Gender Detection From Audio

CS 337 Project

Shiv Kiran Bagathi, 200050019
Bukke Likith Rajesh, 200050025
Deekonda Venkatesh Prasad, 200050030
Jithendra Landa, 200050068

24-11-2022

Problem Statement

- Developing a classifier which identifies if the voice is of a male or a female when given a voice clip.
- We developed two models both end-to-end and one where we did feature engineering based on acoustic properties.

Major Challenges

- Frequency range for gender recognition should be in the range of human vocal range (0-280 hz)
- In general frequencies in the nature frequencies are all over the 0-20k hz.
- If we consider the usual frequencies the data is usually very sparse. We need to perform proper feature analysis and extract out the required features.
- Also speech signals are highly time-varying and have very high randomness.
- After extracting out the required features the problem converts into classification problem.

Model 1 - Based on Feature Engineering

Voice gender dataset (https://www.kaggle.com/primaryobjects/voicegender/home)

This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech.

It consists of 3168 recorded voice samples, collected from male and female speakers.

The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz (human vocal range).

Features extracted from the audio clips are meanfreq, sd, median, Q25, Q75, IQR, skew, kurt, sp.ent, sfm, mode ,centroid, meanfun, minfun , maxfun, meandom , mindom, maxdom, dfrange, modindx, label

Dataset Features Description

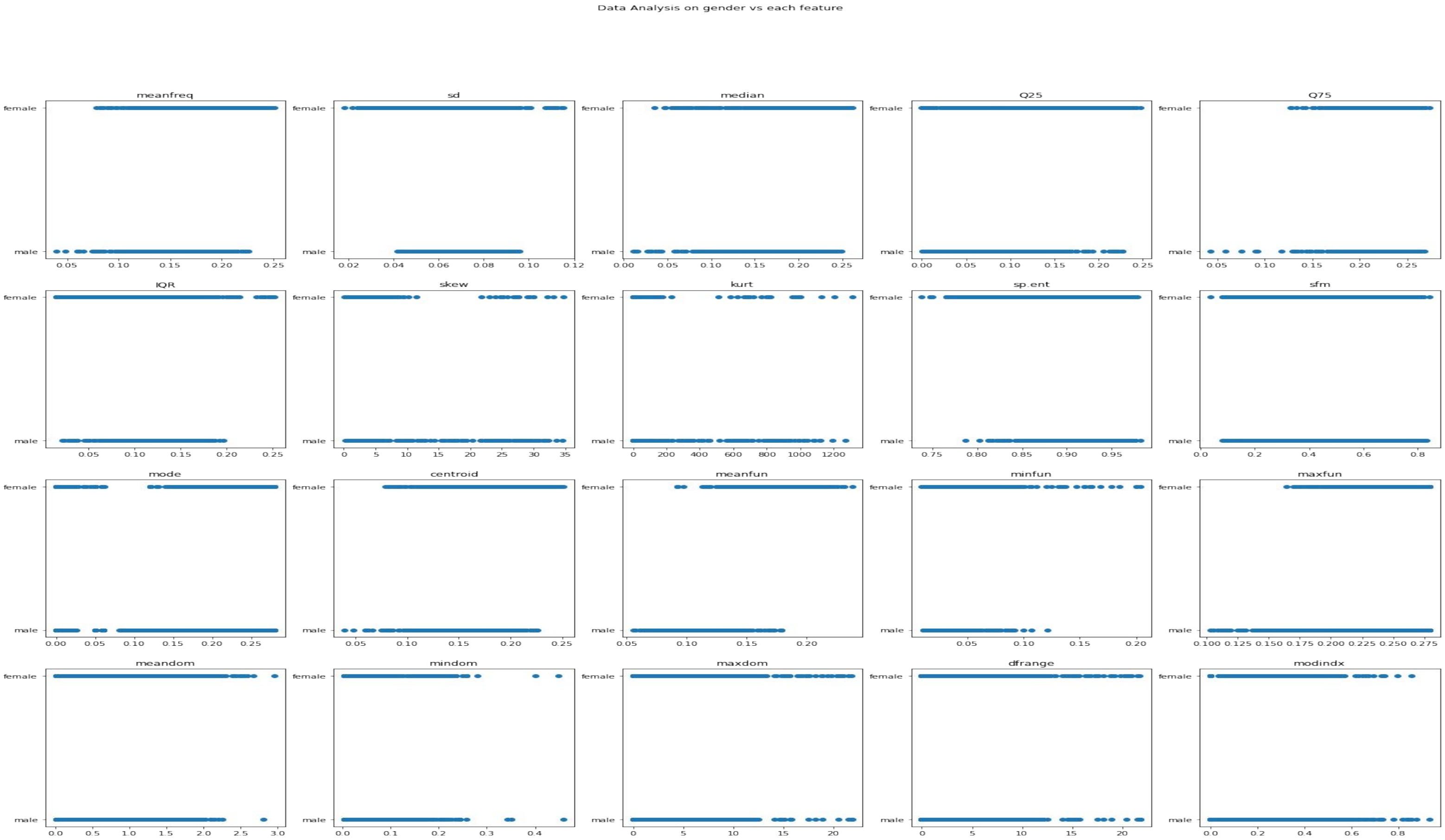
meanfreq, sd, median, Q25, Q75, IQR, skew, kurt, sp.ent, sfm, mode ,centroid, meanfun, minfun , maxfun, meandom , mindom, maxdom, dfrange, modindx, label

