11/06/2020

Report

SPEECH EMOTION RECOGNITION (SER)

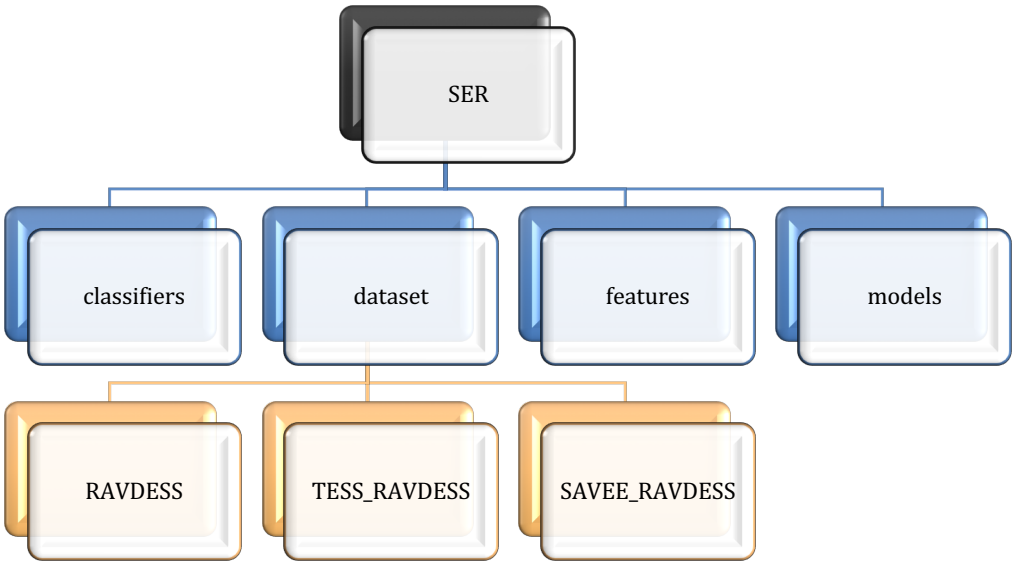
Dewansh Kr Singh

dewansh@iit.ac.in

+91-7906806147

Abstract: To extract out the emotional cues from speech samples on the basis of several spectral feature extraction methods from the librosa library (MFCC, mel, chroma) and using these emotion vectors and various classification algorithm (DT, SVM, MLP and CNN) to do a comparative study on their performances with the aim to achieve and increase accuracy in a speech emotion recognition system.

1. REPOSITORY STRUCTURE



1. classifiers – Contains python notebooks for various classifiers.
2. dataset – Contains the downloaded dataset. Please refer to the file **dataset.pdf** present in the dataset directory before continuing with this report.
3. features – Contains the saved joblib files for the extracted features.
4. models – Contains the saved models (.h5 files) for the CNN classifier.

2. FEATURE EXTRACTION

Feature extraction is an important step towards fetching the useful features from the speech signal.The major gaol is to find a low-resource consuming sequence of feature vectors that provide a compact representation of the input speech signal .

I have made use of the Librosa library for spectral feature extraction.

https://librosa.github.io/librosa/

I have extracted three features MFCC, mel and chroma and used their combination as below

1. MFCC
2. MFCC+mel
3. MFCC+mel+chroma

There were two possible ways to extract the features. Either we could have used a certain duration of the speech sample ,extract those features (get a 2d array) and take mean along the Axis =0 which would have given us a a feature vector whose length would depend on the duration we chose.

Or we could have used the entire length of speech sample and extracted the features , taking mean along the Axis =1 resulting into same feature vector length for all audio samples irrespective of their duration.

Duartions of Different Datasets

Dataset	Sample lenght	Duration chosen for feature extraction
RAVDESS Song Samples	4-5 s	3.5s
RAVDES Speech Samples	3-5s	2.5s
TESS samples	1-3s	1s
SAVEE samples	2-5s	None

Note – SAVEE dataset wasn't suitable for axis =0 because of the range differece in the length of the speech samples. Also it has silence voice in the start and end and thus require additional preprocessing on data such as length normalisation by trimming, zero padding etc.

