VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

PROJECT WORK-4 REPORT
on

"Speech Emotion Recognition"

Submitted by

Prithvi J (1BM19CS122)
S Shree Lakshmi (1BM19CS136)
Sohan R Kumar (1BM19CS159)
Suraj Nair (1BM19CS163)

Under the Guidance of
Prof. Rekha G S
Assistant Professor, BMSCE

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE AND ENGINEERING

B. M. S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)

BENGALURU-560019
April-2022 to July-2022

B. M. S. College of Engineering,

Bull Temple Road, Bangalore 560019
(Affiliated To Visvesvaraya Technological University, Belgaum)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project work entitled "**Speech Emotion Recognition**" was carried out by **Prithvi J (1BM19CS122), S Shree Lakshmi (1BM19CS136), Sohan R Kumar(1BM19CS159), and Suraj Nair(1BM19CS163)** who are Bonafede students of **B. M. S. College of Engineering.** It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visveswaraya Technological University, Belgaum during the year 2021. The project report has been approved as it satisfies the academic requirements in respect of **Project Work-4 (20CS6PWPW4)** work prescribed for the said degree.

Signature of the Guide
Prof. Rekha G S
Assistant Professor
BMSCE, Bengaluru

Signature of the HOD
Dr. Jyothi S Nayak
Professor & Head, Dept. of CSE
BMSCE, Bengaluru

External Viva

Name of the Examiner — Signature with date

1._____ _____

2. _____ _____

B. M. S. COLLEGE OF ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECALARATION

We, Prithvi J (1BM19CS122), S Shree Lakshmi (1BM19CS136), Sohan R Kumar(1BM19CS159), and Suraj Nair**(**1BM19CS163**)**, students of 5th Semester, B.E, Department of Computer Science and Engineering, B. M. S. College of Engineering, Bangalore, hereby declare that, this Project Work-4 entitled " **Speech Emotion Recognition** " has been carried out by us under the guidance of Prof. Rekha G S, Assistant Professor, Department of CSE, B. M. S. College of Engineering, Bangalore during the academic semester Mar-2021-Jun-2021

We also declare that to the best of our knowledge and belief, the development reported here is not from part of any other report by any other students.

Signature

Prithvi J (1BM19CS122)

S Shree Lakshmi (1BM19CS136)

Sohan R Kumar(1BM19CS159)

Suraj Nair(1BM19CS163)

1. Introduction

In recent years, human-machine interactions have become highly representative of a realistic interpersonal interaction. Speech analysis has played an integral role in reducing this gap between physical and digital world. Since a lot of information in human speech is conveyed through emotional cues, there has been a growing interest in the subfield of emotion recognition. However, automatic speech emotion recognition (SER) is a challenging task as it heavily depends on the effectiveness of the features used for classification. The primary objective of this project is to compare the performance of two classifiers: (i) conventional classifiers such as Support Vector Machine (SVM) that rely on the hand-picked features provided as input (ii) deep learning models such as a Recurrent Neural Network (RNN) that can automatically discover emotionally relevant features from speech. The proposed solution was evaluated on the RAVDESS corpus and the classification result verified that RNN model provides better accuracy than that achieved by using conventional machine learning classification methods.

Speech not only serves as the most natural and effective way of communication but also carries of the most expressive modalities for human emotions. Emotions play an important role in the interactions between human beings as it influences most aspects of communication such as facial expressions, body gestures, voice and tonal properties and the linguistic content [1]. For an efficient interaction, we need to recognize and understand the correct emotion of the other person and be able to deliver an appropriate reaction. The role of emotional expressions is not limited only to the human world. Automatic speech emotion recognition has recently become a vast research field and has found applications in areas like psychology, psychiatry, behavioral science, artificial intelligence, computer vision and humanmachine interactions [1-3]. However, one of the fundamental challenges in automatic speech emotion recognition has been the identification and extraction of appropriate features from speech. There have been many endeavors taken to discover speech features that can be indicative of different emotions [4-5]. Though both short-term frame-level as well as long-term utterance level features have been proposed, shown in Table I, there is still no definitive answer for which features are the appropriate descriptors for emotions. In this project, the performance of speech emotion recognition is compared between two methods. Conventional classifiers that uses machine learning algorithms has been used for decades in recognizing emotions from speech. However, in recent years, deep learning methods have taken the center

stage and have gained popularity for their ability to perform well without any input hand-crafted features. Speech emotion sets obtained from RAVDESS corpus is classified using a conventionally used Support Vector Machine (SVM) and its performance is compared to that of a bidirectional long short-term memory (LSTM). The inspiration behind this project comes from the studies done by [6] which couples a recurrent neural network (RNN) with an attention mechanism to enable the model to focus on emotionally salient part of the sentence, significantly improving the accuracy to that of SVM.

