Springboard : Capstone Project 2

Tensorflow Speech Recognition Challenge

Milestone Report

Ravi Maranganti

1. Problem Statement:

Speech recognition is ubiquitous around us. Devices such as Alexa, Siri etc. have permeated our daily lives and speech recognition wherein a device accepts human spoken speech as input to trigger some action is becoming quite the norm. While large companies like Amazon and Apple have the resources to access large amount of speech data and use it to implement speech recognition, this is not so easy for independent makers and entrepreneurs. To address this, TensorFlow recently released the Speech Commands Datasets. It includes 65,000 one-second long utterances of 30 short words, by thousands of different people.

The problem we have at hand is to take this dataset and build a deep learning neural network which can take these audio files as input and classify them into the commands. If a good recognition accuracy can be achieved then we can have a model which can help to improve the effectiveness of voice recognition products .Further, because this is a model built on a dataset which is available freely to download, it can also facilitate accessibility of data and model to entrepreneurs and independent product developers.

The client in this case is anyone who maybe developing a product which needs to take in spoken command or speech as input and uses this input to trigger an action.

2. Data:

2.1 Data Description and Source

Tensorflow group from Google held a Kaggle competition where they provided a rich dataset containing about 64000 audio files which have already been split into training / validation / testing sets. We are then asked to make predictions on about 150000 audio files for which the labels are unknown.

The dataset is available at this URL https://www.kaggle.com/c/tensorflow-speech-recognition-challenge/data

2.2 Data Wrangling

Since we will be using deep learning to solve our problem, it is prudent to convert audio files into images which can then be used as input for training the deep learning model in a supervised learning approach using the labels associated with the audio files. For prediction, the audio files in the test set can also be converted into their image representations and the trained deep learning model can be used to predict their classes. Log Spectrograms or Mel Power Spectrograms of audio are commonly used as input in speech recognition and for this project we can use both and compare the difference in predictions. The dimensions of each audio file needs to be the same to be used for training and prediction for deep learning. While most of the audio files were 1 second long, it was observed that more than 6000

files are less than 1 second long. On the other hand, audio files corresponding to the background noise in the training set are significantly longer. In order to deal with this, shorter audio files were padded with 0s to be 1 second long while background noise files were chopped down to 1 second.

3. EDA

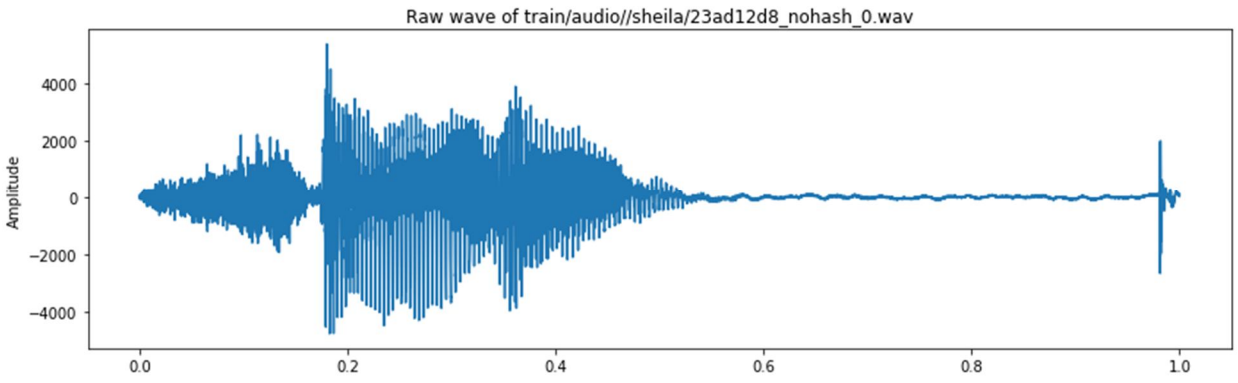
3.1 Preliminary Investigation:

There are 31 labels in the training set. However, we only have to predict if the audio is one among 10 words in a “target list” , noise or of an unknown label. There are 64727 files in the training set while there are 158538 files in the test set which we need to predict. In the training set, 6469 files were seen to be less than 1 second long while 6 files (belonging to background noise label) were seen to be longer than 1 second.

