Machine Learning Capstone - Final Paper

Background Noise Suppression

Asha Cheruvu, Daniel Kweon, Nikhil Shirolkar

May 7, 2020

Abstract

Digital audio is essential for modern communication and can be found everywhere from voice calls to video conferences and hearing aids. Poor quality audio hinders communication; thus, audio technologies have evolved to better capture higher quality sound. These technologies take advantage of microphone placements or incorporate digital signal processing techniques to remove unproductive background noises and extract clean speech. There exists deep learning based speech extraction techniques; however, the computational complexity of deep learning prevents its widespread use in real time. To determine the viability of machine learning based real time denoising, we developed a digital signal processing technique and recurrent neural networks of two different sizes. These models were evaluated on audio quality improvement and computational intensity. The performances of the recurrent neural networks were compared with that of the standard digital signal processing to determine the feasibility of a machine learning based real time background noise suppression system.

Introduction

- Background -

Background noise is sound produced by environmental factors. In the context of voice audio, background noise is irrelevant and typically unwanted, as it can reduce the quality of the recording and make it difficult to hear the speaker.

Due to the prevalence of background noise, various methods of denoising speech have been developed. Typical methods involve strategic mic placements, reduced mic sensitivity, audio frequency smoothing, and noise reduction. These methods, however, are not perfect as they require multiple mics or have difficulty suppressing non-stationary background noises. (Stationary noises are constant and repeating noises e.g. breathing, footsteps)

With the advent of modern machine learning technologies, several companies like 2Hz, Mozilla, and Microsoft have began utilizing deep learning to improve the quality of speech recordings. Through the application of deep learning, these companies have managed to develop new methods of improving their user's audio experiences.

With the increasing prevalence of real time audio, developing a machine learning model that can provide speech enhancement with minimal computation has immense implications. This model can be applied to phone calls, video conversations, and live streams to improve audio experiences for all listeners. Companies that are invested in the audio business (cellular service providers, video conference providers), would be able to provide quality audio for their customers.

- Related Works -

Simple machine learning models tend to under-perform due to the complexity of audio data. From our literature survey, we found that researchers successfully enhanced speech using a hybrid approach consisting of digital signal processing and deep neural networks. (Valin et al.) In this approach, audio clips were broken down into features (pitch, spectral bands) and inputted into a deep neural network with four hidden layers.

Another group of researchers obtained similarly successful results using recurrent neural networks. (Xiph et al.) The recurrent neural networks in our research are based on the architecture from Xiph et al's recurrent neural network. In this architecture, audio features (pitch, first and second derivative coefficient across frames) were calculated and fed into a recurrent neural network.

These approaches display the ability to denoise audio; however, the evaluations of these approaches do not thoroughly investigate or quantify the approach's viability to be applied in real time audio denoising.

- Limitations -

Although these deep learning models are capable of denoising speech, they have limitations when applied real time. Filtering real time audio through machine learning models are computationally intensive and can introduce latency. Latency may be worse for a user's experience than low quality audio. This latency will worsen as audio quality increases. The model will require a higher audio sampling rate, thus needing more computations in shorter time intervals. Due to the issue of latency, background noise suppression in real time audio is currently being studied as researchers strive to build models that minimize computation but optimize quality.

Despite the benefits of machine learning in speech enhancement, there are various risks associated with these models. A model may have difficulty performing in conference calls where multiple people are simultaneously speaking. A speaker's sound volume depends on their distance from the microphone and lower volume speakers may be classified as background noise. If a machine learning model incorrectly filters and eliminates speech, it may provide a poor audio experience and may even induce miscommunication between parties.

Framework

- Approach -

The audio dataset was generated through the Microsoft Scalable Noisy Speech Dataset. The audio data was customized and collected to train recurrent neural networks of varied sizes (87K, 70K hyper parameters). This allowed for the analysis of the quality of denoised audio based on the size of the model. Recurrent neural networks were chosen as our machine learning model architecture to build upon Xiph et al's previous research with recurrent neural networks. These models were evaluated on the quality of background noise suppression and the computational intensity.

Because real time audio requires minimal latency, the effectiveness of the model's application in real time audio must be evaluated. Latency differs between various devices, as devices equipped with accelerators handle matrix operations significantly faster. Instead of evaluating latency, we can measure the number of hyper parameters required by the model. This is a viable means of indirectly evaluating latency, as the number of hyper parameters correlate to the number of computations the model must undertake to filter background noise. However, it is still important to measure the time required for a machine learning model to denoise audio clips, as it allows for comparisons with the performance of digital signal processing.