Motivate and abstractly describe the problem you are addressing and how you are addressing it. What is the problem? Why is it important? What is your basic approach? A short discussion of how it fits into related work in the area is also desirable. Summarize the basic results and conclusions that you will present.

Recently, attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. There are many systems proposed in the literature to identify the emotional state through speech.

In a voice-based system, a computer agent is required to completely comprehend the human's speech percept in order to accurately pick up the commands given to it. This field of study is termed as Speech Processing and consists of three components:

- Speaker Identification
- Speech Recognition
- Speech Emotion Detection

Speech Emotion Detection is challenging to implement among the other components due to its complexity.

1.1 Motivation

Emotion is an integral part of human behavior and inherited property in all mode of communication. We, human is well trained thought your experience reading recognition of various emotions which make us more sensible and understandable. But in case of machine, however, it can easily understand content based information such as information in text, audio or video but still far behind to access the depth

behind the content. It is need of the era that machine should also be trained to understand emotions correctly for better understanding and to avoid any miscommunication. Present study comes into domain of emotion recognition from audio conversation. Moreover, Audio emotion analysis has many applications in various sectors like healthcare, banking, defense and IT. On the other part, text emotions are easy to decode as there is no role of factors like tone and pitch, but in case of audio emotion analysis both the factors need attention for better accuracy. Also there are several factors like noise, disturbance, and various pauses in communication which results in degrading the accuracy. It is a challenging task to make machine to understand the emotion of the respondents.

1.2 Scope of the Project

- Speech is the fast and best normal way of communicating amongst human.
- The recognition of emotional speech aims to recognize the emotional condition of individual utterer by applying his/her voice automatically.
- We are going to develop a machine learning project to automatically classify the given input sample (audio) into a few predefined emotions.
- In this project we transform the audio data into numeric include Mel Spectrograms that visualize audio signals based on their frequency components which can be plotted as an audio wave and fed to train a CNN as an image classifier. We can capture this using Mel-frequency cepstral coefficients (MFCCs)

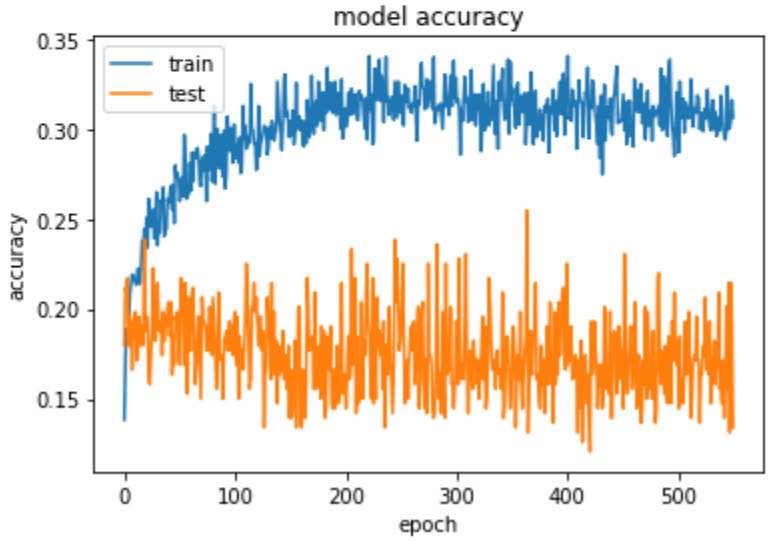
1.2 Problem statement

- In this project, We transform the audio data into numeric include Mel Spectrograms that visualize audio signals based on their frequency components which can be plotted as an audio wave and fed to train a CNN as an image classifier. We can capture this using MFCCs.
- After extracting features from the audio, the popular choice of model architecture has changed over time.

2. Literature Survey

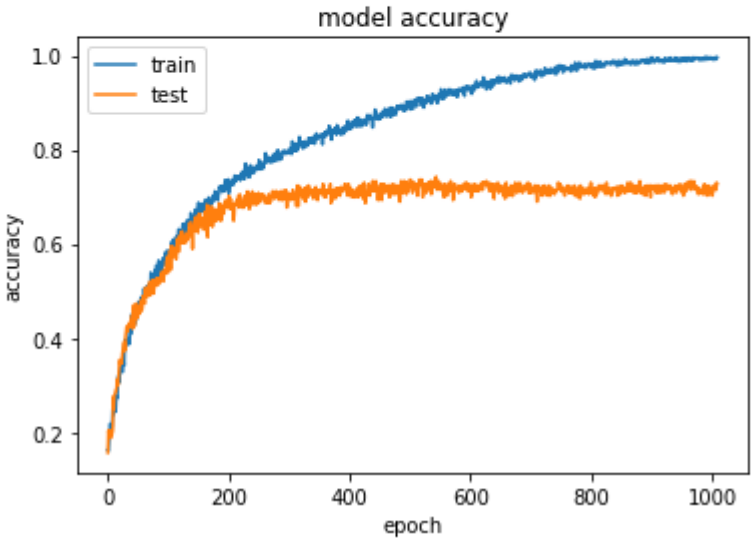
MLP Model: The MLP model we created had a very low validation accuracy of around 25% with 8 layers, softmax function at the output, batch size of 32 and 550 epochs.

```
In [105]: plt.plot(history.history['acc'])
          plt.plot(history.history['val_acc'])
          plt.title('model accuracy')
          plt.ylabel('accuracy')
          plt.xlabel('epoch')
          plt.legend(['train', 'test'], loc='upper left')
          plt.show()
```



