DSP Assignment 4

Mili Goyal (IMT2017513)

Vaishnavi Dhulipalla (IMT2017514)

Introduction

The basic task of this assignment is IViE-corpus_British_Dialects_Classification.
Dialects are variations in the wordings, the grammar and pronunciations in the same language whereas accents are just variations in pronunciations.
The dataset has 9 folders with 67 audio samples each. The audio samples are narrations of a Cinderella passage by both male and female speakers.
Phonemes are what differentiate words from one other in a particular dialect/language. By finding patterns in different phonemes which make up a word in different dialects, we can possibly try classifying them.

Data Processing

We broke down our audio samples into small parts (**frames**) for analysis which approximately contain enough spectral content on how different phonemes make up each word. Length of one frame is taken to be around 25ms to avoid ambiguous nature of the frame.
We also made sure that there is some overlap (**stride**) between the frames for correlation between adjacent phonemes. Length of one stride is taken to be 15ms to avoid overfitting or underfitting.
We applied a **hamming window** to smoothen out the frame endings and found the power spectrum of each frame for spectral analysis and feature extraction.

Feature Extraction

Features we took into consideration here are : MFCC, delta and delta-delta coefficients.

a) **MFCC - Mel Frequency Cepstral Coefficients :**

For a very basic understanding, cepstrum is the information of rate of change in spectral bands. In the conventional analysis of time signals, any periodic component (for eg, echoes) shows up as sharp peaks in the corresponding frequency spectrum (ie, Fourier spectrum. This is obtained by applying a Fourier Transform on the time signal).

On taking the log of the magnitude of this Fourier spectrum, and then again taking the spectrum of this log by a cosine transformation, we observe a peak wherever there is a periodic element in the original time signal. Since we apply a transform on the frequency spectrum itself, the resulting spectrum is neither in the frequency domain nor in the time domain and hence Bogert et al decided to call it the ***quefrency domain.*** And this spectrum of the log of the spectrum of the time signal was named ***cepstrum***.

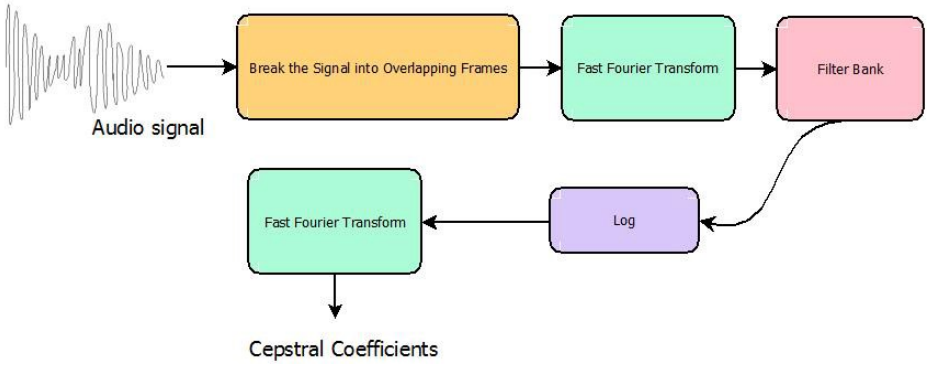
Humans can hear frequencies between 20 to 20KHz, but our hearing is more discriminative at lower frequencies than at higher frequencies. Cochlea cannot discern the difference between two closely spaced frequencies and this effect becomes more phenomenon as frequency increases. Thus we analyze higher frequencies with a wider range and lower frequencies with a smaller range to catch phoneme based features.

To convert our power spectrum range from the generic range to a more perceptual range, we use something called the ***Mel scale***. Mel scale normalizes the frequency scale to match our perceptual frequency distinguishing capabilities. We see that an increment of around 240Hz, from 160Hz to 394Hz is equivalent to 250 Mels and the same jump of 250 Mels at higher frequencies is a jump of 5000Hz from 9000Hz to 14000Hz.

A frequency measured in Hertz (f) can be converted to the Mel scale using :

$$\mathbf{Mel(f) = 2595 \log(1 + (f/700))}$$

Any sound generated by humans is determined by the shape of their vocal tract (including tongue, teeth, etc). If this shape can be determined correctly, any sound produced can be accurately represented. The envelope of the time power spectrum of the speech signal is representative of the vocal tract and MFCC (which is nothing but the coefficients that make up the *Mel-frequency cepstrum*) accurately represents this envelope. The following block diagram is a step-wise summary of how we arrived at MFCCs:



Here, Filter Bank refers to the mel filters (converting to mel scale) and Cepstral Coefficients are nothing but MFCCs.

b) Delta and Delta-Delta Coefficients :

Dialects also have different velocities and acceleration of transition between phonemes. We have created delta coefficients (velocity) and delta-delta coefficients (acceleration) to learn these features.

Training and Observation

We have written down all the coefficient data into pandas DataFrame object which contains 67*9 rows and many columns. We took the mean of each coefficient in

each dialect, which gives us 12 average coefficient values for each dialect. This is done because the spectral content of each frame isn't very important for generalization. It would be more helpful to understand what kind of spectral content is present in the phonemes of each dialect.
After finding the means, our DataFrame is of the size 67x9 rows and 36 columns. In order to get more insight from the coefficients, we also find the min-value, max value, standard deviation, skewness and median of each of our coefficients. This creates a DataFrame of 67x9 rows and 216 + 1(labels) columns.

We then split the dataset into 80% training data and 20% testing data using sklearn's train_test_split.

Then we trained **Logistic Regression Classifier, KNN classifier and SVM classifier.** The accuracy turned out to be around **86-89%, 87-94%, 89-91%** respectively.

The code and the dataset is attached in the zip file along with this report.