meanfreq, sd, median, Q25, Q75, Inter Quartile Range (in kHz) , skew (measure of asymmetry), kurt(measure of tailedness), mode, centroid(frequency centroid)

meanfun, minfun, maxfun, are mean, min, max of fundamental frequency across acoustic signal

meandom, mindom, maxdom, dfrange, are measures of dominant frequency across acoustic signal.

modindx: modulation index. The accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range


Data Analysis on gender vs each feature
meanfreq
sd
median
Q25
Q75
IQR
skew
kurt
sp.ent
sfm
mode
centroid
meanfun
minfun
maxfun
meandom
mindom
maxdom
dfrange
modindx
female
male
0.05 0.10 0.15 0.20 0.25
0.02 0.04 0.06 0.08 0.10 0.12
0.00 0.05 0.10 0.15 0.20 0.25
0.00 0.05 0.10 0.15 0.20 0.25
0.05 0.10 0.15 0.20 0.25
0.05 0.10 0.15 0.20 0.25
0 5 10 15 20 25 30 35
0 200 400 600 800 1000 1200
0.75 0.80 0.85 0.90 0.95
0.0 0.2 0.4 0.6 0.8
0.00 0.05 0.10 0.15 0.20 0.25
0.05 0.10 0.15 0.20 0.25
0.05 0.10 0.15 0.20
0.05 0.10 0.15 0.20
0.100 0.125 0.150 0.175 0.200 0.225 0.250 0.275
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0.0 0.1 0.2 0.3 0.4
0 5 10 15 20
0 5 10 15 20
0.0 0.2 0.4 0.6 0.8

Classification models trained

We trained the features on various models for classifying gender

The Models we trained are

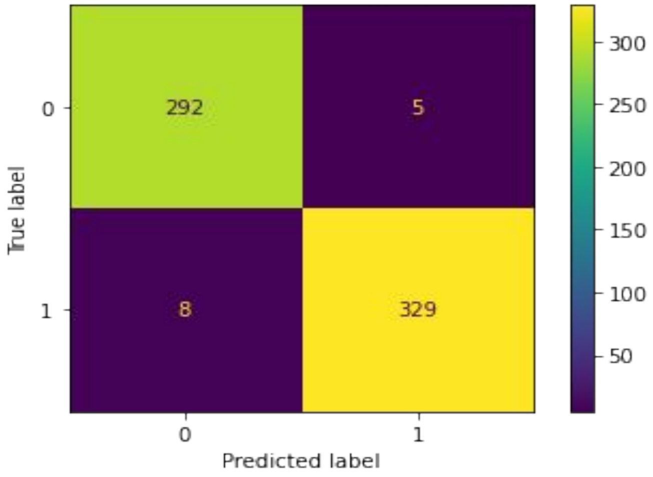
1. Neural Net with 2 hidden layers
2. SVM Classifier
3. KNN Classifier
4. Random Forest Classifier
5. Decision Tree
6. MLP Classifier
7. GMM Model

Results and Analysis from trained models.