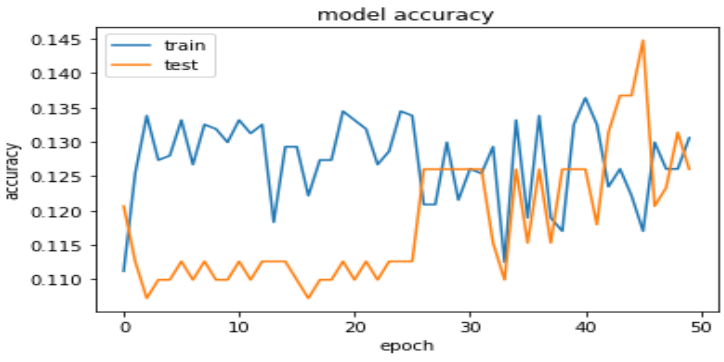
CNN : CNN model was the best for our classification problem. After training numerous models we got the best validation accuracy of 70% with 18 layers, softmax activation function, rmsprop activation function, batch size of 32 and 1000 epochs.

```
In [110]: #sigmoid
plt.plot(cnnhistory.history['acc'])
plt.plot(cnnhistory.history['val_acc'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
```



LSTM: The LSTM model had the lowest training accuracy of around 15% with 5 layers, tan h activation function, batch size of 32 and 50

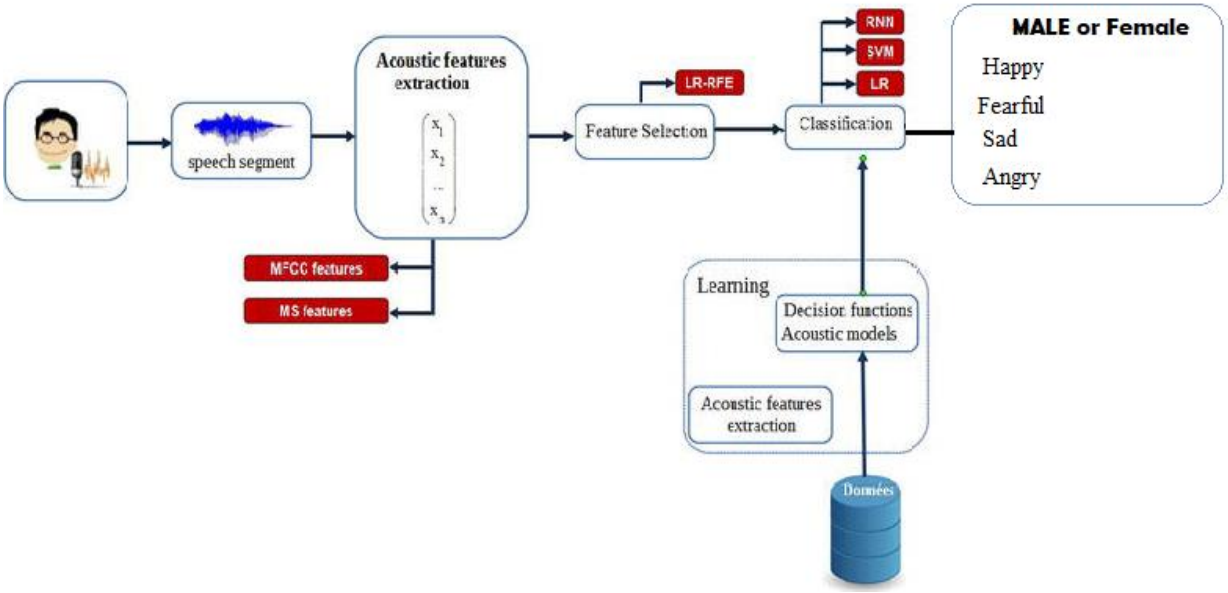
```
In [91]: plt.plot(lstmhistory.history['acc'])
plt.plot(lstmhistory.history['val_acc'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
```



