SPEECH PREDICTIONS - ML USING KERAS -

03-DEC-2019

GROUP 6 : RODOLFO / HAITHAM / ANKUR

AGENDA

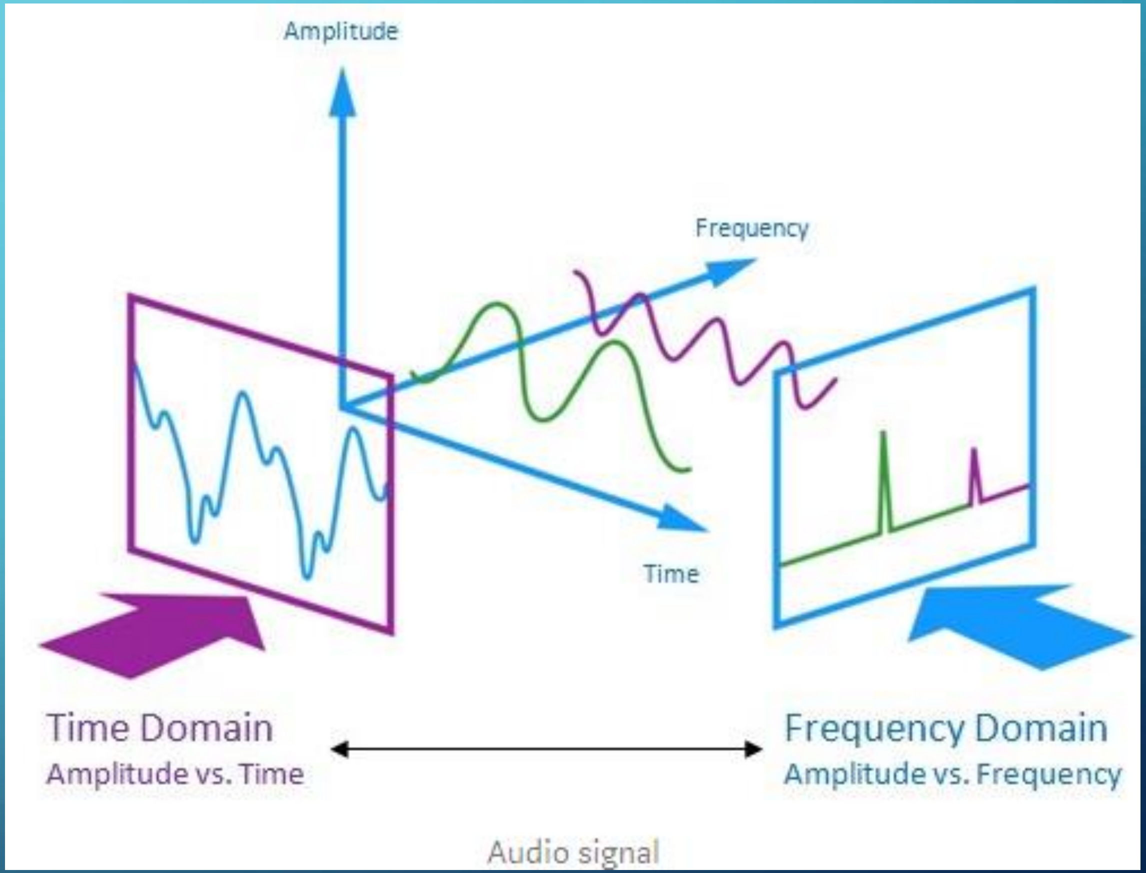
- Problem Statement
- Context - Audio Signal
- Feature extraction from Speech
- Model pipeline
- Results (Speaker Recognition / Spoken Digit Recognition)
- Conclusion
- Industry Trends

