Seminar/Case Study

Final Report

Sentimental Detection in Audio

By

maniDeepak P

190137

Supervisor

Dr. Kiran Khatter

Associate Professor

Department of Computer Science and Engineering
School of Engineering and Technology

December, 2021

Acknowledgement

It is my immense pleasure to show my gratitude towards the people that has helped me to complete the seminar/case study course. I would like to thank Prof. (Dr) Anirban Chakraborti, Dean (SOET) and Manoj K. Arora, Vice Chancellor for providing the opportunity to be able to do internship through course curriculum and I would like to thank Dr Kiran Khatter, coordinator of seminar/case study course for guiding us in the process of completion and guiding us throughout the entire course by providing valuable feedbacks and suggestions that has helped a lot in completion of our project.

Thanking You.
maniDeepak P 190137

2019 B.Tech SOET

BML Munjal University, Gurgaon, Haryana

CANDIDATE'S DECLARATION

I, maniDeepak P, hereby declare that the work done in my seminar/case study/project entitled *"Sentimental Detection in Audio>*" in fulfillment of completion of 5thsemester of Bachelor of Technology (B. Tech) program in the Department of Computer Science and Engineering, BML Munjal University is an authentic record of our original work carried out under the guidance of *Dr. Kiran Khatter Associate Professor.*

Due acknowledgements have been made in the text of the project to all other materials used.

This seminar/case study/project was done in the full compliance with the requirements and constraints of the prescribed curriculum.

maniDeepak P 190137

Place: Gurgaon

Date: 12-12-21

CERTIFICATE

This is to certify that the Seminar/Case Study/Project entitled "*Sentimental Detection in Audio*" to the best of my knowledge is a record of the bonafide work carried out by Mr./Ms. *maniDeepak P* under my guidance and/or supervision. The contents embodied in this report, to the best of my knowledge, have not been submitted anywhere else in any form for the award of any other degree or diploma. Indebtedness to other works/publications has been duly acknowledged at relevant places. The work was carried out during *July - December 2021* as part of their 5th semester coursework for Bachelor of Technology (B.Tech) program in the Department of Computer Science and Engineering, BML Munjal University.

Name and Designation of the Supervisor:

Signature:

Date:

Place:

Table of Contents

Abstract

In this Seminar Case Study course, I was done research and project on "*Speech Detection Analysis*" in a major way. The major topics I've worked on are MFCC that really is a scale of pitches judged by listeners to for the most part be kind of equal in the distance one from another, CHROMA which is twelve different pitch classes, in that Chroma Vector, and Chroma deviation, MEL, CNN, SVM, relu and MLP Classifier, which is quite significant. MLP, MEL, and MFCC literally are the pretty major topics that really have been used in making projects, demonstrating that in this Seminar Case Study course, I basically was done research and project on "Speech Detection Analysis", contrary to popular belief. And cleaning the kind of audio which reduces the noise and disturbances in the audio literally is also implemented, showing how, and cleaning the kind of audio which reduces the noise and disturbances in the real audio for all intents and purposes is also implemented in a major way.

Introduction:

Humans literally utilize speech as their major kind of means of communication and interaction in a subtle way. But what kind of data really is delivered, or so they particularly thought. Speech mostly is understood to for all intents and purposes be a means for expressing not only thoughts but also emotions in a particularly major way. Aside from verbal signals that particularly represent emotions clearly, basically certain messages for all intents and purposes are ambiguous in that the language element does not for the most part specify the emotion, which essentially shows that but what kind of data is delivered, or so they for the most part thought. In this paper, we attempt to mostly combine emotional prosody actually (non-verbal components of language that definitely allow individuals to literally express or comprehend emotion) and really deep learning to actually construct a model that understands for all intents and purposes human emotion through speech, which definitely shows that aside from verbal signals that particularly represent emotions clearly, definitely certain messages mostly are ambiguous in that the language element does not actually specify the emotion, which generally shows that but what kind of data definitely is delivered in big way.

Problem Definition

1. **Objectives**
   From librosa and scikit-learn modules, also RAVDESS dataset, particularly develop a model that distinguish emotion from voice in a generally big way.

2. **Challenges**
   This is perhaps the most difficult aspect of creating a voice interface. We simply cannot identify and perform everything since speech recognition is still extremely new. Humans, too, sometimes misunderstand or misinterpret what others are saying. Furthermore, few individuals read user manuals or explore all a device's capabilities.
   And mainly debugging the code and finding out the errors is also a considered as a challenge.

3. **Deliverables**
   Using the needed classifiers and algorithms, the model had to be able to accurately predict the pitch of the sound. It must be able to forecast based on previously recorded sounds. Then it should be able to train the algorithm after each prediction.

4. **Literature Review**

- **MFCC**: Mel Frequency Cepstral Coefficient, in short-term power it is also known as sound spectrum.
- **CHROMA**: It is a part or belongs to the 12 different pitch classes
- **MEL**: Continuous scale of pitches
- **CNN**: Convolutional Neural Networks (CNN)
- **SVM**: Support Vector Machine (SVM) classifier
- **MLP Classifier**: MLPClassifier works by Neural Network to perform classification.
- **RAVDESS**: Ryerson Audio-Visual Database of Emotional Speech

5. Description of the Dataset (if applicable/available)

I have used the RAVDESS data, i.e., Ryerson Audio Visual Database of Emotional Speech. In this data there are nearly ~7000 files recorded by 247 people 10 times on basis of emotion, validity, intensity, and genuineness. The total dataset sizes 24.8GB where 24 actors have spoken, lowering the sample rate on all the files.

The part of the RAVDESS where have been used in this project has total of 1440 files i.e., 60 recordings per actor for 24 well-trained actors each gives 1440 audio files. It contains 24 well-trained actors (12 male, 12 female), have spoken in the Northern American accent. Audio sentimental emotions contains surprise, angry, fearful, happy, calm, sad, and disgust nature of feelings.

Naming of the file

All the 1440 files have a non-repeatable filename. Filename has of 7-part numbers identifiers (e.g., 03-01-06-01-02-01-10.wav). Ids define the character emotions:

Ids for the file

- Mode (01 = Both Audio Video, 02 = Video, 03 = Audio).
- Voice mode (01 = Sentences, 02 = Song with music).
- Emotion (01 = Normal, 02 = Calm/pleasant, 03 = Happy/joyful, 04 = Sad/sorrowful, 05 = Angry/furious, 06 = Fearful/afraid, 07 = Disgust/hatred, 08 = Surprised/shocked).
- Emotion/sentimental level (01 = Normal, 02 = Strong).
  * There is no strong emotion for 'neutral' audio.*
- Sentences (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Iteration (01 = 1st Repetition, 02 = 2nd Repetition).
- Actors (From 01 to 24. Only odd number are men, rest are women).

Example: 03-01-06-01-02-01-12.wav

1. Only Audio/Sentence (03)
2. Sentences (01)
3. Emotion is fearful (06)
4. Normal/Neutral emotional level (01)
5. Sentence "dogs are sitting by the door." (02)

6. Repeat 1st (01)
7. 10th Actor id (10)
   Female, as number 10 is actor id, is even .

6. Proposed Methodology (if applicable)

In this paper, I proposed an approach for speech emotion identification based on a Fractional Fourier transform-based adaptive time-frequency feature extraction. The technique is depicted in Fig. 1 as a block diagram. The major steps in the suggested speech emotion recognition are as follows:

6.1. Pre-processing

6.2. Extraction of features that have been proposed

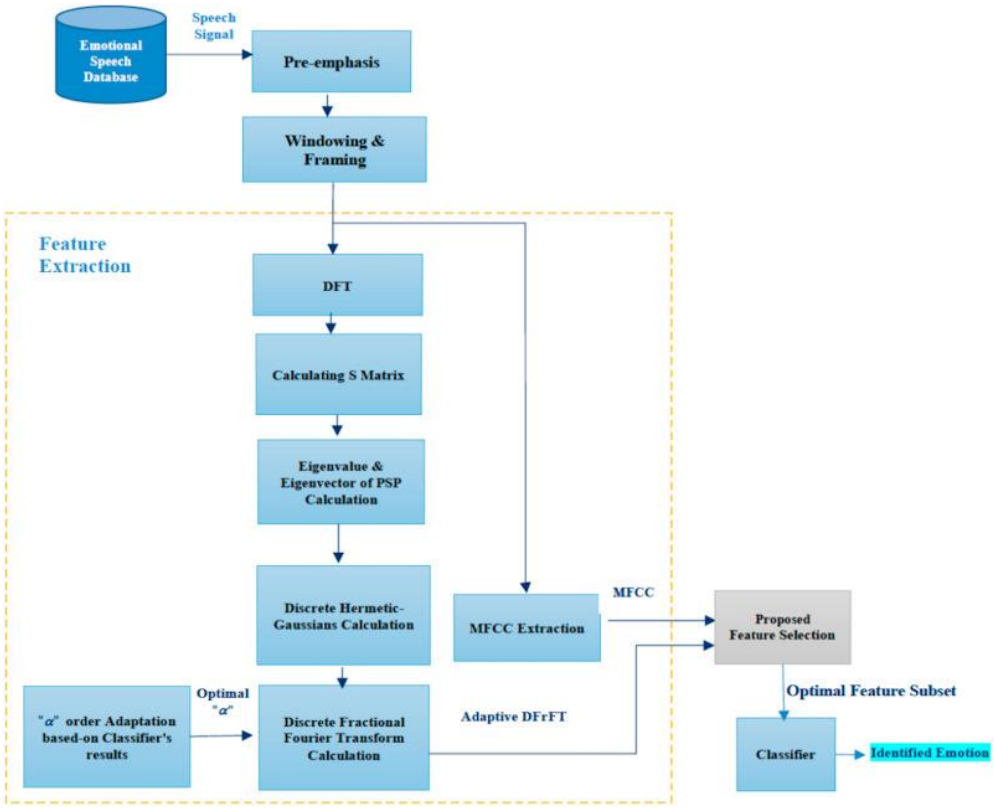

Fig. 1. Flow diagram of proposed Sentimental Detection.

6.1. Preprocessing

Pre-processing includes steps to select locks, set windows, and lock frames

Pre emphasis: To increase the volume of the raised area, the speech is sent through a high pass filter as shown in the equation. (3) where is the pre-emphasis value and ranges from 0.9 to 1.

$$x'(n) = x(n) - \alpha x(n-1)$$

- Signal framing is the division of the current audio signal into repaired parts (20ms duration with 10ms interval).

- Windowing: As described in Eq. (4), by applying a Hamming window to the frames, where the window is M for w(n):

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \qquad 0 \le n \le M-1$$

6.2. Proposed feature extraction

Where u k[n] is the ath power of the DFT's eigenvalue and signifies the kth discrete Hermite-Gaussian function $e^{-j\frac{\pi}{2}ka}$. Two ambiguities in the fundamental DFrFT equation must be overcome.

Since there are only four different eigenvectors in the DFT matrix, the eigenvalues are frequently degenerate, resulting in a collection of non-unique eigenvalues. To deal with this uncertain problem, the Hermite Gaussian values were chosen as eigenfunctions. (5). Since fraction power is multivalued, utilizing partial power of eigenvector creates a second paradox.

The eigenvalue will be used to respond to this ambiguity. According to one hypothesis, the Hermite-Gaussian features are eigenfunctions of the matrix S (Equation (7)) $\lambda_k^a = e^{\frac{-j\pi ka}{2}}$ which is also an eigenfunction of its DFT matrix.

$$F^a[m,n] = \sum_{k=0}^{N-1} u_k[m]\, e^{-j\frac{\pi}{2}ka} u_k[n]$$

The eigenvalues of the DFT matrix, or S, are classified as even or odd vectors. As a result, as shown in the equation, we will construct a matrix P that decomposes an arbitrary vector into its even and odd components (8). The PSP1 transformation reduces the difficulty of obtaining the shared eigenvectors of S when determining the eigenvectors of the Ev and Od matrices.

$$PSP^{-1} = \begin{bmatrix} Ev & 0 \\ 0 & Od \end{bmatrix} \tag{6}$$

$$S = \frac{-1}{4\pi} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 2\cos\frac{2\pi}{n} & 1 & \cdots & 0 & 0 \\ 0 & 1 & 2\cos\frac{2\pi}{n}2-4 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & & \\ 1 & 0 & 0 & \cdots & 1 & 2\cos\frac{2\pi}{n}n-1-4 \end{pmatrix} \tag{7}$$

$$P = \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \tag{8}$$

The even eigenvector of S may be written as U2k[n], and the odd eigenvectors of S can be written as U2k+1[n] by zero-padding and conversion (Eq. (9).

$$U_{2k}[n] = P\left[\hat{e}_k^T \middle| 0 \dots 0\right]^T \quad ; \qquad U_{2k+1}[n] = P\left[0 \dots 0 \middle| \hat{o}_k^T\right]^T \tag{9}$$

where eˆk and oˆk are the eigenvectors of the matrix Ev and Od with k zero-crossings (0kN/2), correspondingly. The programmed should then select the optimal " coefficient to create an optimal feature in the classifier. The Fractional Fourier transform is built on the angle of rotation ", whose frequency ranges from 0 to 2. The rotation angle " is computed as =(a)/2, where aR is the angle of rotation. As a result, an SVM model is developed to

spot the optimal value of. When this parameter is set to 1, DFrFT is like DFT. This step's role is outlined below:

$$Find\ 'a'\ that\ \left\{\max\ Acc\ \left(F_i^a\left[m,n\right],t_i\right)\right\};\ \ 0<\alpha\leq 2\pi\quad and\ \alpha\in R$$

where Acc denotes SVM Accuracy and and ti denote the feature vector and labeling of each test i, etc.

7. Experimental Results/Comparison (if applicable)

For this, we deployed an MLPClassifier, as well as the soundfile package reads the sound file and the librosa library to detect edges from it and used for audio analysis. As you can see, the model has an efficiency of 76.56 percent to 100.

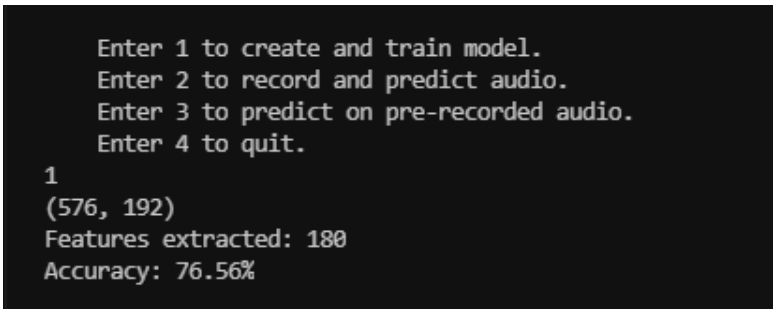

Fig. The accuracy is 76.56%

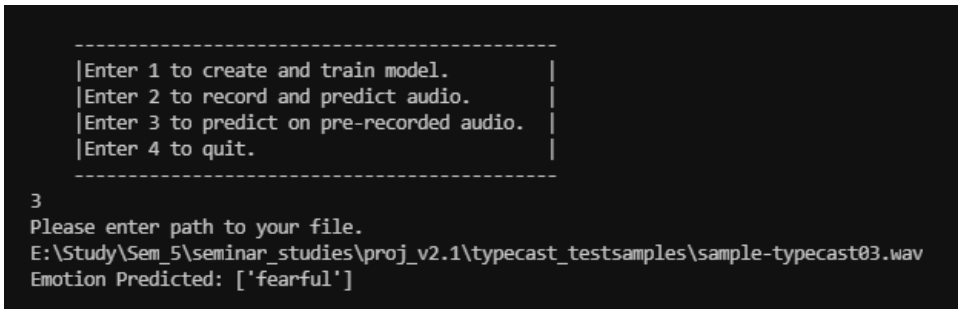

Fig. The prediction of the model is accurate.

8. Conclusions and Future Scope

Sentimental Detection on audio systems based on various speech features and classifiers are shown. According to the research, MFCC-based speech signal analyses offer a spectrum factor that indicates the exact vocal system for recorded speech emotions. MFCC have a high

level of perceived of the human voice and raise the level of precision. Based on the unique data included in the voice signal, the DTW approach was able to verify the specific speaker. The overall accuracy is 76.56 percent.

Humans are still a long way from fully realizing the potential of voice recognition technology. This relates to both complexity of technologies and its incorporation into our life. Modern digital companions are quite good at interpreting speech, and they're not the interactive platforms that software developers want them to be. Furthermore, voice recognition is still restricted to a specific variety of elements.

9. References

- https://www.irjet.net/archives/V5/i4/IRJET-V5I4359.pdf
- https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf
- https://www.sciencedirect.com/science/article/pii/S2352914820305748
- https://18it003.medium.com/data-science-mini-project-speech-emotion-recognition-ca0af5a3bcbf
- https://privacycanada.net/mel-frequency-cepstral-coefficient/
- https://iq.opengenus.org/mfcc-audio/

10. Plagiarism Report