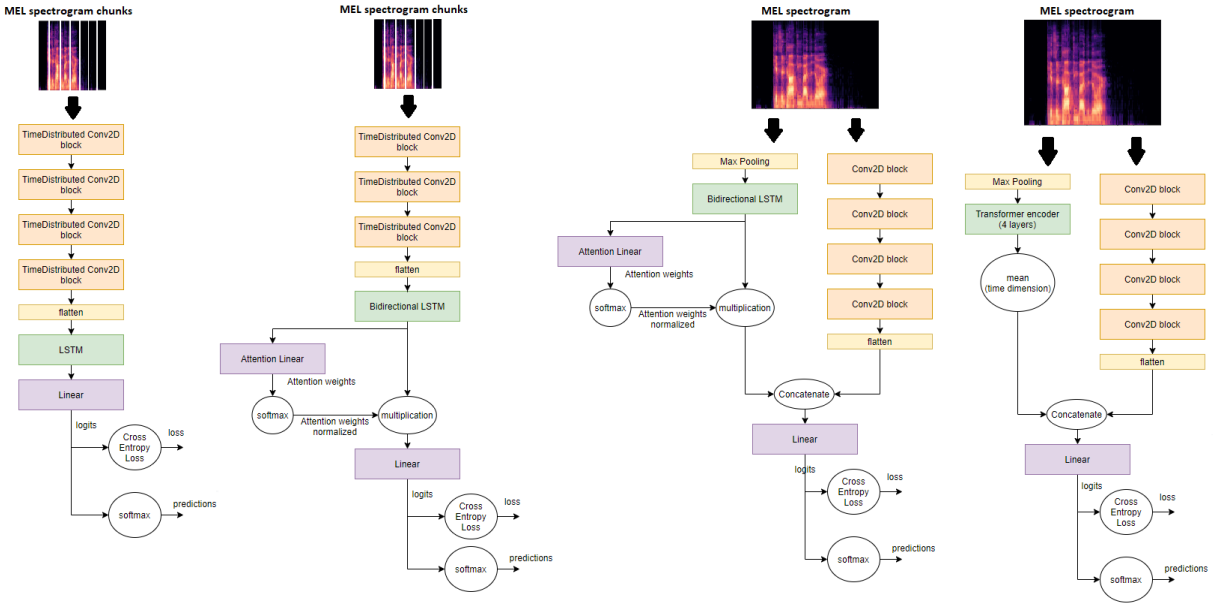
- After comparing 3 different models, we have found our best CNN model for our emotion classification problem. We achieved a validation accuracy of 90% with our existing model.
- Our model could perform better if we have more data to work on. What's more surprised is that the model performed excellent when distinguishing between a males and females voice.

