EEE 598/SHS 598
Speech and Audio Production, Processing and Perception

Speech and Speaker Recognition System

(Recognize the word and the speaker)

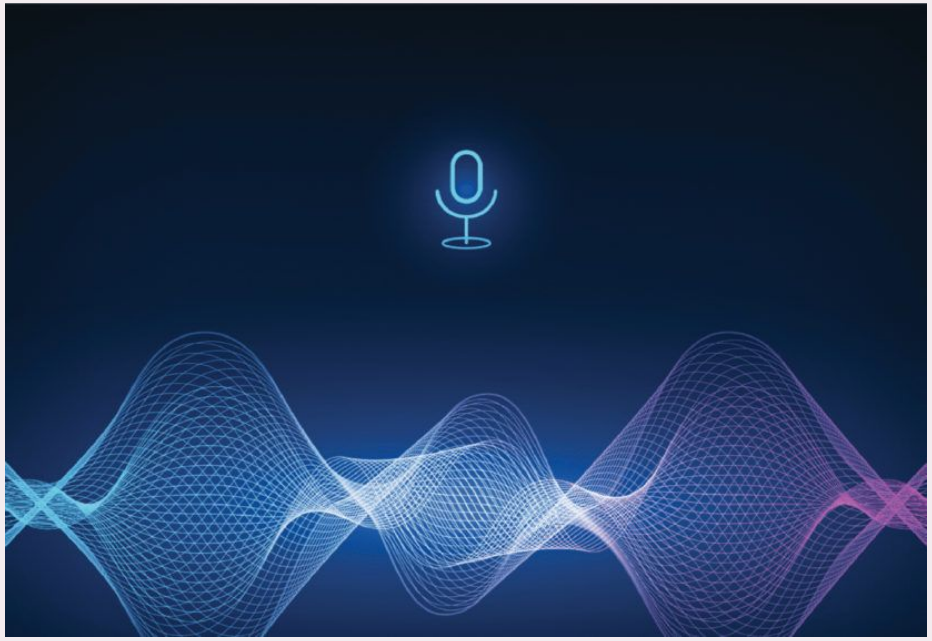
By :
Aradhita Sharma, Carole Pearsall

Supervisor : Visar Berisha

Contents

1 Introduction

This project aims to build a speech recognition system along with speaker recognition and identification system using speaker specific speech features.



Source : https://recfaces.com/articles/what-is-voice-recognition

2 Motivation

Problem Statement

The main motivation for this project is to implement voice fingerprinting, a biometric way of identification, along with recognizing the word said.

This project is oriented towards learning the algorithm of speech and voice recognition.

Recognize the digits said by speakers, and identify the speaker.

3 Reference and Dataset

1. Reference :

Reference Paper : Fang-Yie Leu, Guan-Liang Lin, “An MFCC-based Speaker Identification System”, IEEE 31st International Conference on Advanced Information Networking and Applications, pp 1055-62, 2017 (https://ieeexplore.ieee.org/document/7921023)

2. Dataset :

- The dataset used consists of 10 digits (0-9) spoken by Aradhita, Carole (50 samples each). Similarly, audio files of the digits by George, Nicolas, Jackson, Theo and Yweweler are taken from the dataset present in the link below..
- Data is segregated into different training and testing folders.

