CS 5500 Capstone: Applications in Data Science
Phase – 2 Project

Audio Recognition using Spoken Digits

By: Simran Girdhar Kodwani, Aishwarya Mathur

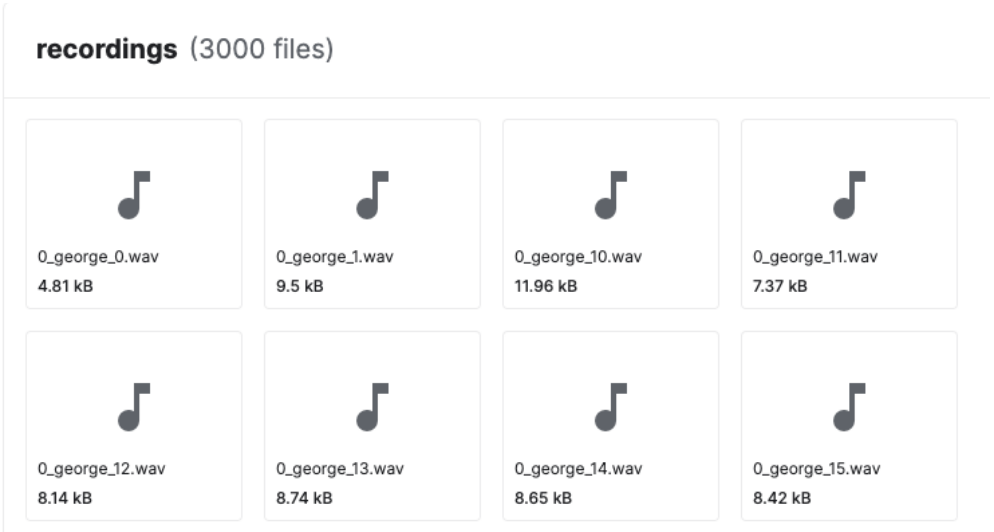
I. Introduction

Sound classification is a widely used application in Audio Deep Learning which involves classifying sounds to predict category of the sound. The project is based on Deep Learning algorithms using audio data. The idea of the project is to take audios of spoken digits and classify the digits. Artificial Neural Networks (ANN) is used to implement a classification model which can be a strong baseline network for audio classification. The model will be evaluated using metrics such as accuracy, precision, f1 score and confusion matrix which can show if the digits were predicted correctly.

There are several applications which can be based on audio recognition such as translation of spoken words to text, generating captions for hearing impaired people, voice commands for devices like Alexa, Siri, Google Assistant, and even hands-free communication. Therefore, implementing a good audio recognition model can help in advanced technology as well.

II. Dataset

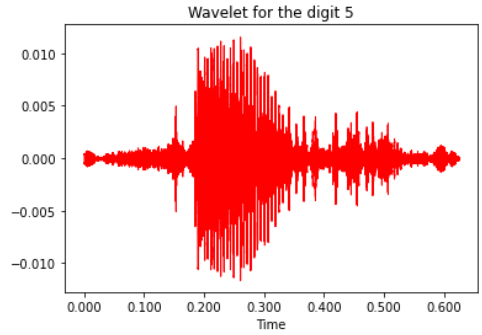
The dataset used for the project is Audio MNIST dataset from Kaggle. The data contains 30,000 audio samples of spoken digits (0-9) of 60 different speakers saying the digits out loud where each digit has 3000 audio samples. The recordings are one-second long and are trimmed so that they have near minimal silence at the beginnings and ends. The relevant features will be extracted from the audio signals to perform speech recognition.



III. Methods

1. Exploratory Data Analysis

The audio data is processed into numerical data to perform exploratory data analysis on audio samples using python library Librosa. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. It helps visualize audio signals and do feature extraction using different signal processing techniques.



2. Pre-processing Audio Data

The audio data was taken as input and features were extracted using Mel frequency cepstral coefficients. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.

50 features were extracted and then scaled. The features extracted looked like the following,

```
dataset['features'][100]
```

```
array([-6.1741779e+02,  8.9915787e+01, -3.9582746e+00,  5.1702354e+01,
        3.2447968e+01, -3.9643118e+00, -6.6765246e+00, -1.2234051e+01,
        1.1816156e+00,  5.2324114e+00, -6.9705386e+00, -5.9391969e-01,
        1.2659647e+00,  7.0302029e+00, -1.2036294e+00,  1.2792585e+01,
       -3.5641048e+00,  1.2185696e+01, -1.1814073e+01,  6.3068676e+00,
       -4.9240613e+00,  3.3437555e+00, -2.1692574e+00,  1.3817080e+00,
       -4.3500113e+00,  4.0364945e-01, -5.4696403e+00,  4.4703975e+00,
        3.3066843e-02, -1.8820531e+00,  3.5944901e+00,  1.5745727e+00,
        9.3516058e-01,  1.8571689e+00,  4.4885021e-01,  1.4613475e+00,
       -2.6095958e+00,  3.4116940e+00, -8.6650878e-01,  1.9677911e+00],
      dtype=float32)
```

3. Artificial Neural Network Model

A feed forward neural network was used to classify digits 0-9. The model had one input layer, two hidden layers and one output layer. It used Rectified Linear Unit, '*relu*' as an activation function for the input and hidden layers and '*softmax*' activation function in the output layer since it is a classification problem. The overall model structure was as below,