3. Design

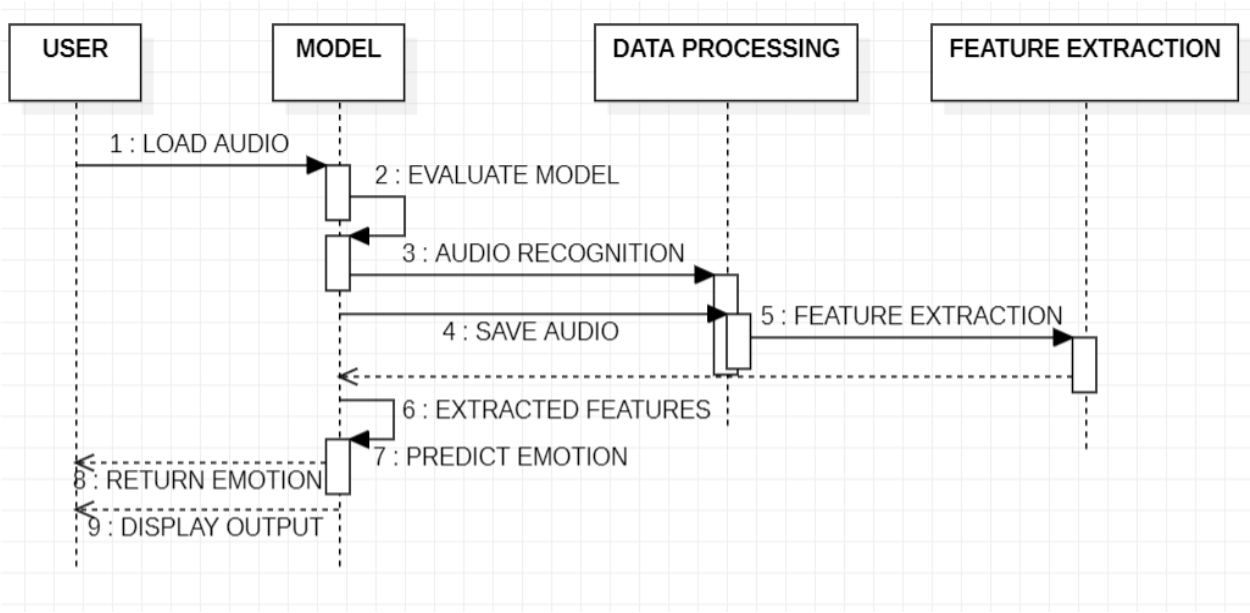
3.1 High Level Design



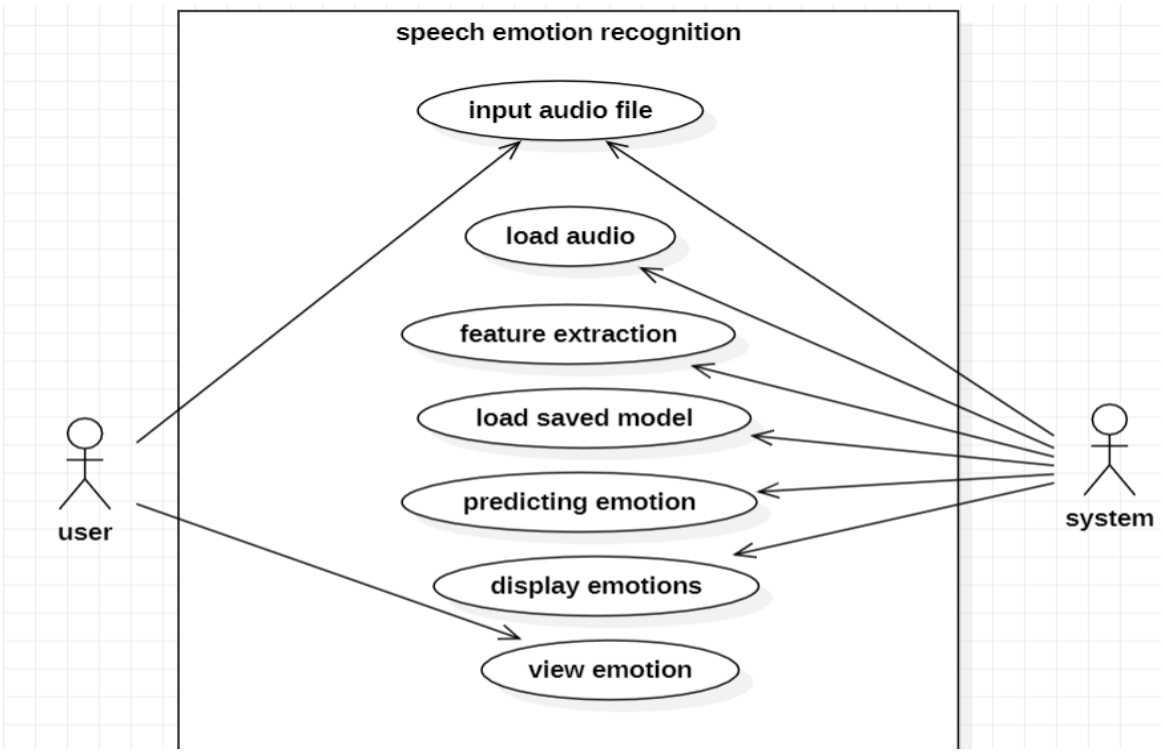
3.2 Detailed Design



3.3 Sequence Diagram



3.4 Use Case Diagram



4. Implementation

4.1 Proposed methodology

The speech emotion detection system is implemented as a Machine Learning (ML) model.The flowchart represents a pictorial overview of the process.
1.data collection:The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data.
2.feature engineering:is a collection of several machine learning tasks that are executed over the collected data.
3.algorithmic based model is developed. This model uses an ML algorithm to learn about the data and train itself to
respond to any new data it is exposed to.
4.is to evaluate the functioning of the built model.
Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms.

4.2 Algorithm used for implementation

This repository can be used to build machine learning classifiers as well as regressors for the case of 3 emotions {'sad': 0, 'neutral': 1, 'happy': 2} and the case of 5 emotions {'angry': 1, 'sad': 2, 'neutral': 3, 'ps': 4, 'happy': 5}

Classifiers

- SVC
- RandomForestClassifier
- GradientBoostingClassifier
- KNeighborsClassifier
- MLPClassifier
- BaggingClassifier• Recurrent Neural Networks (Keras)

Regressors

- SVR
- RandomForestRegressor
- GradientBoostingRegressor
- KNeighborsRegressor
- MLPRegressor
- BaggingRegressor
- Recurrent Neural Networks (Keras)

4.3 Tools and technologies sued

- OS-Windows 10/11
- Programming Language-Python
- jupyter Notebook
- Librosa
- Dataset-RAVDESS

Python Packages

- tensorflow
- librosa==0.6.3
- numpy
- pandas
- soundfile==0.9.0
- wave
- scikit-learn==0.24.2

- tqdm==4.28.1
- matplotlib==2.2.3
- pyaudio==0.2.11
- ffmpeg (optional): used if you want to add more sample audio by converting to 16000Hz sample rate and mono channel which is provided in convert_wavs.py

Dataset

This repository used 4 datasets (including this repo's custom dataset) which are downloaded and formatted already in data folder:

- **RAVDESS :** The Ryson Audio-Visual Database of Emotional Speech and Song that contains 24 actors (12 male, 12 female), vocalizing two lexically-matched statements in a neutral North American accent.