PROBLEM STATEMENT

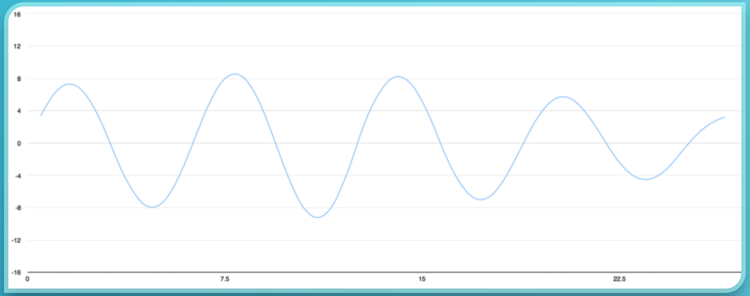
- Business Problem :
  - Biometric authentication using the speech dataset – Speaker Recognition (usage : Service Centers)
  - Inference of the digits as said by users on phone – Spoken Digit Recognition (usage : Navigate Menu or input digit datasets)

CONTEXT – AUDIO SIGNAL

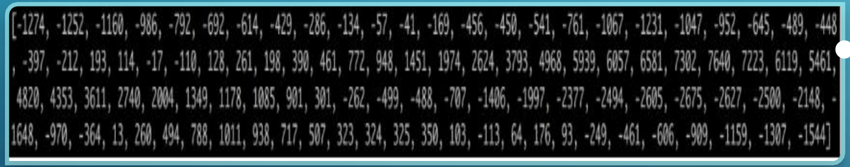
- Audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency.
- Problems :
  - “Ooonnnnneeeee” vs “One!”
  - Align audio files of various length to a fixed piece
  - Convert audio wave into set of numbers – record the height of wave at equally-spaced points.



AUDIO WAVE SAMPLING – CONVERTING WAVE TO NUMBERS



- Reading thousands of times a second and recording number representing the height of sound wave at that time
- Speech recognition can be done at a sampling rate of 16Khz (16000 times per second

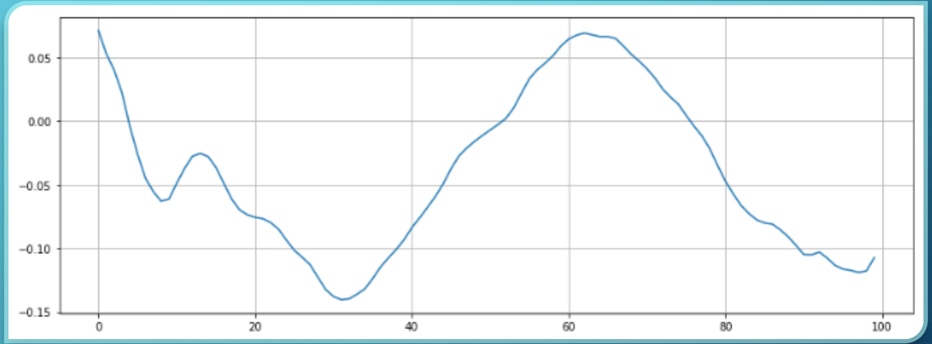


- Here is an example of sample for “Hello”

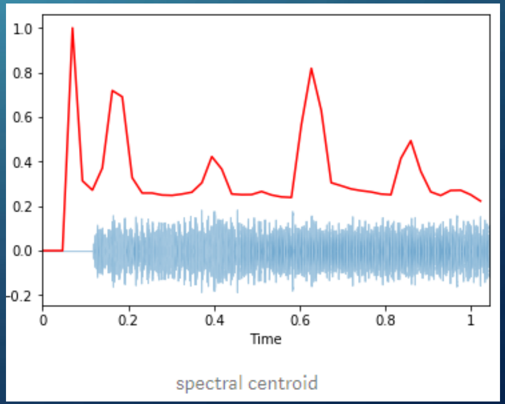
FEATURE ENGINEERING

- **Energy / RMSE:** The of a signal corresponds to the total magnitude of the signal and roughly corresponds to how loud the signal is. The energy in a signal is defined by ∑n∣x(n)∣2. The root mean square of this formula is called RMSE
- **Zero Crossing Rate :** Rate of sign changes along a signal (ie the rate at which the signal changes from positive to negative or back.
- **Spectral Centroid :** It indicates where the “center of mass” for a sound is located and is calculated as weighted mean of the frequencies present in the sound.

