Emotion Detection in Speech

by

Milad Shirani

August 2022

Contents

- Project Overview
- Data
- Modeling and Results
- Q & A

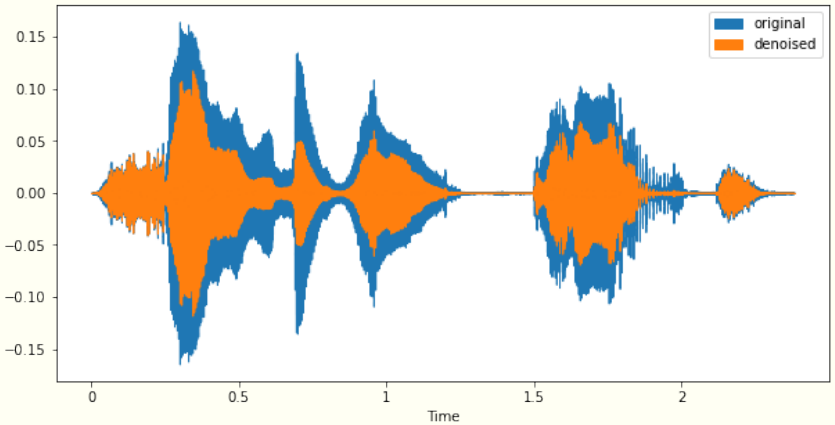
Project Overview

1. In order for an AI to understand a speech, it is important to understand both speech text and the the way each word is said, i.e., the emotion of the sentence.
2. In this work, we want to introduce a new model to detect the emotion of an audio file.
3. We used several categorical models such as Logistic Regression, Decision Tree, Random Forese, XGBoost, LightGBM, as well as Convolutional Neural Network and Transfer Learning such as EfficientNetB3 and EfficientNetB7

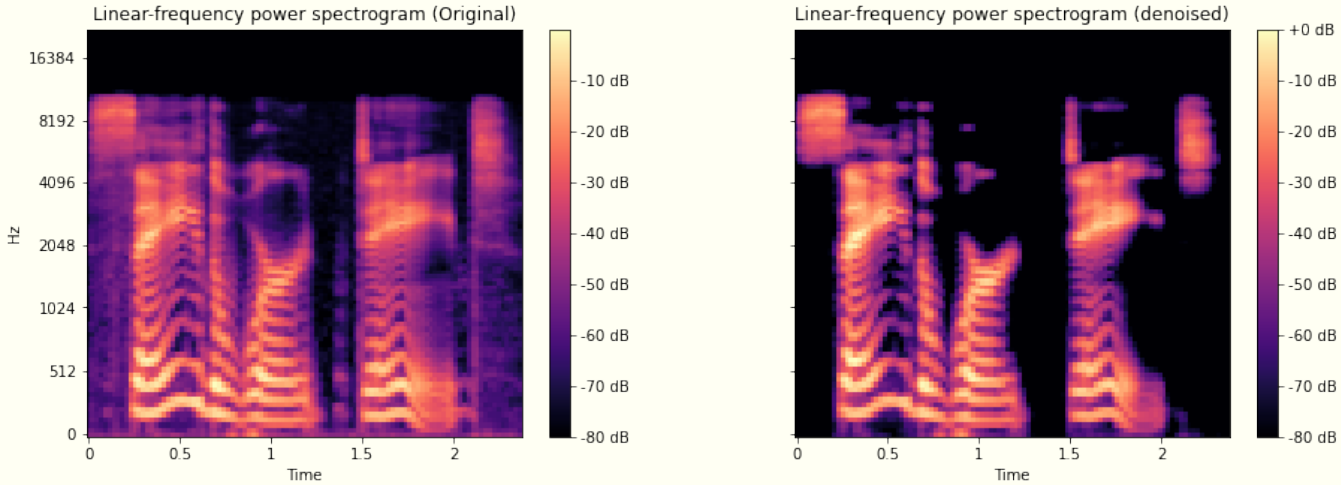
Data and Method

- We use 2800 audio files for our analysis provided by University of Toronto and it contains 7 categories (emotions), which are disgust, surprise, happy, sad, neutral, fear, and angry.
- We converted audio files to numerical values and denoised them and produced mel-spectrograms (from denoised values) using which, we train and evaluate the models.

