SPEECH RECOGNITION - MODELLING USING KERAS -

03-DEC-2019

AGENDA

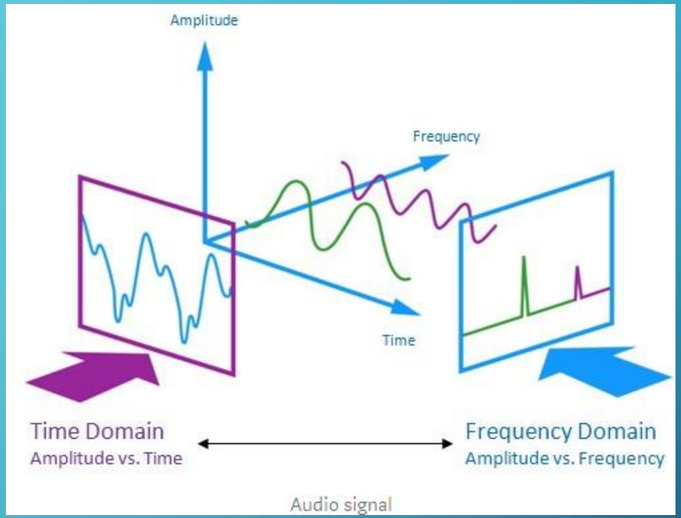
- Problem Statement
- Dataset
- Feature extraction from Speech
- Model used for learning
- Results
- Conclusion
- Further studies

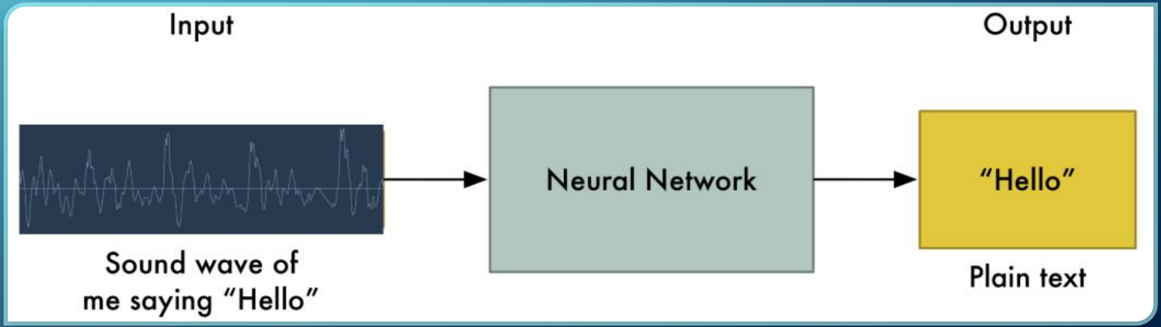
PROBLEM STATEMENT

- Business Problem :
  - Inference of the digits as said by users on phone
  - Biometric authentication using the speech dataset

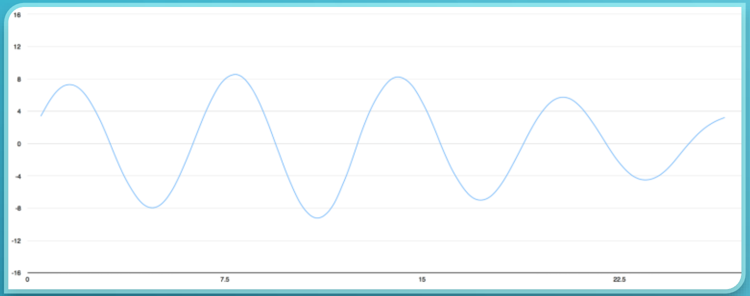
CONTEXT – SPEECH DATA PROCESSING

- Audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency.
- Feed sound recordings into neural network and train to produce text
- Problems :
  - “Heeeellllllllooooooo” vs “Hello!”
  - Align audio files of various length to a fixed piece
  - Convert wave into numbers – record the height of wave at equally-spaced points.





