Audio classification using Gaussian Mixture Models

Bharat Mohan
bmohan@iitg.ac.in

February 21, 2020

1 Methodology

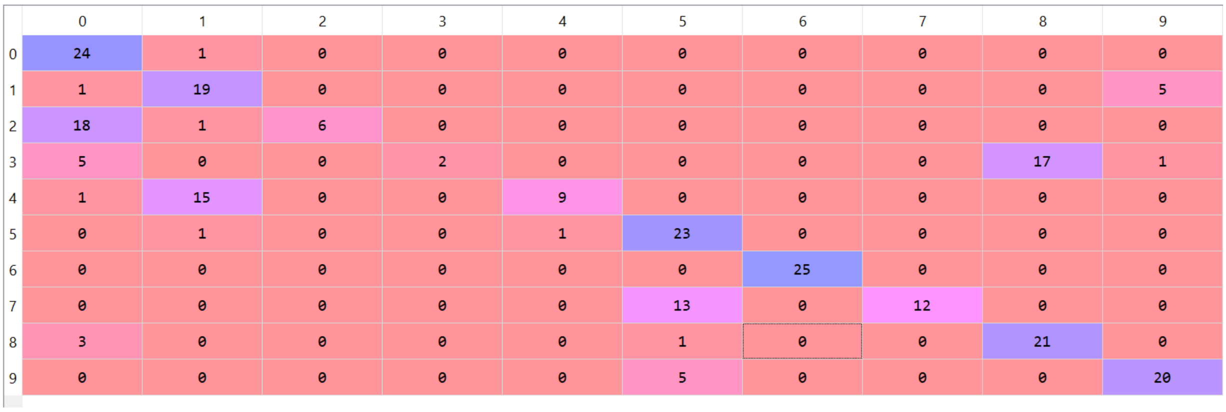
The dataset is split into train and test sets. The training dataset comprising of digits recorded by different speakers are used to train GMM models,one for each digit. The classification is to be based on the digit uttered. Each training audio is normalised,trimmed to remove silences, split into frames of 20ms each and mfcc,delta and double delta(13 dimensions each) is computed in each frame and appended to form a feature vector.The feature vectors of a particular digit spread over frames and audio files are appended to form a 39*N matrix (where N is the total number of feature vectors for a digit) which is fed to the GMM model for training. Separate GMM models (comprising of 16 modes each) are trained for each digit. The average log likelihood values(averaged over frames) for each testdata is computed against all the GMM models, out of which the maximum is evaluated and the test digit is classified as the digit associated with that particular model. A corresponding confusion matrix is generated for the test dataset.

2 Library Functions Used

- librosa.feature.mfcc
- sklearn.mixture.GaussianMixture.fit
- sklearn.mixture.GaussianMixture.score

3 Result

The following confusion matrix is generated.



	0	1	2	3	4	5	6	7	8	9
0	24	1	0	0	0	0	0	0	0	0
1	1	19	0	0	0	0	0	0	0	5
2	18	1	6	0	0	0	0	0	0	0
3	5	0	0	2	0	0	0	0	17	1
4	1	15	0	0	9	0	0	0	0	0
5	0	1	0	0	1	23	0	0	0	0
6	0	0	0	0	0	0	25	0	0	0
7	0	0	0	0	0	13	0	12	0	0
8	3	0	0	0	0	1	0	0	21	0
9	0	0	0	0	0	5	0	0	0	20