Title of the Project:

SPEECH EMOTION ANALYSIS PROJECT.

A MINI-PROJECT REPORT

*By*

GAVIN CORRIEA (8325)

RIA D’MELLO (8330)

CLEONA PEREIRA (8477)

*Under the guidance of*

Prof. Dr.Sujata Deshmukh.



DEPARTMENT OF COMPUTER ENGINEERING

FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

FR. AGNEL ASHRAM, BANDRA (W),

MUMBAI - 400 050.

UNIVERSITY OF MUMBAI

(2019 – 2020)

FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

FR. AGNEL ASHRAM, BANDRA (W),

MUMBAI - 400 050.



CERTIFICATE

This is to certify that the following students working on the project “Speech Emotion Analysis ” have satisfactorily completed the requirements of the project in fulfillment of the course T.E in Computer Engineering of the University of Mumbai during academic year 2019-2020 under the guidance of “Prof. Dr.Sujata Deshmukh.’’

Submitted By: GAVIN CORRIEA (8325)

RIA D’MELLO (8330)

CLEONA PEREIRA (8477)

|  |  |
| --- | --- |
| Prof. Dr.Sujata Deshmukh.  Guide | Dr. B.S.Daga  Head of the Department |

\_\_\_\_\_\_\_\_

Principal

|  |
| --- |
|  |
|  |

CERTIFICATE

This is to certify that the project synopsis entitled “Speech Emotion Analysis” submitted by the following students is found to be satisfactory and the report has been approved as it satisfies the academic requirements in respect of mini-project work prescribed for the course.

NAME: GAVIN CORRIEA (8325)

RIA D’MELLO (8330)

CLEONA PEREIRA (8477)

|  |  |
| --- | --- |
| Internal Examiner  (Signature)  Name:  Date: | External Examiner  (Signature)  Name:  Date: |

Seal of the Institute

DECLARATION OF THE STUDENT

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources.

We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea / data / fact / source in my submission.

We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature of the student

Date:

GAVIN CORRIEA

(8325)

Signature of the student

Date:

RIA D’MELLO

(8330)

Signature of the student

Date:

CLEONA PEREIRA

(8477)

ACKNOWLEDGEMENT

We have taken efforts in this project. However ,it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend my sincere thanks to all of them.

Our special thanks to Prof Dr.Sujata Deshmukh for her guidance and constant supervision as well as for providing necessary information regarding the project.

We would like to thank our parents and members of FR.CRCE, Bandra, for their kind cooperation and encouragement which helped us in completion of this project.

ABSTRACT

Introduction

With rapid developments in the design of deep architecture models and learning algorithms, methods referred to as deep learning have come to be widely used in a variety of research areas such as pattern recognition, classification, and signal processing. Deep learning methods are being applied in various recognition tasks such as image, speech, and music recognition. Convolutional Neural Networks (CNNs) especially show remarkable recognition performance for computer vision tasks. In this project, we investigate the result of the Speech Emotion Analysis (SEA) algorithm based on CNNs.

Problem Statement

Speech Emotion Analysis using convolutional neural networks.

Table of Contents

Chapter 1: Introduction........................................................................................................ 3

Chapter 2: Literature Review................................................................................................ 4

Chapter 3: Proposed System............................................................................................... .. 6

3.1: Problem statement Analysis............................................................................. 6

3.2: Design and Methodology of proposed system................................................. 7

3.3: Algorithms

Chapter 4: Hardware software requirements and Implementation....................................... 12

4.1: Results

4.2: Conclusion

Appendix / Appendices......................................................................................................... 14

References(IEEE Format.......................................................................................................14

1. Introduction

Multimedia pattern recognition is an emerging technology that can extract and analyze large amounts of multimedia information from video and audio sources. In recent years, there has been a drastic growth in the application of machine learning technology using deep learning to solve various recognition problems. Speech Emotion Recognition (SER) is an especially significant task in understanding the characteristics of speech in the media. However, recognizing emotions from speech is a very challenging problem because people express emotions in different ways, and the features are unclear to distinguish the emotions . Conventional techniques for solving this problem are extracting low-level descriptors and training the machine appropriately through learning those features. These methods have been accepted as state of the art for many years in machine learning.However, selecting good features to extract is difficult, and optimization is even more difficult, often being significantly time-consuming in research, development, and validation. Because of this, the traditional trend in speech/audio information retrieval is to focus on the use of powerful strategies for semantic analysis, often relying on model selection to optimize the results . However, deep neural architectures can share low-level representations and naturally progress from low-level to high-level structures.

2. Review of Literature

Despite the great progress made in artificial intelligence, we are still far from being able to naturally interact with machines,partly because machines do not understand our emotion states.

Recently, speech emotion recognition, which aims to recognize emotion states from speech signals, has been drawing increasing attention.Researchers have developed various methods for emotion recognition in the speech signal. Traditional speech recognition methods are proposed in many research studies . However, these conventional approaches have the drawback of using hand-crafted features. More recently, deep neural networks have taken center stage in speech and language analysis. CNN-based image, video, speech, audio and music recognition methods have been proposed in several research studies . From these studies, we already know that CNN-based analysis can be applied in one-dimensional signals such as speech and audio . In particular, a CNNbased SER method has been proposed that learns salient features of SER using semi-CNNs. This method, however, has the drawback of feeding the learned features into a support vector machine classifier.The SER system using RNNs was proposed in , which accounts for long contextual effect in emotional speech and the uncertainty of emotional labels. In addition, audio/video based multimodal emotion recognition approaches were experimented in . Furthermore, the cross-corpus emotion recognition method was explored in , which investigates the ability of an emotion recognition detector to be applied to other databases.

