accuracy of 97.1%. In the same way, utilizing our familiar machine-learning principles, Sule et al. (2023) approached adaptive and selective signal processing for optimizing noise cancellation results using k-nearest neighbors (KNN) and logistic regression. As the highlight of their study, instead of focusing on training the model using audio signals, they focused on training the model using noisy signals. This approach combined active noise cancellation and adaptive filtering techniques to cancel out the noise while preserving essential noises. While other works had more general machine-learning approaches, Kejalakshmi et al. (2022) discussed active noise cancellation (ANC) using deep learning which is one of the most effective ways of reducing noise. This study chose to train a CRN model to estimate the canceling signals and attenuate the primary noises. As an effective system with the intrinsic ability to model nonlinearities, the deep ANC system performed well in wideband noise removal. Another study also tackled the nonlinear problem with the deep ANC method(Zhang & Wang, 2021). Besides seeking solutions to the fundamental noise canceling problem, there is also research starting to explore flaws and technical gaps. For instance, advanced deep learning-based feedback ANC was proposed to overcome the limitations of traditional ANC and address acoustic delay(Cha et al., 2023). To cancel various high nonlinear and nonstationary noises, the model integrated advanced Conv, AConv, PW, ReLU, ASC module, RNN, and FCL. In addition to conducting research on various machine learning methods, we also found studies focusing on utilizing noise reduction with ML methods in specific fields practically. For instance, the deep learning model could be trained to filter out barking, typing, and other noise from video calls in Microsoft Teams(Sharma & Dash, 2023); integrating the noise reduction model into a music genre classifier may increase the probability of getting higher accuracy compared to just using it(Chavan et al., 2019). Overall, our literature reviews on current machine learning methods not only confirmed the credibility and potential of machine learning in noise canceling but also inspired us to build up the neural adaptation process based on previous efficient methodologies.

3. METHODS

The datasets we used are LibriSpeech (Panayotov et al., 2015) and UrbanSound8K (Salamon et al., 2014). LibriSpeech is a corpus of read English speech, derived from audiobooks, containing 1000 hours of speech sampled at 16kHz. UrbanSound8K is a dataset that contains 8732 sound excerpts of around 4s of urban sounds derived from various conditions such as air conditioners, and car horns. Both of these datasets are publicly available. We aim to use LibriSpeech as the clean audio and to use UrbanSound8K as the noise signals mixed in. The mixing process is done through randomly selecting samples from the pool of speech and urban sound data. When the speech sample is shorter or longer than the urban sound, the urban sound is either repeated or truncated to reach the same length.

Due to the computational restraints of our project, we only used the development set of clean speeches of LibriSpeech. We implemented some baseline signal processing algorithms to compare with our machine learning algorithms. Both classes of algorithms were evaluated on the same portion of data for comparison. The signal processing algorithms that we implemented and tested are the spectral subtraction algorithm and linear state space model algorithm that is based on Kalman filters. For the spectral subtraction algorithm, the magnitude of the noise signal was estimated using a short-time fourier transform (STFT), and subtracted from the frequency domain representation of the noisy signal, and then reconstructed. We consider the reconstructed signal to be our clean signal. For the linear state space model algorithm, the latent state representation of the noisy signal was extracted from the linear Kalman filter, where the system parameters of the Kalman filter were estimated using the expectation-maximization (EM) algorithm. We consider the filtered observation as our clean signal.

The machine learning models that we implemented and tested are the 2-layer long-short-term memory (LSTM) model and the modified version of the convolutional recurrent network (CRN) model presented in Zhang & Wang (2021), which is the state-of-the-art. In Figure 1, we show the model architecture of the LSTM model, and in Figure 2, we show the model architecture of the CRN model.