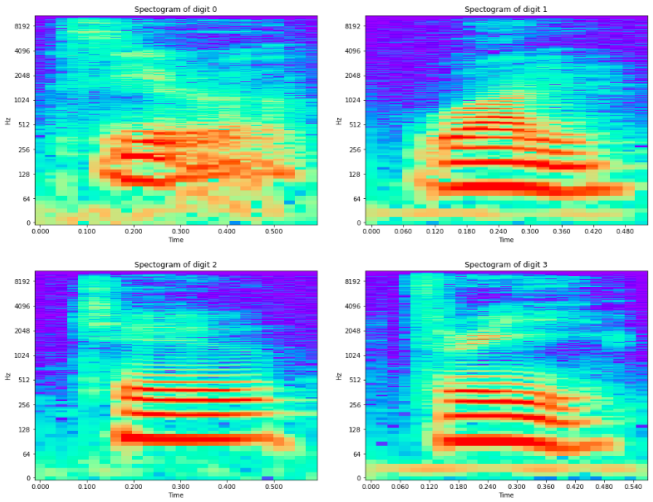
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 100)               4100

 dense_1 (Dense)             (None, 100)               10100

 dense_2 (Dense)             (None, 100)               10100

 dense_3 (Dense)             (None, 10)                1010

=================================================================
Total params: 25,310
Trainable params: 25,310
Non-trainable params: 0
_________________________________________________________________
```

4. Evaluations

The results were evaluated using the testing data and metrics such as precision, recall, f1 score and confusion matrix which helped in to evaluating if the ANN model correctly predicted the digits spoken in the audio input.

IV. Results and Discussion

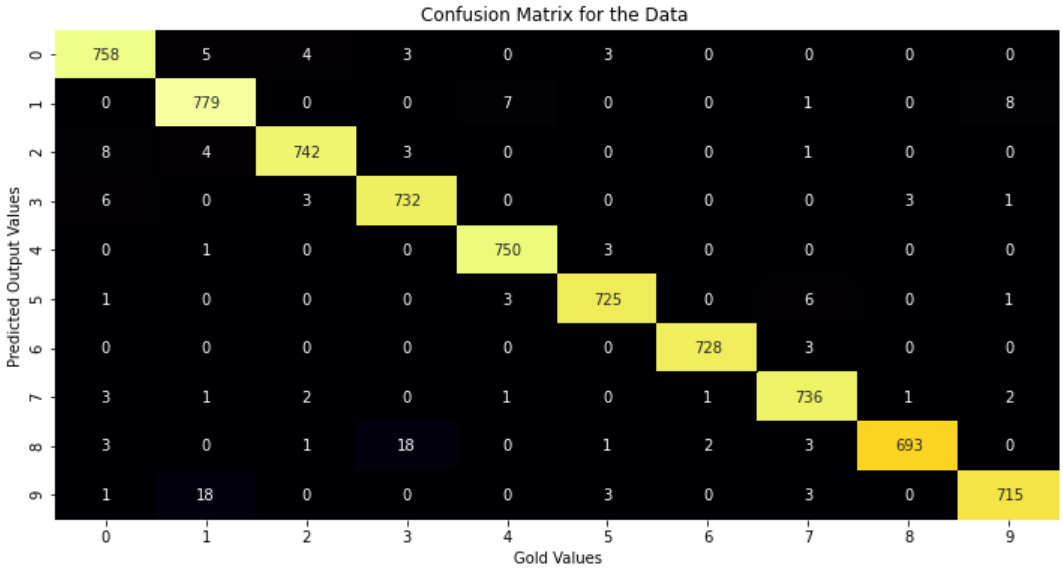
Spectrogram, a visual representation of waveform of signal strength or "loudness" overtime at various frequencies was analysed. The below figures show the spectrogram for digits 0,1,2 and 3. Spectroram is a visual representation if the spectrum of frequencies of a signal as it varies with time and can be depicted with heat map. It can help in distinguish sound elements in a recording and their harmonic structure. Louder events in the audio are indicated with brighter colors in the plot and quieter events are indicated by darker colors in the plot.



After training the model on the training data with a batch size of 128 data points and 100 epochs, the model achieved an accuracy of 100% on the training set and 98.27% on the validation set. The classification report of the model was the following,

	precision	recall	f1-score	support
0	0.97	0.98	0.98	773
1	0.96	0.98	0.97	795
2	0.99	0.98	0.98	758
3	0.97	0.98	0.98	745
4	0.99	0.99	0.99	754
5	0.99	0.99	0.99	736
6	1.00	1.00	1.00	731
7	0.98	0.99	0.98	747
8	0.99	0.96	0.98	721
9	0.98	0.97	0.97	740
accuracy			0.98	7500
macro avg	0.98	0.98	0.98	7500
weighted avg	0.98	0.98	0.98	7500

The below confusion matrix shows the actual and predicted values by the model for all the digits. We can see that digits 8 and 3, and 1 and 9 were wrongally predicted. One of the causes of wrong prediction can be diffence of accent of various speakers.



V. References

[1] Si, Shijing, et al. "Variational Information Bottleneck for Effective Low-Resource Audio Classification." *Interspeech 2021*, ISCA, Aug. 2021.

[2] Korkmaz, Yunus, and Aytuğ Boyacı. "milVAD: A Bag-level MNIST Modelling of Voice Activity Detection Using Deep Multiple Instance Learning." *Biomedical Signal Processing and Control*, vol. 74, Elsevier BV, Apr. 2022, p. 103520.

[3] Sören Becker, et al. "Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals." *arXiv: Sound*, July 2018.

[4] Khandelwal, Renu. "Deep Learning Audio Classification - Analytics Vidhya." *Medium*, 22 Dec. 2021, *medium.com/analytics-vidhya/deep-learning-audio-classification-fcbed546a2dd.*