**Zephyrus**

**Speech Emotion Recognition**

**A PROJECT REPORT**

###### ***Submitted by***

##### Shubhan Tejani (20BAI10152)

**Parv Paliwal (20BAI10228)**

**Bhavyangna Kanthed** **(20BAI10380)**

**Nayan Kumar (20BAI10176)**

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

*in*

# *CSE With Specialization in Artificial Intelligence and Machine Learning*

****

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

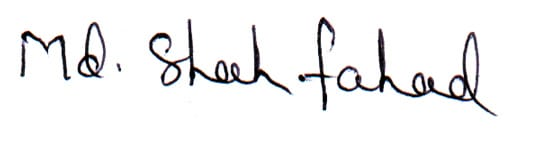
**VIT BHOPAL UNIVERSITY**

**KOTHRIKALAN, SEHORE**

**MADHYA PRADESH - 466114**

**BONAFIDE CERTIFICATE**

Certified that this project report titled **“Zephyrus - Speech Emotion Recognition”** is the Bonafide work of “**Bhavyangana Kanthed - 20BAI10380, Shubham Tejani - 20BAI10152, Nayan Kumar - 20BAI10386, Parv Paliwal - 20BAI10228”** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.



**PROGRAM CHAIR PROJECT GUIDE**

Dr. S.Sountharrajan Mr. MD Shah Fahad

School of Computer Science and Engineering School of Computer Science and Engineering

VIT BHOPAL UNIVERSITY VIT BHOPAL UNIVERSITY

The Project Exhibition II Examination is held on 22nd April 2022

**ACKNOWLEDGEMENT**

First and foremost, I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr. S.Sountharrajan, Head of the Department, CSE Artificial Intelligence and Machine Learning for much of his valuable support and encouragement in carrying out this work.

I would like to thank my internal guide Mr.MD Shah Fahad, for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of AI Division, who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| Abbv. | Full form |
| MLP | Multi-Layer Perceptron |
| MFCC | Mel-frequency cepstrum coefficients |
| ML | Machine Learning |
| IPYNB | python notebook |
| SVM |  |
| CSV | Comma Separated Values |

**LIST OF FIGURES AND GRAPHS**

|  |  |  |
| --- | --- | --- |
| FIGURE  NO | TITLE | PAGE NO |
| 1 |  | 18 |
| 2 |  | 19 |
| 3 |  | 22 |
| 4 |  | 23 |

**ABSTRACT**

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and the associated affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch.  
Emotion recognition is a rapidly growing research domain in recent years. Unlike humans, machines lack the ability to perceive and show emotions. But human-computer interaction can be improved by implementing automated emotion recognition, thereby reducing the need for human intervention. In this project, basic emotions like calm, happy, fear, disgust, etc. are analyzed from emotional speech signals. We use machine learning techniques like Multilayer perceptron Classifier (MLP Classifier) which is used to categorize the given data into respective groups which are non-linearly separated. Mel-frequency cepstrum coefficients (MFCC), chroma and Mel features are extracted from the speech signals and used to train the MLP classifier. For achieving this objective, we use python libraries like Librosa, sklearn, pyaudio, NumPy, and soundfile to analyze the speech modulations and recognize the emotion.

Keywords: Speech emotion recognition, Mel cepstral coefficient, artificial neural network, multilayer perceptron, mlp classifier, python.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  | List of Abbreviations  List of Figures and Graphs  Abstract | 4  4  5 |
| 1 | **CHAPTER-1:**  **PROJECT DESCRIPTION AND OUTLINE** Introduction 1.2 Motivation for the work  1.3 [About Introduction to the project  including techniques]  1.4 Problem Statement  1.5 Objective of the work  1.6 Organization of the project | 9  9  10  10  10  11 |
| 2 | **CHAPTER-2:**  **RELATED WORK INVESTIGATION**  2.1 Introduction  2.2 Core area of the project  2.3 Existing Approaches/Methods  2.3.1 Approaches/Methods -1  2.3.2 Approaches/Methods -2  2.4 Pros and cons of the stated Approaches/Methods  2.5 Issues/observations from investigation | 12  12  12  13  13  13 |
| 3 | **CHAPTER-3:**  **REQUIREMENT ARTIFACTS**  3.1 Introduction  3.2 Hardware and Software requirements  3.3 Specific Project requirements  3.3.1 Data requirement  3.3.2 Functions requirement  3.3.3 Performance and security requirements. | 14  14  15  15  15 |
| 4 | **CHAPTER-4:**  **DESIGN METHODOLOGY AND ITS NOVELTY**  4.1 Methodology and goal  4.2 Functional modules design and analysis  4.3 Software Architectural designs | 16  17  18 |
| 5 | **CHAPTER-5:**  **TECHNICAL IMPLEMENTATION & ANALYSIS**  5.1Outline  5.2 Technical coding and code solutions  5.3 Prototype submission  5.4 Test and validation  5.5 Performance Analysis (Graphs/Charts) | 19  20  22  22  23  23 |
| 6 | **CHAPTER-6:**  **PROJECT OUTCOME AND APPLICABILITY**  6.1Outline  6.2 Significant project outcomes  6.3 Project applicability on Real-world applications | 24  24  24 |
| 7 | **CHAPTER-7:**  **CONCLUSIONS AND RECOMMENDATION**  7.1Outline  7.2 Limitations of the System  7.3 Future Enhancements | 25  25  26 |
|  | Appendix A  References | 27  30 |

**CHAPTER-1:**

**PROJECT DESCRIPTION AND OUTLINE**

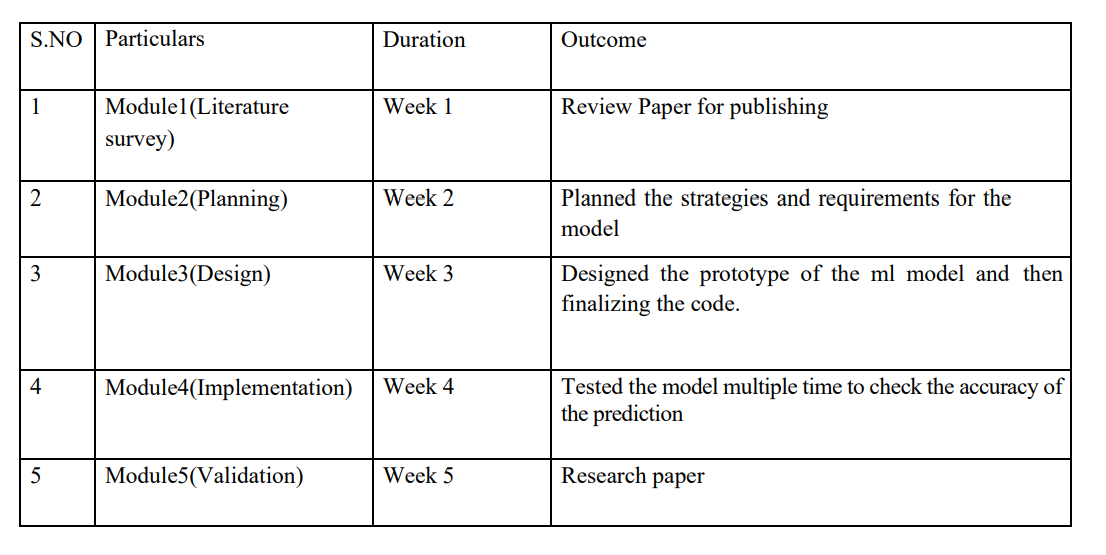
# 1.1 Introduction

1.2 Motivation for the work

1.3 About Introduction to the project including techniques

1.4 Problem Statement

1.5 Objective of the work

1.6 Organization of the project 

# CHAPTER-2:

# RELATED WORK INVESTIGATION

# 

# 2.1 Introduction

2.2 Core area of the project

2.3 Existing Approaches/Methods

2.3.1 Approaches/Methods -1

2.4 Pros and cons of the stated Approaches/Methods

2.5 Observations from investigation

.

**CHAPTER-3:**

**REQUIREMENT ARTIFACTS**

3.1 Introduction

Here we are going to explore the system requirements that are needed if someone wants to create this kind of ML project that can smoothly run on their computer and save it securely in its Memory. The system requirements mentioned here would help the user or a developer to analyze what kind of architecture he/she has and what it takes to build the project. We require certain hardware and software and also you should have a certain kind of IDE to code or run the code.

3.2 Hardware and Software requirements

Hardware requirements

PC with 250 GB or more Storage Space

PC with 4 GB RAM. PC with i3 and above.

Software Requirements

Operating system - Windows 7 and above. Language - Python

IDE - Google Colab or Jupyter Notebook

Browser - Google Chrome

3.3 Specific Project requirements

3.3.1 Data requirement

The analyzed dataset was collected from

The RAVDESS *(The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English)* is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats.

The set of 7356 recordings was each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. Prominent levels of emotional validity and test-retest inter-rater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli.

3.3.2 Performance and security requirements

As we run our project on Google Colab.

the specifications of different runtimes given by Google Colab:

* CPU: Model name: Intel(R) Xeon(R) CPU @ 2.30GHz
* Address sizes: 46 bits physical, 48 bits virtual
* Cache size: 46080 KB
* GPU: single 12GB NVIDIA Tesla K80 GPU

It can be used up to 12 hours continuously

Validate input. Validate input from all untrusted data sources. Proper input validation can eliminate most software vulnerabilities. Be suspicious of most external data sources, including command line arguments, network interfaces, environmental variables, and user controlled files.

**CHAPTER-4:**

**DESIGN METHODOLOGY AND ITS NOVELTY**

4.1 Methodology and goal

4.2 Functional modules design and analysis

4.3 Software Architectural designs

**CHAPTER-5:**

**TECHNICAL IMPLEMENTATION & ANALYSIS**

5.1Outline

5.2 Technical coding and code solutions

5.3 Prototype submission

5.4 Test and validation

5.5 Performance Analysis (Graphs

**CHAPTER-6:**

**PROJECT OUTCOME AND APPLICABILITY**

6.1Outline

The purpose of this project was to perform parameterization of audio data for the purpose of automatic recognition of emotions in speech. A collection of audio data from several videos related to human emotional expressions were gathered and turned into a data set. The feature selection increases the efficiency of the accuracy and the recall. The feature selection also allows reduction of the dimensionality of the data in turn leading to less computation processes in the robot memory. After the selection of features, a group of experiments to select the best classifiers were conducted. Multilayer Perceptron have achieved the best results. Support Vector Machine and Bayes Net could be good candidates to build the emotional recognition system of a robot, because of their easily implementation and the less computational complexity. This simple system with the classifiers is easy to understand and implement because of the utilization from a small group of features would work remarkably well on real-world data, making it possible to develop a real-time system in which the robot can make a fast decision in accordance with the emotional feedback provided from humans.

6.2 Significant project outcomes

This paper shows that MLPs are powerful in classifying speech signals. Even with simplified models, a limited set of characters can be easily identified. We have obtained higher accuracies as compared to other approaches for individual emotions. The performance of a module is highly dependent on the quality of pre-processing. Mel Frequency Cepstrum Coefficients are very dependable. Every human emotion has been thoroughly studied, analysed and the accuracy has been checked. The results obtained in this study demonstrate that speech recognition is feasible, and that MLPs can be used for any task concerning recognizing of speech and demonstrating the accuracy of each emotion present in the speech.

6.3 Project applicability on Real-world applications

This project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights into the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc.

As a real application, it could be considered a real-time system that can serve like a motor of emotional knowledge to understand the autistic children, to describe accurately their internal state and show the real content of their emotions. The system is not only applied to companion robots it could also be applicable to diverse smart sources (smart devices), this could be the case of healthcare, telemedicine or smart well-being systems that can be seen more often. This type of emotional devices working with emotional feedback will have the potential to reveal more about emotional state and the early detection of crisis, balanced lifestyle including and regulated stress level.

**CHAPTER-7:**

**CONCLUSIONS AND RECOMMENDATION**

7.1Outline

We have identified and detailed the parts that make up a [speech emotion recognition](https://www.sciencedirect.com/topics/computer-science/speech-emotion-recognition) system. These systems require training data provided by speech databases that are created using either acted, elicited, or natural sources. The signals are then preprocessed to make them fit for feature extraction. SER systems most commonly use prosodic and [spectral features](https://www.sciencedirect.com/topics/computer-science/spectral-feature) since they support a wider range of emotions and yield better results. The results can further be improved by adding features from other modalities, such as the ones that depend on visual or linguistic features.

Once all the features are extracted, SER systems have a wide range of [classification algorithms](https://www.sciencedirect.com/topics/computer-science/classification-algorithm) to choose from. While most use classical approaches, there are an increasing number of studies that incorporate recent advances, such as Convolutional or Recurrent [Neural Networks](https://www.sciencedirect.com/topics/social-sciences/neural-network).

All these preprocessing and feature extractions are done to detect the emotion in the speech signal, yet emotions are still an open problem in psychology. There are several models that define them. SER systems use manual labeling for their training data, which, as mentioned earlier, is not always correct.

Although there are systems and realizations of real-time emotion recognition, SER systems are not yet part of our everyday life, unlike speech recognition systems that are now easily accessible even with [mobile devices](https://www.sciencedirect.com/topics/computer-science/mobile-device). To reach this goal, SER systems need more powerful hardware so that processing can be done faster; more correctly labeled data so that the training is more accurate; and more powerful algorithms so that the recognition rates increase. We believe that the research will continue towards solutions that apply deep learning algorithms, and since they require more data and more powerful processors, these advances are likely to follow.

We believe that, as SER systems become more part of our daily lives, there will be more data available to learn from, which will improve their performance, even when at times humans can fail. The subtle differences which may not be registered by humans can be picked up by these networks that will improve the areas where emotion recognition is applicable, such as human-computer [interaction](https://www.sciencedirect.com/topics/computer-science/human-computer-interaction), healthcare, and alike.

7.2 Limitation/Constraints of the System

Although there are many advancements in speech emotion recognition systems, there are still several obstacles that need to be removed for successful recognition.

One of the most important problems is the generation of the dataset that is used for the learning process. Most of the data sets used for SER are acted or elicited that are recorded in special silent rooms. However, the real-life data is noisy and has far more distinctive characteristics than the others. Although natural data sets are also available, they are fewer in number. There are legal and ethical problems to record and use natural emotions. Most of the utterances in natural data sets are taken from talk shows, call-center recordings, and similar cases where the involved parties are informed of the recording. These data sets do not contain all emotions and may not reflect the emotions that are felt. In addition, there are problems during the labeling of the utterances. There are human annotators labeling the speech data after the utterances are recorded. The actual emotion felt by the speaker and emotions perceived by human annotators may show differences. Even the recognition rates of human annotators are not over 90%. In favor of humans, however, we believe that we also depend on the content and the context of the speech as we are evaluating.

There are also cultural and language effects on SER. There are several studies available working on cross-language SER. However, the results show that the current systems and features used are not sufficient for it. The intonation of emotions in speech among various languages may show differences for example.

An overlooked challenge is the case of multiple speech signals, where the SER system has to decide which signal to focus on. Although it can be handled via a speech separation algorithm in the preprocessing stage, current systems fail to notice this problem.

7.3 Future Enhancements

A few possible steps that can be implemented to make the models more robust and accurate are the following

● An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.

● Figuring out a way to clear random silence from the audio clip.

● Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP, or Harmonic cepstrum.

● Following a lexical features-based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.

● Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

**APPENDIX**

**REFERENCES**

[1] Iqbal, A. and Barua, K. A real-time emotion recognition from speech using  
gradient boosting. In 2019 International Conference on Electrical, Computer and  
Communication Engineering (ECCE) (2019), IEEE, pp. 1–5  
[2] Jannat, R., Tynes, I., Lime, L. L., Adorno, J., and Canavan, S. Ubiquitous emotion  
recognition using audio and video data. In Proceedings of the 2018 ACM  
International Joint Conference and 2018 International Symposium on Pervasive and  
Ubiquitous Computing and Wearable Computers (2018), ACM, pp. 956–959.  
[3] LIVINGSTONE, S. R., AND RUSSO, F. A.: The Ryerson audio-visual database  
of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and  
vocal expressions in north American English. PloS one 13, 5 (2018), e0196391.  
[4] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In  
ISMIR (2000), vol. 270, pp. 1–11.  
[5] Muda, L., Begam, M., and Elamvazuthi, I. : Voice recognition algorithms using  
Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW)  
techniques. arXiv preprint arXiv:1003.4083 (2010).  
[6] Nair, V., and Hinton, G. E. : Rectified linear units improve restricted Boltzmann  
machines. In Proceedings of the 27th international conference on machine learning  
(ICML-10) (2010), pp. 807–814  
[7] Platt, J. C., Cristianna, N. and Shawe-Taylor, J. : Large margin dags for multiclass  
classification. In Advances in Neural Information Processing Systems 12, S. A. Solla,  
T. K. Leen, and K. Muller, Eds. MIT Press, 2000, pp. 547–553  
[8]Toronto emotional speech set (TESS)  
(https://tspace.library.utoronto.ca/handle/1807/24487)