SAMPLING

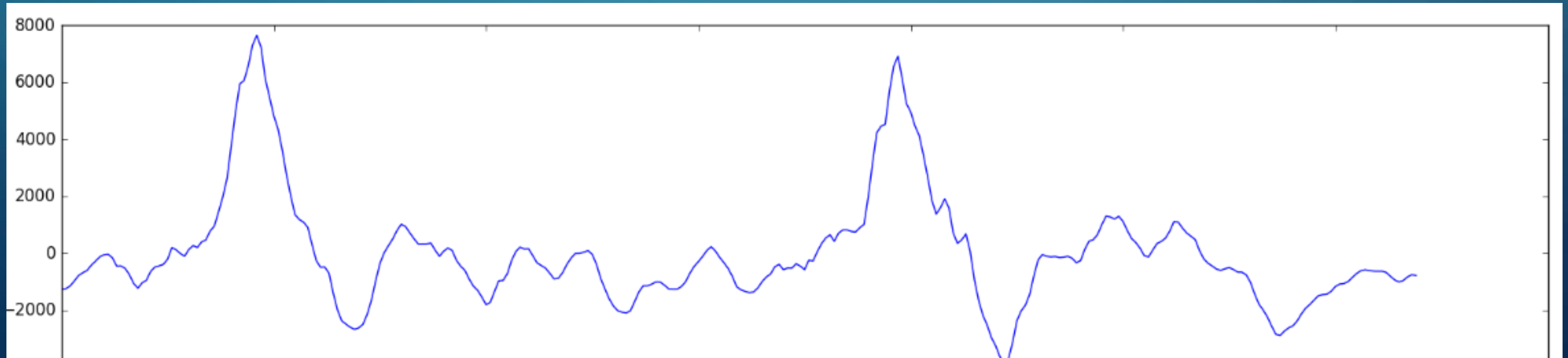



```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544]
```


- Reading thousands of times a second and recording number representing the height of sound wave at that time
- CD quality is sampled at 44.1khz (44,100 readings per second)
- Speech recognition can be done at a sampling rate of 16Khz (16000 times per second)
- Here is an example of sample for “Hello”

PREPROCESSING

- Each number in array represents sound wave amplitude at 1/16,000th of a second
- Instead we start grouping samples into 20 ms chunks (it would consists of 320 samples)
- Still different frequencies mix together to make up complex sound of human speech



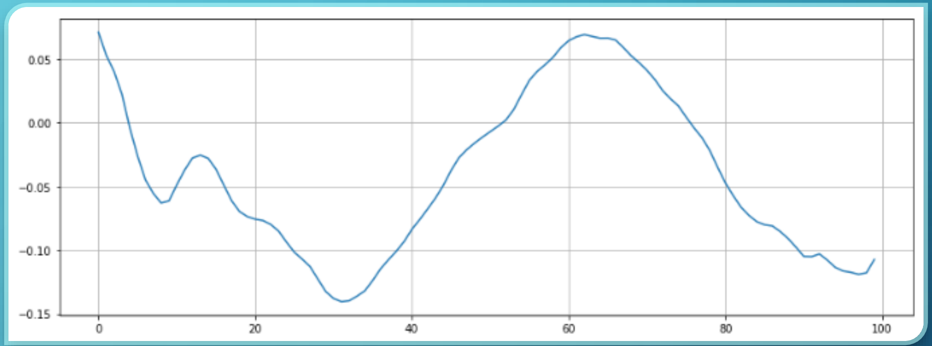
PREPROCESSING

- To make it easier, break apart this complex sound wave into it's component part.
- Break into low-pitched paths, the next-lowest-pitched-parts and so on.
- Then add up how much energy is in each of those frequency bands (from low to high), we create a fingerpart of sorts from this audio.
- Use 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.
- End result is a score of how important each frequency range is. Each number below represents energy in each of 50hz band of our 20 ms audio clip