Complete review on the speech emotion recognition is explained in [1] which reviews properties of dataset, speech emotion recognition study classifier choice. Various acoustic features of speech are investigated and some of the classifier methods are analyzed in [2] which is helpful in the further investigation of modern methods of emotion recognition. This paper [3] investigated the prediction of the next reactions from emotional vocal signals based on the recognition of emotions, using different categories of classifiers. Some of the classification algorithms like K-NN, Random Forest are used in [3] to classify emotion accordingly. Recurrent Neural network arises enormously which tries to solve many problems in the filed of data science. Deep RNN like LSTM, Bi-directional LSTM trained for acoustic features are used in [4]. Various range of CNN are being implemented and trained for speech emotion recognition are evaluated in [5]. Emotion is inferred from speech signals using filter banks and Deep CNN[6] which shows high accuracy rate which gives an inference that deep learning can also be used for emotion detection. Speech emotion recognition can be also performed using image spectrograms with deep convolutional networks which is implemented in [7]

[1]. S. G. Koolagudi and S. R. Krothapalli, “Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features,” Int. J. Speech Technol., vol. 15, no. 4, pp. 495–511, 2012.

[2]. J. Rong, G. Li, and Y. P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, 2009.

[3]. F. Noroozi, N. Akrami, and G. Anbarjafari, “Speech-based emotion recognition and next reaction prediction,” 2017 25th Signal Process. Commun. Appl. Conf. SIU 2017, no. 1, 2017. [4]. A. Graves, A. Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[5]. C.-W. Huang and S. S. Narayanan, “Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition,” pp. 1–19, 2017.

[6]. H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” Neural Networks, vol. 92, pp. 60–68, 2017.

[7]. A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network,” 2017 Int. Conf. Platf. Technol. Serv., pp. 1–5, 2017.

3. Report on the proposed system

Speech Emotion Recognition

SPEECH RECOGNITION: Speech Recognition is the technology that deals with techniques and methodologies to recognize the speech from the speech signals.

EMOTION RECOGNITION : Emotion Recognition deals with the study of inferring emotions, methods used for inferring. Emotion can be recognized from facial expressions, speech signals. Various techniques have been developed to find the emotions such as signal processing, machine learning, neural networks, computer vision.

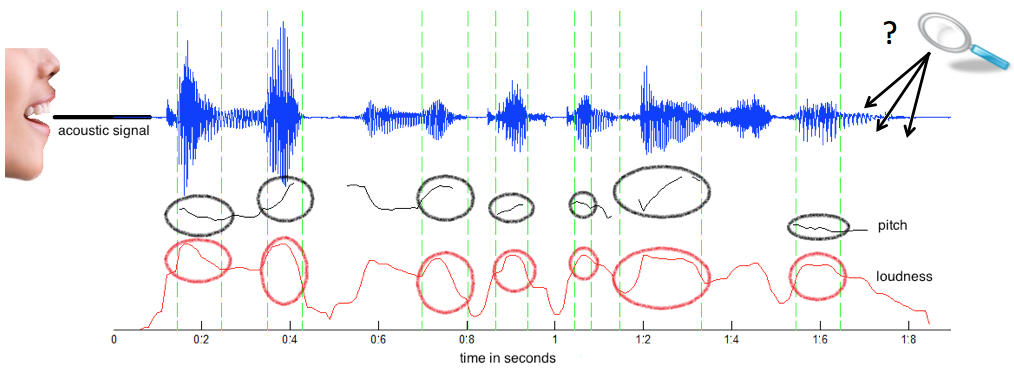
SPEECH EMOTION RECOGNITION : Speech Emotion Recognition is a research area problem which tries to infer the emotion from the speech signals. Various surveys state that advancement in emotion detection will make a lot of systems easier and hence making the world a better place to live.

The idea behind creating this project was to build a machine learning model that could detect emotions from the speech we have with each other all the time. Nowadays personalization is something that is needed in all the things we experience everyday. So why not have a emotion detector that will guage your emotions and in the future recommend you different things based on your mood. This can be used by multiple industries to offer different services like marketing company suggesting you to buy products based on your emotions, automotive industry can detect the persons emotions and adjust the speed of autonomous cars as required to avoid any collisions etc.

3.1. Problem Statement Analysis

Detecting emotions is one of the most important marketing strategy in today’s world. You could personalize different things for an individual specifically to suit their interest. For this reason, we decided to do a project where we could detect a person’s emotions just by their voice which will let us manage many AI related applications. Some examples could be including call centers to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful. As a result this type of application has much potential in the world that would benefit companies and also even safety to consumers.