1. Neural Net with 2 Layers

   This model is based on simple sigmoid activation and final softmax layer with hidden layer of 6 nodes.

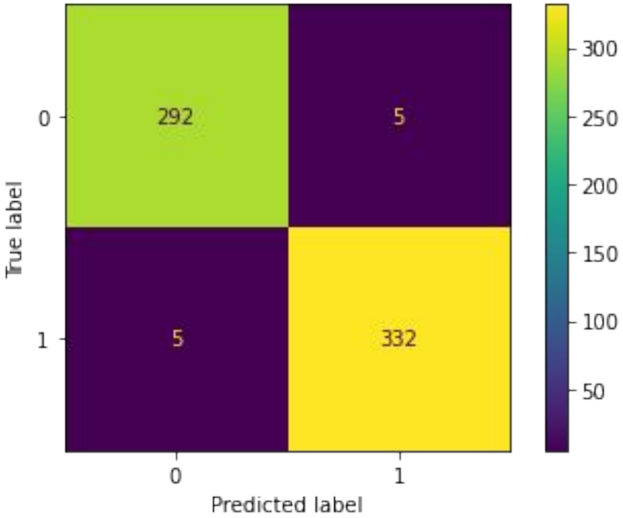
   Neural Net accuracy: 0.98, precision: 0.98, recall: 0.979, f1_score: 0.98



Results and Analysis from trained models.

2. SVM Classifier with rbf kernel

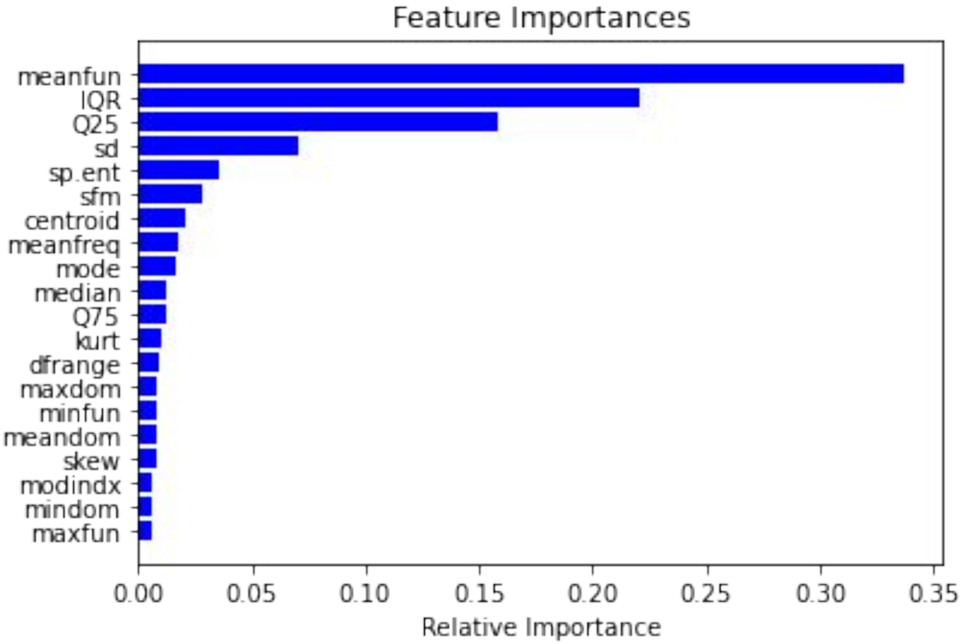
SVM accuracy: 0.984 precision: 0.985 recall: 0.985 f1_score: 0.985



Results and Analysis from trained models.

3. Random Forest Classifier

Random Forest accuracy: 0.981 precision: 0.987 recall: 0.976 f1_score: 0.982



Results and Analysis from trained models.