This approach addresses the limitations of past research by thoroughly evaluating audio quality improvement and computational intensity of recurrent neural network denoising. Two different sizes of recurrent neural networks were compared to understand the effect the hyper parameter count has on audio quality improvement and computational intensity. Given a quantitative measurement of the performance and computational intensity, these models were compared with the standard digital signal processing technique to quantify the feasibility of machine learning based real time denoising.

- Real Time Audio Denoising -

Real time audio denoising relies on inexpensive and fast computations. Research indicates that humans can tolerate between 20 to 200 milliseconds audio delay in conversation. Thus, 20 to 200 milliseconds is the acceptable latency for real time audio denoising. Current real time denoising approaches are able to achieve this range by reducing the sampling rate. The trade off is the reduced quality of the denoised audio. There is research being done to leverage GPUs to scale up machine learning models to work with incoming audio more efficiently. While the original intent was to apply these methods and develop a real time audio denoising system, due to time constraints we focused on finding a machine learning model that minimizes computational intensity.

Experiment

- Data -

The training and testing audio data were generated using Microsoft Scalable Noisy Speech Dataset. This dataset consisted of various sound clips of clean speech and noise. Clean speech audio clips are isolated recording of voices speaking from various accents, ages, and genders. Noise audio clips are recording of various environmental noises that may occur in the background. These noise clips include background noises generated by air conditioning, printers, crowds, traffic, airport announcements, and restaurants.

The data for clean speeches consisted of 2.7 Gigabytes of audio clips (0.42 Megabytes of which was specific for testing). Data for noise consisted of 0.45 Gigabytes of audio clips (0.76 Megabytes of which was specific for testing). The Microsoft Scalable Noisy Speech Dataset combines various clean speech clips and noise clips. These clean speech clips and noise clips were then combined to generate the noisy speech dataset.

The Microsoft Scalable Noisy Speech Dataset allowed for customized configurations regarding sampling rate, audio format, and audio length. For our experiment, we choose a 16 kHz sampling rate, .wav audio format, 16 bits per sample, and a 1 hour audio length. We focused on audio with a speech to noise ratio of 10 db and 1 second intervals between different voices. The rest of the configurations were left as default.

- Methodology -

Digital Signal Processing

Digital signal processing is a non-machine learning based approach for background noise reduction. In denoising, digital audio is edited to remove background noise. This digital signal processing technique was used as a baseline for real time audio denoising performance. The digital signal processing algorithm takes the noisy speech and the prototypical noise audio. The prototypical noise audio represents sample noise audio to be removed. The algorithm uses the fast fourier transformation on the noisy speech and the prototypical noise audio, then creates a threshold through the prototypical noise audio. The noisy speech is removed of audio outside these threshold frequencies and smoothed.

Pipeline

Using the Microsoft Scalable Noisy Speech Dataset, a large .wav file is generated for training. This .wav file is converted to a .pcm to remove the unnecessary headers and footers. The .pcm file is then converted to a .h5 file, as the .h5 format is more compatible for intensive matrix computations. The recurrent neural network trains on the .h5 file and the hyper parameters are collected. The hyper parameters recreate the model and predict the denoised audio data. The output of the denoised test data is a .pcm file, which is converted to a .wav file. This is a playable and hearable format. The audio clip was then evaluated using the M.O.S. and P.E.S.Q. metrics, in addition to its computational intensity. The diagram below depicts our workflow to achieve denoised speech.