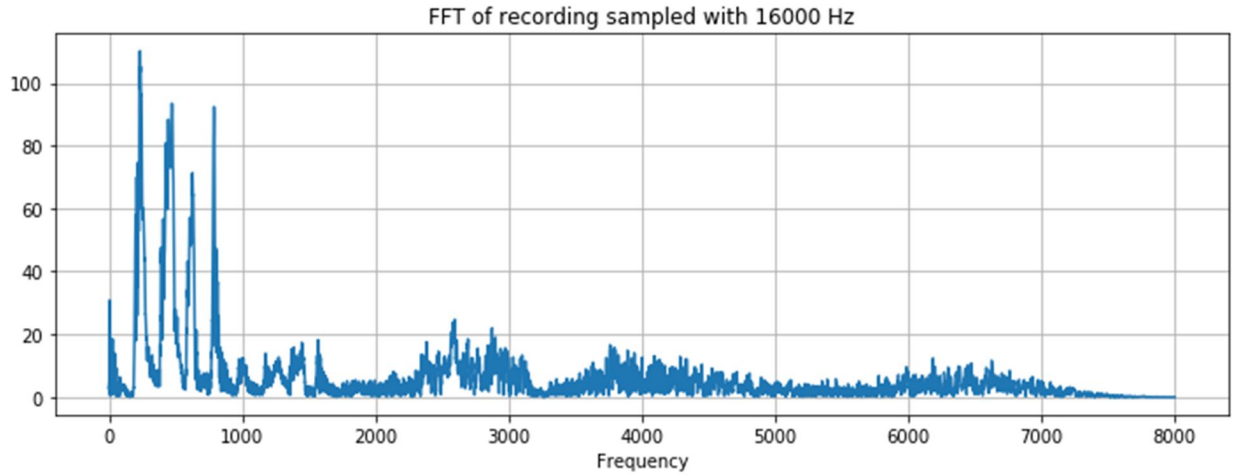
3.2 Visualization:

A file was randomly obtained from the training set and was visualized in different formats.