Audio Signals:

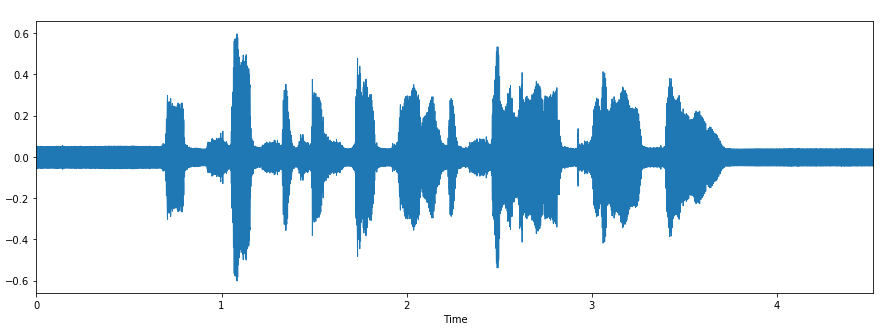


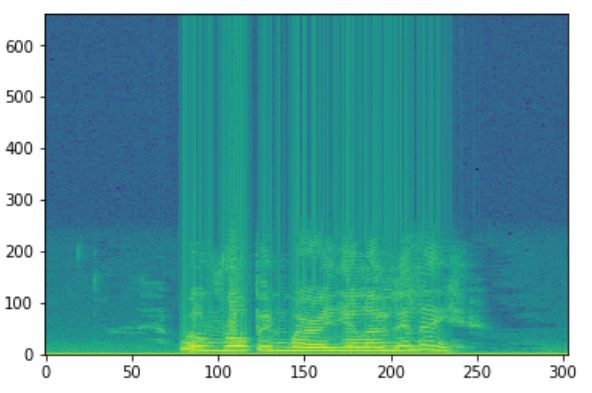
**Data Used:** We got audio datasets with around 2000 audio files which were in the wav format from the following website:

<https://zenodo.org/record/1188976#.XrehDmgzaUk>

|  |  |
| --- | --- |
| |  | | --- | |  | |

We went with the Audio only zip file because we are dealing with finding emotions from speech. The zip file consisted of around 1500 audio files which were in wav format. We tested out one of the audio file to know its features by plotting its waveform and spectrogram.





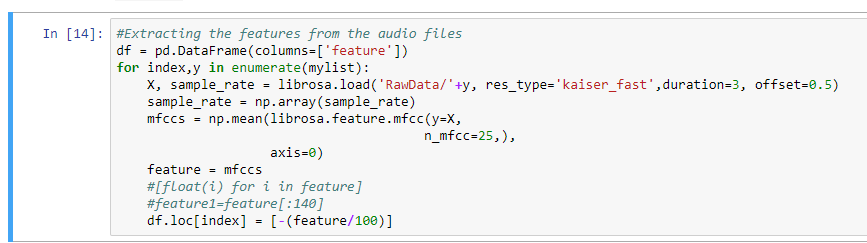
The next step involves organizing the audio files. Each audio file has a unique identifier at the 6th position of the file name which can be used to determine the emotion the audio file consists. We have 5 different emotions in our dataset.

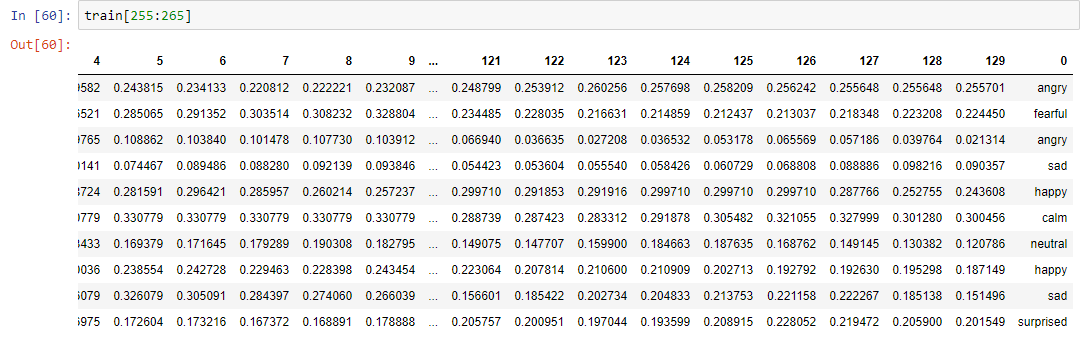
1. Calm
2. Happy
3. Sad
4. Angry
5. Fearful

We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Using the librosa library we were able to extract features i.e MFCC(Mel Frequency Cepstral Coefficient). MFCCs are a feature widely used in automatic speech and speaker recognition. We also separated out the females and males voice by the using the identifiers . This was because as experiment we found out that separating male and female voices increased by 15%. It could be because of the pitch of the voice was affecting the results.

Each audio file gave us many features which were basically array of many values. These features were then appended by the labels which we created in the previous step.

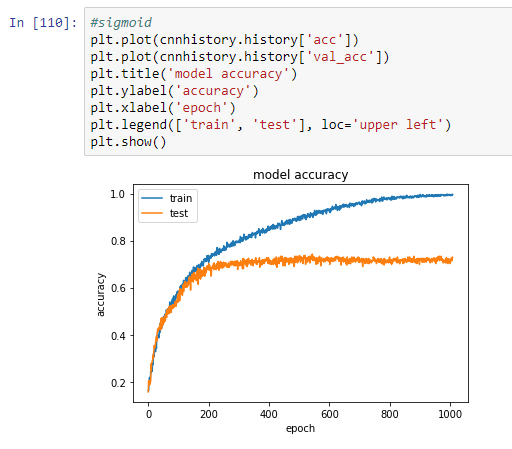
The next step involved dealing with the missing features for some audio files which were shorter in length. We increased the sampling rate by twice to get the unique features of each emotional speech. We didn’t increase the sampling frequency even more since it might collect noise thus affecting the results



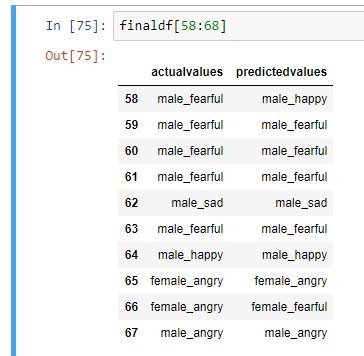


The next steps involve shuffling the data, splitting into train and test and then building a model to train our data.

**CNN :** CNN model was the best for our classification problem.



After training the model we had to predict the emotions on our test data. The following picture shows our prediction with the actual values.



 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. We can also see above how the model predicted against the actual values. In the future we could build a sequence to sequence model to generate voice based on different emotions. E.g. A happy voice, A surprised one etc.

3.2. Design and Methodology of proposed system

The whole pipeline is as follows (as same as any machine learning pipeline):

* Preparing the Dataset: Here, we download and convert the dataset to be suited for extraction.
* Loading the Dataset: This process is about loading the dataset in Python which involves extracting audio features, such as obtaining different features such as power, pitch and vocal tract configuration from the speech signal, we will use librosa library to do that.
* Training the Model: After we prepare and load the dataset, we simply train it on a suited sklearn model.
* Testing the Model: Measuring how good our model is doing.

The block diagram of the proposed work is shown below in fig 3.2.1.It consists of four main steps namely dataset collection,speech preprocessing,feature extraction and , training and testing.The various features are extracted from the given voice .These Features are given as input and classified based on which we get the emotion of that person.

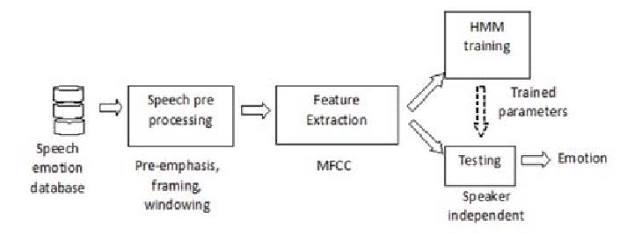


Figure 3.2.1. Block Diagram of Proposed system

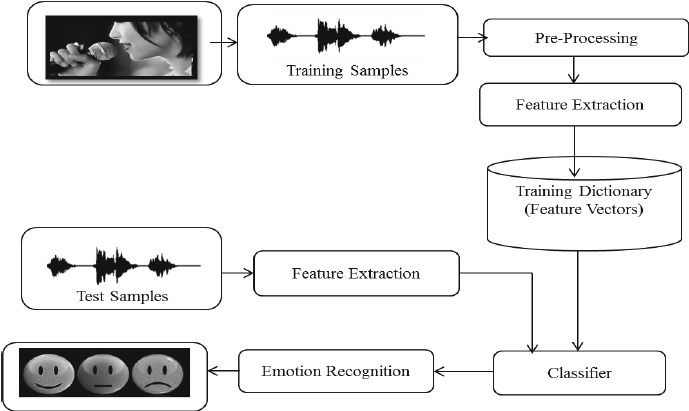


Figure 3.2.2. Architecture Diagram of Proposed system

4. Hardware software requirements and Implementation

Hardware requirements: The system must be capable of running machine learning model and must have one microphone input for recording voice.

Software requirements: The system must have python with version 3.6x or higher

and the following libraries

1. Librosa : Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.
2. Numpy : is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
3. Matplotlib : is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. ...
4. Keras : Keras is a open source neural network library written in Python. It is capable of running on top of  Tensorflow.  Designed to enable fast experimentation with deep neural network , it focuses on being user-friendly, modular, and extensible.
5. Tensorflow : TensorFlow is a free  and open source software library  for dataflow and differentiable  programming across a range of tasks. It is a symbolic math library, and is also used for machine leraning applications such as netural networks.
6. Pandas:  pandas is a software library written for the python programming language  for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
7. scikit-learn: is a free software machine learning library for the python programming language. It features various classification ,regression  and clustering  algorithms including support vector machines, random forest,gradient boosting ,k-means  and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries [NumPy](https://en.wikipedia.org/wiki/NumPy) and [SciPy](https://en.wikipedia.org/wiki/SciPy).

8.**Json**: JavaScript Object Notation is an open standard file format , and data interchange format,

that uses human-readable  text to store and transmit data objects consisting of attribute-value pair

and array data types.. It is a very common data format, with a diverse range of applications, such

as serving as a replacement for [XML](https://en.wikipedia.org/wiki/XML) in [AJAX](https://en.wikipedia.org/wiki/Ajax_(programming)) systems.

JSON is a language -independent  data format. It was derived from [JavaScript](https://en.wikipedia.org/wiki/JavaScript), but many modern

programming langauages  include code to generate and parse  JSON-format data. The official

Internet media type  for JSON is application/json. JSON filenames use the

extension .json.

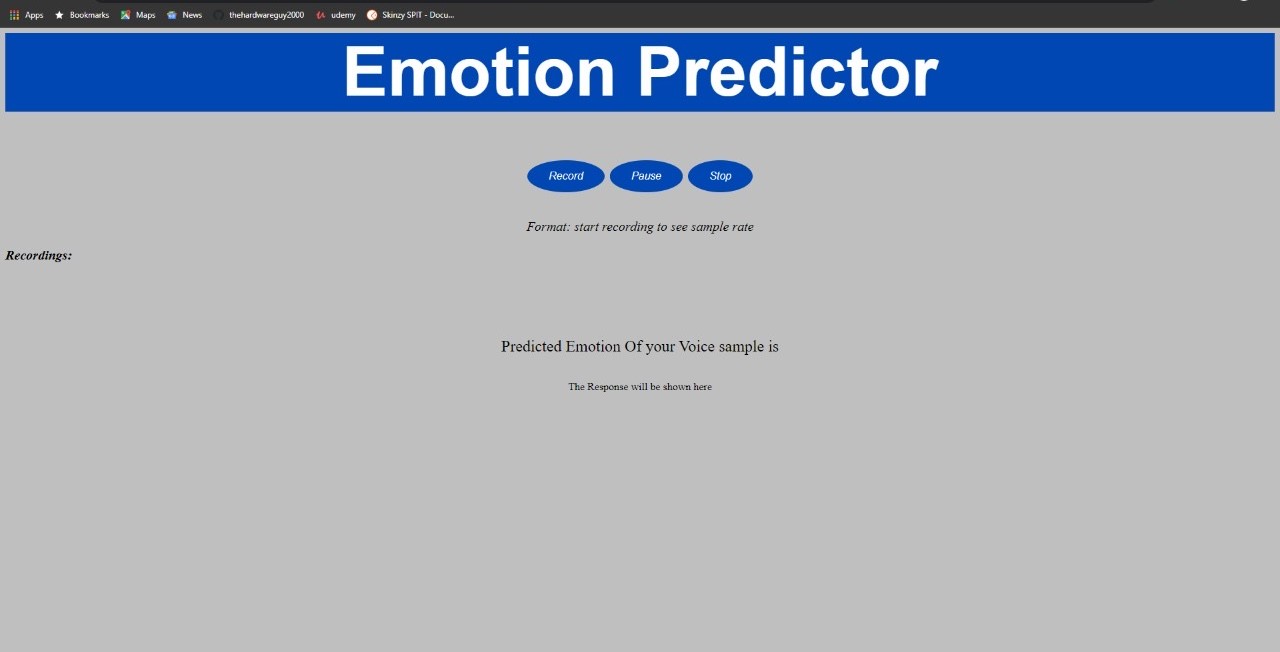
9.**pyaudio 10.wave**: PyAudio provides Python bindings for PortAudio, the cross-platform audio I/O library. With PyAudio, you can easily use Python to play and record audio on a variety of platforms. PyAudio is inspired by:

* pyPortAudio/fastaudio: Python bindings for PortAudio v18 API.
* tkSnack: cross-platform sound toolkit for Tcl/Tk and Python.

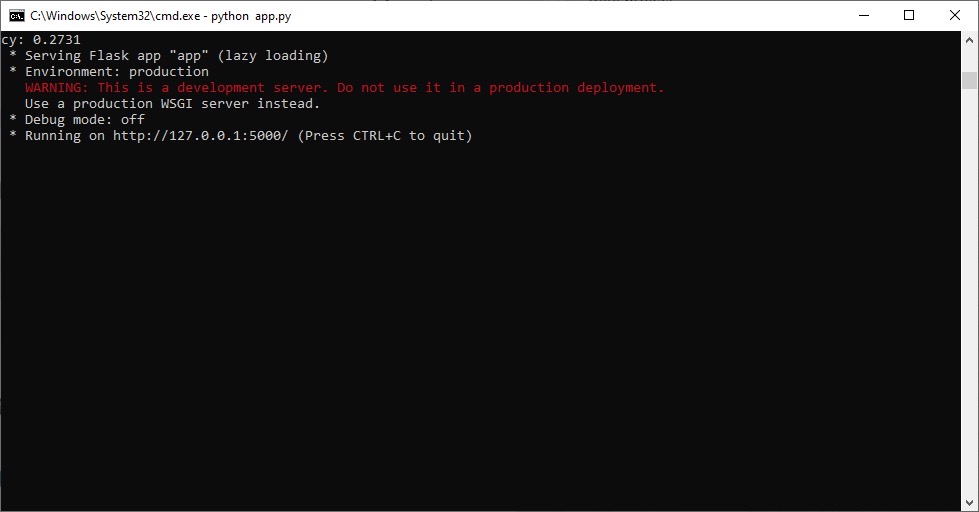
4.1 Results

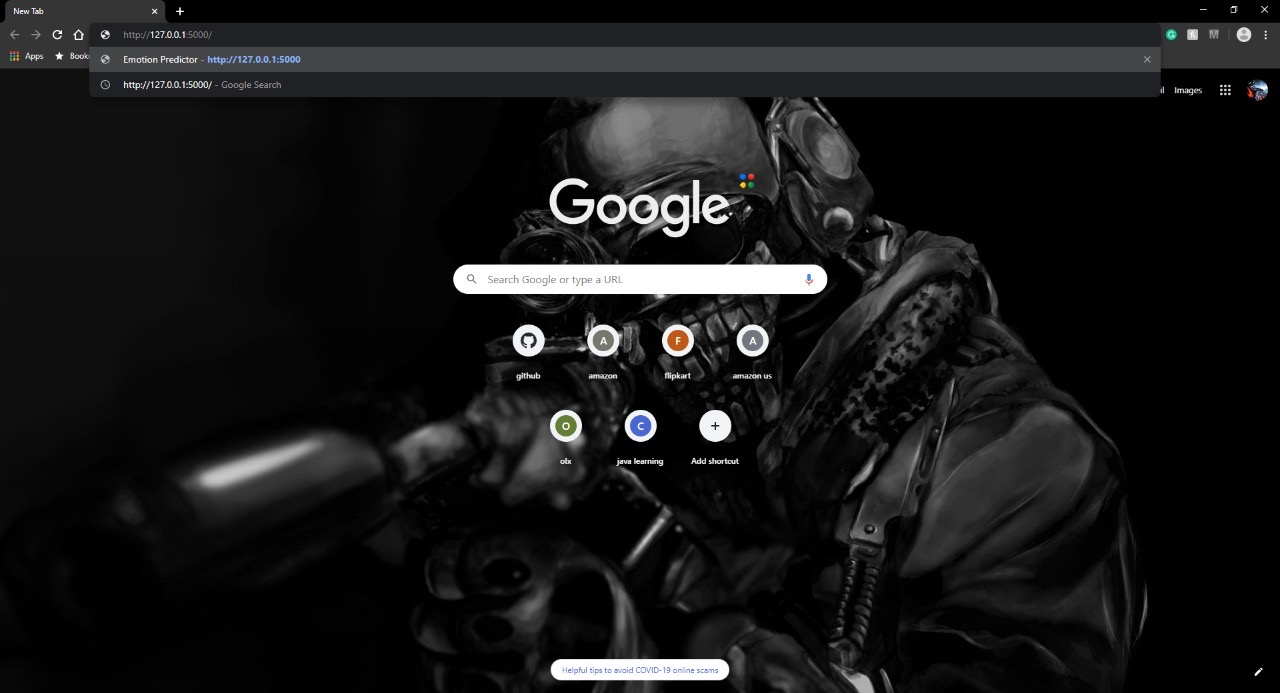
SPEECH EMOTION ANALYSIS

FRONT END:

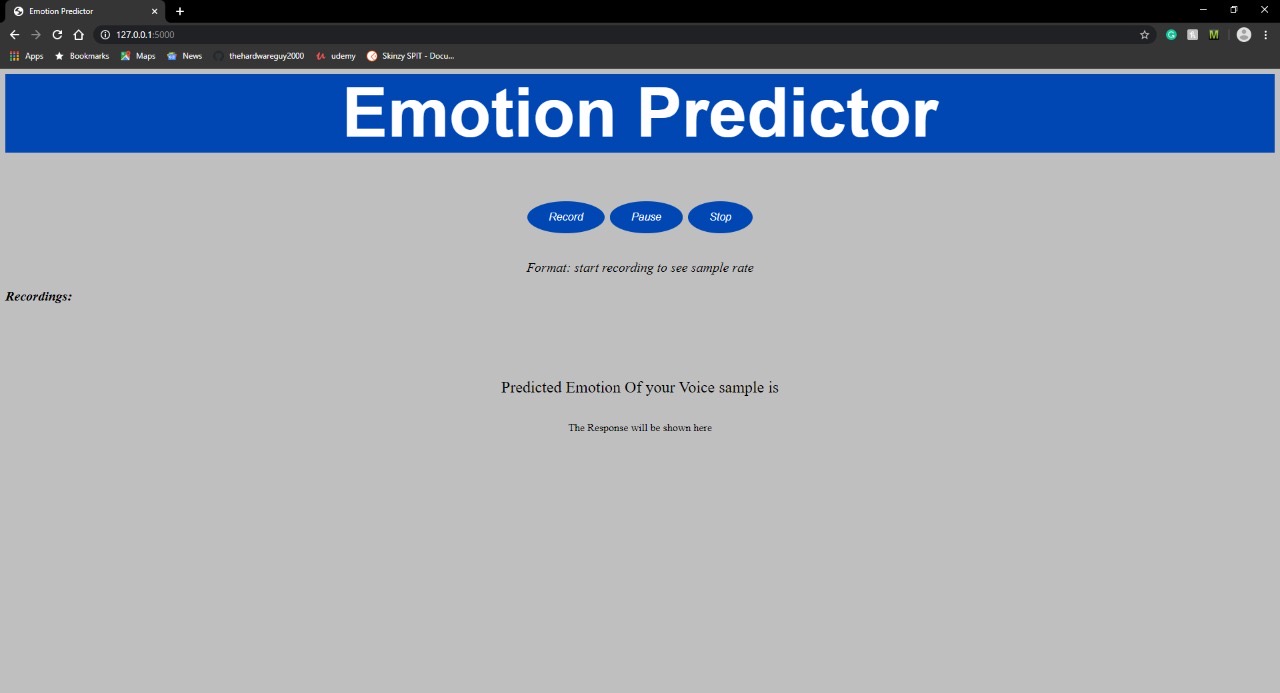


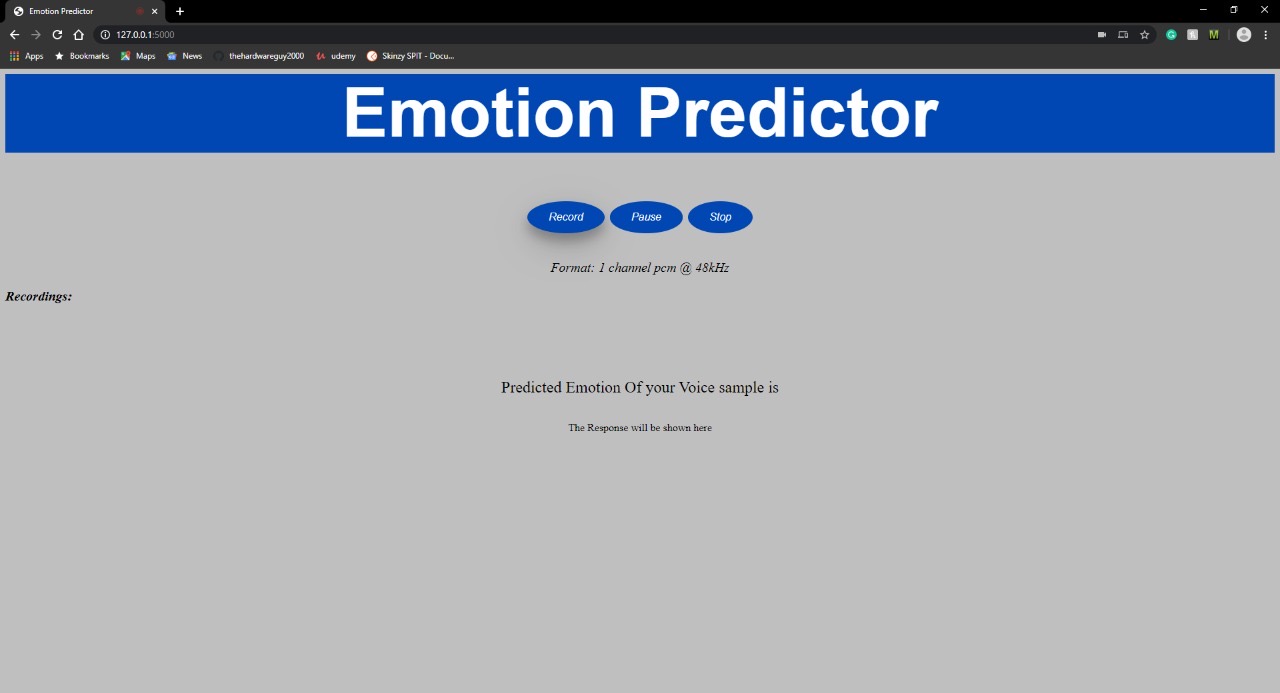
flask Server running on local host:



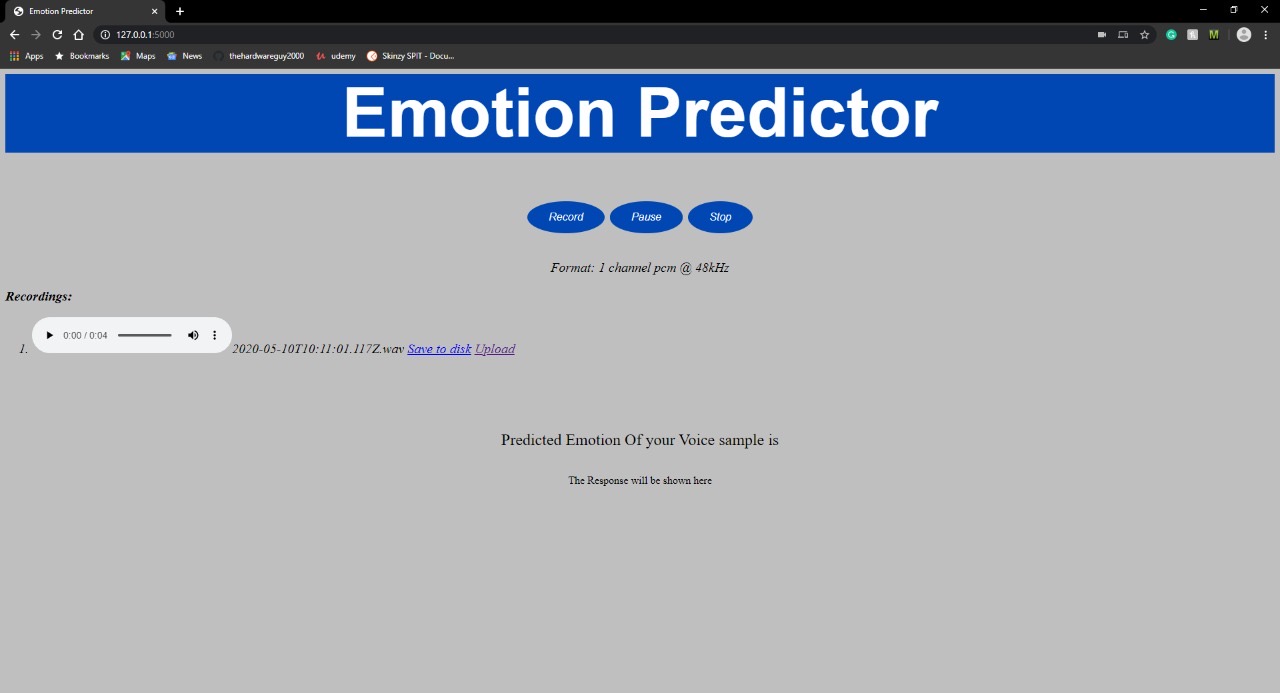


Frontend connected to Flask server:

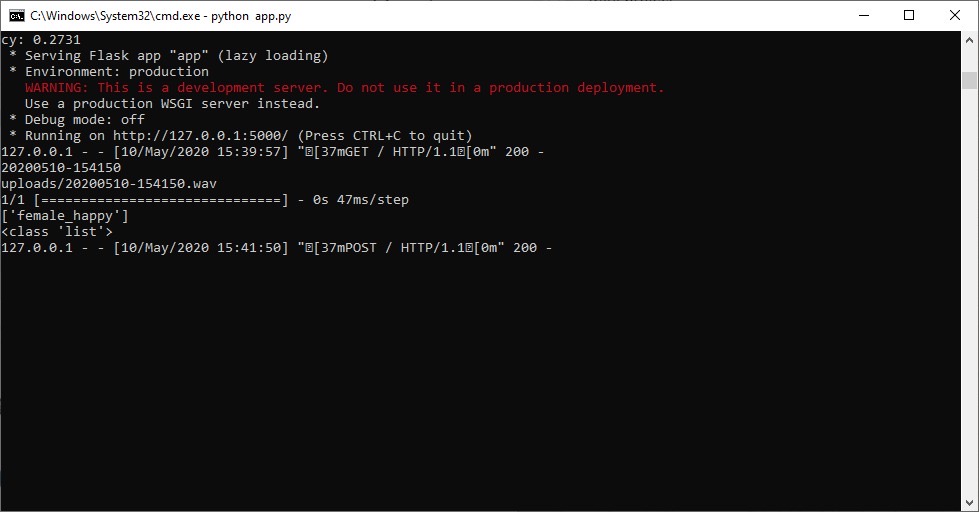




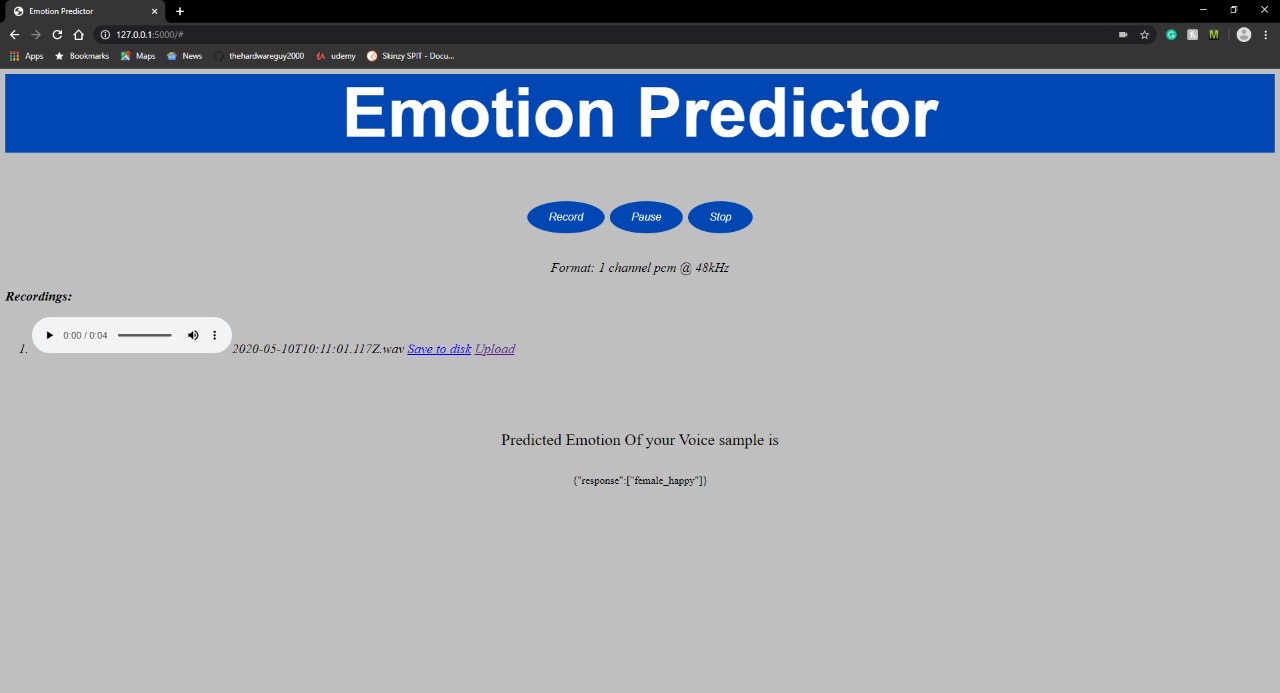
Recorded audio on frontend:



File uploaded on Flask and predicted:



Response printed on screen:



Accuracy 64.78 %

4.2 Conclusion

The field of machine learning is sufficiently new to still be rapidly expanding, often from innovation in new formalizations of machine learning problems driven by practical applications. However, recognizing emotions from speech is still a challenging problem. In our project we used CNN based network without using any traditional hand-crafted features to classify emotional speech.Our experimental results indicate that the efficiency of the model can be increased by randomizing data and increasing data.

5. Appendix (if applicable)

6. References

[1]Speech Emotion Recognition Using Deep Neural Network and Extreme

Learning Machine, Kun Han1, Dong Yu2, Ivan Tashev2

Department of Computer Science and Engineering,

The Ohio State University, Columbus, 43210, OH, USA

Microsoft Research, One Microsoft Way, Redmond, 98052, WA, USA

[2]Speech Emotion Recognition using Convolutional and Recurrent Neural Networks Wootaek Lim, Daeyoung Jang and Taejin Lee Audio and Acoustics Research Section, ETRI, Daejeon, Korea