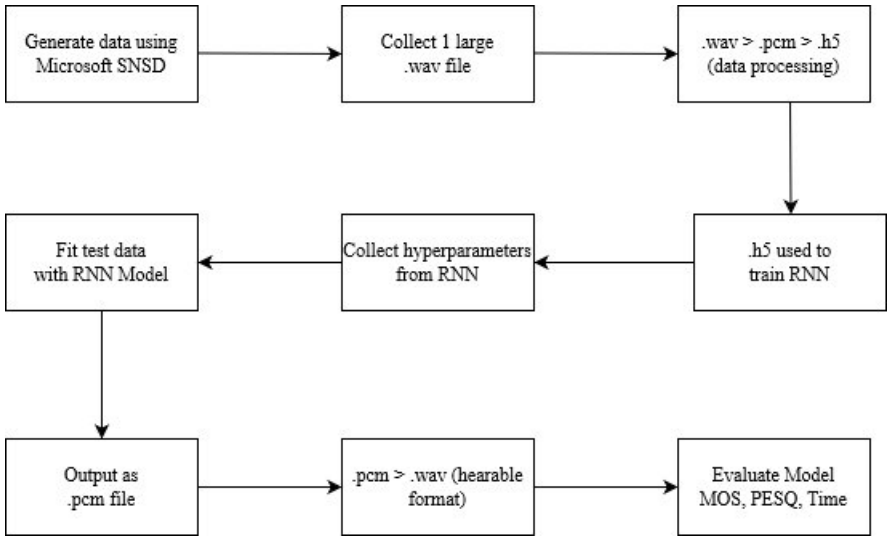

Figure 1: Research Pipeline

Data Processing

The data processing stage first requires the generated .wav file to be converted into a .pcm file. This is a necessary step as .wav files contain headers and footers that make it difficult to process the audio blocks that we want denoised. The .pcm file contains the audio data of a .wav file without the headers. However, this file format is not optimal for processing. The script we use to train our model relies on the .h5 file format. The data in the .pcm file is reformatted to fit an optimal .h5 structure that the training script can handle. Following training and testing, the denoised audio data is outputted as a .pcm file. This file solely contains audio data, thus is not quite playable or hearable. This .pcm file is converted to a .wav file using the same headers and footers. This .wav allows us to hear the audio and run our evaluations.

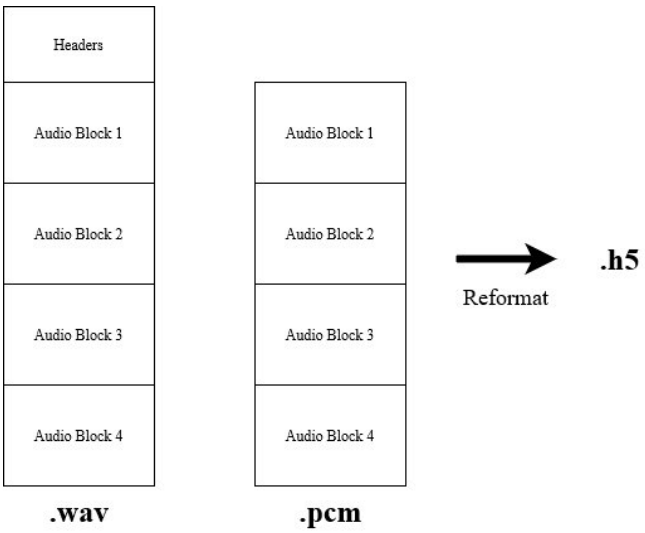

Figure 2: Audio Format Conversion

Recurrent Neural Networks

Our recurrent neural networks are based on the architecture of Xiph et al's research. The recurrent neural network takes as inputs pitch, frequency, and the first and second derivative coefficients across frames. It then outputs 22 bands that represent spectral bands of audio. These bands are used to reconstruct the denoised audio.

Two recurrent neural networks of 87k and 70k hyper parameters were trained. These recurrent neural networks used gated recurrent unit for their recurrent connections. Gated recurrent units were used to ensure long term memorization of patterns, compared to single connections. In addition, gated recurrent units use 2 parameters instead of 3 parameters in the long short-term memory units, while performing similarly well. This allows for a lighter recurrent neural network without foregoing performance.

Within the architecture of our recurrent neural network, there exists a subroutine that determines the presence of voice activity. When there is no voice activity detected, the sound clips are classified as background noise and subtracted from the original audio to generate a deniosed audio clip. There is no way to prove that the recurrent neural network works in this manner; however, the architecture of the recurrent neural network was designed with this intent.

