SPOKEN LANGUAGE CLASSIFICATION

GROUP NO. 20

Abhishek Agarwal
2016126
Raghav Sood
2016259
Surabhi S Nath
2016271

Introduction

With the increase in the amount of audio content, we need proper mechanisms to analyze and study them. Sound analysis has been performed for genre identification, but not much research has been done for language detection. The speech to text applications need prior information of the language. Efficient language identification systems can hence assist this and other such applications.

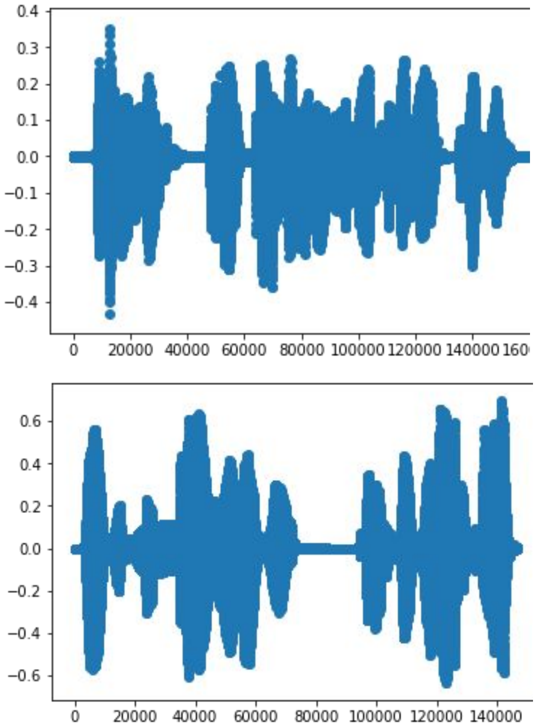
We aim to build classifiers to predict language for a set of selected regional Indian languages. As of now, we have successfully implemented efficient models to differentiate between English and Hindi samples.

Dataset

- English Data collected from OpenLRS dataset
  - 1000 samples of audio
  - Average length of 15s
  - flac format
  - 40 speakers

- Hindi Data collected from TopCoder Spoken Languages Dataset
  - 380 samples of audio
  - Average length of 10s
  - mp3 format
  - Multiple speakers

- Regional Languages - Contacted Linguistic Data Consortium for Indian Languages for Speech Corpora dataset containing 15 regional languages



Sample Audio for English and Hindi

Intermediate Results

- Segregated Hindi data from entire TopCoder dataset
- Converted format of Hindi audio to .flac format
- Read and plotted the audio data

PREPROCESSING

- Split the audio into frames of 25 ms each with 10ms overlap in consecutive frames
- Performed hamming window smoothening for each frame to remove extra noise

FEATURE EXTRACTION

- Extracted MFCC features
- Calculated Delta features on the basis of MFCC features
- Normalised the features to bring them to the same scale

TRAINING

- Split the data into training: 80%, testing: 20%
- Number of samples used for training = 5,000
- Trained our baseline SVM classifier on this data

Intermediate Results

- Using only MFCC features
  - Accuracy = 0.51

- Added Delta features
  - Accuracy = 0.55

- Performed Grid Search to choose hyperparameters
  - 5 fold Cross Validation
  - Kernel = rbf
  - C = 1
  - Gamma = 0.05
  - Accuracy = 0.6

- Normalized the data
  - Accuracy = 0.96

- Trained using 50,000 samples
  - Accuracy = 0.97

Next Steps

- Varied sources of audio samples
  - Sound quality
  - Accent
  - Gender
  - Age
- Multiple Regional Languages of India
  - 10-15 languages
- Train other models and compare performance
  - Hidden Markov Model
  - Neural Networks
  - Gaussian Mixture Models
- Segmentation of Audio Samples
  - Detect Multiple Languages in same audio clip