Dataset : https://github.com/Jakobovski/free-spoken-digit-dataset

4 Contributions

1

Carole Pearsall

Organised dataset and processing

Feature Extraction

2

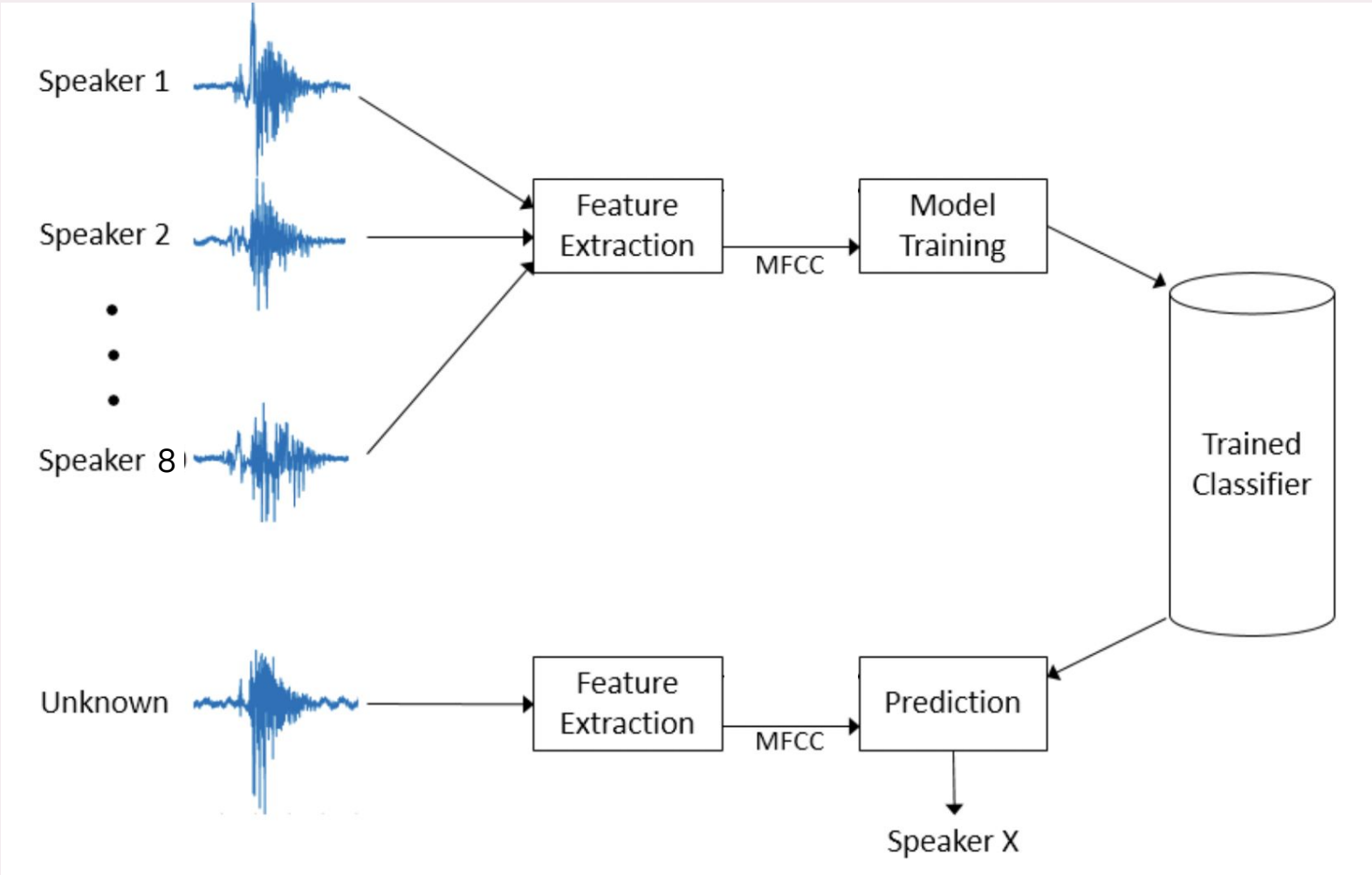
Aradhita Sharma

GMM model and DNN model

Classifier results

Methodology

Block Diagram :



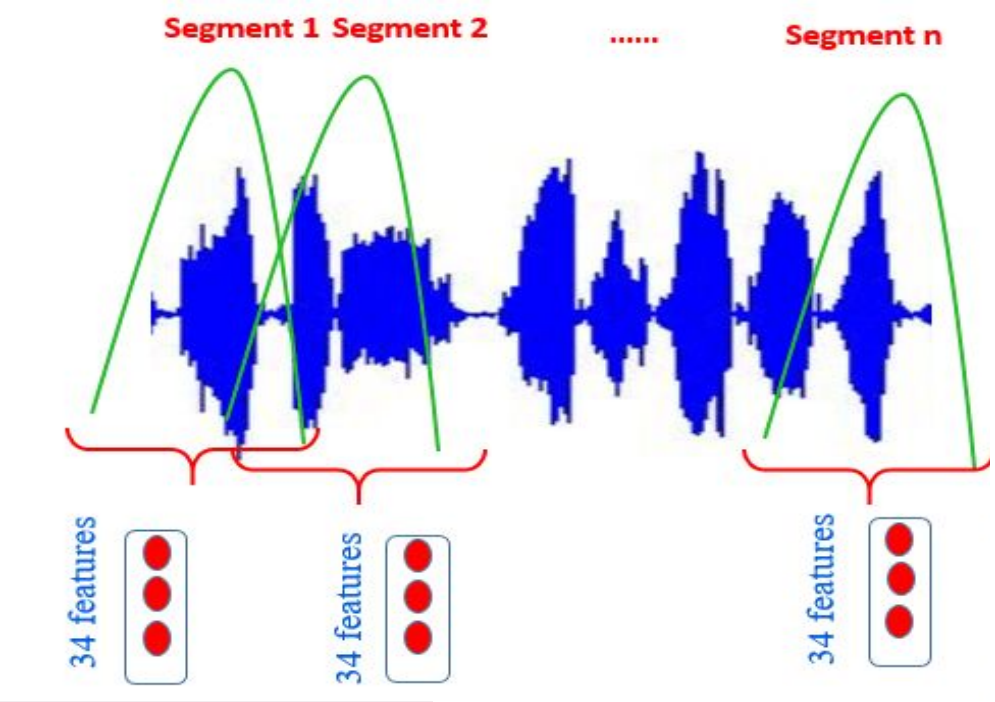
Source : https://www.mathworks.com/help/audio/ug/speaker-identification-using-pitch-and-mfcc.html