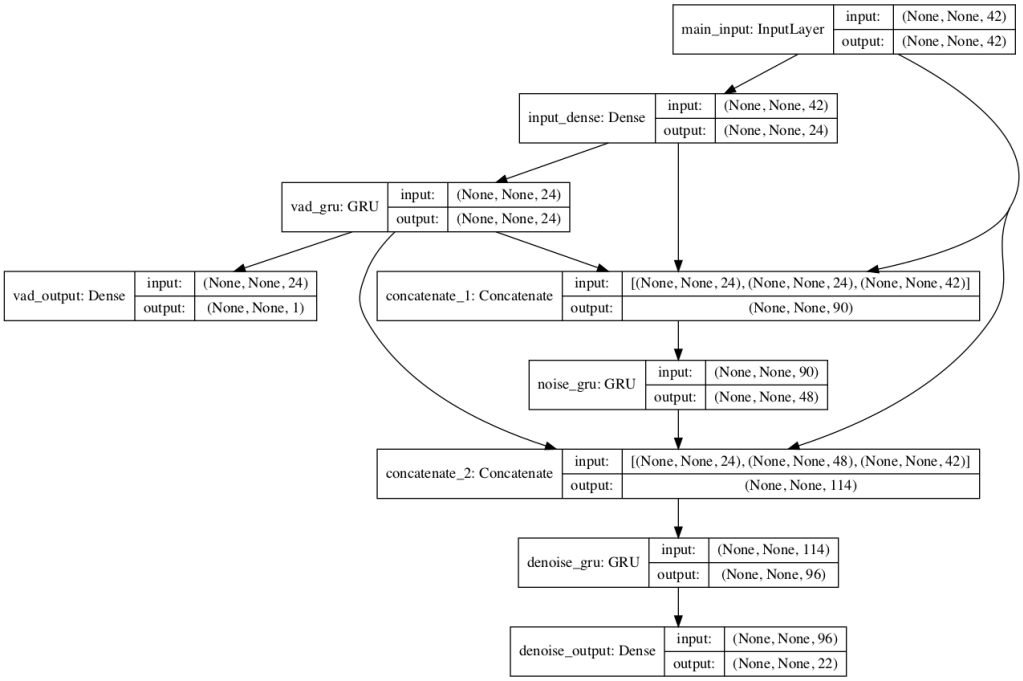

Figure 3: Recurrent Neural Network (87K)

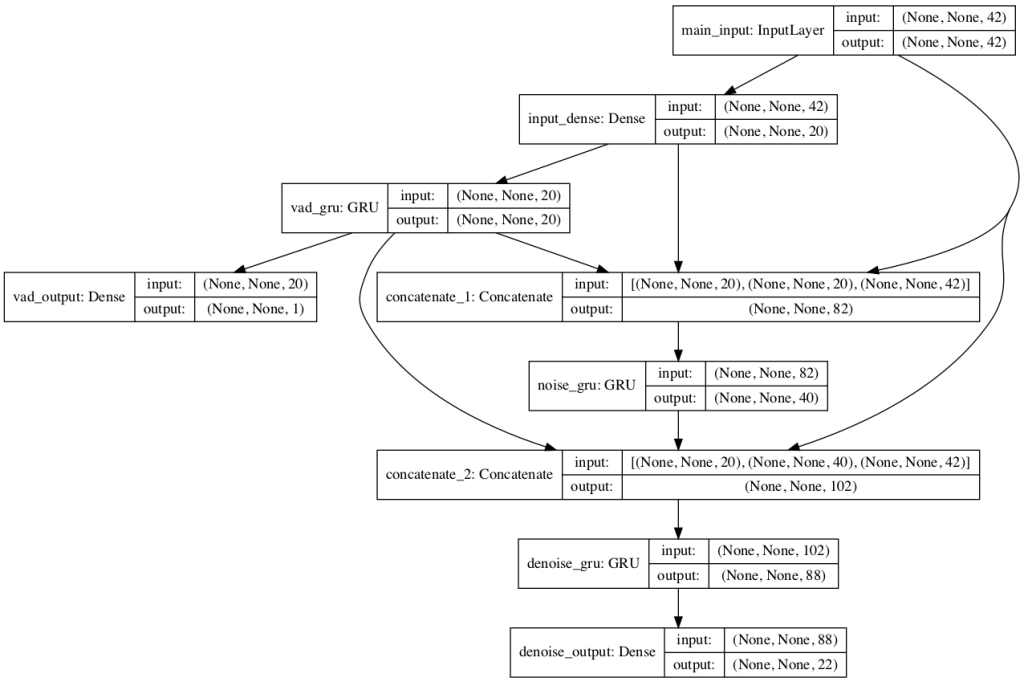

Figure 4: Recurrent Neural Network (70K)

Quality Evaluation

The quality of the background noise suppression was quantified through two evaluation method: mean opinion score (M.O.S.) and perceptual evaluation of speech quality (P.E.S.Q).

M.O.S. acknowledges the subjective nature of assessing the quality of speech audio. This evaluation methods requires individually grading the performance of a model between 0 (terrible) and 5 (excellent). The mean of all opinions are then used to score the performance of the model.

