Machine Learning Capstone Proposal

Background Noise Suppression Using ML

Asha Cheruvu, Daniel Kweon, Nikhil Shirolkar

February 20, 2020

Introduction

Problem Overview

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 de-noising 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)

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.

Current Research

Current state-of-the-art technologies use deep learning techniques. Simple learning models tend to under perform with the complexity of audio data. From our brief survey of papers, researchers were able successfully enhance speech using deep neural network models. Other researchers have incorporated deep learning into digital signal processing techniques to achieve successful results as well.

Although these deep learning models are capable of de-noising 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.

Research Proposal

Proposed Study

We propose to study the performance of various machine learning models in regards to background noise suppression and effectiveness in real-time audio. If we have time remaining, we would like to formulate a model that maintains the best audio quality with minimized computational intensity.

Impact Of Study

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.

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.

Evaluation

Our solution will be evaluated in two categories: the quality of background noise suppression and the computational intensity.

Background Noise Suppression Evaluation: The quality of the background noise can be quantified through two evaluation method: mean opinion score (MOS) and perceptual evaluation of speech quality (PESQ). MOS acknowledges the subjective nature of assessing the quality of speech audio. This evaluation methods requires the individually grading the performance of a model between 0 (terrible) and 5 (excellent). The mean of all opinions will then be used to score the performance of the model.

PESQ is an industry standard for testing voice quality. The software that runs the PESQ evaluation can be downloaded from the International Telecommunication Union website and be ran through the command line. Original audio clips and version that has been removed of background noise can be ran through PESQ, and a mean of several audio clips can be used as an evaluation. PESQ returns a score between -0.5 and 4.5, where higher scores indicate better sound quality.

Computational Intensity Evaluation: Because real time audio requires minimal latency to be effective, the effectiveness of the model's application at real time audio must be evaluated. The latency will differ between various devices, as devices equipped with accelerators will handle matrix multiplication much faster. Thus, instead of evaluating latency we can measure the number of hyper-parameters required by the model. This is a viable means of evaluating latency, as the number hyper-parameters correlate to the number of computations the model must undertake to filter background noise.

Background Papers

Relevant research papers to understand as background:

1. A Scalable Noisy Speech Dataset and Online Subjective Test Framework (2019)
   `https://www.isca-speech.org/archive/Interspeech_2019/pdfs/3087.pdf`

2. Speech Enhancement In Multiple-Noise Conditions using Deep Neural Networks (2016)
   `http://www.cs.cmu.edu/~alnu/sefiles/SE_4.pdf`

3. A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement (2017)
   `https://arxiv.org/pdf/1709.08243.pdf`
   `https://people.xiph.org/~jm/demo/rnnoise/`

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.

Project Management

Methods

Approach: The most commonly used machine learning model in our literature was the deep neural network. We will begin by reproducing the research described in literature. After successfully reproducing the results of previous research we will attempt to improve our model's performance so that it can be applied to real-time audio. If time permits, we would like to train our model on a larger and more diverse data set so that it may handle a wider variety of background noises and circumstances (e.g. conference calls).

The primary language we will be coding in is Python and code collaboration will be through the Col-lab environment. We will also use Git for version control and Trello for our agile project management. After an initial survey of available Python libraries, we've identified the noise-reduce and the logmmse libraries to aid us in audio handling.

Microsoft has conducted similar research and has generously provided open data as long as credit is given. This data has a large quantity of cleaned and noisy voice audio that can be used to train our models.

Responsibilities: In the initial stages of our research project, every team member is responsible for becoming familiar with the existing literature. Because this is a new domain (in regards to subject matter and data modality), we believe that cooperation on the majority of the initial tasks is essential. As we become more familiar with the domain and project, we will begin using our project management tools to divide various portions of the project between team members.

Timeline

Steps:

Literature Review
Data Retrieval & Manipulation
Model Training
Model Evaluation
Improve Model Performance

Mar 6 - Milestone 1: By Milestone 1, we would like to have a more detailed outline of the steps we need to successfully reproduce existing research. In order to do this, we would need to have read and understood all literature in addition to the technologies and approaches that were mentioned.

Apr 1 - Mid-Term Check: By the Mid-Term Check, we would like to have reproduced results from existing research. This means we would have built a machine learning model to suppress background noise from various audio sources with a performance similar to those achieved by past research.

Apr 17 - Milestone 2: By Milestone 2, we would like to evaluate various machine learning model performances. After exploring various models, we would like to attempt to formulate a method of determining the best model for real-time audio.

Apr 24 - Final Check: By the Final Check, we would like to have an improved model with sufficient performance for real-time audio.

Weekly Work	
Feb 17	Project Proposal Write Up
Feb 24	Literature Review
Mar 2	Data Retrieval and Data Manipulation
Mar 9	- Spring Break -
Mar 16	Model Training
Mar 23	Model Training and Evaluation
Mar 30	Midterm Presentation
Apr 6	Research for Performance Improvement
Apr 13	Model Training and Evaluation
Apr 20	Project Presentation
Apr 27	Poster Session
May 4	Final Paper Write Up