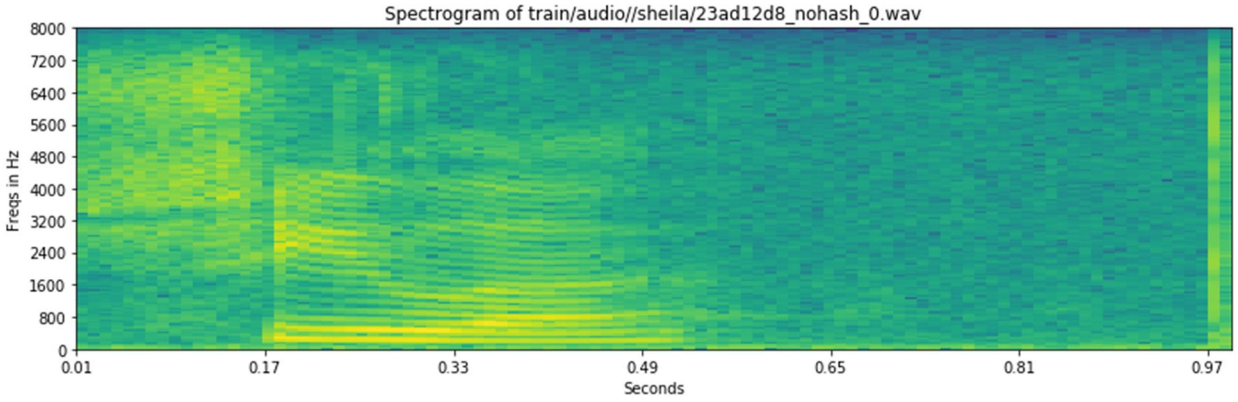
Raw-Wave



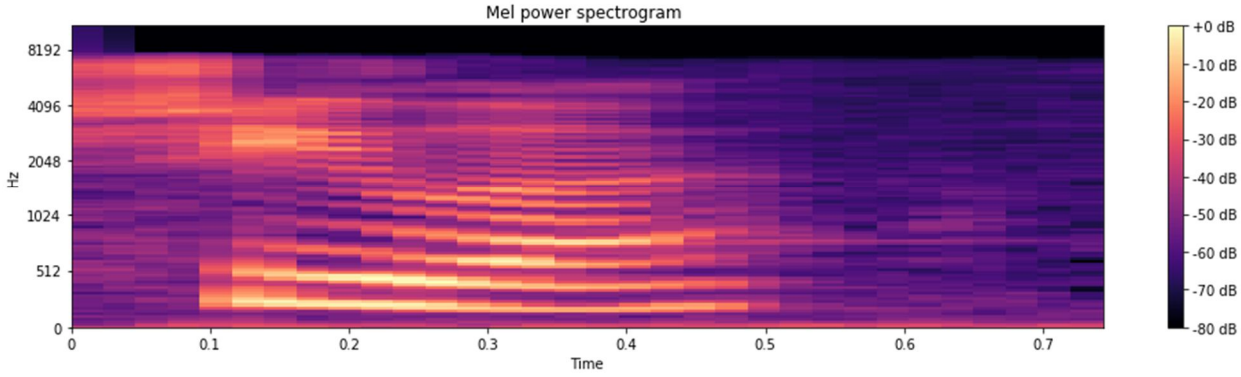
FFT:



Log Spectrogram:



Mel Power Spectrogram:

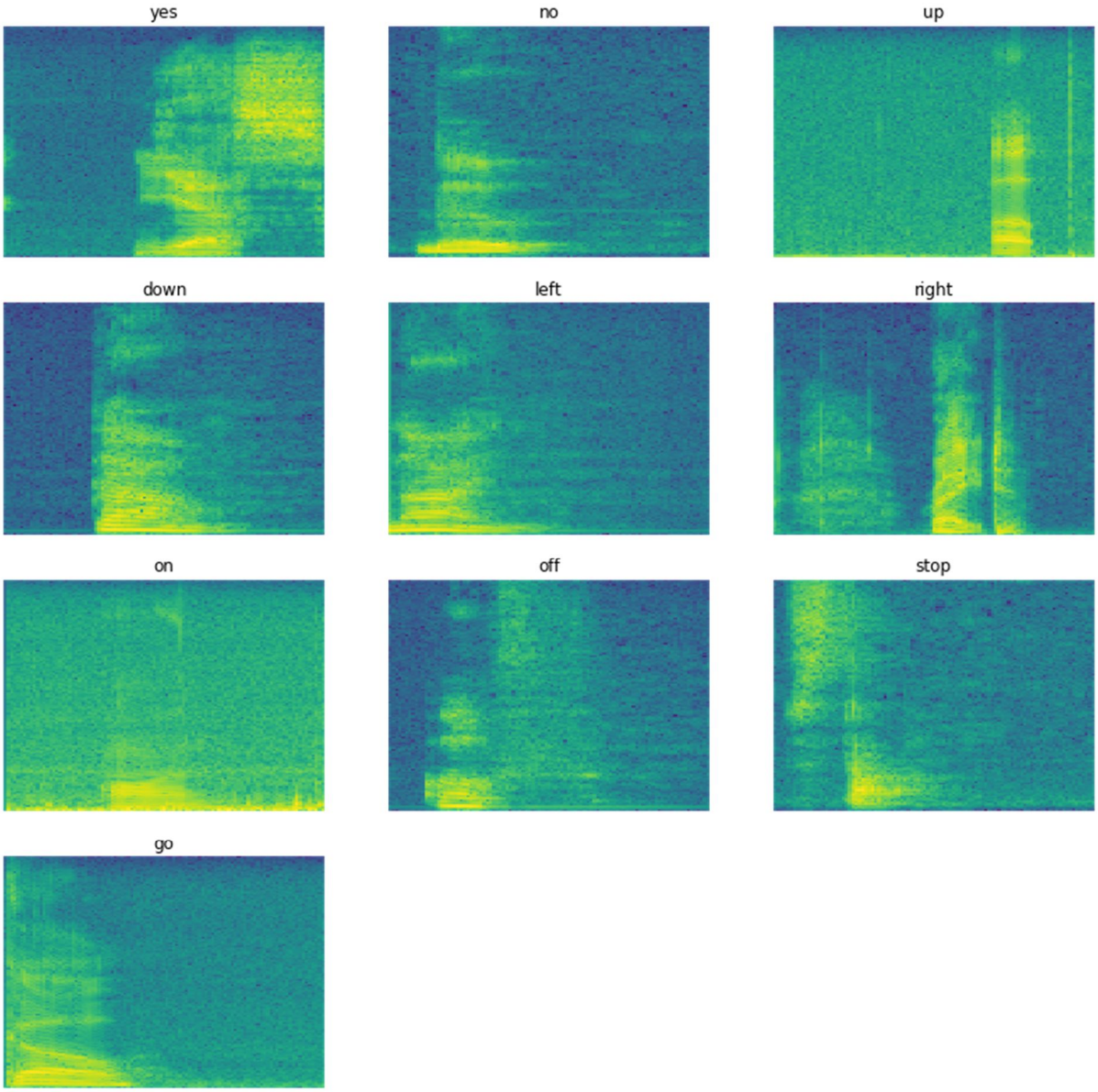


Raw temporal data is almost never used for speech recognition. Log Spectrograms and Mel Power Spectrograms do a better job of representing how human hearing works.