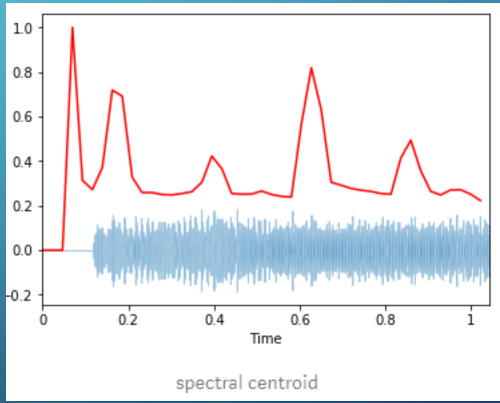

```
[110.97481594791122, 166.61537247955155, 180.43561044211469, 175.09309469913353, 180.0168691095916, 176.00619977472167, 179.79737781786582, 173.53025213548219, 176.87177119846058, 170.42684732853121
, 159.26023828556598, 163.24469810981628, 149.15527353931867, 154.34196586290136, 151.46179061113972, 152.99674239973979, 143.98878156117371, 156.6033737693738, 155.78237530428544, 157.1793094101783
8, 146.28632297509679, 164.37233032929228, 158.1282656446088, 147.23266451005145, 133.26597973863801, 116.5170100028831, 116.85501120577126, 115.40519005123537, 120.85619013711488, 112.4840612316109
1, 111.80244759457571, 92.590676871856431, 105.75863927434719, 95.673146446282971, 90.391748128064208, 79.355818055314899, 86.080143147713926, 84.748200268709567, 83.050569583779065, 86.207180262242
758, 90.252031938154076, 89.361567351948437, 90.917307309643206, 90.746777849123049, 86.726552726337033, 85.709412745066928, 95.938840816664865, 99.09254575917069, 96.632437741434885, 103.2396123166
6669, 105.80328302591124, 109.53029281234707, 116.46408227060996, 129.20890691592615, 130.43460361780441, 138.15581799444712, 128.25056761852832, 138.14492240466387, 140.0352714810314, 128.151381394
29752, 123.93018478493934, 121.19289035588113, 119.03159255422509, 114.23027889344033, 119.1717342154997, 101.02560719093093, 110.91192243698025, 106.04872005953503, 100.86977927980999, 92.123301579
000341, 94.376766266598295, 97.850709698634489, 113.37126364077845, 110.24526597732718, 113.72249347908021, 120.63960942628063, 122.06482553759932, 117.96716716036715, 120.87682744817975, 125.060973
81947157, 111.57319012901624, 115.54483708595507, 116.99850750130265, 114.40659619324526, 79.869543980883975, 104.83111191845597, 104.66218602004588, 104.91691734582642, 97.143620527536072, 78.43459
781117835, 82.214144782667248, 67.246072805959614, 66.578937262360313, 74.100307226086798, 64.861423011415653, 59.167561212002269, 62.479712687304911, 63.568362396107467, 55.906096471453267, 42.7908
02909362839, 55.693923524361097, 50.776364877715011, 41.196111220671298, 51.062413666348945, 58.493563858289065, 53.081835042922769, 73.060663128159547, 68.21625202122361, 66.7701034934517, 59.76625
124915202, 35.413635503802389, 22.705615809958832, 16.458048045346381, 44.910670465379937, 59.282513769840705, 69.241393677323856, 81.778634874076346, 88.409923803546008, 94.688033733251245, 96.6408
67526244051, 91.806226496828543, 94.570526932206619, 99.250924315589074, 97.899164767741183, 75.176507616277235, 80.947474423758905, 71.859103451990862, 93.863684037461738, 96.757146539348298, 96.52
8614354976241, 99.366456533638413, 102.18717608176904, 102.06596663023235, 101.78493139911082, 103.7883358299547, 99.915220403870748, 107.43478470929935, 104.46449552620618, 105.70789868195298, 101.
10596541338749, 100.75737831526195, 91.742897073196886, 88.307278943069093, 90.936627732905492, 71.134275744339803, 72.504304977841457, 76.233185506299705, 63.281284410272761, 45.380164336858961, 43
.018963766250437, 49.133789791276826, 53.507751009532953, 48.586423555688746, -4.4730776113028883, 50.833000650183408, 51.003802143009629, 39.577356593427531, 47.096919248906332, 55.442197175664383,
 56.967128095484341, 49.383247263177985]
```

FEATURE ENGINEERING

- **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.**

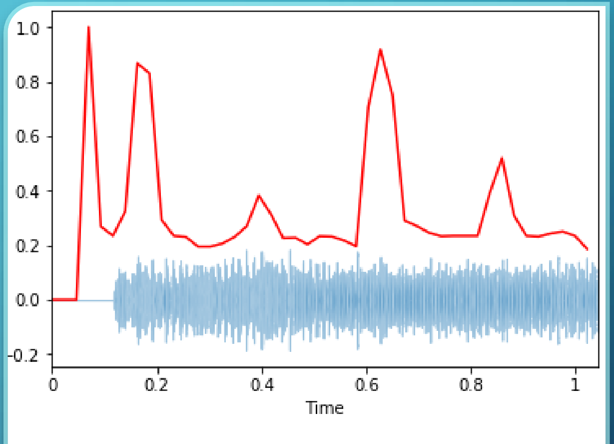


FEATURE ENGINEERING

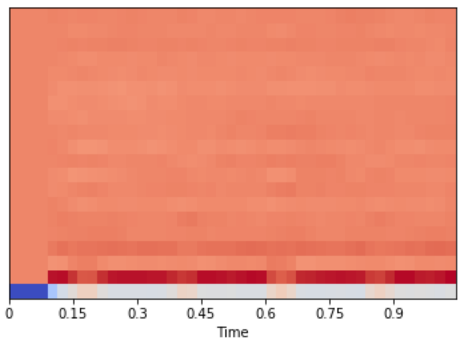
- Spectral Roll off : It's the frequency below which a specified percentage of the total spectral energy, e.g. 85%, lies.

- Chroma : It's a typically 12-element feature vector indicating how much energy of each pitch class is present in the signal

- 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.