Zero Crossing Rate



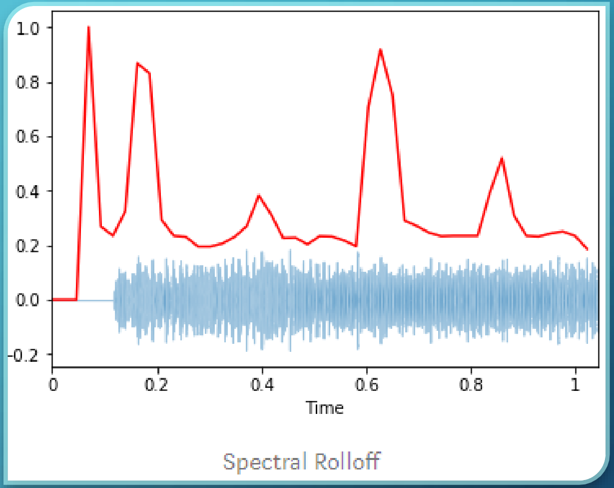
Spectral centroid



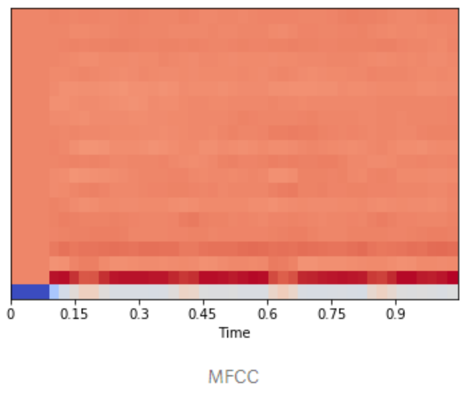
FEATURE ENGINEERING

- **Spectral Roll off** : It's the frequency below which a specified percentage of the total spectral energy, e.g. 85%, lies.
- **Chroma** : It's a typically 12-element feature vector indicating how much energy of each pitch class is present in the signal
- **Spectral Bandwidth** : It's the p'th-order spectral bandwidth.
- **MFCC (Mel-Frequency Cepstral Coefficients)** : These are small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope.
  - PS: It uses Fourier transform , to break apart the complex sound wave into the simple sound waves and then add up to get a sum of energy contained in each one.

Spectral Rolloff



MFCC



MODEL PIPELINE – SPEAKER RECOGNITION / SPOKEN DIGIT RECOGNITION

Extract sound features

Preprocess and transform the dataset

Apply the Neural Network (Keras)

Tensorflow backend

Prediction

Calculate accuracy matrix

DATA PREPROCESSING STEPS

- Step 1: Sound features extraction using librosa
- Step 2: Reading through the files in training set and appending to train matrix
- Step 3: Data Classification and labelling
- Step 4: Data Scaling of the feature set
- Step 5: Divide the dataset into training, validation and test set

NEURAL NETWORK EXPLAINED

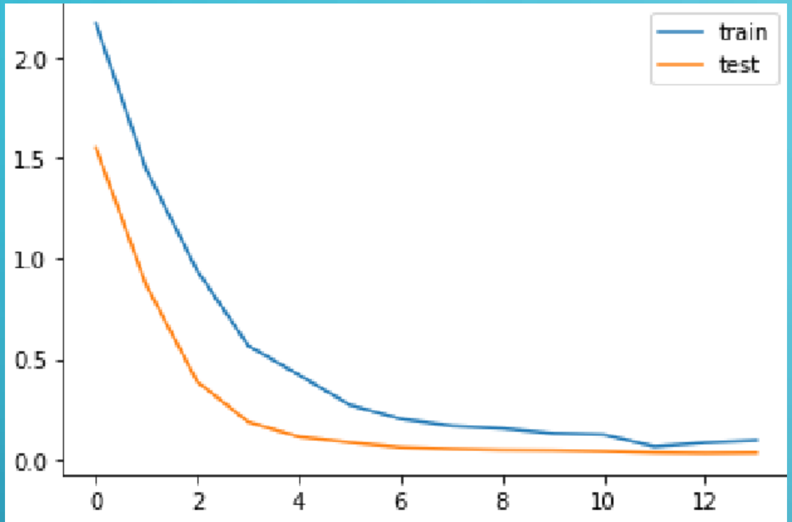
Denselayer(256, relu)
Dropout(0.5)
Denselayer(128, relu)
Dropout(0.5)
Denselayer(64m relu)
Dropout(0.5)
Denselayer(10,softmax)

TEST DATA EXPLAINED

- Experiment 1 : Got the speech dataset for 3 users (Jackson, Theo, Nicholas) for digits 0-9 repeated 49 different (training) and use the last one set as the test sample.