3.3 Comparing Visualization across different Labels:

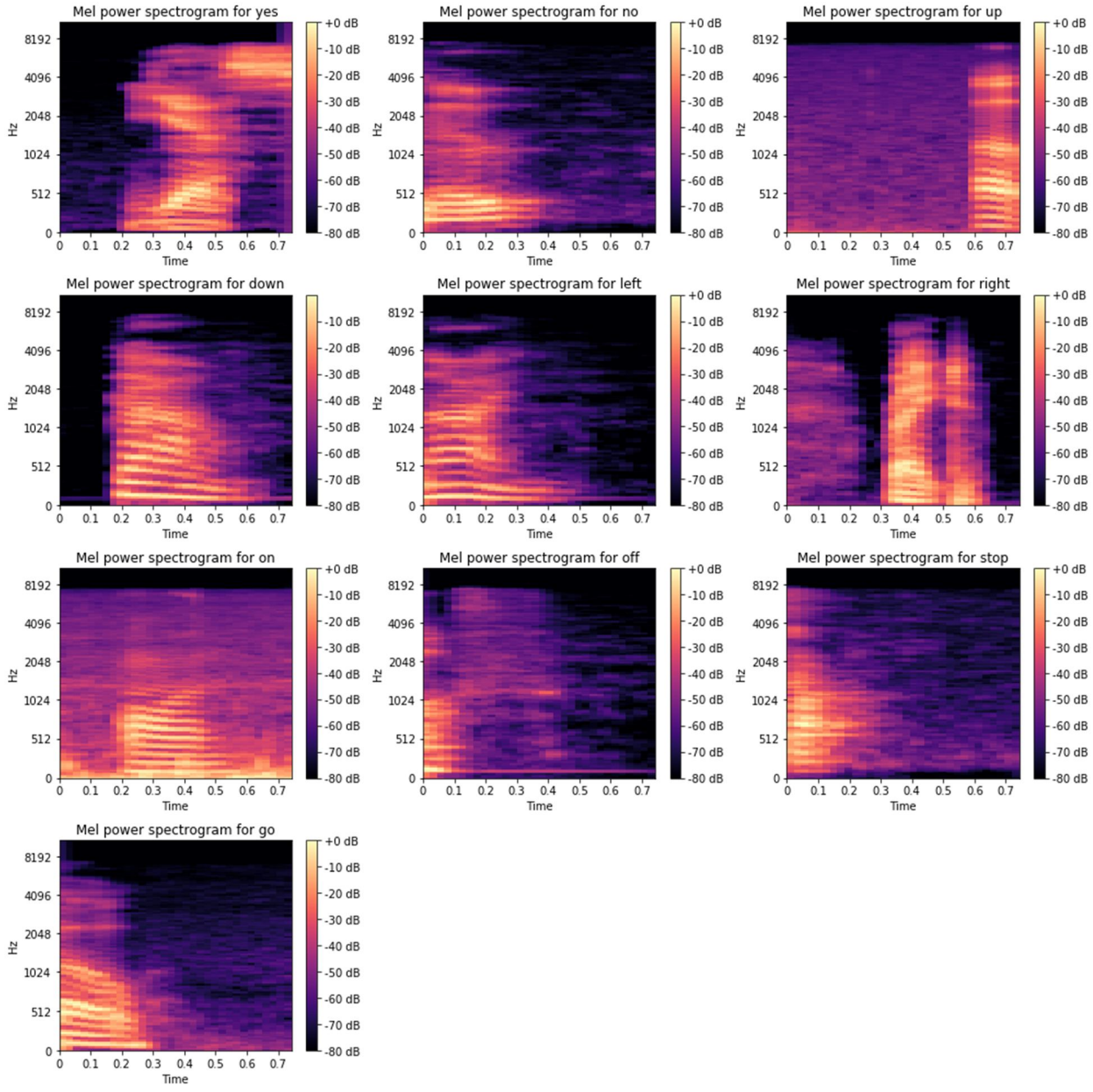
We compare log power spectrograms for 10 randomly picked files from the 10 different classes in the target list.

