Final Year Project Phase 1 Report
Classification of Whisper and Normal speech signals

Software Tools Used: MATLAB, Python, Teachable machine (online software)

Primary Analysis

From the Wtimit Audio Dataset the whisper and normal audio recordings of one speaker (101) was taken for primary analysis.

Spectrogram images were generated for both whisper and normal audio recordings of this speaker using MATLAB.

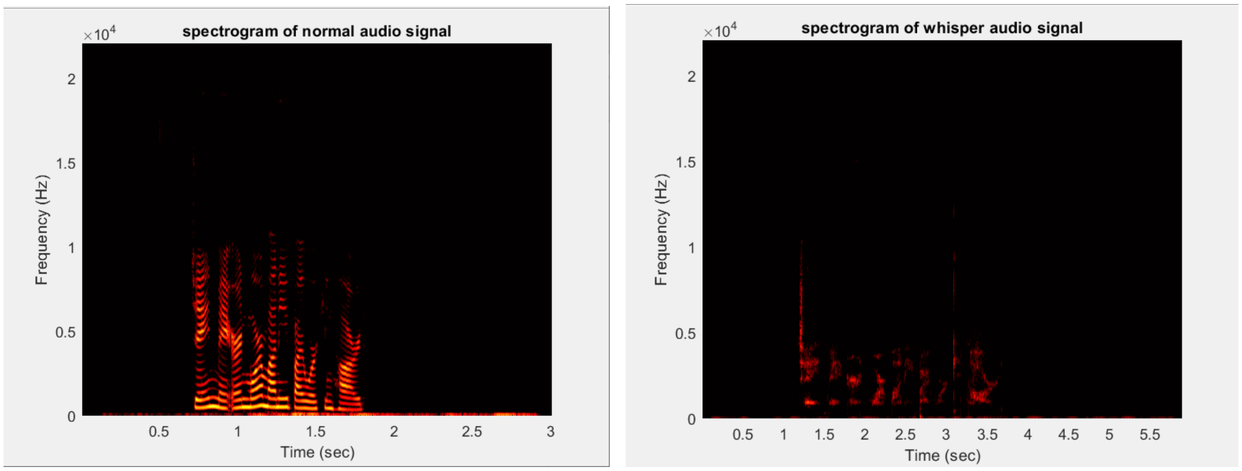
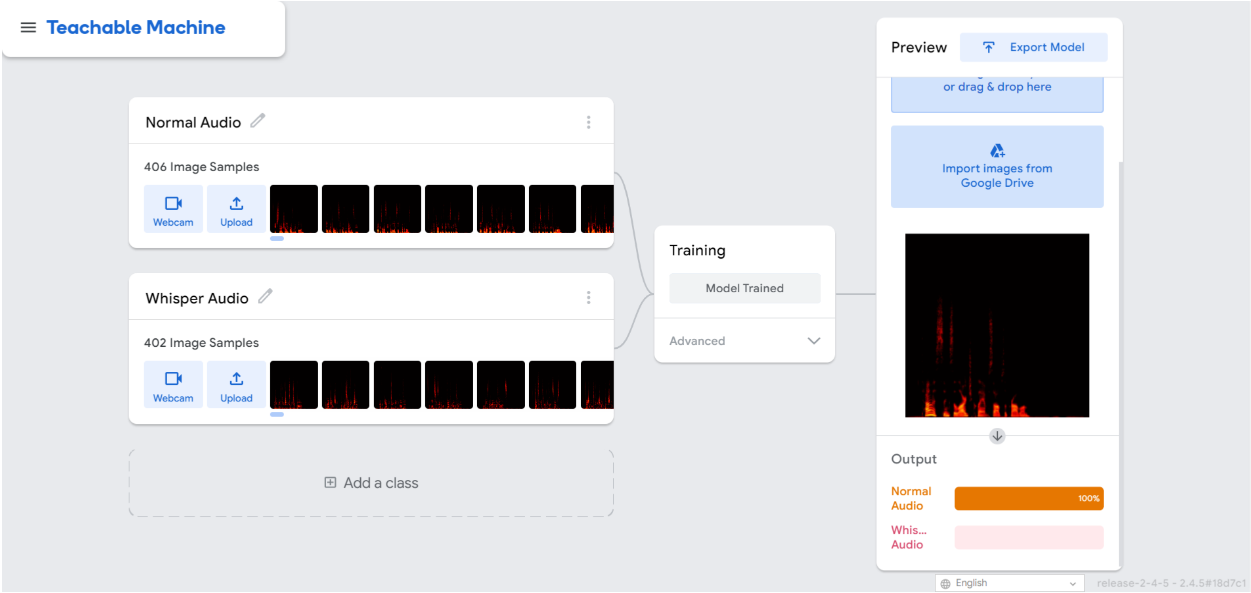