4. Decision Tree

   Decision Tree accuracy: 0.974 precision: 0.979 recall: 0.973, f1_score: 0.976

5. KNN Classification

   KNN accuracy: 0.979 precision: 0.979 recall: 0.982 f1: 0.98

6. MLP Classification

   MLP accuracy: 0.982, precision: 0.985 recall: 0.982 f1_score: 0.983

7. GMM Model

   GMM accuracy: 0.96 precision: 0.972 recall: 0.952 f1_score: 0.962

Key Observations

- All models give relatively high accuracy on the model as the selected features are sufficient to properly differentiate between genders.
- Neural Networks give relatively better performance but take more time.
- Here we trained model on extracted features from the frequency distribution of the fourier transform of audio clip.
- Next we will train a model to extract out the required features when we will give entire distribution as input.

Model 2 - End-to-End Model

Common Voice dataset (https://www.kaggle.com/datasets/mozillaorg/common-voice)

Common Voice is a corpus of speech data based upon text from a number of public domain sources.

The corpus is labelled with 3 labels age, gender and accent.

It consists of 380k recorded voice samples, collected from male and female speakers.

We used this dataset to extract out the features from audio files based on different extraction techniques.

Extracted features are passed through a series of 5 neural layers.

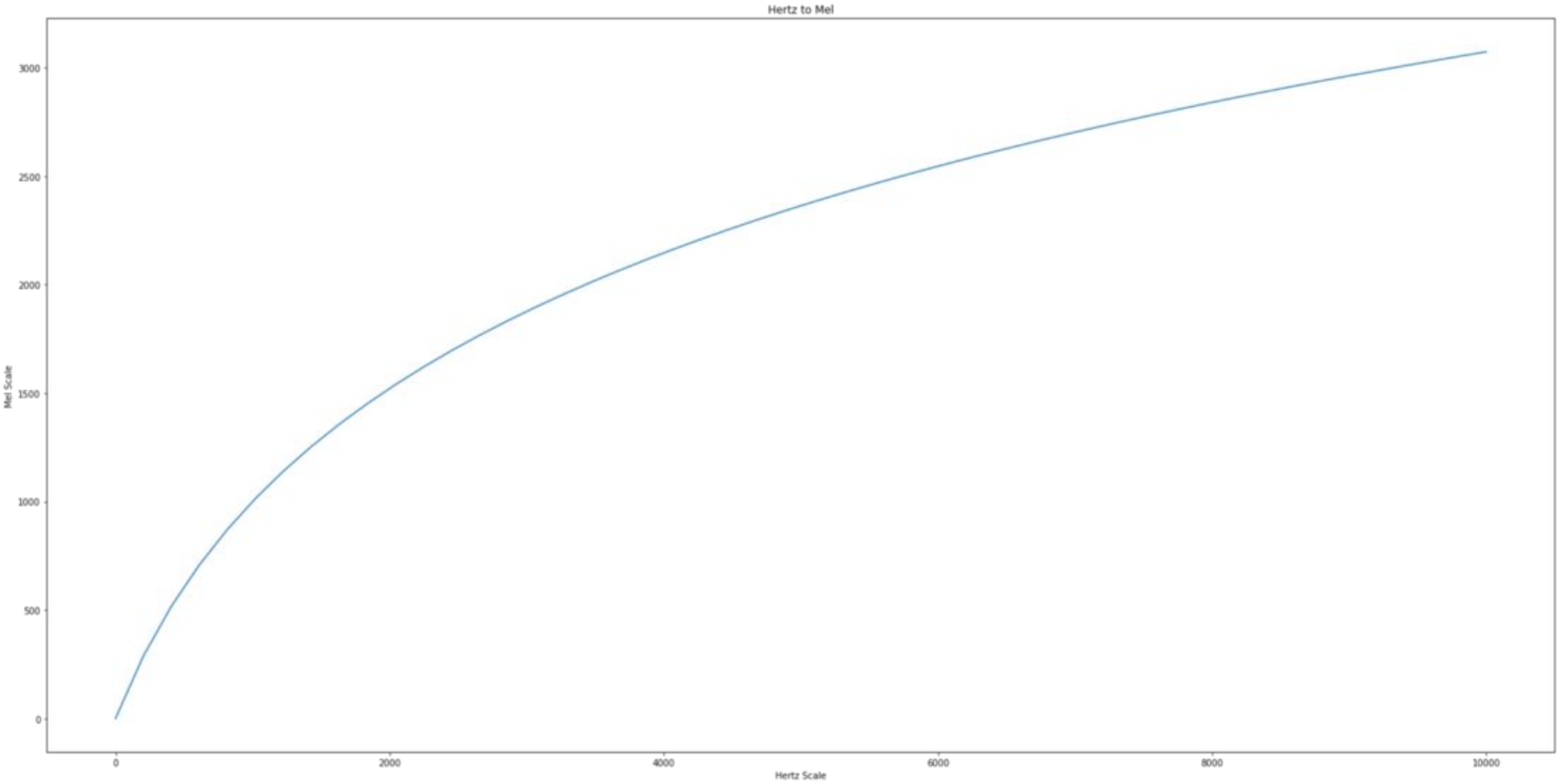
Features Extraction

- We designed our features using Mel Spectogram Frequency

- Mel Spectrograms are spectrograms that visualize sounds on the Mel scale as opposed to the usual frequency domain.

- The Mel Scale is a logarithmic transformation of a signal’s frequency

$$m = 1127 \cdot \log\left(1 + \frac{f}{700}\right)$$


Hertz to Mel
Mel Scale
3000
2500
2000
1500
1000
500
0
0
2000
4000
6000
8000
10000
Hertz Scale

Features Extraction

- We also implemented various other feature Extraction techniques
- Mel Frequency Cepstral Coefficients
- Tonnetz
- Chroma
- Each of the features Extraction technique is explained in report.

Model Architecture

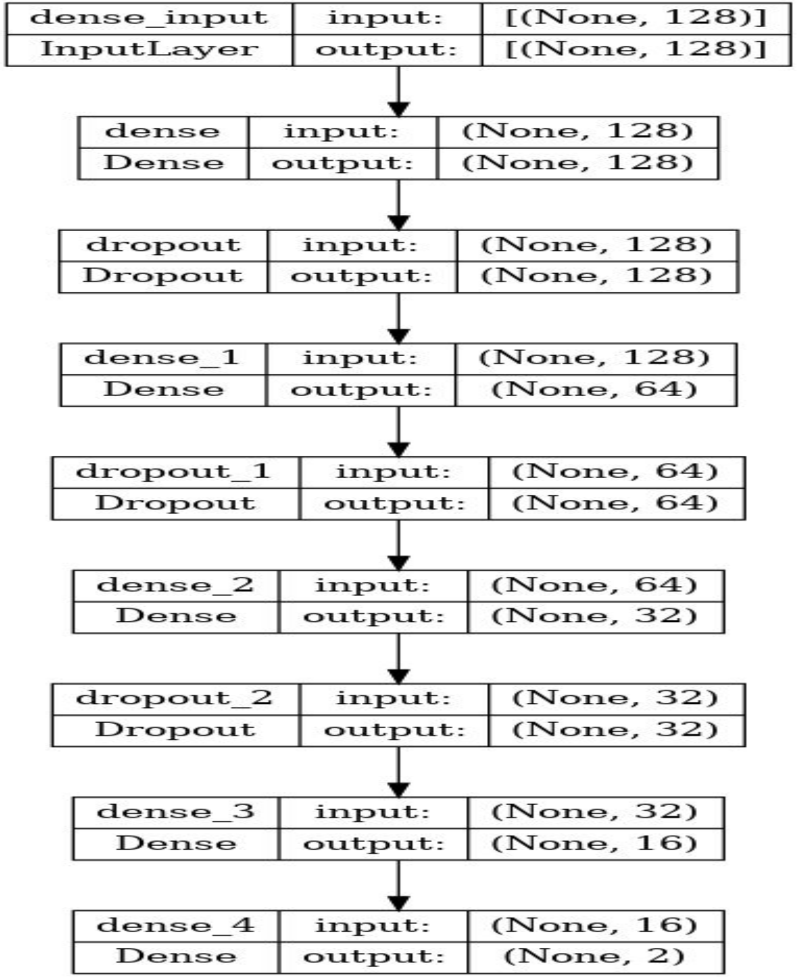
We designed our Model with 4 hidden with reLu as activation function.