1. Feature Extraction

- Various lengths of audio signals are aligned in one piece
- Speech file is converted into numbers using sampling

Features of Speech signal are :

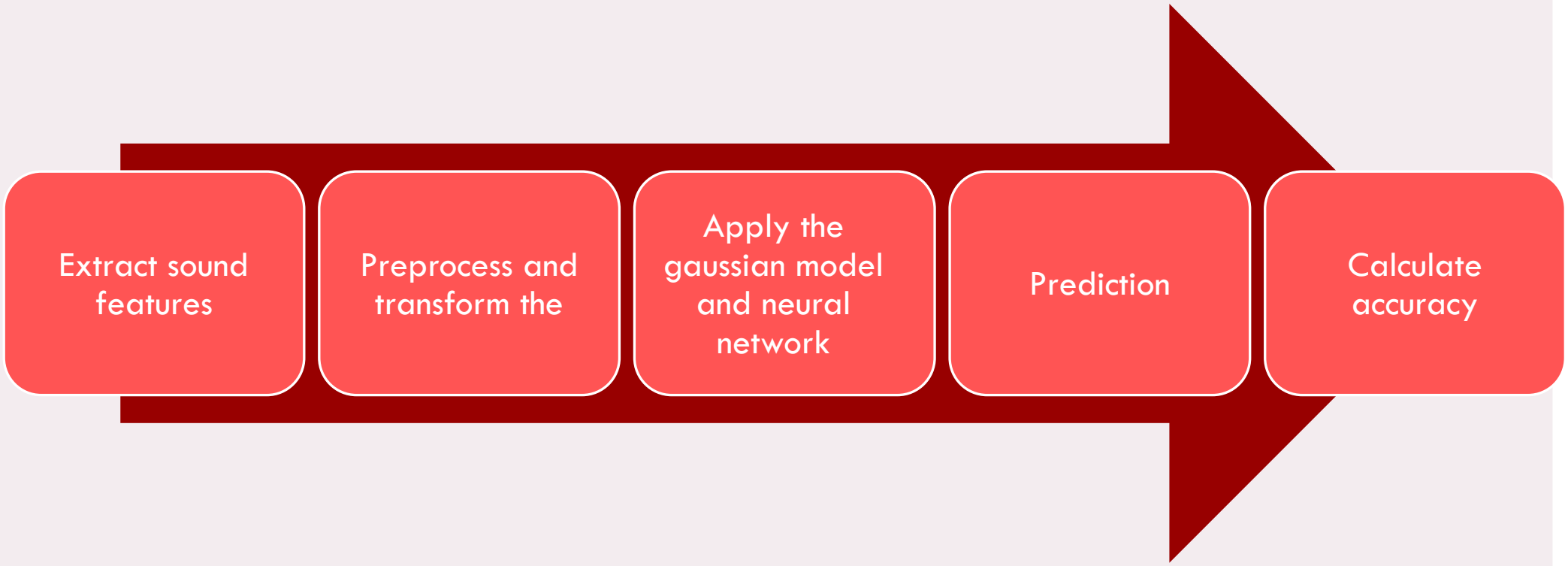
- Energy / RMSE
- Zero Crossing Rate
- Spectral Centroid
- Spectral Roll off
- MFCC (Mel-Frequency Cepstral Coefficients)