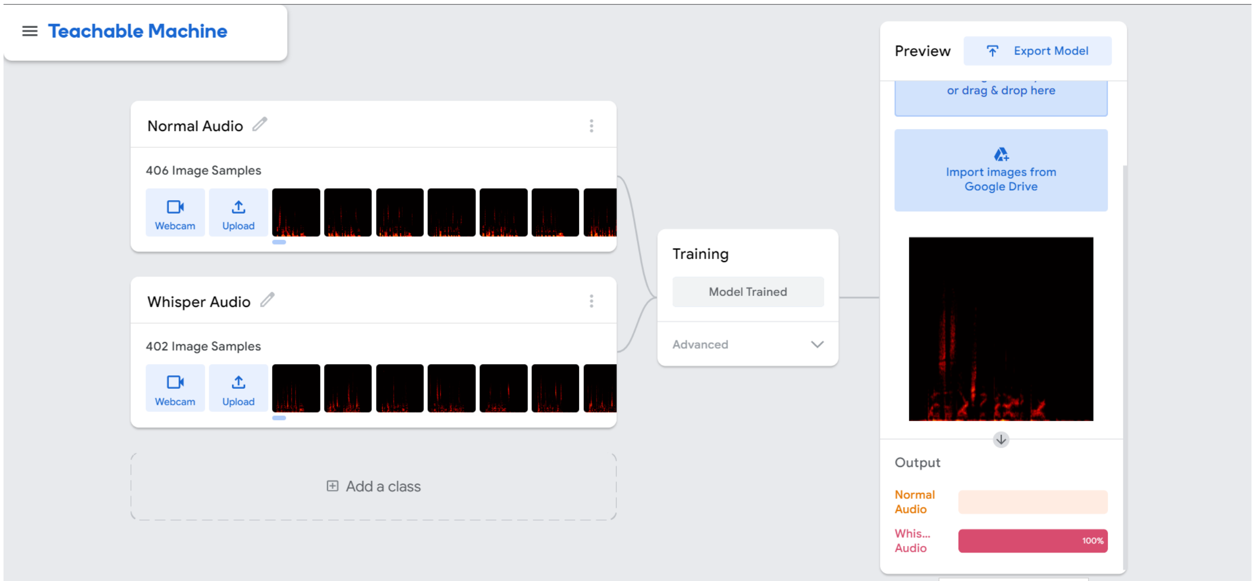
Fig1 spectrogram of normal and whisper audio signals (audio files s101u004n.wav and s101u004w.wav)

Speaker 101 had 422 normal audio files and 423 whisper audio files.
The corresponding spectrogram images of all the normal and whisper audio signals were generated and stored locally and every image had a uniform dimension of 682X539.
These images will now be used to classify normal and whisper audio signals.

To test the actual feasibility of this method, an initial classification was done using the teachable machine online software platform.

Results for teachable machine online platform



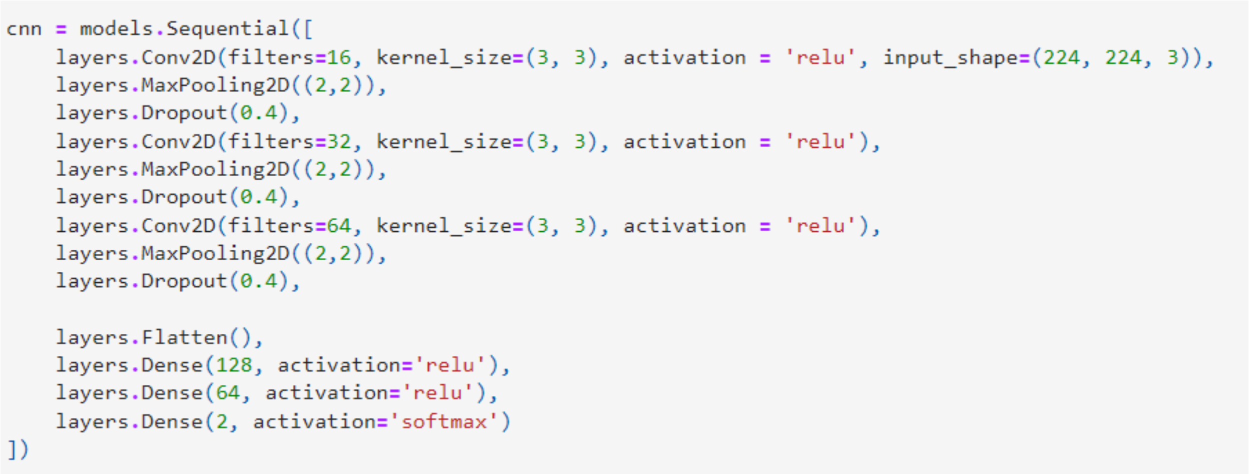


The trained model that was generated by the online platform was able to differentiate between a normal and whisper audio spectrograms.

Building a basic CNN classifier model to train and test the images

A basic CNN classifier model was created on python to train and test the images.