We kept reducing number of nodes in each layer by a factor of 2.

Each layer has drop out value of 0.2

Total number of trainable parameters used = 27,410

Model is optimized with adam optimizer and trained with sparse categorical cross entropy with accuracy as metric.



dense_input	input:	[(None, 128)]
InputLayer	output:	[(None, 128)]

dense	input:	(None, 128)
Dense	output:	(None, 128)

dropout	input:	(None, 128)
Dropout	output:	(None, 128)

dense_1	input:	(None, 128)
Dense	output:	(None, 64)

dropout_1	input:	(None, 64)
Dropout	output:	(None, 64)

dense_2	input:	(None, 64)
Dense	output:	(None, 32)

dropout_2	input:	(None, 32)
Dropout	output:	(None, 32)

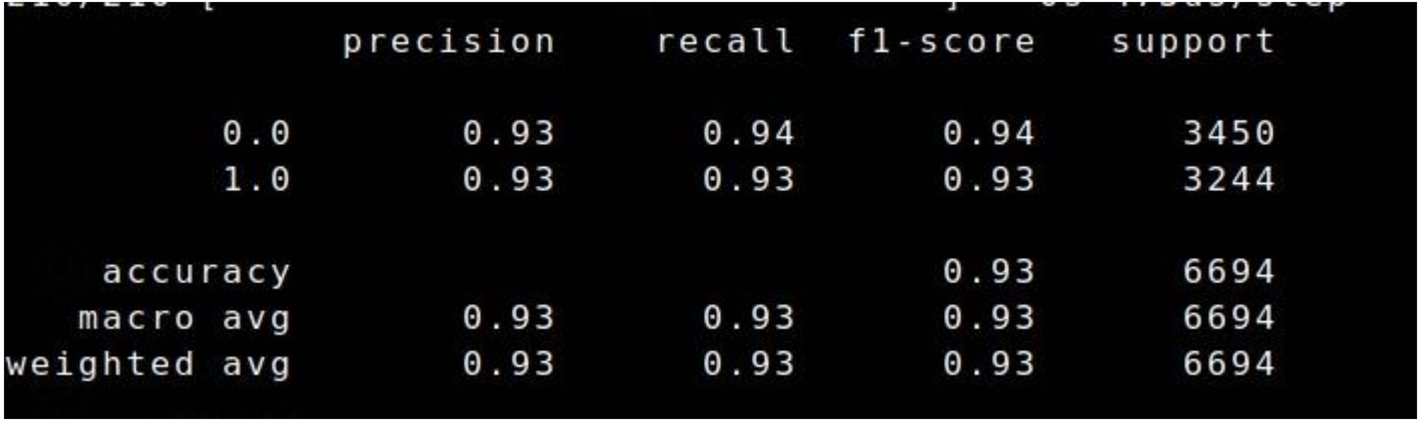
dense_3	input:	(None, 32)
Dense	output:	(None, 16)

dense_4	input:	(None, 16)
Dense	output:	(None, 2)

Results and Analysis

We trained our model with 100 epochs and of batch size 64.

The value of loss of validation reduced from 0.39 to 0.16 when implemented with Early Stopping and reach accuracy of 94%.



```
              precision    recall  f1-score   support

         0.0       0.93      0.94      0.94      3450
         1.0       0.93      0.93      0.93      3244

    accuracy                           0.93      6694
   macro avg       0.93      0.93      0.93      6694
weighted avg       0.93      0.93      0.93      6694
```

Observations

- The value of accuracy is comparatively lower than previous data. This may be due to selective analysis of particular values which are more favorable for detection of gender such as lower mean fundamental frequency corresponds to man.

- This is model is better for extended to voice recognition and sound recognition as the features constitute of entire distribution.

- First Model is specially tailored for this problem statement.