Programming for AI Section J/K

Combined Project for Course Work and Lab

Deadline 10th June

Due Date: Thursday, June 10th by 11:59pm.

Project is to be done individually. No late submissions will be accepted.

Submissions that do not comply with the specifications given in this document will not be marked and a zero grade will be assigned.

You are required to submit a single .ipnyb file containing all the cell outputs. You should name your .ipnyb file as i19-XXXX.zip where i19-XXXX represents your student id.

Speech Emotions Recognition SER

1 Introduction

In this project, you will implement a classifier that can determine the emotions concealed in speech signals. There are many potential applications of SER such as in call centres, health care and human resources, it will help them to improve their businesses by better understanding of customer emotional needs.

You are free to use scikit-learn or other classification libraries along with spaCy if you are comfortable doing so. You can use linear or non-linear classifiers for SER model.

2 Background

Audio speech signals contain a large number of parameters that reflecting emotional characteristics. Feature selection and feature extraction are vital steps toward building a successful classification models. In SER, various types of features can be extracted from the speech signals. However, extracting the correct feature sets is a challenging task.There are three prominent categories in speech features used in SER: (i) the prosodic features, (ii) the spectral or vocal tract features, and (iii) the excitation source features.

Prosody features are the characteristics of sound generated by the human speech,for example, pitch or funda- mental frequency (F0),duration and energy.

Spectral features are the characteristics of various sound components generated from different cavities of the vocal tract system. Some spectrum features are such as linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC), and modulation spectral features.

The features used to represent glottal activity, mainly the vibration of glottal folds, are known as excitation source features. These are also called voice quality features because glottal folds determine the characteristics of voice. Voice quality measures for a speech signal includes harshness, breathiness, and tenseness.

3 Project Task

The task is to build a model to recognize emotions from speech and calculate the accuracy of the predictions made by your system.

Follow the below steps to implement the SER system.

- Load the Data Source for the project
- The dataset provided for this project contains emotional utterances of Urdu speech gathered from Urdu talk shows. It contains utterances of four basic emotions: Angry, Happy, Neutral, and Emotion. There are 38 speakers (27 male and 11 female).
- The dataset provided for this is already split into training and test sets and provided as separate train and test folders.
- Extract the features from the provided audio files.
- Select the classification model.
- Evaluate the model to check the accuracy of the model.

Requirements

1- In the notebook open a cell and list all the features you used for classification. For example

- MFCC
- Chroma
- Mel Spectrogram
- Contrast Spectrogram
- Tonnetz

2- Train a model that is able to classify between **Angry**, **Happy**, **Neutral**, & **Sad**

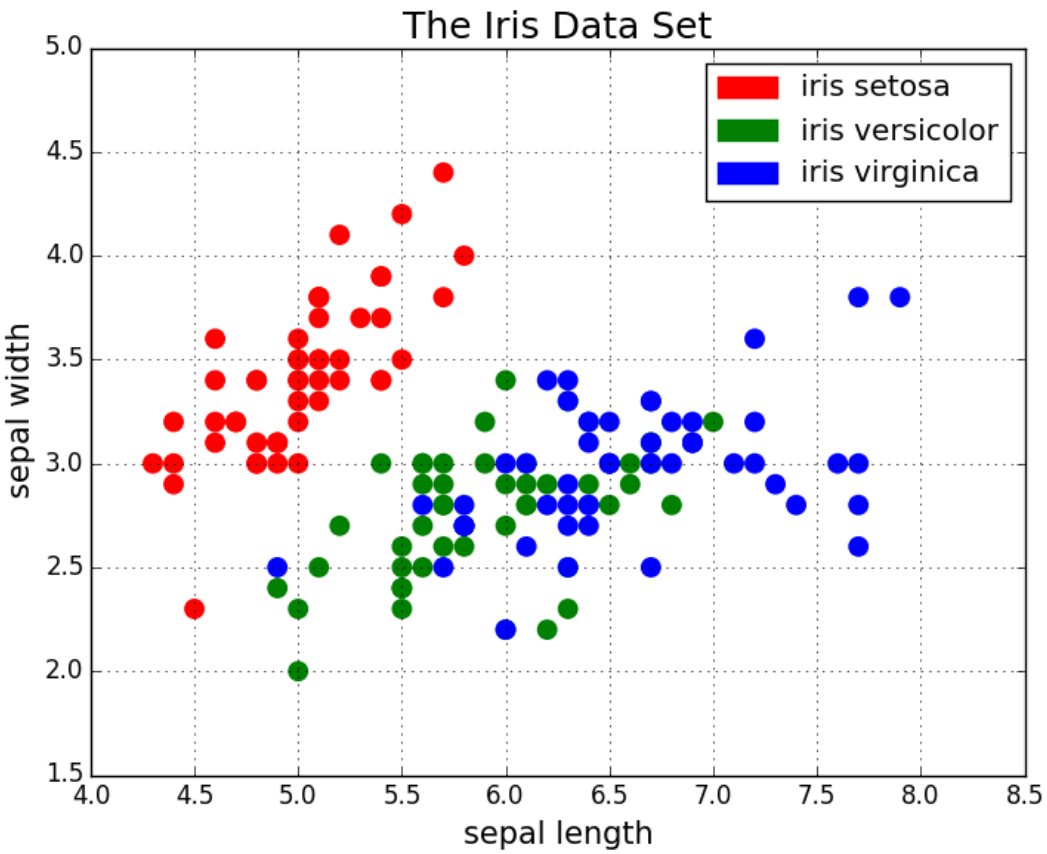
You can use SVM classifier or any other that you are comfortable with for classification

```
from sklearn.svm import SVC # "Support vector classifier"

model = SVC(kernel='linear', C=1E10)


model.fit(X, y)
```

You can use librosa library to extract these features and stack them together horizontally to create one set of features against one audio file. extract at least three features or all the above listed features (bonus). For each audio you can get its label from the folder name.

3- Take all the audio files and their features that you have selected and divide the features into two sections first half and second half. If odd number e.g. 33 then 15 features in first half and 16 features will go to the second half then take the average of the first half and the second half separately for each audio file and generate a scatterplot which looks something

like



The x and y labels will be first half and second half respectively. The color of the audio will be based on the class to which it belongs e.g., Sad, Happy, ... will have different colors

4- Output your accuracy in the notebook output cell

```
In[17]: from sklearn.metrics import accuracy_score

accuracy_score(ytest, y_model)

Out[17]: 0.97368421052631582
```

4- Take one audio Test > Sad > SM25_F34_S084.wav and pass it through your model and output the predicted emotion

Note:

The notebook output cells should be saved. There would be 4 sections the first section output is features you used. The second section output is the .fit statement output. The third section will be your graph. The fourth section output is the accuracy of your model. The fifth section output is the prediction of your model for Test > Sad >

SM25_F34_S084.wav. **Failure to follow any of these requirements could result in Zero Marks. You can write the print statements for debugging while experimentation and when everything works then comment these out and run the cell again so that only the required outputs are shown.**

This is what Model.fit ( ) output can look like:

```
In [11]: 1 Model = SVC(kernel='linear', gamma=0.001, C=1)
         2 Model.fit(X_train, y_train)

Out[11]: SVC(C=1, cache_size=200, class_weight=None, coef0=0.0,
             decision_function_shape='ovr', degree=3, gamma=0.001, kernel='linear',
             max_iter=-1, probability=False, random_state=None, shrinking=True,
             tol=0.001, verbose=False)
```