- **TESS :** Toronto Emotional Speech Set that was modeled on the Northwestern University Auditory Test No. 6 (NU-6; Tillman & Carhart, 1966). A set of 200 target words were spoken in the carrier phrase "Say the word _____' by two actresses (aged 26 and 64 years).

- **EMO-DB :** As a part of the DFG funded research project SE462/3-1 in 1997 and 1999 we recorded a database of emotional utterances spoken by actors. The recordings took place in the anechoic chamber of the Technical University Berlin, department of Technical Acoustics..

- **Custom :** Some unbalanced noisy dataset that is located in data/train-custom for training and data/test-custom for testing in which you can add/remove recording samples easily by converting the raw audio to 16000 sample rate, mono channel (this is provided in create_wavs.py script in convert_audio(audio_path) method which requires ffmpeg to be installed and in PATH) and adding the emotion to the end of audio file name separated with '_' (e.g "20190616_125714_happy.wav" will be parsed automatically as happy)

Emotions available There are 9 emotions available: "neutral", "calm", "happy" "sad", "angry", "fear", "disgust", "ps" (pleasant surprise) and "boredom".

4.4 Testing

Determining the best model

```
# loads the best estimators from `grid` folder that was searched by GridSearchCV in `grid_search.py`,

# and set the model to the best in terms of test score, and then train it

rec.determine_best_model()

# get the determined sklearn model name

print(rec.model.__class__.__name__, "is the best")

# get the test accuracy score for the best estimator

print("Test score:", rec.test_score())
```

output

```
MLPClassifier is the best

Test Score: 0.8958333333333334
```

Predicting

```
 is a neutral speech from emo-db from the testing set

print("Prediction:", rec.predict("data/emodb/wav/15a04Nc.wav"))

# this is a sad speech from TESS from the testing set

print("Prediction:",
rec.predict("data/validation/Actor_25/25_01_01_01_back_sad.wav"))
```

output

```
Prediction: neutral Prediction: sad
```

Using RNNs For Emotions

```
from deep_emotion_recognition import DeepEmotionRecognizer
```

```
# initialize instance

# inherited from emotion_recognition.EmotionRecognizer

# default parameters (LSTM: 128x2, Dense:128x2)

deeprec = DeepEmotionRecognizer(emotions=['angry', 'sad', 'neutral', 'ps', 'happy'], n_rnn_layers=2, n_dense_layers=2, rnn_units=128, dense_units=128)

# train the model

deeprec.train()

# get the accuracy

print(deeprec.test_score())

# predict angry audio sample

prediction = deeprec.predict('data/validation/Actor_10/03-02-05-02-02-02-10_angry.wav')

print(f"Prediction: {prediction}")
```

output

0.7717948717948718

Prediction: angry

5. Results and Discussion

Appropriate graphs must be included to portray the results.

