

ECE1901 – Technical Answers for Real World Problems (TARP)

A Project Report

titled

SPEECH EMOTION RECOGNITION

Submitted by

MADHAN KUMAR S

(20BEC1112)

YASWANTH KANNAN G

(20BEC1201)

SANTHAKUMAR M

(20BEC1334)

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION
ENGINEERING**



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

**Vandalur – Kelambakkam Road
Chennai – 600127**

April 2023

SCHOOL OF ELECTRONICS ENGINEERING

DECLARATION BY THE CANDIDATE

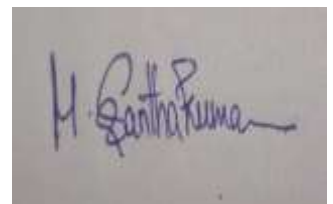
I hereby declare that the Report entitled “**SPEECH EMOTION RECOGNITION**” submitted by me to VIT Chennai is a record of bonafide work undertaken by me under the supervision of **Dr. ASHISH KUMAR**, Professor, SENSE, VIT Chennai.



Madhan Kumar S
20BEC1112



Yaswanth Kannan G
20BEC1201



SanthaKumar M
20BEC1334

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our TARP project guide, **Dr. Thiripurasundari D**, School of Electronics Engineering for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We are extremely grateful to **Dr. Susan Elias**, Dean of the School of Electronics Engineering (SENSE), VIT University Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our Head of The Department **Dr. MOHANAPRASAD K**, B.Tech-ECE for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses till date.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

BONAFIDE CERTIFICATE

Certified that this project report titled “**SPEECH EMOTION DETECTION**” is the bonafide work of “**Madhan Kumar S (20BEC1112), Yaswanth Kannan G (20BEC1201) & Santhakumar M (20BEC1334)**” who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. ASHISH KUMAR ,

TARP Supervisor

Assistant Professor,

School of Electronics Engineering,

VIT Chennai.

ABSTRACT

This project report presents an investigation into speech emotion recognition, which is the process of identifying emotional states based on audio recordings of human speech. The proposed system employs machine learning techniques to extract relevant features from speech signals, and then utilizes a deep neural network model to classify emotions. The dataset used for training and evaluation consists of audio recordings of spoken sentences in different emotional states, including anger, happiness, sadness, and neutral. The performance of the proposed system is evaluated using various metrics, such as accuracy, precision, recall, and F1-score. The experimental results demonstrate the effectiveness of the proposed system in accurately recognizing emotions from speech signals, with an overall accuracy of 70.8% on the test set. The report concludes with a discussion of potential applications of the proposed system and suggestions for future research directions.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	5
	LIST OF TABLES	8
	LIST OF FIGURES	9
1	INTRODUCTION	10
	1.1 PROJECT OVERVIEW	11
	1.2 TECHNOLOGY STACK USED	11
	1.3 OBJECTIVES	13
2	REVIEW OF LITERATURE	14
3	POTENTIAL COMPETITORS	16
4	METHODOLOGY	17
	4.1 PROCEDURE	17
	4.2 PROCESS OVERVIEW	18
	4.3 PROCESS NOVELTY	19
5	PROPOSED SYSTEM'S ATTRIBUTES	20
	5.1 WORKING PRINCIPLE	21
	5.2 FEATURES	22
	5.3 PROPOSED ALGORITHM WORKFLOW CHART	23
	5.4 PROPOSED CNN MODEL ARCHITECTURE	24

6	RESULTS AND DISCUSSION		25
	6.1	CNN	25
	6.2	CONFUSION MATRIX	26
	6.3	APPLICATION INTERFACE	27
	6.4	TARGET AUDIENCE	28
7	CONCLUSION & FUTURE SCOPE		29
	7.1	CONCLUSION	29
	7.2	FUTURE SCOPE	29
8	REFERENCES		31

LIST OF TABLES

TABLE NO.	TITLE	PAGE NO.
3.1	Potential Competitors	6
6.1	CNN Results	25

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
4.1	Process Overview	18
5.1	Proposed Algorithm Workflow chart	22
5.2	Proposed CNN Model Architecture	23
5.3	Proposed CNN model developed	24
6.1	Confusion Matrix for SER using CNN	26
6.2	Output Screenshot-1	27
6.3	Output Screenshot-2	27

CHAPTER 1

INTRODUCTION

The emotional content communicated through speech, such as tone of voice, inflection, and other vocal characteristics, is referred to as speech emotion. Emotion recognition has grown in popularity in recent years as a result of its potential applications in a variety of disciplines, including healthcare, education, marketing, and human-computer interaction.