128 MFCC , 128 mel and 12 chroma features are extracted using the librosa library feature extraction.

A brief summary of the features extracted along with their name and shape is presented in a tabular form in features/**FeatureSummary.png**

Various plots and waveforms for the sample_speech can be seen with the **feature_plot.ipynb** file present in the main directory.

---

3. CHALLENGES

Emotions are subjective thing and thus it requires multiple person to label emotions in a speech sample. It is an extensive task and thus there is clearly a lack of emotion labbled audio samples and even if you find a few resources chances are that:

- They are recorded in different environments with different parameters, such as no of actors, whether they are male or female.
- Length of the speech samples is varying.
- Different language
- Presence or absence of some additional emotions.
- Some are sentences, some recordings and some mere utterances.

Hence accumulating a good dataset is a complicated tast in itself. Also defining a definite set/collection of features for emotional identification varies widely.

All these factors contribute to the varying accuracies for the various classification models

---

4. RESULTS

I made use of four classification algorithms namely:

- Decision Tree (DT)
- State Vector Machine (SVM)
- Multi Layer Perceptron (MLP)
- Convolutional Neural Network (CNN)

The code for each classifier along with its expexted output/accuracy can been seen in the directory **classifiers**

There are two files corresponding to each classifier, one in which the features are extracted and mean is taken along the axis=0 and the other with axis =1.

Here I'll summarise the results for a quick reference and comparison

First define the feature sets as follows:

1. **Feature Set A = {MFCC}**
2. **feature Set B = {MFCC , mel}**
3. **Feature Set C = {MFCC, mel, chroma}**

AXIS = 0

Feature Set A:

Dataset		Classification Algorithm			
Train	Test	DT	SVM	MLP	CNN
Ravdess_split1_song 528 samples	Ravdess_split2_song 484 samples	24%	30%	23%	36%
Ravdess_split1_speech 720 samples	Ravdess_split2_speech 720 samples	19%	22%	23%	23%
0.8Ravdess_song 809 samples	0.2Ravdess_song 203 samples	30%	27%	19%	41%
0.8Ravdess_speech 1152 samples	0.8Ravdess_speech 288 samples	23%	22%	23%	25%
0.8Tess 2240 samples	0.2Tess 560 samples	44%	42%	60%	58%

Feature Set B:

Dataset		Classification Algorithm			
Train	Test	DT	SVM	MLP	CNN
Ravdess_split1_song 528 samples	Ravdess_split2_song 484 samples	28%	35%	24%	39%
Ravdess_split1_speech 720 samples	Ravdess_split2_speech 720 samples	22%	22%	24%	26%
0.8Ravdess_song 809 samples	0.2Ravdess_song 203 samples	38%	34%	63%	48%
0.8Ravdess_speech 1152 samples	0.8Ravdess_speech 288 samples	24%	28%	33%	36%
0.8Tess 2240 samples	0.2Tess 560 samples	75%	68%	88%	89%

Feature Set C:

Dataset		Classification Algorithm			
Train	Test	DT	SVM	MLP	CNN
Ravdess_split1_song 528 samples	Ravdess_split2_song 484 samples	30%	34%	28%	35%
Ravdess_split1_speech 720 samples	Ravdess_split2_speech 720 samples	25%	25%	28%	32%
0.8Ravdess_song 809 samples	0.2Ravdess_song 203 samples	36%	46%	49%	42%
0.8Ravdess_speech 1152 samples	0.8Ravdess_speech 288 samples	27%	27%	34%	35%
0.8Tess 2240 samples	0.2Tess 560 samples	78%	84%	95%	93%

AXIS = 1

Feature Set A:

Dataset		Classification Algorithm			
Train	Test	DT	SVM	MLP	CNN
Ravdess_split1_song 528 samples	Ravdess_split2_song 484 samples	46%	76%	71%	70%
Ravdess_split1_speech 720 samples	Ravdess_split2_speech 720 samples	20%	37%	33%	34%
0.8Ravdess_song 809 samples	0.2Ravdess_song 203 samples	66%	84%	86%	85%
0.8Ravdess_speech 1152 samples	0.8Ravdess_speech 288 samples	32%	52%	64%	59%
0.8Tess 2240 samples	0.2Tess 560 samples	88%	99%	99%	99%
0.8Ravdess 1961 samples	0.2Ravdess 461 samples	41%	56%	6%	63%
0.8Ravdess+Tess 4201 samples	0.2Ravdess+Tess 1051 samples	71%	74%	82%	85%
0.8Ravdess+Tess+Savee 4585 samples	0.2Ravdess+Tess+Savee 1147 samples	64%	70%	84%	84%

Feature Set B:

Dataset		Classification Algorithm			
Train	Test	DT	SVM	MLP	CNN
Ravdess_split1_song 528 samples	Ravdess_split2_song 484 samples	50%	73%	76%	73%
Ravdess_split1_speech 720 samples	Ravdess_split2_speech 720 samples	26%	38%	34%	38%
0.8Ravdess_song 809 samples	0.2Ravdess_song 203 samples	62%	87%	91%	83%
0.8Ravdess_speech 1152 samples	0.8Ravdess_speech 288 samples	36%	53%	63%	48%
0.8Tess 2240 samples	0.2Tess 560 samples	91%	99%	99%	99%
0.8Ravdess 1961 samples	0.2Ravdess 461 samples	40%	60%	72%	65%
0.8Ravdess+Tess 4201 samples	0.2Ravdess+Tess 1051 samples	65%	73%	87%	84%
0.8Ravdess+Tess+Savee 4585 samples	0.2Ravdess+Tess+Savee 1147 samples	67%	73%	83%	85%

Feature Set C:

Dataset		Classification Algorithm			
Train	Test	DT	SVM	MLP	CNN
Ravdess_split1_song 528 samples	Ravdess_split2_song 484 samples	48%	73%	75%	74%
Ravdess_split1_speech 720 samples	Ravdess_split2_speech 720 samples	27%	38%	34%	38%
0.8Ravdess_song 809 samples	0.2Ravdess_song 203 samples	68%	83%	92%	89%
0.8Ravdess_speech 1152 samples	0.8Ravdess_speech 288 samples	35%	48%	56%	48%
0.8Tess 2240 samples	0.2Tess 560 samples	91%	99%	99%	99%
0.8Ravdess 1961 samples	0.2Ravdess 461 samples	43%	57%	67%	69%
0.8Ravdess+Tess 4201 samples	0.2Ravdess+Tess 1051 samples	70%	77%	86%	84%
0.8Ravdess+Tess+Savee 4585 samples	0.2Ravdess+Tess+Savee 1147 samples	66%	70%	82%	85%

5. FUTURE WORK

- Using on of the trained model to predict on live speech samples throigh microphone
- Preprocess the data with methods like length normalisation, speaker normalisation, zero padding etc
- Apply a Recurrent Neural Network approach.
- Implement Feature Selection Techniques to reduce the Size of the overall feature joblib files which would decrease the model training time and thus would also be beneficial in live prediction.
- Using the reduced feature set for targeted emotional conversion of speech signals through Generative Adversarial Networks (GANs)

Resources

ReferenceArticles-

[1]HTTPS://CORE.AC.UK/DOWNLOAD/PDF/82526915.PDF

[2]HTTP://WWW.JATIT.ORG/VOLUMES/VOL79NO1/5VOL79NO1.PDF

[3]HTTPS://LINK.SPRINGER.COM/CONTENT/PDF/10.3758/S13428-017-0873-Y.PDF

[4]HTTPS://WWW.SCSS.TCD.IE/~CABRALJ/WEB/ARTIGOS/JPC_INTERSPEECH_2006.PDF

[5] HUMAN-CENTRIC INTERFACES FOR AMBIENT INTELLIGENCE

[6] HTTPS://LIBROSA.GITHUB.IO/LIBROSA/FEATURE.HTML#SPECTRAL-FEATURES

GithubRepos-

[1] MITESHPUTHRANNEU

[2] REZACHU

[3] MARCOGDEPINTO

[4] X4NTH055