2. Data Pre Processing

- Sound features extraction using librosa
- Reading through the files in training set and appending to train matrix
- Data Classification and labelling (0_aradhita_12.wav)
- Data Scaling of the feature set
- Divide the dataset into training, validation and test set
- Text dependent speaker identification

Work Flow :


Extract sound features
Preprocess and transform the
Apply the gaussian model and neural network
Prediction
Calculate accuracy

5 Models :

GMM

Gaussian Mixture Model

Gmm =
mixture.GaussianMixture(
n_components = 10,
covariance_type = “full”)
gmm.fit(features)
Stored as pickle file

Two Approaches For Acoustic model

DNN

Deep Neural Networks

Denselayer(256, relu)
Dropout(0.5)
Denselayer(128, relu)
Dropout(0.5)
Denselayer(64m relu)
Dropout(0.5)
Denselayer(10,softmax)
Adam optimizer, sparse categorical entropy loss

7 Results

When training dataset is 49 samples of each digit spoken by each speaker, we get 100% accuracy for the speaker's case, and better accuracy for digit recognition in neural network model case.

https://colab.research.google.com/drive/13x2DKGbfDliTb85b-eZXGUzfe3e8Llae?usp=sharing
(ACspeakerRecognition_fulldata)

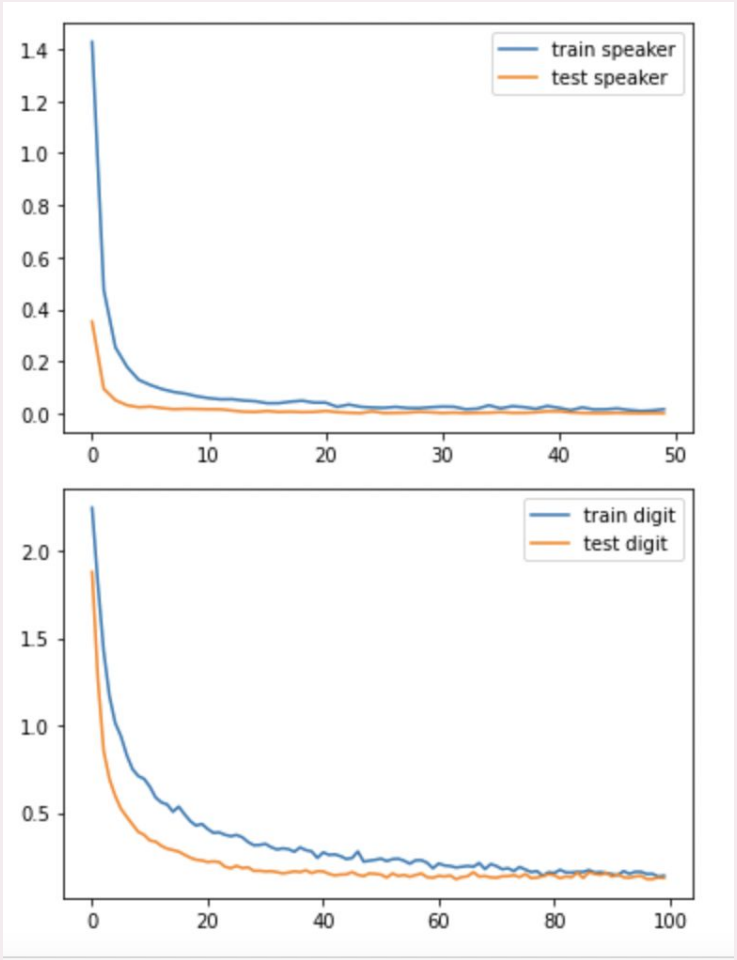
```
percentageCorrect_speaker = (nCorrect_speaker/(len(X_test)))*100
print('Accuracy of speaker with GMM-MFCC is ', percentageCorrect

percentageCorrect_digit = (nCorrect_digit/(len(X_test)))*100
print('Accuracy of digit with GMM-MFCC is ', percentageCorrect_d
```

```
Accuracy of speaker with GMM-MFCC is  100.0 %
Accuracy of digit with GMM-MFCC is  90.0 %
```

Accuracy results with GMM model

```
loss: 0.0162 - accuracy: 0.9958 - val_loss: 7.8918e-04 - val_accuracy: 1.0
```




```
loss: 0.1403 - accuracy: 0.9589 - val_loss: 0.1268 - val_accuracy: 0.9717
```

```
# DNN-Accuracy for Speaker #

3/3 [==============================] - 0s 4ms/step - loss: 0.0040 - accuracy: 1.0000
accuracy: 100.00%

# DNN-Accuracy for Digit #

3/3 [==============================] - 0s 4ms/step - loss: 0.1709 - accuracy: 0.9625
accuracy: 96.25%
```

Accuracy results with DNN model

7 Results

When training dataset is reduced to six samples of each speaker's each digit, we get reduced accuracy.