Spectral Rolloff



MFCC

MODEL PIPELINE – SPEAKER RECOGNITION

Extract sound features

Preprocess and transform the dataset

Apply the Neural Network (Keras)

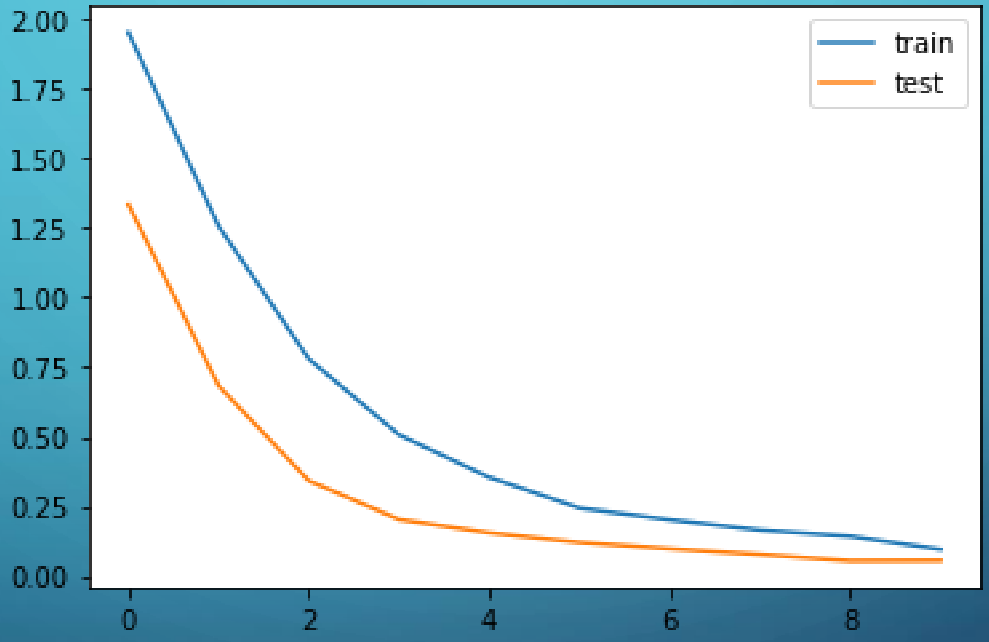
Tensorflow backend

Prediction

Calculate accuracy matrix

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

SPEAKER RECOGNITION RESULTS



MODEL PIPELINE – DIGIT RECOGNITION

Extract sound features

Preprocess and transform the dataset

Apply the Neural Network (Keras)

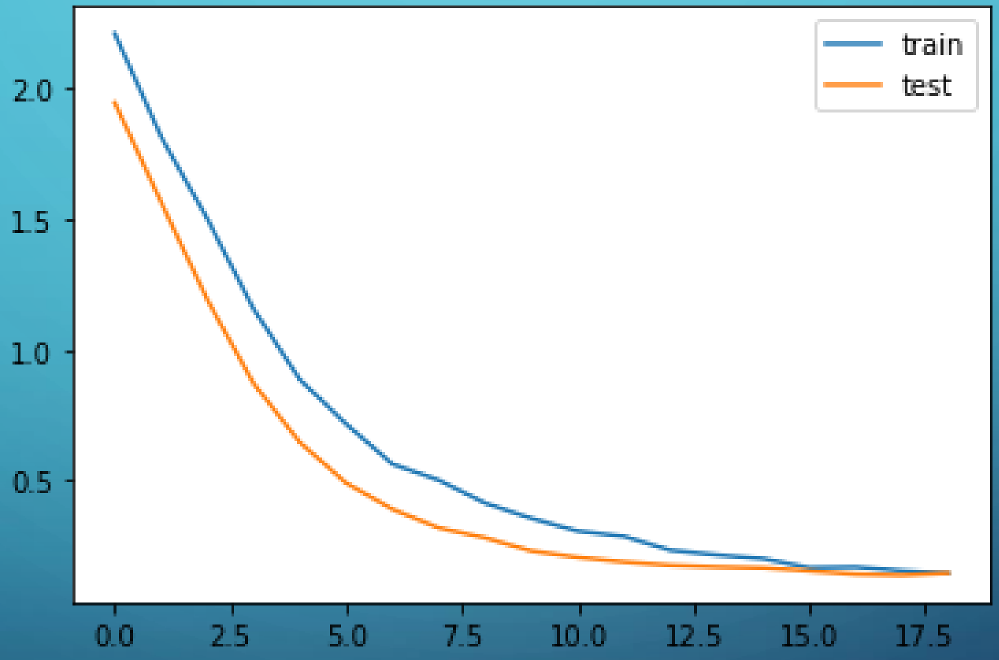
Tensorflow backend

Prediction

Calculate accuracy matrix

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

DIGIT RECOGNITION RESULTS



CONCLUSION

FURTHER RECOMMENDATION