Speech emotion recognition is the process of identifying the emotional state of a speaker through their speech signal. The process of emotion recognition typically involves using machine learning algorithms to analyze and classify the relevant cues. These algorithms can be trained using datasets that contain labeled examples of emotional states. Deep learning techniques such as convolutional neural networks and recurrent neural networks have been found to be particularly effective for emotion recognition tasks. Speech emotion recognition is a task that trains a machine learning model to recognise the emotional state of a speaker from their speech. This can be accomplished through the use of various methods, such as convolutional neural networks (CNNs) and machine learning algorithms. CNNs are a type of deep learning algorithm that excels at image recognition jobs.

1.1 Project Overview

Data Collection: Collecting a large dataset of speech signals labeled with different emotions is essential for training and testing the SER model. The dataset can be obtained from publicly available sources or through recording speech signals.

Data Preprocessing: Preprocessing involves filtering the speech signal, removing noise, and extracting relevant features from the signal. The features extracted can be statistical measures such as mean, variance, or spectral features such as Mel-frequency cepstral coefficients (MFCCs).

Model Selection: The next step involves selecting an appropriate machine learning algorithm that can classify the emotional state of the speaker based on the extracted features. Several algorithms such as Support Vector Machines (SVM), Random Forest, and Neural Networks can be used.

Model Training and Testing: Once the machine learning algorithm is selected, the next step is to train the model on the labeled dataset. The model's performance can be evaluated using various metrics such as accuracy, precision, recall, and F1-score.

Integration with API: Finally, the trained SER model can be integrated with an API to enable real-time emotion recognition from speech signals. The API can be designed to accept input in the form of speech signals and return the emotional state of the speaker as an output.

1.2 Technology Stack Used

Machine Learning is a subset of artificial intelligence in which computers are taught to learn from data without being specifically programmed. It is a data analysis technique that automates the creation of analytical models. Machine learning algorithms employ statistical methods to identify and learn from patterns in data, which can then be used to make predictions or decisions.

Deep Learning is a subfield of machine learning that models and solves complex issues using artificial neural networks. These neural networks are made up of multiple layers of interconnected nodes or neurons that process and transform the input data until it produces the intended output.

Convolutional Neural Network can be used to recognise speech emotions by transforming audio signals into spectrograms and performing convolutional operations to extract pertinent features. For classification, the CNN can be built with numerous convolutional layers, followed by pooling layers and fully connected layers.

Python is a high-level, interpreted computer language that is widely used by developers due to its simplicity, flexibility, and readability. It is compatible with a variety of computer paradigms, including procedural, object-oriented, and functional programming.

Flask is a famous Python web framework that is frequently used to create RESTful APIs. RESTful APIs allow clients to access and manipulate resources over the web. It offers a straightforward and flexible method for mapping URLs to Python functions, making it simple to define API endpoints.

Mel-frequency cepstral coefficients (MFCCs) are a technique for extracting features that is extensively used in audio signal processing and speech recognition. They are frequently used in speech recognition systems to compactly and efficiently represent speech signals.

TensorFlow is a Google open-source software framework for developing and training machine learning models. It includes support for neural networks, deep learning, and other kinds of machine learning algorithms, as well as a variety of tools and APIs for creating and training machine learning models.

Keras is a high-level neural network API that was originally developed as a user-friendly interface to TensorFlow. It allows developers to easily build and train neural networks using a simple and intuitive API, while also providing the flexibility and power of TensorFlow in the backend.

Theme: Machine Learning/ Deep Learning/ Feature Extraction/Speech recognition

1.3 Objectives

1. Short Term objective: To correctly identify the emotional state of a speaker based on their spoken words or voice characteristics. This can be accomplished by employing machine learning algorithms that analyse various characteristics of the speech signal, such as pitch, intensity, and spectral content, to identify the speaker's emotional state. The ultimate aim is to enable more natural and intuitive human-machine interaction by enabling machines to recognise and respond appropriately to the emotional state of the user.

Long-term Objective: Speech emotion recognition can be used in healthcare to track the emotional state of individuals suffering from mental health conditions such as depression and anxiety. It can alert healthcare workers to potential changes in an emotional state of a patient by detecting changes in speech patterns, allowing for early intervention and treatment.

It can be also used in education to monitor pupil engagement and emotional state in real-time. Teachers can spot when students are struggling or disengaged by analysing speech patterns and adjusting their teaching approach accordingly.

CHAPTER 2

REVIEW OF LITERATURE

Sharmeen M Saleem, *et al.* (2021), in the paper titled, “Multimodal Emotion Recognition using Deep Learning”, examined the use of deep learning to recognize emotional signs in multimodal data and compares their applicability based on current research. Multimodal affective computing systems are compared to unimodal solutions since they have a better classification accuracy. The number of emotions seen, characteristics collected, categorization method, and database consistency all affect accuracy.

Emotion recognition using prosodic features" by Carlos Busso et al. (2008): This paper proposed a system for SER based on prosodic features such as pitch, energy, and duration. The system achieved an accuracy of 55% in classifying four emotional states (anger, happiness, sadness, and neutrality) in the IEMOCAP database.

Speech emotion recognition: A review" by Rita Singh et al. (2019): This paper provides a comprehensive review of the various techniques used for SER, including feature extraction, classification algorithms, and databases. It also highlights the challenges associated with SER, such as the lack of standardization in the annotation of emotional states and the need for large, diverse databases.

A survey of speech emotion recognition: Techniques and databases" by Minhao Xia et al. (2021): This paper provides a comprehensive survey of the various techniques used for SER, including traditional machine learning algorithms, deep learning approaches, and multimodal techniques. It also provides a detailed analysis of several benchmark databases for SER and highlights their strengths and weaknesses.



Md. Shah Fahad *et al.* (2020), in the paper titled, “DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features” used robust epoch recognition from emotional speech to extract emotion-specific epoch-based characteristics, such as immediate pitch, phase, and excitation strength. The combined feature set outperforms the MFCC characteristics, which have been used as a baseline for SER systems in the literature, by 5.07 percent, and state-of-the-art methods by 7.13 percent. The suggested model outperforms state-of-the-art methods by 2.06% when just MFCC characteristics are used.

CHAPTER 3

POTENTIAL COMPETITORS

There are potential competitors to the proposed deep learning-based speech emotion recognition system, namely the manual interrogation or the chatting process which may be time consuming, often leads to wrong conclusions and violates privacy. The other alternative is the usage of polygraph machine, which is prone to tampering of data. A summary of the potential competitors and their drawbacks are presented in Table 3.1.

Table 3.1 Potential Competitors

Topic	Picture	Drawbacks
Manual Interrogation /Chatting Process		<ul style="list-style-type: none">1) Time Consuming2) Often leads to wrong Conclusion3) Violates privacy
Polygraph machine		<ul style="list-style-type: none">1) Subjects can bypass the system2) Tampering of Data is possible

CHAPTER 4

METHODOLOGY

4.1 Procedure

The proposed deep learning-based speech recognition system works by the following steps:

Step 1: Initially the subject would upload the audio file recording of his/ her speech.

Step 2: The speech signal present in the file would be processed into an audio signal that would be sent to the model which is embedded in the User Interface for validation.

Step 3: Upon completion of input validation, the proposed autoencoder model would extract the best speech features from the input speech features such as Mel, Chroma & Mel Frequency Cepstral Coefficients (MFCC), followed by which the best features are supplied as input to the proposed Super Learner Model, which would classify the emotion corresponding to the audio signal. Appropriate action may be taken by the user for their wellbeing, based on the emotion classification report.

4.2 Process Overview

The process flow of the project has been explained in Figure 4.1.

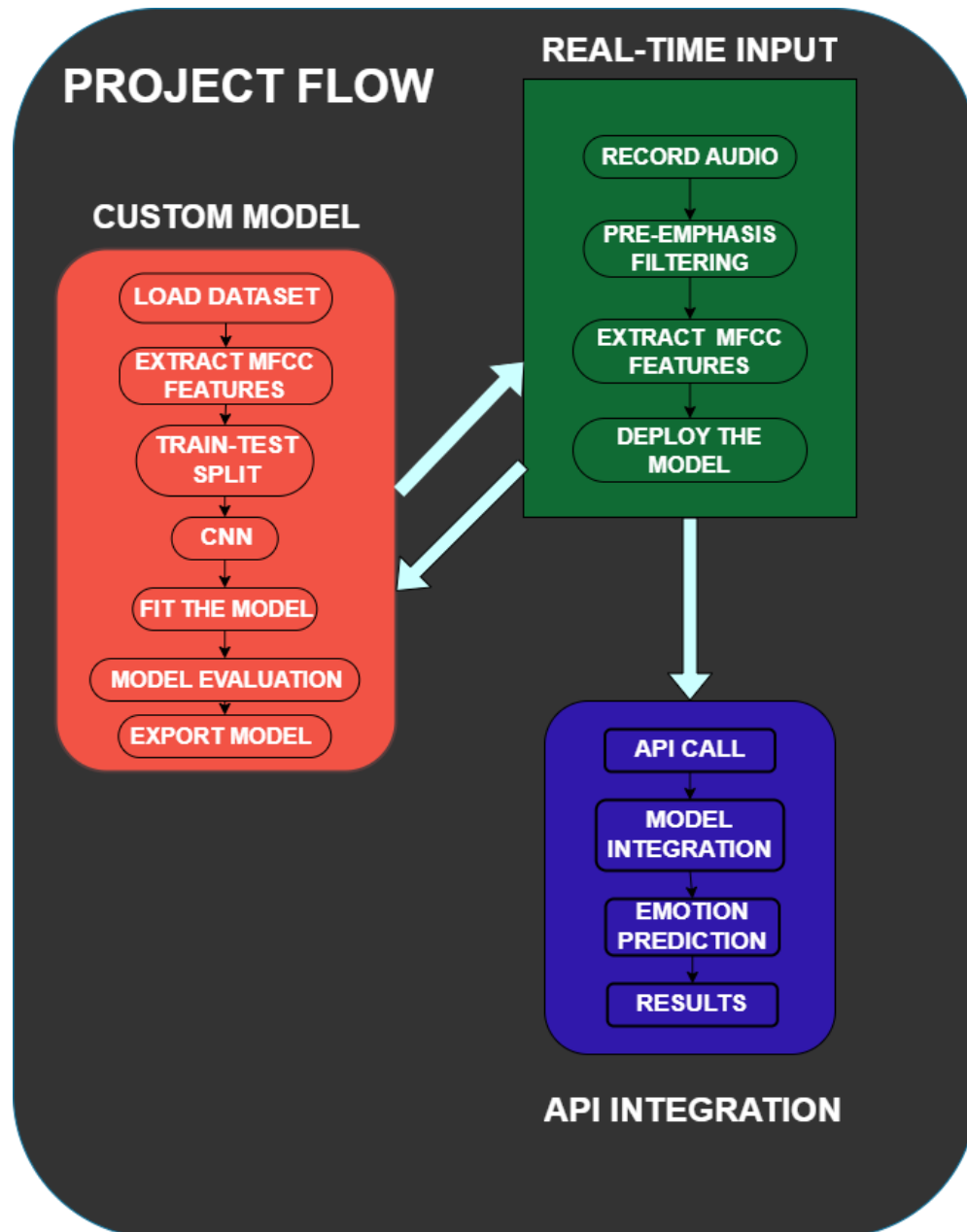


Fig 4.1 Process Overview

4.3 Process Novelty

The unique features of the proposed model are as follows:

1) Multi Label Classification:

The proposed classification model can classify more than 4 unique emotions.

2) Feature Engineering:

Features extracted by a customized neural network framework, with a one-hot encoder is incorporated

3) Custom CNN model:

The CNN model has been constructed with a stack of eight different unique hyper tuned models. The model yields precise classifications with high accuracy. The proposed working environment would be Jupyter Notebook or python.

CHAPTER 5

PROPOSED SYSTEM'S ATTRIBUTES

5.1 Working Principle

In our proposed project, we aim to improve the performance of CNN-based speech emotion recognition by exploring different techniques for feature extraction, such as Mel-frequency spectrogram and pitch-based features, using data augmentation to increase the size of the training dataset, and fine-tuning pre-trained CNN models to improve classification accuracy.

5.1.1 Librosa

In our project, we use the librosa library to pre-process and extract features from the audio files, such as MFCCs and Mel-spectrograms, which are then fed into the CNN model for training and classification. We also use librosa for data augmentation techniques such as time-stretching, pitch-shifting, and adding noise to the audio files, which helps to improve the model's ability to generalize to new, unseen data. Additionally, we use librosa for visualizing the audio signals and the extracted features to gain insights into the data and model performance.

5.1.2 CNN

The CNN model chosen for training is a deep learning architecture that consists of multiple convolutional layers followed by pooling layers and fully connected layers for classification. It learns to extract relevant features from the audio input and map them to the corresponding emotions. The proposed model architecture includes techniques such as batch normalization, dropout, and early stopping to prevent overfitting and improve generalization. The CNN model is

trained using backpropagation with an optimization algorithm such as Adam or RMSprop.

5.2 Features

1) Accessibility

The proposed system incorporates a unique interactive website, which could be easily accessed by users across the globe with the aid of the internet.

2) Security

The proposed system restricts access to the designated user alone, making the data provided by the user feel safe.

3) Precision

The proposed system as a whole yield better results with high efficiency when compared to results yielded by other models.

4) Compatibility

The proposed system is highly compatible and can be accessed across a wide range of devices.

5) Low Time / Memory Consumption

The proposed system would consume low time and space while processing results with high efficiency.

6) Customizable

The proposed system can be customized based on consumer requirements

5.3. Proposed Algorithm Workflow chart

Figure 5.1 illustrates the proposed algorithm flowchart for the CNN-based speech emotion recognition project. Firstly, the audio signals are pre-processed and the features such as MFCC and spectral contrast are extracted from the audio data. The MFCC values are concatenated to form a feature matrix. One-hot encoding is then applied to the emotion labels for classification. The model is trained using the concatenated feature matrix and the one-hot encoded labels. The trained model is then used to predict the emotion felt by the speaker.

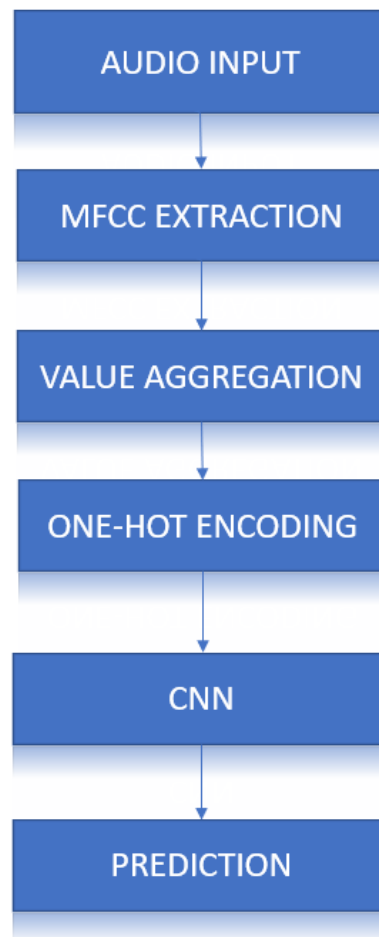


Fig 5.1 Proposed Algorithm Workflow chart

5.4 Proposed CNN Architecture

In this project, we propose a CNN-based architecture for speech emotion recognition that consists of multiple convolutional layers followed by pooling layers and fully connected layers for classification. The model is trained using backpropagation with an optimization algorithm such as Adam or RMSprop. The proposed model architecture includes techniques such as batch normalization, dropout, and early stopping to prevent overfitting and improve generalization. The final output layer of the model predicts the emotion label corresponding to the input audio signal.

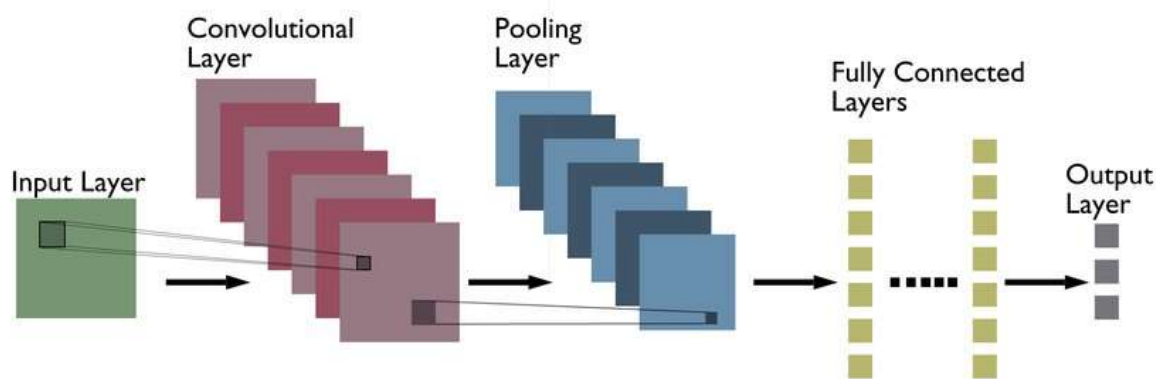


Fig 5.2 Proposed CNN Model Architecture

Layer (type)	Output Shape	Param #
reshape_3 (Reshape)	(None, 20, 1)	0
conv1d_10 (Conv1D)	(None, 18, 32)	128
max_pooling1d_12 (MaxPooling1D)	(None, 9, 32)	0
dropout_11 (Dropout)	(None, 9, 32)	0
conv1d_11 (Conv1D)	(None, 9, 64)	6208
max_pooling1d_13 (MaxPooling1D)	(None, 4, 64)	0
conv1d_12 (Conv1D)	(None, 4, 256)	49408
max_pooling1d_14 (MaxPooling1D)	(None, 1, 256)	0
flatten_4 (Flatten)	(None, 256)	0
dense_12 (Dense)	(None, 256)	65792
dropout_12 (Dropout)	(None, 256)	0
dense_13 (Dense)	(None, 256)	65792
dropout_13 (Dropout)	(None, 256)	0
dense_14 (Dense)	(None, 8)	2056

Fig 5.3 Proposed CNN model developed

CHAPTER 6

RESULTS AND DISCUSSION

6.1 CNN Model results

Metric	CNN OUTPUT
Accuracy Score	0.7166
Balanced Accuracy Score	0.7343
Cohen Kappa Score	0.6747
F1 Score (Macro)	0.6879
F1 Score (Micro)	0.7166
F1 Score (Weighted)	0.6671
Jaccard Score (Macro)	0.6044
Jaccard Score (Micro)	0.5584
Jaccard Score (Weighted)	0.5780
Hamming Loss	0.2833

Table 6.1 CNN Results

6.2 Confusion Matrix

A confusion matrix is a table that is commonly used to evaluate the performance of a CNN-based speech emotion recognition model on a test dataset, where the true emotion labels are known. The matrix shows the number of true positives, true negatives, false positives, and false negatives for each emotion class. The model's performance can be evaluated based on metrics such as accuracy, precision, recall, and F1-score, which can be calculated using the values in the

confusion matrix. The confusion matrix for the CNN-based model can be visualized using tools such as matplotlib or seaborn.

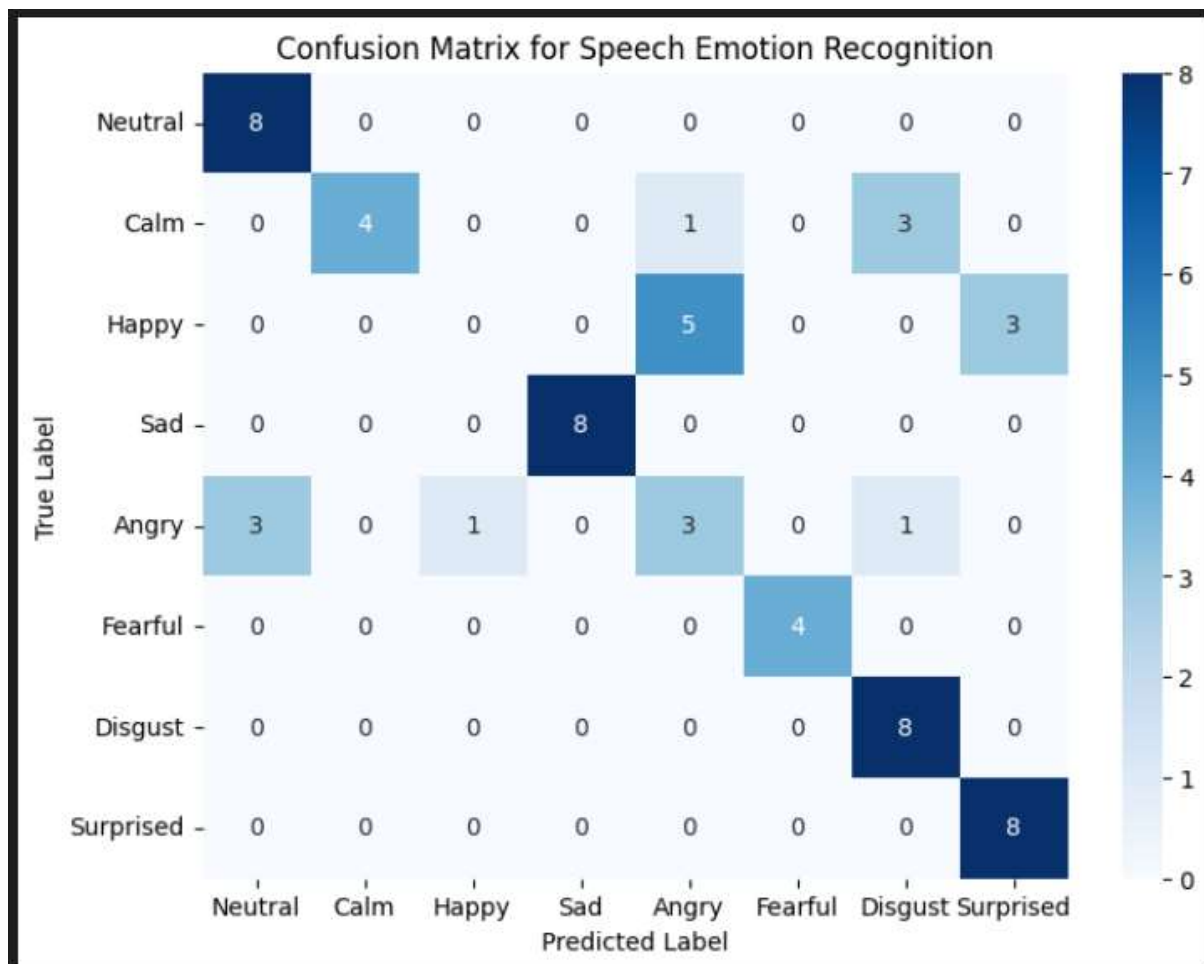


Fig 6.1 Confusion Matrix for SER using CNN

6.3 Application Interface:

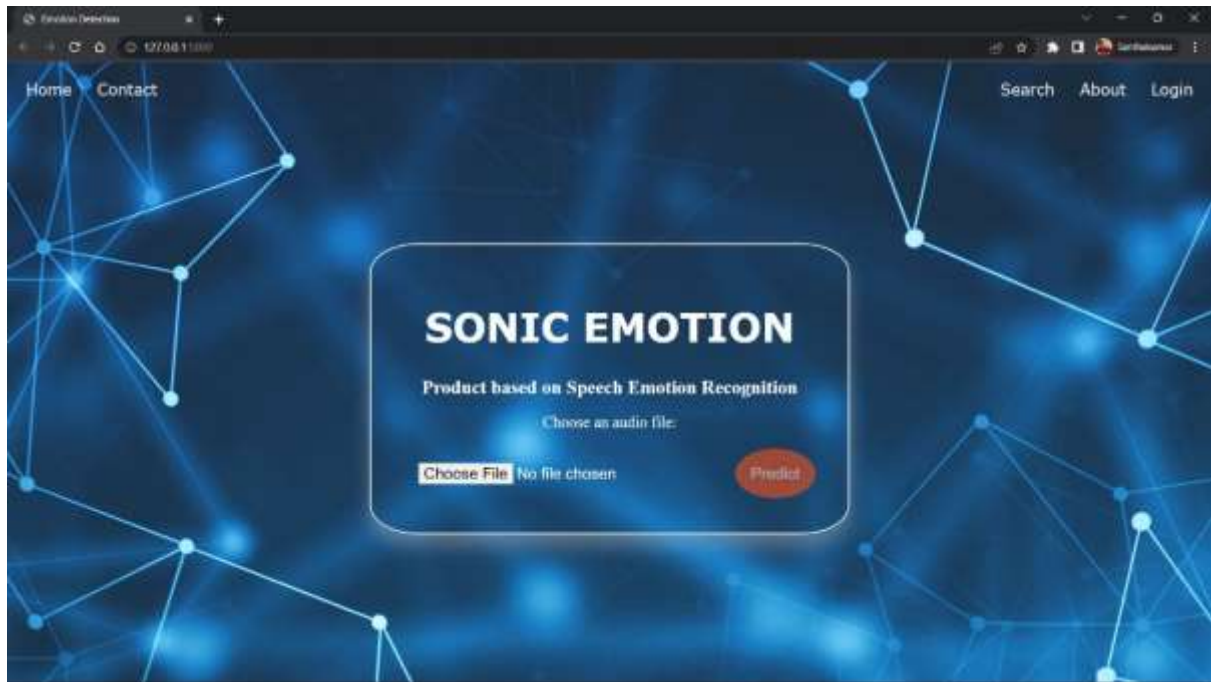


Fig 6.2 Output Screenshot-1

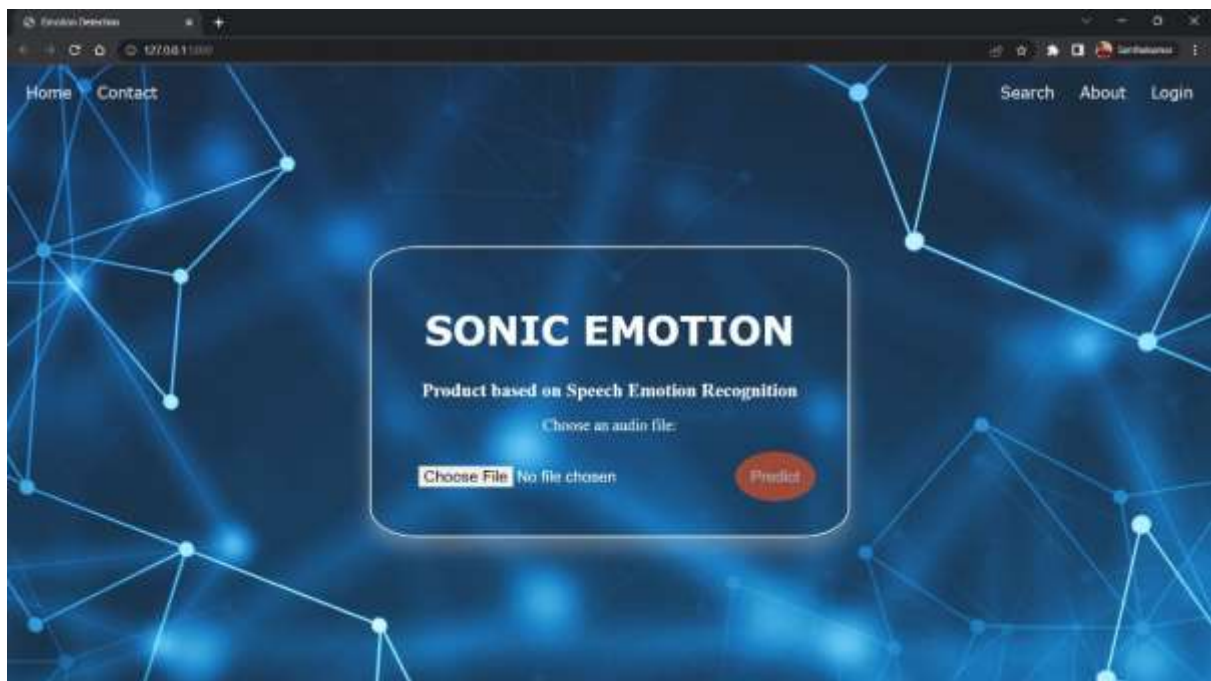


Fig 6.3 Output Screenshot-2

6.4 Target Audience

The target audience for speech emotion recognition can vary depending on the specific application.

Mental Health Professionals: Speech emotion recognition can be used in mental health diagnosis and treatment to assess patients' emotional states and monitor treatment progress.

Designers of Human-Robot Interaction: Speech emotion recognition can be used to create intelligent systems that engage with people in a more natural and empathetic manner.

Customer Service Providers: Call centres and online customer service can use speech emotion recognition to recognise customers' emotional states and provide personalised responses.

CHAPTER 7

CONCLUSION & FUTURE SCOPE

7.1 Conclusion

In conclusion, speech emotion recognition has gained significant attention from researchers due to its potential applications in various fields. The growing availability of large datasets and advances in machine learning algorithms have aided in the development of speech emotion detection systems. In recent years, speech emotion recognition using machine learning and CNN has yielded encouraging results. CNNs have been successful in extracting meaningful features from speech signals, while machine learning algorithms have been used to classify emotions accurately.

7.2 Future Scope

The future scope of speech emotion recognition is vast and holds great potential for a variety of applications in different fields.

Multimodal Speech Emotion Recognition: By combining various modalities such as facial expressions, physiological signals, and textual information, speech emotion recognition systems can improve their accuracy and robustness in real-world situations.

Real-time Speech Emotion Recognition: Systems for real-time speech emotion recognition can be helpful in a variety of scenarios, including video conferencing, online customer service, and virtual reality environments. Creating systems that can process real-time speech signals is a significant task that necessitates efficient algorithms and hardware resources.

Personalized Speech Emotion Recognition: Systems for recognising emotions in speech can be used in mental health diagnosis and therapy, personalised marketing, and human-robot interaction. An important research path is the development of systems that can adapt to individual differences in emotional expression and perception.

REFERENCES

- [1] Stuti Juyal, Chirag Killa, Gurvinder Pal Singh, Nishant Gupta, Vedika Gupta 'Emotion Recognition from Speech Using Deep Neural Network' https://link.springer.com/chapter/10.1007/978-3-030-76167-7_1
- [2] Huang, F., Zhang, J., Zhou, C. et al. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. *Landslides* 17, 217–229 (2020). DOI: 10.1007/s10346-019-01274-9 <https://link.springer.com/article/10.1007%2Fs10346-019-01274-9>
- [3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in *IEEE Access*, vol. 7, pp. 117327-117345, 2019, DOI: 10.1109/ACCESS.2019.2936124 <https://ieeexplore.ieee.org/document/8805181>
- [4] Meheebub Sahana, Binh Thai Pham, Manas Shukla, Romulus Costache, Do Xuan Thu, Rabin Chakraborty, Neelima Satyam, Huu Duy Nguyen, Tran Van Phong, Hiep Van Le, Subodh Chandra Pal, G. Areendran, Kashif Imdad & Indra Prakash (2020) Rainfall induced landslide susceptibility mapping using novel hybrid soft computing methods based on multi-layer perceptron neural network classifier, *Geocarto International*, DOI: 10.1080/10106049.2020.1837262
- [5] Sheena Christabel Pravin, Palanivelan, M, 'A Hybrid Deep Ensemble for Speech Disfluency Classification', *Circuits, Systems, and Signal Processing*, Springer, vol. 40, no.8, pp. 3968-3995, July 2021. https://www.researchgate.net/publication/349228170_A_Hybrid_Deep_Ensemble_for_Speech_Disfluency_Classification
- [6] Krishnan, P.T., Joseph Raj, A.N. & Rajangam, V. Emotion classification from speech signal based on empirical mode decomposition and non-linear features.

Complex Intell. Syst. 7, 1919–1934(2021).DOI:10.1007/s40747-021-00295-z
<https://link.springer.com/article/10.1007%2Fs40747-021-00295-z>

[7] Mohamad Nezami, O., Jamshid Lou, P. & Karami, M. ShEMO: a large-scale validated database for Persian speech emotion detection. Lang Resources & Evaluation 53, 1–16 (2019).DOI:10.1007/s10579-018-9427-x
<https://link.springer.com/article/10.1007%2Fs10579-018-9427-x>

[8] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 858-862, DOI: 10.1109/ISS1.2017.8389299
<https://ieeexplore.ieee.org/document/8389299>

[9] M. N. Stolar, M. Lech, R. S. Bolia and M. Skinner, "Real-time speech emotion recognition using RGB image classification and transfer learning," 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS), 2017, pp. 1-8, DOI: 10.1109/ICSPCS.2017.8270472 .
<https://ieeexplore.IEEE.org/document/8270472>

[10] J. D. Arias-Londono, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, On combining information from modulation spectra and Mel-frequency cepstral coefficients for automatic detection of pathological voices, Logoped. Phoniatr. Vocol. 36(2) (2011) 60–69. <https://pubmed.ncbi.nlm.nih.gov/21073260/>