Effects of Denoising

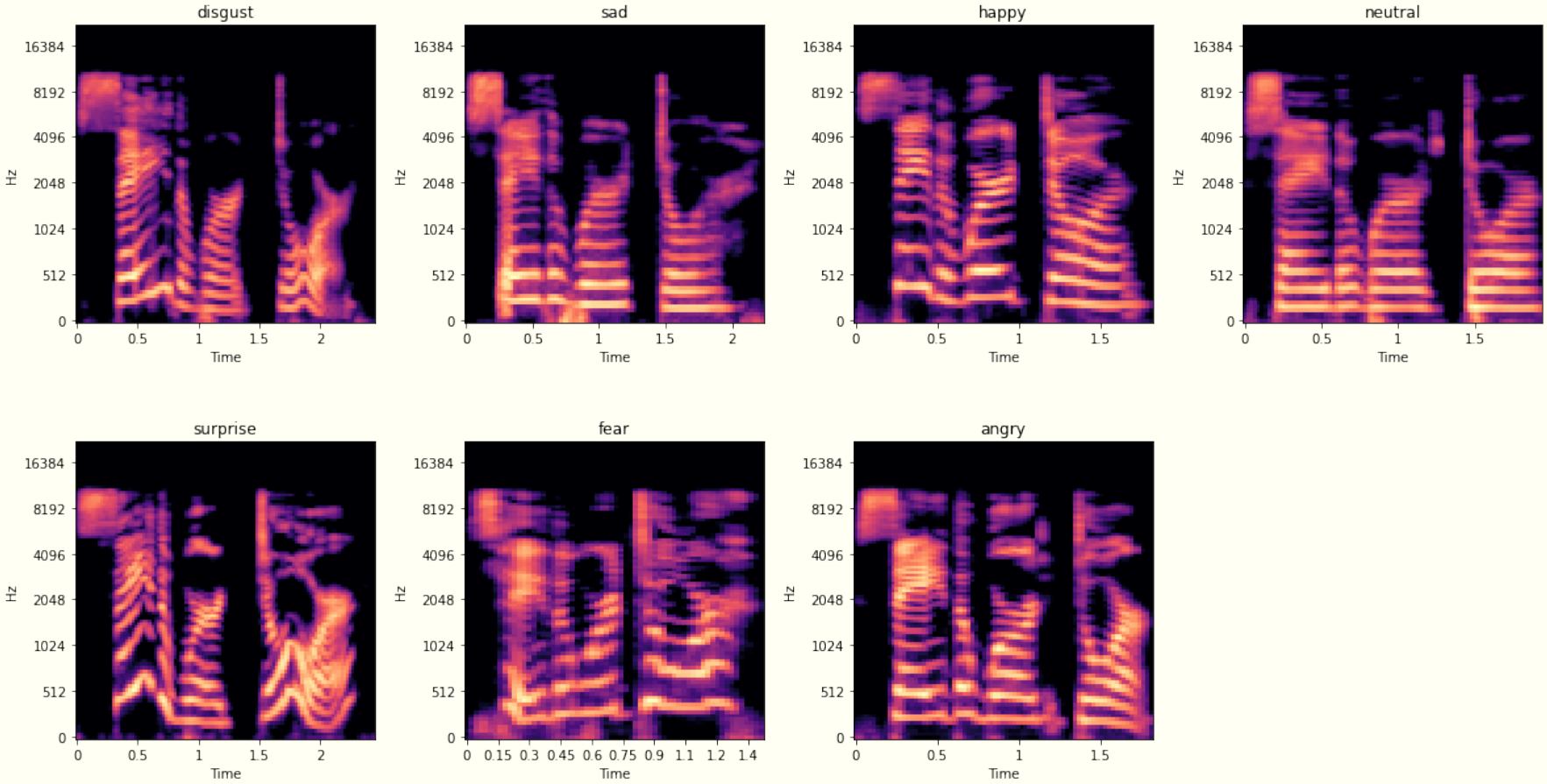


The Original and Denoised of an Audio File



Mel-Spectrograms of an Original and Denoised Audio File

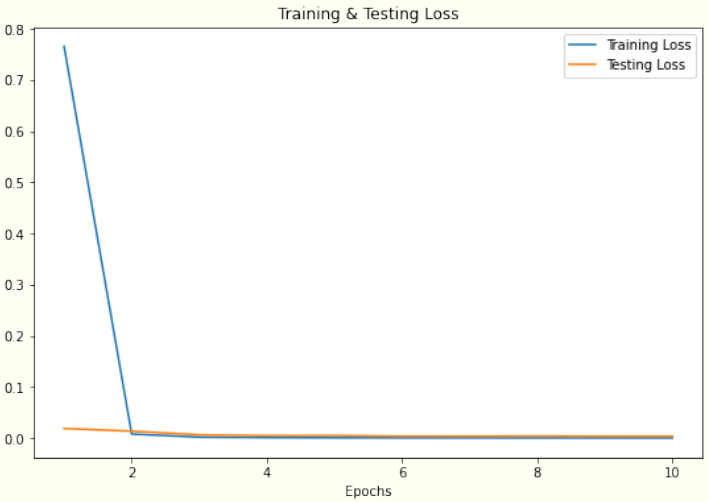
Effects of emotion

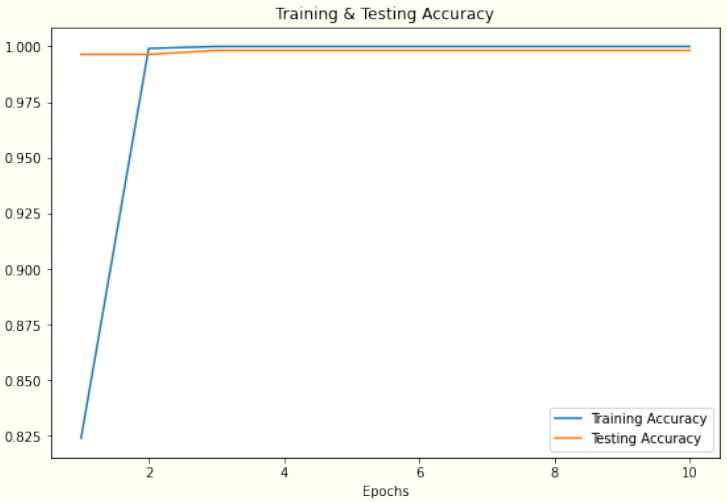


Effects of emotion in saying the word "door"

Modeling and Results

1. All the deep learning models performed well with test accuracy of 99% (expect EfficientNetB7 which has the lowest test accuracy which is about 0.95% after 35 epochs).
2. We would recommend the first CNN model (link to the model) because it has the simplest structure and converges after 2 epochs as shown below





Conclusion

1. The information in speech is conveyed through words and emotion.
2. Depending on how one pronounces a word, we can understand different meaning.
3. It is important that an AI understands both the words in a speech and the emotion of the speech.
4. The final model we introduce has the simplest structure and converges after 2 epochs. It constructed by using Convolutional layer followed by a max pooling layer, followed by a neural network. (link to the model)
5. This model can be implemented by virtual assistant such as Amazon Alexa or Siri.

Next Steps

1. Gathering more data points for training.
2. Deploying neural networks by using LSTM or Conv1d layers and train them on numerical values obtained from audio files.
3. Trying using MFCCs (Mel Frequency Cepstral Coefficients) to train machine learning models.

Q and A