P.E.S.Q is an industry standard for testing audio quality. The software that runs the P.E.S.Q evaluation can be downloaded from the International Telecommunication Union website and be run through the command line. Original audio clips and denoised audio can be run through P.E.S.Q. A mean score of several audio clips can be used as an evaluation. P.E.S.Q returns a score between -0.5 and 4.5, where higher scores indicate better sound quality.

Every denoising techniques was evaluated by the P.E.S.Q and M.O.S. methods. A total of 4 audio clips with 5 voices each were evaluated.

Computational Intensity Evaluation

The number of hyper parameters were recorded to compare sizes between recurrent networks. In addition, the computation time required to denoise an audio clip of 1 minute length was recorded to compare the recurrent neural networks with the digital signal processing technique.

- Results -

Digital Signal Processing

Audio Quality Evaluation

	Voice 1	Voice 2	Voice 3	Voice 4	Voice 5	M.O.S. Avg.	P.E.S.Q.
Audio 1	2, 1, 2.5	2.5, 1, 1.5	2, 1, 2.5	2.5, 2, 3	3, 1, 1.5	1.93	2.897
Audio 2	3.5, 2, 1.5	2, 0, 1	2, 2, 2.5	2.5, 1, 2	2.5, 2, 2	1.90	1.108
Audio 3	2.5, 1, 2	2.5, 1, 2.5	2.5, 0, 2.5	2, 0, 2	2, 0, 2	1.63	2.829
Audio 4	3, 4, 2	2.5, 4, 2	2.5, 1, 3	2.5, 3, 3	3, 4, 3	2.83	1.872

Overall Evaluation

M.O.S. [0, 5]	P.E.S.Q. [-0.5, 4.5]	Computation Time	Training Time
2.194	2.176	0.532 sec	N/A

Recurrent Neural Network Size 87,503 hyper parameters

Audio Quality Evaluation

	Voice 1	Voice 2	Voice 3	Voice 4	Voice 5	M.O.S. Avg.	P.E.S.Q.
Audio 1	3, 2, 3	2.5, 3, 2	3, 2, 2	3, 2, 2.5	2.5, 2, 2	2.43	3.301
Audio 2	3, 1, 1	2.5, 1, 3	2.5, 3, 3.5	3, 3, 3	2.5, 2, 3	2.46	3.260
Audio 3	3.5, 3, 3.5	2.5, 2, 2	2, 1, 2.5	2.5, 2, 3	2.5, 3, 3	2.53	3.017
Audio 4	3.5, 4, 4	3.5, 4, 4	3.5, 2, 4	4, 3, 3	3.5, 3, 4	3.53	3.562

Overall Evaluation

M.O.S. [0, 5]	P.E.S.Q. [-0.5, 4.5]	Computation Time	Training Time
2.742	3.285	1.2 sec	4.5 hr

Recurrent Neural Network Size 70,483 hyper parameters

Audio Quality Evaluation

	Voice 1	Voice 2	Voice 3	Voice 4	Voice 5	M.O.S. Avg.	P.E.S.Q.
Audio 1	1, 0, 1.5	2.5, 3, 3	2, 2, 2	2, 1, 1	2, 1, 1.5	1.7	3.296
Audio 2	2, 3, 3	1, 1, 2	2, 1, 2.5	3, 1, 3	2, 3, 1	2.03	3.328
Audio 3	3, 3, 2.5	1.5, 1, 2	1, 1, 2	2, 2, 1.5	2, 2, 1	1.83	3.134
Audio 4	4, 4, 4	4, 3, 3.5	3.5, 2, 2.5	4, 3, 4	4, 3, 3	3.43	3.587

Overall Evaluation

M.O.S. [0, 5]	P.E.S.Q. [-0.5, 4.5]	Computation Time	Training Time
2.249	3.336	1.2 sec	3.5 hr

Conclusion

- Summary -

Background noise suppression is a critical area with applications in cellular service, video calls, and live streams. While there are several methods available that mitigate irrelevant background noise, these methods have difficulty well isolating clean speech. Due to the difficulty of audio denoising, modern method have begun incorporating machine learning techniques. However, the computational complexity of machine learning models prevents its widespread use in real time audio processing.