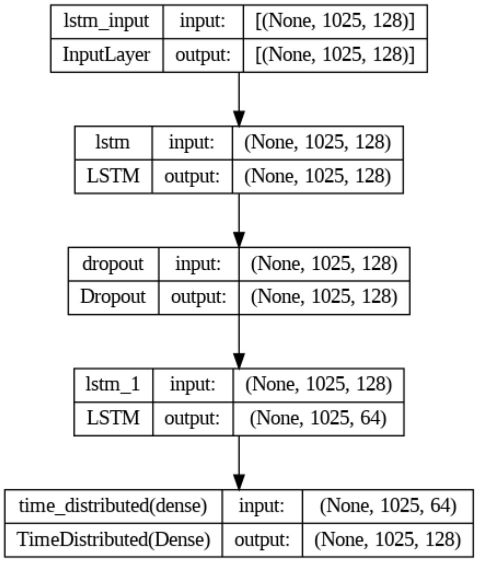
lstm_input | input: | [(None, 1025, 128)]
InputLayer | output: | [(None, 1025, 128)]
lstm | input: | (None, 1025, 128)
LSTM | output: | (None, 1025, 128)
dropout | input: | (None, 1025, 128)
Dropout | output: | (None, 1025, 128)
lstm_1 | input: | (None, 1025, 128)
LSTM | output: | (None, 1025, 64)
time_distributed(dense) | input: | (None, 1025, 64)
TimeDistributed(Dense) | output: | (None, 1025, 128)

Figure 1: LSTM model architecture

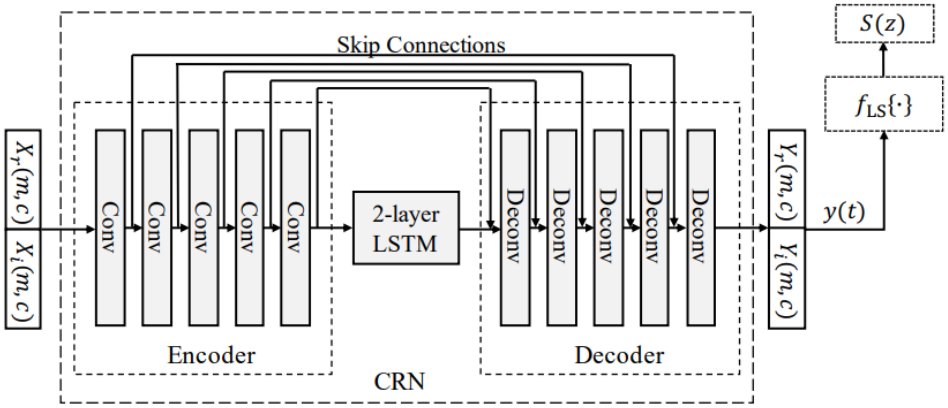

Figure 2: CRN model architecture

The goal is to create a model that excels beyond traditional noise reduction methods, adeptly adjusting to the dynamic nature of real-world sounds. By training on the LibriSpeech dataset, augmented with both Gaussian and real-world noises, the model is designed to tackle a wide spectrum of auditory challenges. This advancement aims to enhance safety in various environments, improve user experience in VR and AR, and upgrade audio quality in entertainment and gaming.

The models were implemented in Python, and the codes are available in the following github repository https://github.com/cheekw/CSCI567-project.

4. RESULTS

Both LSTM, and CRN models introduced in the previous section were trained using the noisy samples obtained from mixing the LibriSpeech and the UrbanSound8k datasets. Due to technical difficulties associated with the available resources on the Google Colab environment, we trained and evaluated on 1024 samples of 1025 timesteps with 128 frequency bins. 80% of the available samples were used to train the models, and the remaining 20% of the samples were used for evaluations. For both models, the same train and evaluation sets were used.

The models were optimized using the Adam optimizer with the objective of minimizing the mean squared error between the output of the model with noisy samples, and the clean samples. The evaluation metric is the mean squared error (MSE) between the outputs and the clean samples. In Figure 3, we show the training and the validation errors for the LSTM model, and in Figure 4, we show the training and the validation errors for the CRN model. Since they are both evaluated on the same dataset, the numbers are comparable.