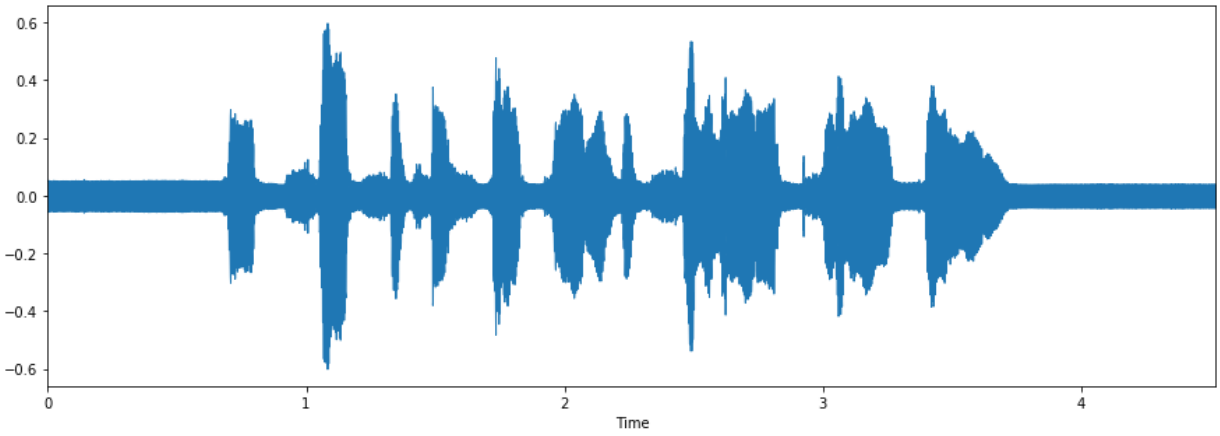


Figure:waveform of the audiofile

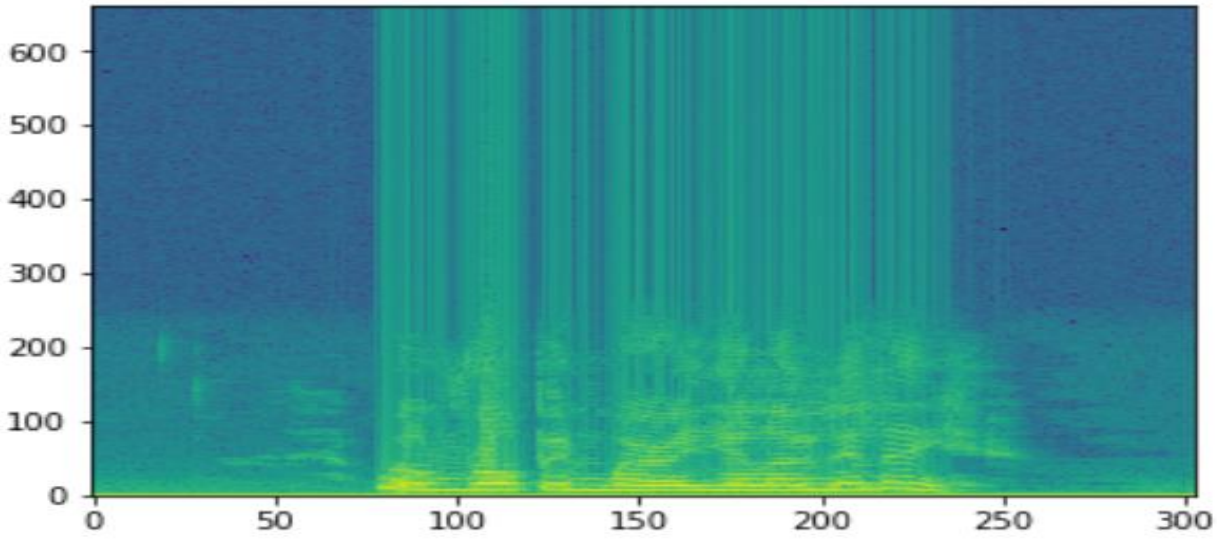


figure:spectometer

```
In [75]: finaldf[58:68]
```

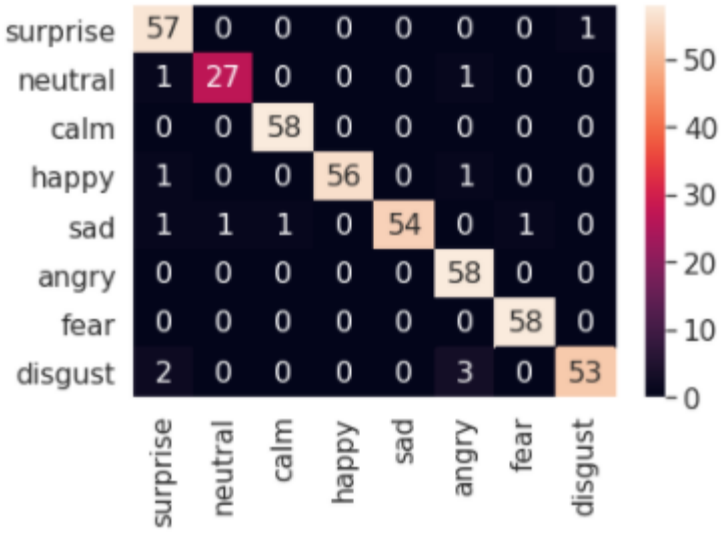
Out[75]:

	actualvalues	predictedvalues
58	male_fearful	male_happy
59	male_fearful	male_fearful
60	male_fearful	male_fearful
61	male_fearful	male_fearful
62	male_sad	male_sad
63	male_fearful	male_fearful
64	male_happy	male_happy
65	female_angry	female_angry
66	female_angry	female_fearful
67	male_angry	male_angry

Predicted output

Confusion matrix

	predicted_angry	predicted_sad	predicted_neutral	predicted_ps	predicted_happy
true_angry	80.769226	7.692308	3.846154	5.128205	2.564103
true_sad	12.820514	73.076920	3.846154	6.410257	3.846154
true_neutral	1.282051	1.282051	79.487183	1.282051	16.666668
true_ps	10.256411	3.846154	1.282051	79.487183	5.128205
true_happy	5.128205	8.974360	7.692308	8.974360	69.230774