Our research began by comparing the performances of recurrent neural networks of different sizes. The larger recurrent neural network obtained a lower P.E.S.Q. score than the smaller recurrent neural network; however, the smaller recurrent neural network obtained a lower M.O.S. score than the larger recurrent neural network. The smaller recurrent neural network theoretically takes less time to denoise audio than the larger recurrent neural network; however, both models took approximately the same time when run on a personal laptop. The difference in the M.O.S. score is significantly larger than the difference in the P.E.S.Q. score, thus we came to the conclusion that it was not necessary, and if anything harmful, to forego additional hyper parameters in a recurrent neural network. There was no observed improvement in denoising time and audio quality was significantly reduced with smaller hyper parameters.

To investigate the viability of a machine learning based real time background audio suppression system, we constructed recurrent neural networks and compared their evaluations with that of the standard signal processing technique. From our experiments, we found that the machine learning based background noise suppression system outperforms the signal processing technique in regards to sound quality. However, the machine learning models demand more computational intensity than digital signal processing. Although this was expected, we were able to thoroughly evaluate and quantify these differences. From this comparison, we conclude that a machine learning based background noise suppression system is viable, given sufficient computational power.

- Contributions -

Our project was inspired by previous works that addressed building machine learning models to denoise audio. However, the main contribution of our project that many other works lacked was a significant degree of evaluation on the different methods. We constructed on the research of Xiph et al to quantify the degree to which a recurrent neural network would be viable for real time background noise suppression. We used both subjective and objective standards of evaluation for audio as a means of quantifying the extent to which a particular method improved the final quality

audio. This quantification used M.O.S. and P.E.S.Q. to evaluate audio quality improvement and hyper parameters count with denoising time to evaluate computational intensity.

In addition, two recurrent neural networks of different sizes were developed and evaluated to determine the effect of the hyper parameter count on the audio quality and computational intensity. Furthermore, the evaluations of the recurrent neural networks were compared with the evaluations of digital signal processing. This established a baseline performance by which we could compare the recurrent neural networks and determine their feasibility to be applied in a real time audio denoising situation.

Through this research project, we successfully quantified the feasibility of a machine learning-based real time background noise suppression system.

- Future Works -

One machine learning model we began investigating was the deep neural network. Our motive for investigating this architecture was to quantify the affect of recurrent units and different data feature extractions. We wanted to attempt to transform the audio data into segments of short-term fourier transforms. The fourier transform may reduce the data preprocessing needed in Xiph et al's architecture and, thus, reduce the computation time needed to denoise audio. The deep neural network would return the appropriate frequencies that could then be used to construct the denoised audio. Despite our progress, we ran into technical and mathematical issues with fourier transforms, as it was our first time working with these concepts. Due to time constraints, we weren't able to complete the experiment with a deep neural network. In the future, we would want to complete this deep neural network and analysis its performance.

There are various directions in which this project can pursued for future works. Despite the work we have done, building and evaluating various machine learning models, these models were not applied real time. Regardless of the computational intensity of these machine learning models, these models are viable options for real time audio processing given sufficient computational power. By constructing various models and testing in a real time situation, the optimal models given the computation constraints is derivable. A mobile phone may not have the same computational power as a multi-core computer, thus making certain machine learning models viable option for real time audio processing on a computer rather than a phone.

This research project can also be expanded through exploration of different machine learning models, like convolutional neural networks. Our data was generated through Microsoft's Scalable Noisy Speech Dataset that, despite it large quantity and diversity of audio clips, is limited. It is important to gather more data from various environmental sources to train a more robust model that can handle all kinds of noisy audio. In addition, our research did not focus on processing audio clips with multiple distinct voices. For future works, it is important to explore the effects

multiple voices would have on the performance and computational intensity of various machine learning models.

References

[1] Baghdasaryan, Davit. "Real-Time Noise Suppression Using Deep Learning." Medium, Towards Data Science, 22 Dec. 2018, towardsdatascience.com/real-time-noise-suppression-using-deep-learning-38719819e051.

[2] Baghdasaryan, Davit, et al. "Real-Time Noise Suppression Using Deep Learning." NVIDIA Developer Blog, 13 Nov. 2019, devblogs.nvidia.com/nvidia-real-time-noise-suppression-deep-learning/.

[3] Reddy, Chandan K.a., et al. "A Scalable Noisy Speech Dataset and Online Subjective Test Framework." Interspeech 2019, 2019, doi:10.21437/interspeech.2019-3087.

[4] Kumar, Anurag, and Dinei Florencio. "Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks." Interspeech 2016, Aug. 2016, doi:10.21437/interspeech.2016-88.

[5] Valin, Jean-Marc. "A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement." 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), 2018, doi:10.1109/mmsp.2018.8547084.