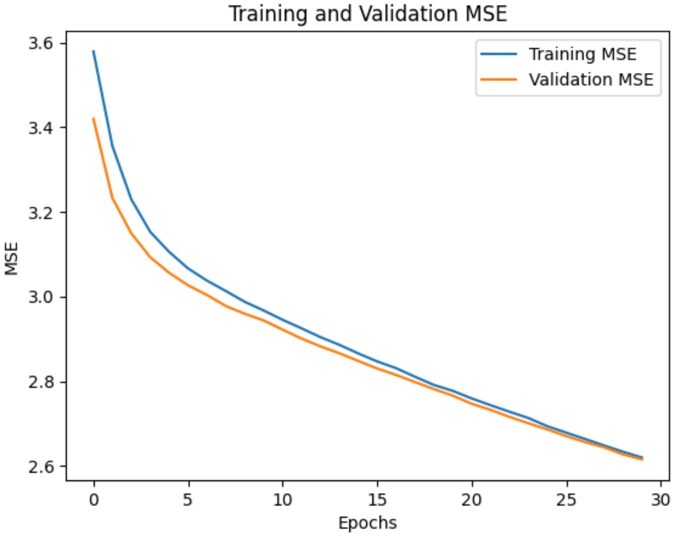

Figure 3: LSTM model MSE

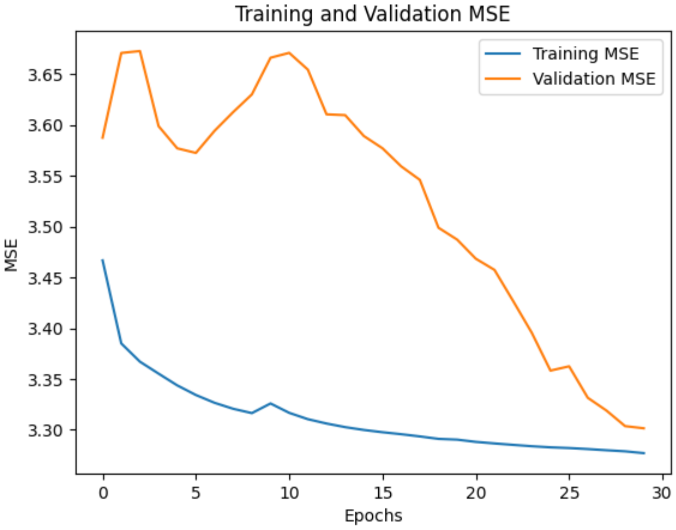

Figure 4: CRN model MSE

The training and validation MSEs of the LSTM model decrease every epoch. On the other hand, training and validation MSEs of the CRN model fluctuate over epochs, and converge to a much higher error compared to the LSTM model. The final validation MSE for the LSTM model was around 2.6, and the final validation MSE for the CRN model was around 3.3.

5. DISCUSSION

The LSTM model outperforms the CRN model for our dataset. There are some unexplored challenges associated with the CRN model. For example, we have no explanation on why the MSE fluctuates, and why the validation MSE quickly catches up to the training MSE after the 10th epoch. We suspect that this was due to the lack of training samples.

This project suffered from the lack of computational power or sufficient time to thoroughly investigate the machine learning models. The machine learning models in consideration need a large amount of resources to fine-tune. Investigating which model works the best also requires the fine-tuning of all different configurations of various models. Without sophisticated fine-tuning, machine learning models will just be as good as classical signal processing models from the results that we have observed.

In addition to the lack of resources, machine learning methods are also prone to overfitting as they hugely rely on the quality and the characteristics of the given training samples. While machine learning models may excel in denoising dynamic distortions in the domain of data that they are trained upon, they may require retraining if we want to utilize the same model in different environments and maintain the same level of performance.

6. FUTURE WORK

Ultimately, we want to create a model that excels beyond traditional noise reduction methods, adeptly adjusting to the dynamic nature of real-world sounds with temporal attenuation effect. However, our results have not shown this feature to be prominent, and we suspect this effect is subject to tuning for it to be noticeable.

We also want to generalize the neural adaptation machine learning model beyond domains of auditory data. This could be achievable if we can find a way to generalize and label different media data so that they can be fed to the model with the same dimensions. This way the model can be trained and improved to handle multiple media types, as a human brain does, and accomplish the denoising effect that closely resembles human neural adaptation.

Additionally, we attempted to build our own model that is similar to the CRN model, but with attention heads instead of LSTM layers to model the features associated with recurrence. Our idea was instead of recurrent neural networks, usage of attention based layers might be better at extracting relevant information about the signals. However, we were not able to implement a working version of this model in time, and we would pursue in this direction given more time.

7. CONCLUSION

In this paper, we explored the concept of applying neural adaptation in signal denoising, where in particular we implemented a neural adaptation-inspired machine learning model for audio signal noise cancellation. Within the limit of our computational power, we see promising signs of a CRN mode with 2 LSTM layers performing well in the limited available data we obtained and tested upon. In spite of necessary comprehensive training and fine-tuning for ML models with LSTM layers to perform significantly better than existing methods, the benefits obtained from neural adaptation are unique and exciting. With increasing average computational power in electronic devices, we believe the aforementioned model along with other ML implementations will become increasingly popular in applications to decrease dynamic noise in various media in the near future.

REFERENCES

Kwon S, Kim BS, Park J. Active Noise Reduction with Filtered Least-Mean-Square Algorithm Improved by Long Short-Term Memory Models for Radiation Noise of Diesel Engine. Applied

Sciences. 2022 Oct 12;12(20):10248.

Benda J. Neural adaptation. Current Biology. 2021 Feb;31(3):R110–6.

Rideaux R, West RK, Rangelov D, Mattingley JB. Distinct early and late neural mechanisms regulate feature-specific sensory adaptation in the human visual system. Proceedings of the National Academy of Sciences. 2023;120(6):e2216192120.

Upadhyay N, Karmakar A. Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study. Procedia Computer Science. 2015;54:574–84.

Dixit S, Nagaria D. LMS Adaptive Filters for Noise Cancellation: A Review. IJECE. 2017 Oct 1;7(5):2520.

Westhausen NL, Meyer BT. Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression. 2020 [cited 2023 Dec 7]; Available from: https://arxiv.org/abs/2005.07551

Zhang H, Wang D. Deep ANC: A deep learning approach to active noise control. Neural Networks. 2021 Sep;141:1–10.

Cha YJ, Mostafavi A, Benipal SS. DNoiseNet: Deep learning-based feedback active noise control in various noisy environments. Engineering Applications of Artificial Intelligence. 2023 May;121:105971.

Nossier SA, Wall J, Moniri M, Glackin C, Cannings N. An Experimental Analysis of Deep Learning Architectures for Supervised Speech Enhancement. Electronics. 2020 Dec 24;10(1):17.

Dogra M, Borwankar S, Domala J. Noise Removal from Audio Using CNN and Denoiser. Advances in Speech and Music Technology [Internet]. 2021; Available from: https://api.semanticscholar.org/CorpusID:236720441

Murugendrappa N, Ananth AG, Mohanesh KM. Adaptive Noise Cancellation Using Kalman Filter for Non-Stationary Signals. IOP Conf Ser: Mater Sci Eng. 2020 Sep 1;925(1):012061.

Izquierdo MAG, Hernández MG, Graullera O, Ullate LG. Time–frequency Wiener filtering for structural noise reduction. Ultrasonics. 2002 May;40(1–8):259–61.

Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [Internet]. South Brisbane, Queensland, Australia: IEEE; 2015 [cited 2023 Dec 7]. p. 5206–10. Available from: http://ieeexplore.ieee.org/document/7178964/

Salamon J, Jacoby C, Bello JP. A Dataset and Taxonomy for Urban Sound Research. In: 22nd ACM International Conference on Multimedia (ACM-MM'14). Orlando, FL, USA; 2014. p. 1041–4.

Sule R, Kolekar A, Patel K, Tandle A. Selective Noise Cancellation using Machine Learning. In: 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT) [Internet]. Bhubaneswar, India: IEEE; 2023 [cited 2023 Dec 7]. p. 645–50. Available from: https://ieeexplore.ieee.org/document/10201773/

Kejalakshmi V, Kamatchi A, Abysyta M. Active Noise Cancellation using Deep learning. IJSDR. 2022 Jul;7(7):133–9.

Cha YJ, Mostafavi A, Benipal SS. DNoiseNet: Deep learning-based feedback active noise control in various noisy environments. Engineering Applications of Artificial Intelligence [Internet]. 2023 May 1 [cited 2023 Dec 8];121:105971. Available from:https://www.sciencedirect.com/science/article/abs/pii/S0952197623001550

Sharma P, Dash B. Using Artificial Intelligence to Filter out Barking, Typing, and other Noise from Video Calls in Microsoft Teams. International Journal on Cybernetics & Informatics. 2023

Jan 30;12(1):1–11.

Chavan O. Machine Learning and Noise Reduction Techniques for Music Genre Classification. wwwacademiaedu [Internet]. 2019 Jan 1 [cited 2023 Dec 8]; Available from: https://www.academia.edu/109029638/Machine_Learning_and_Noise_Reduction_Techniques_for_Music_Genre_Classification