Log Spectrograms for 10 files belonging to 10 different labels



We do the same for the Mel Power Spectrogram

Mel Power Spectrograms for same 10 files belonging to 10 different labels

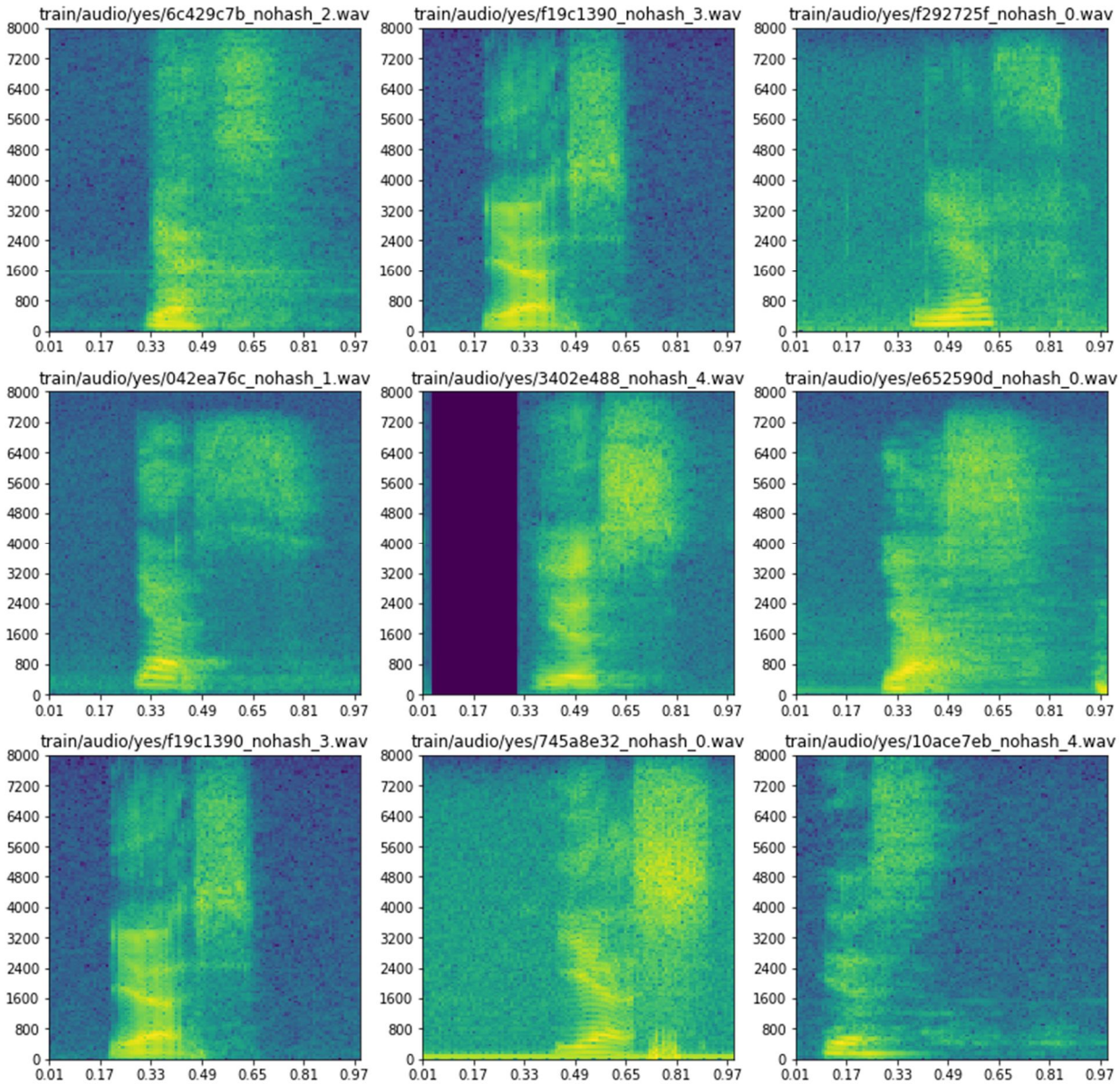


We can see that both these representations capture similar looking patterns.

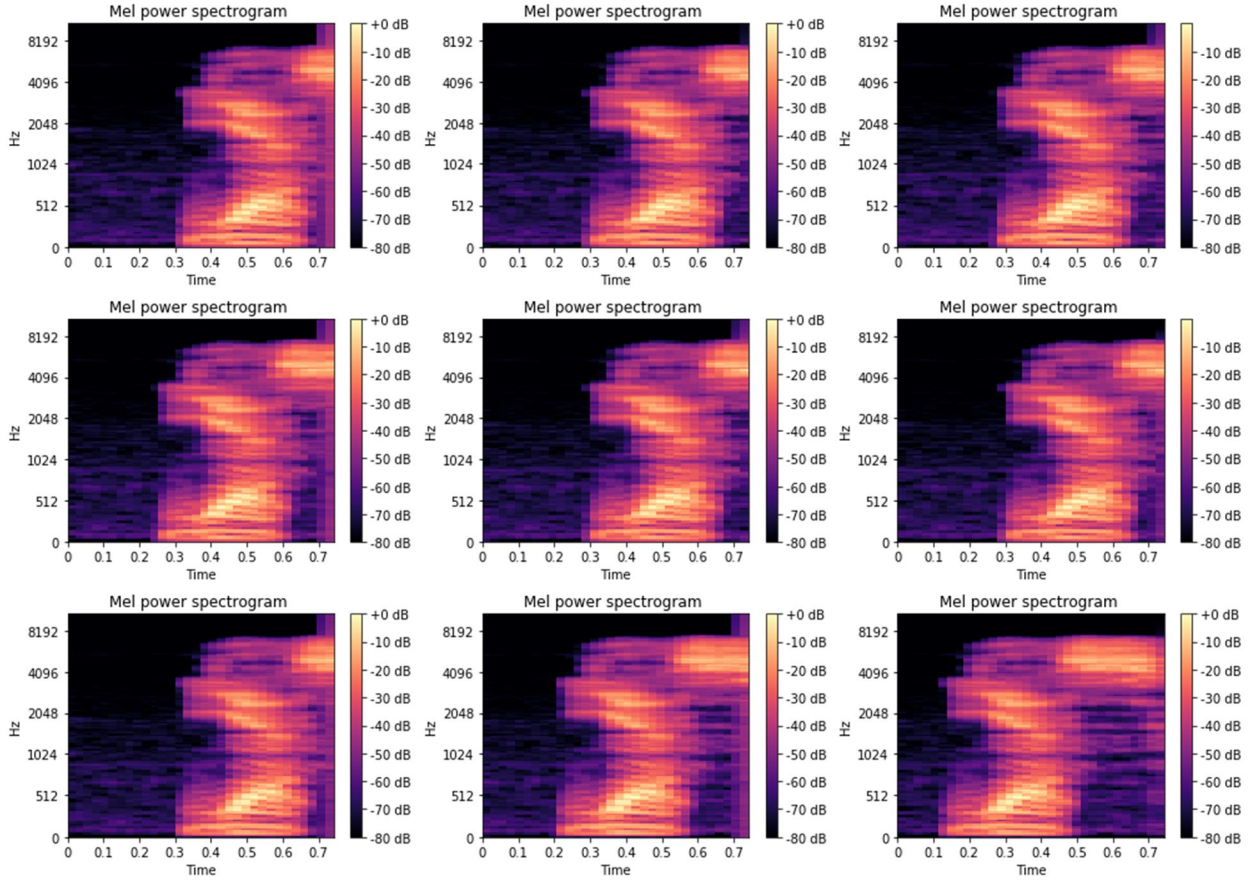
3.4 Comparing Visualization across different files belonging to same label :

Now let us compare the visualizations for 9 randomly chosen audio files with the same label and see how similar (or dissimilar) they look.

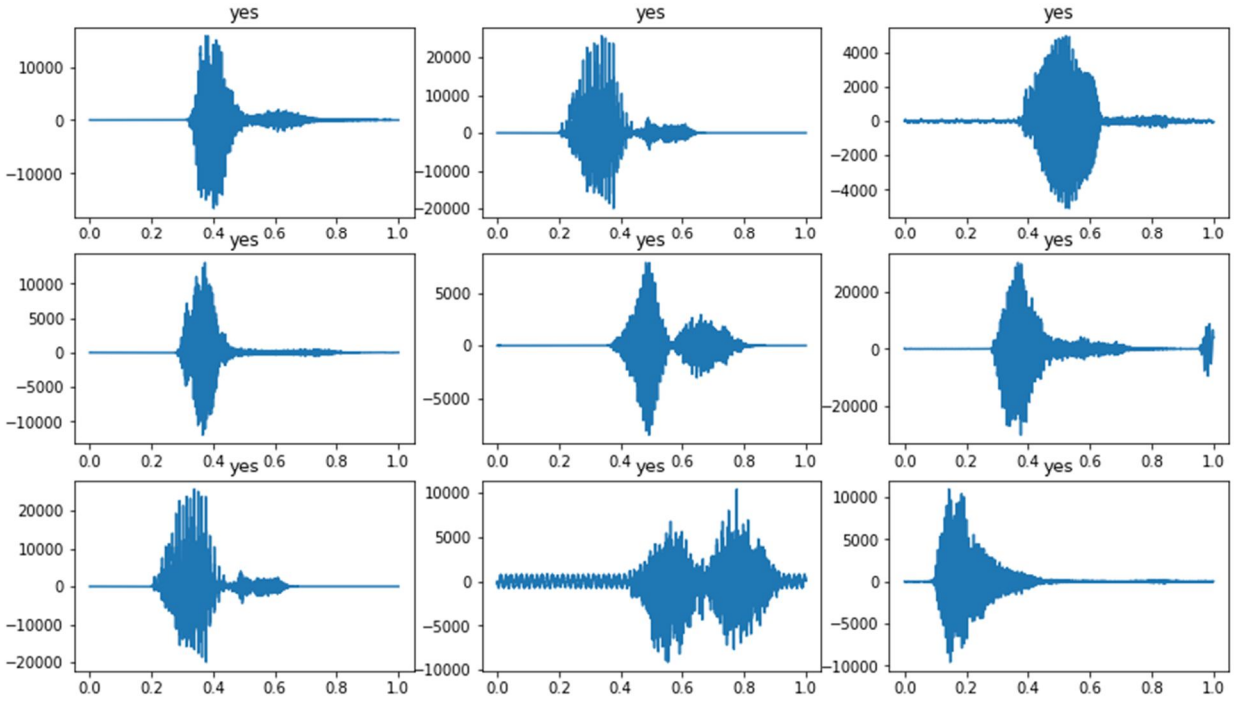
Log Spectrograms of 9 randomly picked files under label “yes”



Mel Power Spectrograms of the same 9 files belonging to label “yes”



Raw wave representations of the same 9 files



4. Conclusion: We can see that the Mel Power Spectrograms seem to do the best and may prove to be the best candidates as input for training our deep learning network.

5. Jupyter Notebook with EDA: https://github.com/ravimaranganti/SpringBoard-CapstoneProject2/blob/master/speech_representation_data_exploration_0829.ipynb

5. References:

1. https://www.kaggle.com/davids1992/speech-representation-and-data-exploration

2. https://en.wikipedia.org/wiki/Place_theory_(hearing
3. https://en.wikipedia.org/wiki/Temporal_theory_(hearing
4. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/