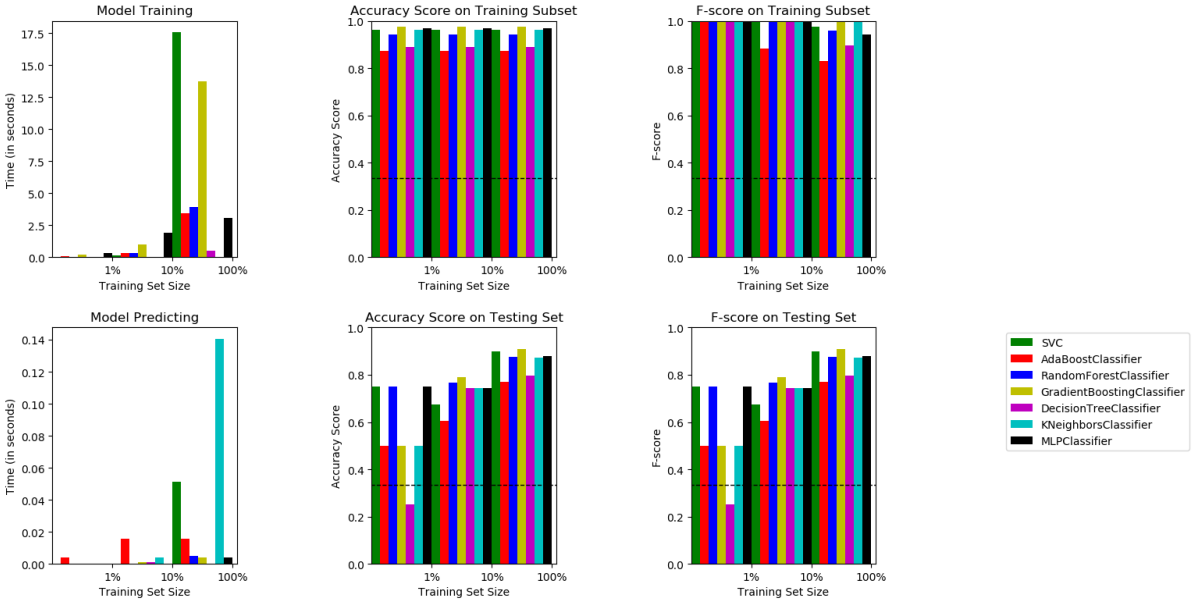
{'angry': 0.99878675, 'sad': 0.0009922335, 'neutral': 7.959707e-06, 'ps': 0.00021298956, 'happy': 8.3598025e-08}

output

0.7717948717948718

Prediction: angry

Plotting histograms



6. Conclusion and Future Work

Conclusion

The use of three features (MFCC's, Mel Spectrograms and chroma STFT) gave impressive accuracy in most of the models, reiterating the importance of feature selection. As with many data science projects, different features could be used and/or engineered. Tonnetz was originally used in modeling, however it led to decreased performance and was removed. Some other possible features to explore concerning audio would be MFCC Filterbanks or features extracted using the perceptual linear predictive (PLP) technique. These features could affect the performance of models in the emotion classification task.

Using feature extraction methods by itself did not achieve a high accuracy score within my CNN model, but using data augmentation methods did improve the accuracy score to 53% however it was overfitting the data. This model needs to be improved upon before being applied towards making an app to detect emotion in real time. Fine tuning the VGG-16 architecture with image augmentation improved the overall model accuracy to 90%/

Future work

An alternate approach that could be explored for this problem is splitting the classifying task into two distinct problems. A separate model could be used to classify gender and then separate models for each gender to classify emotion could be utilized. This could possibly lead to a performance improvement by segregating the task of emotion classification by gender.

It would be interesting to see how a human classifying the audio would measure up to these models, however, finding someone willing to listen to more than 2,400 audio clips may be a challenge in of itself because a person can only listen to "the children are talking by the door" or "the dogs are sitting by the door" so many times.

Next steps for this project include building a front-end for user interaction, then work towards building an app to detect emotion. Afterwards, I would like to be able to build system that can recognize emotion in real time and then calculate degree of affection such as love, truthfulness, and friendship of the person you are talking to.

- The training data I list above (Berlin) may insufficient, the validation accuracy and loss can't be improved while the training result is not good.

- Given sufficient training examples, the parameters of short-term characterization, long-term aggregation, and the attention model can be jointly optimized for best performance.

- Update the current network architecture to improve the accuracy (already in progress).

References:

Hasan, M., Islam, M.M., Zarif, M.I.I. and Hashem, M.M.A., “Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches”, Internet of Things Journal, Volume 7, Page no.10-15.

M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification

schemes, and databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.

R. Banse and K. R. Scherer, “Acoustic profiles in vocal emotion expression.,” Pers. Soc. Psychol, vol. 70, no.

, pp. 572–587, 1996.

V. Hozjan and Z. Kačič, “Context-Independent Multilingual Emotion Recognition from Speech Signals,” Int.

J. Speech Technol., vol. 6, no. 3, pp. 311–320, 2003..

F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional

speech.,” in Interspeech, 2005, vol. 5, pp. 1517–1520.

A. Schuller, B. , Steid, S. l, and Batliner, “The interspeech 2009emotion challengee,” Interspeech, pp. 312–

315, 2009.

W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion Recognition and Affective Computing on Vocal Social Media,”

Inf. Manag., Feb. 2015

S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral

features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and

spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary

decision tree approach,” Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.

T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun.,

APPENDIX: