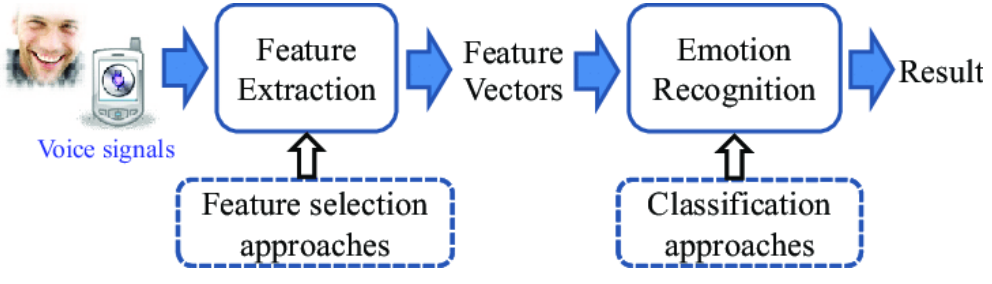
Speech Emotion Recogniser(SER)

Abhishek
2017EEB1121

Indian Institute of Technology
Ropar, Punjab
India

1 Introduction

Emotions in speech play an extremely important role in human mental life. It is a medium of expression of one's perspective or one's mental state to others. Speech Emotion Recognition(SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions- including Neutral, Anger, Happiness, Sadness in which any intelligent system with finite computational resources can be trained to identify or synthesize as required. In this work spectral and prosodic features are used for speech emotion recognition because both of these features contain the emotional information. Mel-frequency cepstral coefficients (MFCC) is one of the spectral features . Fundamental frequency, loudness, pitch and speech intensity and glottal parameters are the prosodic features which are used to model different emotions. The potential features are extracted from each utterance for the computational mapping between emotions and speech patterns. Pitch can be detected from the selected features, using which gender can be classified. Back Propagation Network is used to recognize the emotions based on the selected features. SER is tough because emotions are subjective and annotating audio is challenging.



flow chat of SER

image copied from : https://www.researchgate.net/figure/Standard-speech-emotion-recognition-process$_fig1_2$99373144

2 Literature Review

2.1 Features extraction

I am using Librosa Library for features extraction. Librosa is a Python library for analyzing audio and music. It has a flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code. This library play main role in the project because this library extract features in array foam which help in model my data. There are many function in librosa like MFCC, Choroma sift,Tonnetz,Contrast.

2.2 Emotion recognition

After training our model. I just have to check its accuracy on our test set of our data. I will recognise only four emotion which are Happy, Sad, Neutral, Angry - based on the feature's threshold.

2.3 About data set

I took only Speech audio-only files (16bit, 48kHz .wav) from the RAVDESS. This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav).

Filename identifiers:

Modality (01 = full-AV, 02 = video-only, 03 = audio-only).

Vocal channel (01 = speech, 02 = song).

Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).

Emotional intensity (01 = normal, 02 = strong).

Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").

Repetition (01 = 1st repetition, 02 = 2nd repetition).

Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

3 classifiers

3.1 Multi-Layer Perceptron(MLP)

I will use Multi-Layer perceptrons(MLP) model. Multilayer perceptrons train on a set of input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs. Training involves adjusting the parameters, or the weights and biases, of the model in order to minimize error. Backpropagation is used to make those weigh and bias adjustments relative to the error, and the error itself can be measured in a variety of

ways,including by root mean squared error (RMSE). Feedforward networks such as MLPs are just like ping-pong: they are mainly involved in two motions, a constant back and forth (forward and backward passes). This is a feedforward ANN model.

3.2 Support Vector Machine(SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an Ndimensional space(N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

3.3 Naive Bayes

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem. Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

3.4 K- Nearest Neighbour(KNN)

The intuition behind the KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. The distance can be of any type e.g Euclidean or Manhattan etc. It then selects the K-nearest data points, where K can be any integer. Finally it assigns the data point to the class to which the majority of the K data points belong. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc. here k =5

3.5 Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

4 About Libraries Function

These libraries function help us to extract features from sound file. later on the basis of thses features we classify the speech in angry,sad,netural and happy emotion.

4.1 MFCC

MFCC are popular features extracted from speech signals for use in recognition tasks.In the source-filter model of speech, MFCC are understood to represent the filter (vocal).The frequency response of the vocal tract is relatively smooth, whereas the source of voiced speech can be modeled as an impulse train. The result is that the vocal tract can be estimated by the spectral envelope of a speech segment. The motivating idea of MFCC is to compress information about the vocal tract (smoothed spectrum) into a small number of coefficients.

4.2 Tonnetz

It will Compute the tonal centroid features. The Tonal Centroids (or Tonnetz) contain harmonic content of a given audio signal. Constructor of the class. Object containing the file paths from where to extract/read the features

4.3 Chroma

The chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form. Therefore chroma features can be considered as important prerequisite for high-level semantic analysis, like chord recognition or harmonic similarity estimation. A better quality of the extracted chroma feature enables much better results in these high-level tasks. Short Time Fourier Transforms and Constant Q Transforms are used for chroma feature extraction.

4.4 mel-scaled spectrogram

If a time-series input y, sr is provided, then its magnitude spectrogram S is first computed, and then mapped onto the mel scale by $\text{mel}_f.dot.Byde fault, power = 2operatesonapowerspectrun$

4.5 spectral contrast

Each frame of a spectrogram S is divided into sub-bands. For each sub-band, the energy contrast is estimated by comparing the mean energy in the top quantile (peak energy) to that of the bottom quantile (valley energy). High contrast values generally correspond to clear, narrow-band signals, while low contrast values correspond to broad-band noise.

5 Functions Result

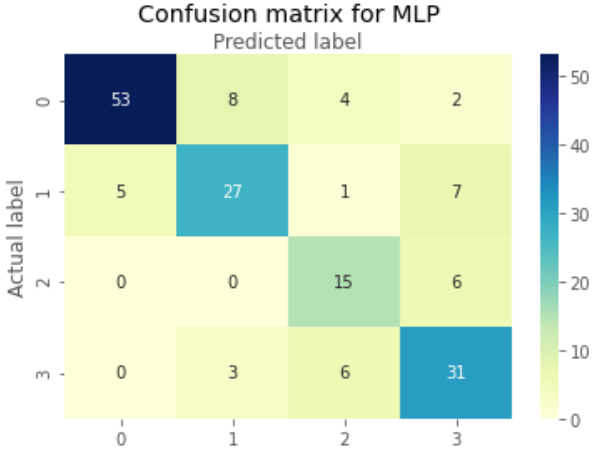
Table 1: Result by using various function

Case	MFCC	Mel	chroma	tonnetz	contrast	no. of features
1	1	0	0	0	0	40
2	0	1	0	0	0	128
3	0	0	1	0	0	12
4	0	0	0	1	0	7
5	0	0	0	0	1	6

6 classification report and confusion matrix of all classifier

classification report for MLP

	precision	recall	f1-score	support
angry	0.91	0.79	0.85	67
sad	0.71	0.68	0.69	40
neutral	0.58	0.71	0.64	21
happy	0.67	0.78	0.72	40
accuracy			0.75	168
macro avg	0.72	0.74	0.72	168
weighted avg	0.77	0.75	0.75	168



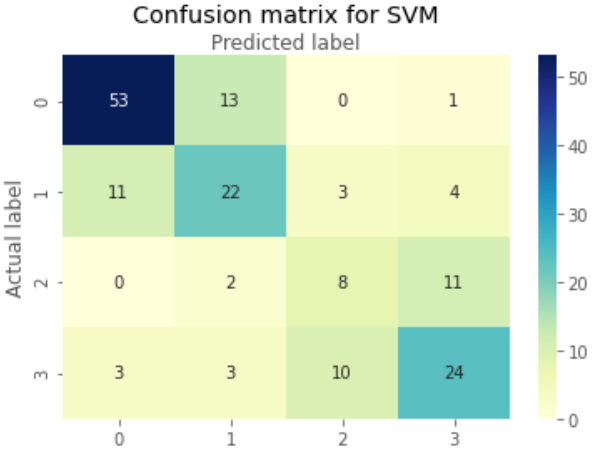
classification report for SVM

	precision	recall	f1-score	support
angry	0.79	0.79	0.79	67
sad	0.55	0.55	0.55	40
neutral	0.38	0.38	0.38	21
happy	0.60	0.60	0.60	40
accuracy			0.64	168
macro avg	0.58	0.58	0.58	168
weighted avg	0.64	0.64	0.64	168

7 Observation

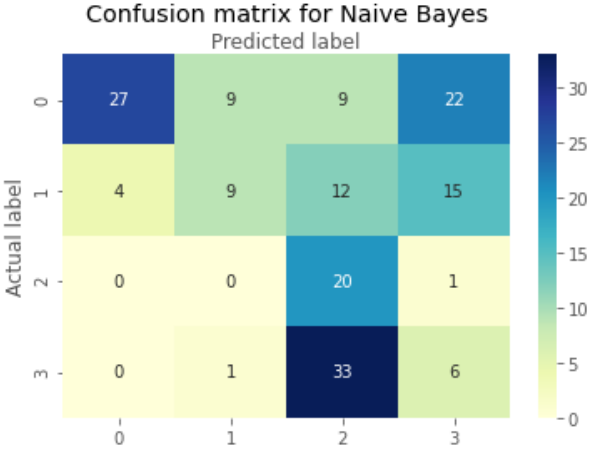
The function Result table represent which function will give how many features. If we notice there MFCC and MEL are providing almost 80 percent of the features to aur data.

The Best Result is given by Multi-Layer Perceptron this is because this classifier used neural network.It learn to model the correlation (or dependencies) between those in-puts and out-



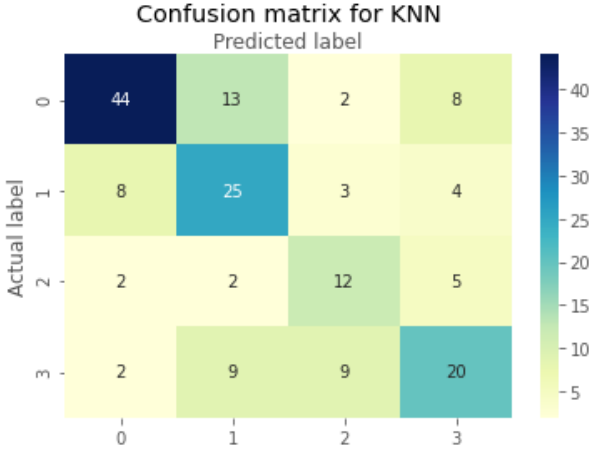
classification report for Naive Bayes:

	precision	recall	f1-score	support
angry	0.87	0.40	0.55	67
sad	0.47	0.23	0.31	40
neutral	0.27	0.95	0.42	21
happy	0.14	0.15	0.14	40
accuracy			0.37	168
macro avg	0.44	0.43	0.36	168
weighted avg	0.53	0.37	0.38	168



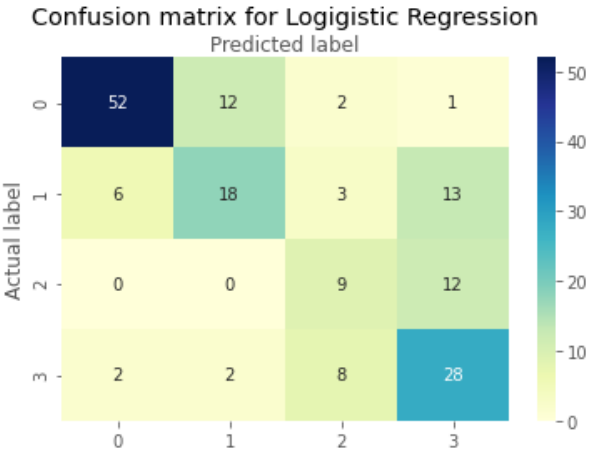
classification report for KNN

	precision	recall	f1-score	support
angry	0.79	0.66	0.72	67
happy	0.51	0.62	0.56	40
neutral	0.46	0.57	0.51	21
sad	0.54	0.50	0.52	40
accuracy			0.60	168
macro avg	0.57	0.59	0.58	168
weighted avg	0.62	0.60	0.61	168



classification report for Logistic

	precision	recall	f1-score	support
angry	0.87	0.78	0.82	67
sad	0.56	0.45	0.50	40
neutral	0.41	0.43	0.42	21
happy	0.52	0.70	0.60	40
accuracy			0.64	168
macro avg	0.59	0.59	0.58	168
weighted avg	0.65	0.64	0.64	168



puts. Training involves adjusting the parameters, or the weights and biases, ofthe model in order to minimize error.

Naive Bayes performed worse among all since because it assumes all the features to be independent but the classification of speech is dependent on the order of the speech features. Although other algorithms also don't consider the ordering but still Naive Bayes showed the worst performance.

if we take a look at logistic and svm both have almost same accuracy this is because i m using linear kernel in svm.

confusion matrix of all classifiers will of show how many of a particular class speech classified wrong and classified in which class out of 4 class.

8 Conclusion

In this project I detect the 4 emotions angry, sad, happy and neutral.First extract the features from the Ravdess dataset with the help of librosa library functions then i train those features and then test them on 5 different classifier with test size of 0.25. I compared the result of each classifier and found maximum accuracy with ANN model of Multi-Layer perceptrons((MLP)