- Experiment 2 : Got the speech dataset for 3 additional users (Ankur, Rodolfo, Caroline) for digits 0-9 (training) and use the another speech set as the test sample.

MODEL RESULTS : SPEAKER RECOGNITION EXP - 1



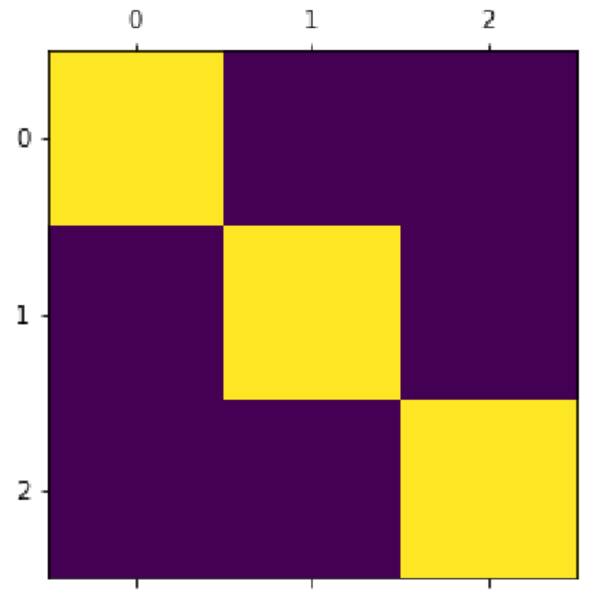
- Train Data Set 1 : : (Jackson, Theo, Nicholas) 48 samples of each digit
- Test Data 1: (Jackson, Theo, Nicholas) 49th recording
- Accuracy : 100% (predict speaker)

Classification Report

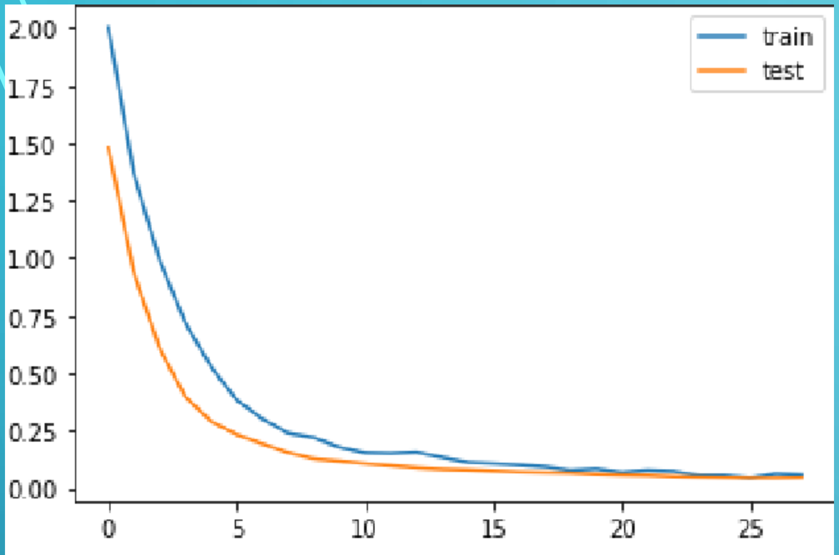
	precision	recall	f1-score	support
0	1.00	1.00	1.00	10
1	1.00	1.00	1.00	10
2	1.00	1.00	1.00	10
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

Classification Report for Test Data

```
[[10  0  0]
 [ 0 10  0]
 [ 0  0 10]]
```