However, it can be seen that accuracy with neural networks is more than GMM model.

https://colab.research.google.com/drive/1MysVBYTzSLMFwpmla2bydT7pO-61QB5p?usp=sharing
(ACspeakerRecognition_reduceddata)

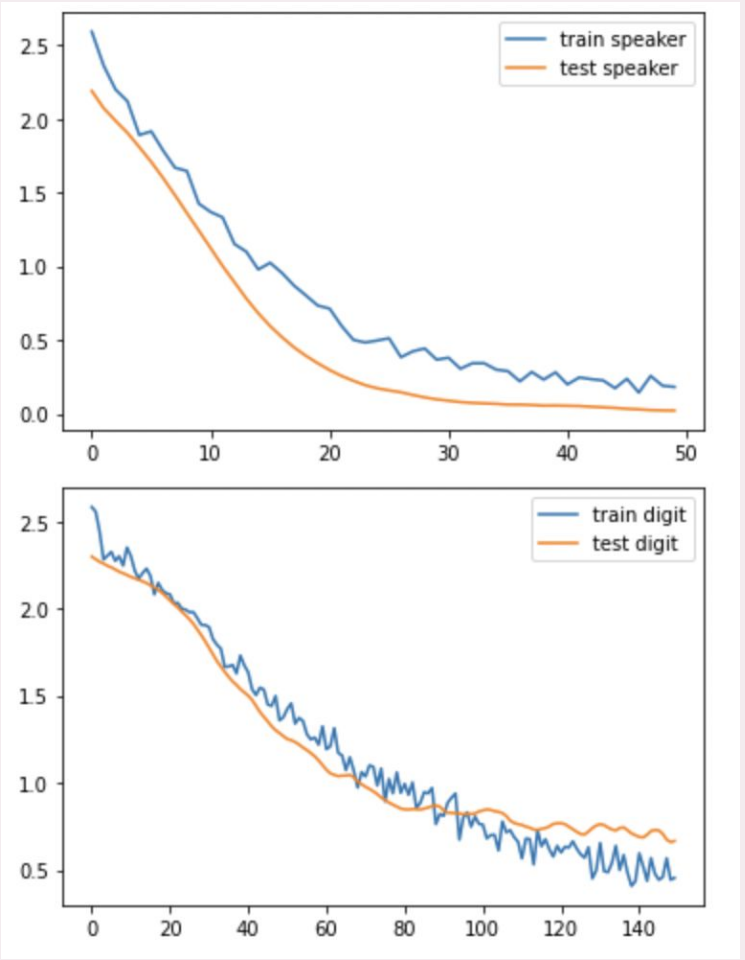
```
percentageCorrect_speaker = (nCorrect_speaker/len(X_test))*1
print('Accuracy of speaker with GMM-MFCC is ', percentageCor

percentageCorrect_digit = (nCorrect_digit/len(X_test))*100
print('Accuracy of digit with GMM-MFCC is ', percentageCorre
```

```
Accuracy of speaker with GMM-MFCC is  96.25 %
Accuracy of digit with GMM-MFCC is  75.0 %
```

Accuracy results with GMM model

```
loss: 0.1815 - accuracy: 0.9431 - val_loss: 0.0211 - val_accuracy: 1.0000
```




```
loss: 0.4545 - accuracy: 0.8293 - val_loss: 0.6662 - val_accuracy: 0.8065
```

```
# DNN-Accuracy for Speaker #

3/3 [==============================] - 0s 4ms/step - loss: 0.0532
accuracy: 97.50%

# DNN-Accuracy for Digit #

3/3 [==============================] - 0s 4ms/step - loss: 0.6659
accuracy: 81.25%
```

Accuracy results with DNN model

8

Conclusions

Our project has included the text dependent speaker recognition and speech recognition implementation.

From the results, it is clear that neural network model yield better results as compared to just MFCC features comparison with the GMM model stored.

More training dataset yield better accuracy.

Future work may include implementation of the system in the real time, that is, identifying the person's voice and word spoken in real time.

Thank you!