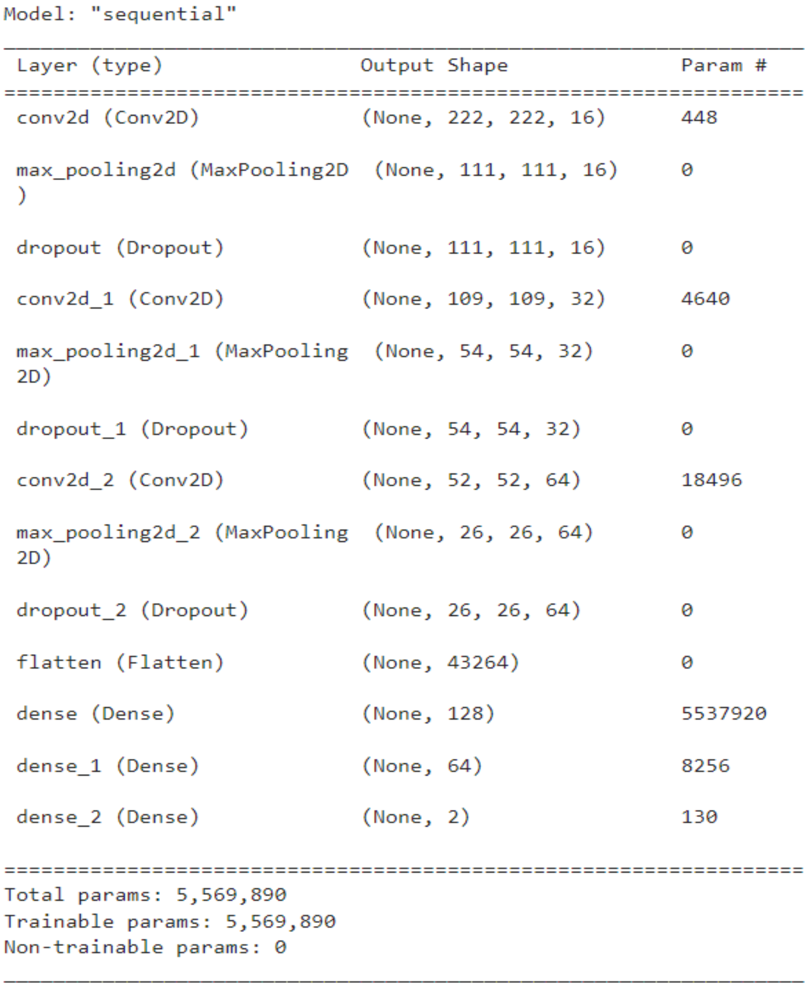
The structure of the model is given below:



```
cnn = models.Sequential([
    layers.Conv2D(filters=16, kernel_size=(3, 3), activation = 'relu', input_shape=(224, 224, 3)),
    layers.MaxPooling2D((2,2)),
    layers.Dropout(0.4),
    layers.Conv2D(filters=32, kernel_size=(3, 3), activation = 'relu'),
    layers.MaxPooling2D((2,2)),
    layers.Dropout(0.4),
    layers.Conv2D(filters=64, kernel_size=(3, 3), activation = 'relu'),
    layers.MaxPooling2D((2,2)),
    layers.Dropout(0.4),

    layers.Flatten(),
    layers.Dense(128, activation='relu'),
    layers.Dense(64, activation='relu'),
    layers.Dense(2, activation='softmax')
])
```

The model accepts an input image of size 224X224 with 3 layers.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 conv2d (Conv2D)             (None, 222, 222, 16)      448       
                                                                 
 max_pooling2d (MaxPooling2D  (None, 111, 111, 16)     0         
 )                                                               
                                                                 
 dropout (Dropout)           (None, 111, 111, 16)      0         
                                                                 
 conv2d_1 (Conv2D)           (None, 109, 109, 32)      4640      
                                                                 
 max_pooling2d_1 (MaxPooling  (None, 54, 54, 32)       0         
 2D)                                                             
                                                                 
 dropout_1 (Dropout)         (None, 54, 54, 32)        0         
                                                                 
 conv2d_2 (Conv2D)           (None, 52, 52, 64)        18496     
                                                                 
 max_pooling2d_2 (MaxPooling  (None, 26, 26, 64)       0         
 2D)                                                             
                                                                 
 dropout_2 (Dropout)         (None, 26, 26, 64)        0         
                                                                 
 flatten (Flatten)           (None, 43264)             0         
                                                                 
 dense (Dense)               (None, 128)               5537920   
                                                                 
 dense_1 (Dense)             (None, 64)                8256      
                                                                 
 dense_2 (Dense)             (None, 2)                 130       
                                                                 
=================================================================
Total params: 5,569,890
Trainable params: 5,569,890
Non-trainable params: 0
_________________________________________________________________
```

From a total of 845 images (422 normal audio spectrograms + 423 whisper audio spectrograms) 760 images were used for training and 85 images were used for testing (90 - 10 percentage split)

The model produced a 100% training and testing accuracy without overfitting.
(All 85 images were predicted accurately)