MODEL RESULTS : SPEAKER RECOGNITION EXP - 2



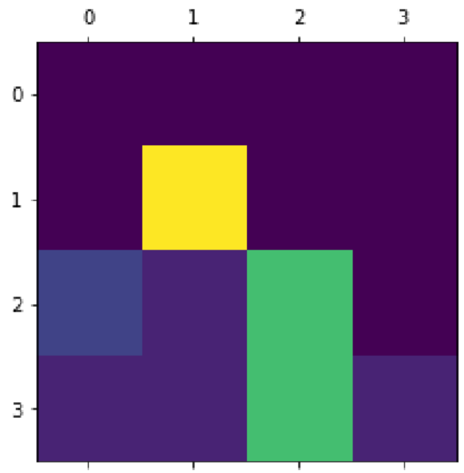
- Train Data Set 2 : : (Jackson, Theo, Nicholas) 48 samples of each digit + (Ankur, Rodolfo, Caroline) 1 sample for each digit
- Test Data 2: (Ankur, Rodolfo, Caroline) one recording
- Accuracy : 60% (Predict speaker)

Classification Report

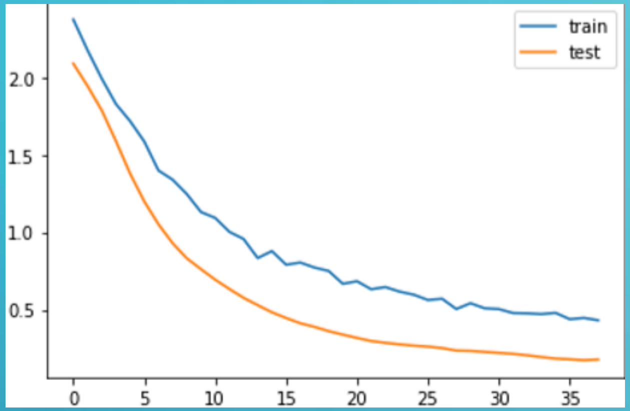
	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
3	0.83	1.00	0.91	10
4	0.50	0.70	0.58	10
5	1.00	0.10	0.18	10
accuracy			0.60	30
macro avg	0.58	0.45	0.42	30
weighted avg	0.78	0.60	0.56	30

Classification Report for Other Speakers

```
[[ 0  0  0  0]
 [ 0 10  0  0]
 [ 2  1  7  0]
 [ 1  1  7  1]]
```



MODEL RESULTS : SPOKEN DIGIT RECOGNITION EXP - 1

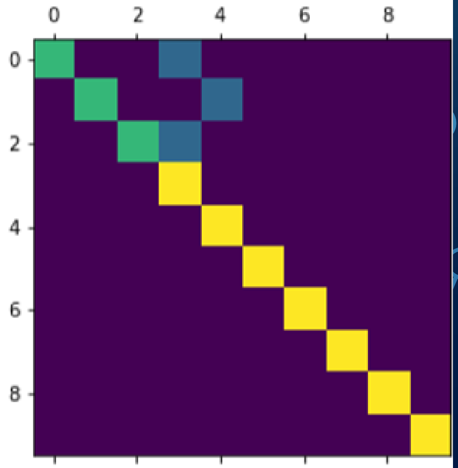


- Train Data Set 1 : : (Jackson, Theo, Nicholas) 48 samples of each digit
- Test Data 1: (Jackson, Theo, Nicholas) 49th recording
- Accuracy : 90% (predict digit from audio)

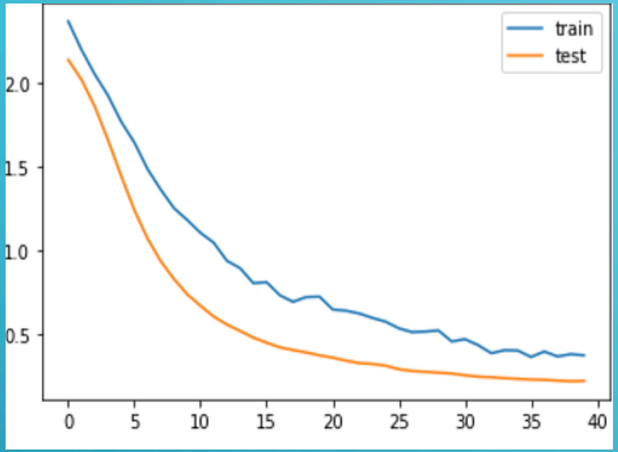
Classification Report

	precision	recall	f1-score	support
0	1.00	0.67	0.80	3
1	1.00	0.67	0.80	3
2	1.00	0.67	0.80	3
3	0.60	1.00	0.75	3
4	0.75	1.00	0.86	3
5	1.00	1.00	1.00	3
6	1.00	1.00	1.00	3
7	1.00	1.00	1.00	3
8	1.00	1.00	1.00	3
9	1.00	1.00	1.00	3
accuracy			0.90	30
macro avg	0.93	0.90	0.90	30
weighted avg	0.94	0.90	0.90	30

```
Confusion Matrix
[[2 0 0 1 0 0 0 0 0 0]
 [0 2 0 0 1 0 0 0 0 0]
 [0 0 2 1 0 0 0 0 0 0]
 [0 0 0 3 0 0 0 0 0 0]
 [0 0 0 0 3 0 0 0 0 0]
 [0 0 0 0 0 3 0 0 0 0]
 [0 0 0 0 0 0 3 0 0 0]
 [0 0 0 0 0 0 0 3 0 0]
 [0 0 0 0 0 0 0 0 3 0]
 [0 0 0 0 0 0 0 0 0 3]]
```



MODEL RESULTS : SPOKEN DIGIT RECOGNITION EXP - 2

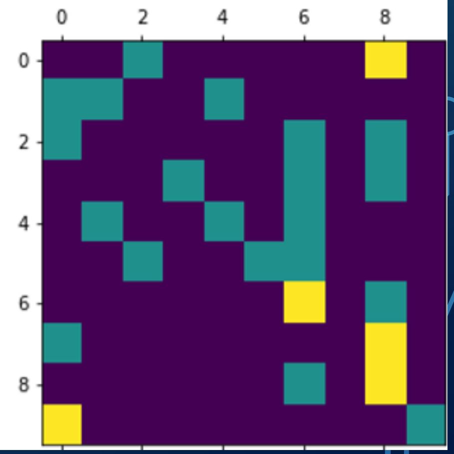


- Train Data Set 2 : : (Jackson, Theo, Nicholas) 48 samples of each digit + (Ankur, Rodolfo, Caroline) 1 sample for each digit
- Test Data 2: (Ankur, Rodolfo, Caroline) one recording
- Accuracy : 30% (predict digit from audio)

Classification Report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	3
1	0.50	0.33	0.40	3
2	0.00	0.00	0.00	3
3	1.00	0.33	0.50	3
4	0.50	0.33	0.40	3
5	1.00	0.33	0.50	3
6	0.29	0.67	0.40	3
7	0.00	0.00	0.00	3
8	0.22	0.67	0.33	3
9	1.00	0.33	0.50	3
accuracy			0.30	30
macro avg	0.45	0.30	0.30	30
weighted avg	0.45	0.30	0.30	30

```
Confusion Matrix
[[0 0 1 0 0 0 0 0 2 0]
 [1 1 0 0 1 0 0 0 0 0]
 [1 0 0 0 0 0 1 0 1 0]
 [0 0 0 1 0 0 1 0 1 0]
 [0 1 0 0 1 0 1 0 0 0]
 [0 0 1 0 0 1 1 0 0 0]
 [0 0 0 0 0 0 2 0 1 0]
 [1 0 0 0 0 0 0 0 2 0]
 [0 0 0 0 0 0 1 0 2 0]
 [2 0 0 0 0 0 0 0 0 1]]
```



CONCLUSION

Neural networks can help predict both spoken digits and speaker to a great extent with an accuracy of more than 90%

Lots of training dataset is required in order to get a good working model.

The models fails to give good accuracy results for new speakers.

Possible Reasons:

More noise signals

Ethnicity and origin of country can change the features of audio.

INDUSTRY TRENDS

Good

Alexa / Siri / Google : Able to understand majority of our voice commands

Multifactor authentication includes some speech analysis to confirm user identity.

Bad

Privacy concerns around speech datasets and requirements to store a lot of datasets in order to predict correctly.

Increase in Voice Deepfakes attacks : https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/#1e4979a52241

PROJECT DETAILS

- Github location :
  - https://github.com/ravasconcelos/spoken-digits-recognition
- Youtube video : https://www.youtube.com/watch?v=_FXqysbYVGs