iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61278

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Face and Speech Emotion Recognition System

Shweta Sondawale[1], Batul Chinikamwala[2], Srushti Dangde[3], Sakshi Salaskar[4], Arpita Shinde[5]
Computer Science Department, Sinhgad Academy of Engineering

Abstract: ***Emotions serve as the cornerstone of human communication, facilitating the expression of one's inner thoughts and feelings to others. Speech Emotion Recognition (SER) represents a pivotal endeavour aimed at deciphering the emotional nuances embedded within a speaker's voice signal. Universal emotions such as neutrality, anger, happiness, and sadness form the basis of this recognition process, allowing for the identification of fundamental emotional states. To achieve this, spectral and prosodic features are leveraged, each offering unique insights into the emotional content of speech. Spectral features, exemplified by the Mel Frequency Cepstral Coefficient (MFCC), provide a detailed analysis of the frequency distribution within speech signals, while prosodic features encompass elements like fundamental frequency, volume, pitch, speech intensity, and glottal parameters, capturing the rhythmic and tonal variations indicative of different emotional states. Through the integration of these features, SER systems can effectively simulate and classify a diverse range of emotional expressions, paving the way for enhanced human-computer interaction and communication technologies.***

Keywords: Emotions, Speech Emotion Recognition (SER), speaker's emotional state, fundamental frequency, speech intensity

I. INTRODUCTION

Emotion recognition through speech analysis has become increasingly important across various sectors, reflecting a growing interest over the years. Speech Emotion Recognition (SER) has emerged as a pivotal research field within Human-Computer Interaction (HCI), finding applications in diverse domains such as robotics, banking, education, customer service, transportation, and entertainment. Understanding the emotional states of students during the learning process, whether in traditional classrooms or through e-learning platforms, is particularly significant, offering immense potential to elevate the quality of education. By employing SER tools, educators can identify areas of strength and weakness in students' comprehension and engagement, thus tailoring teaching strategies to address specific emotional needs and fostering a conducive learning environment. Recognizing and addressing students' emotional well-being in educational settings is crucial for nurturing holistic development.

Practically, SER presents a computational challenge characterized by two core components: extracting relevant features from speech signals and accurately classifying emotional states. Depression, with its symptoms of fatigue and disinterest, is intricately intertwined with emotional experiences. Although contemporary research often treats depression and emotional states separately, integrating assessments of depression with dimensional emotion analysis offers a promising avenue towards a more comprehensive understanding of mental health. By bridging the gap between emotional recognition and mental health assessment, researchers aim to provide more effective interventions and support systems for individuals experiencing emotional distress.

Additionally, as you mentioned, there is a growing interest in exploring the relationship between depression and affective states. By integrating depression estimation with dimensional affective analysis, researchers aim to develop more comprehensive models for understanding and predicting depressive symptoms based on speech cues. This interdisciplinary approach could lead to more effective screening and monitoring tools for mental health disorders, ultimately facilitating early intervention and treatment.

In terms of technological advancements, leveraging asymmetric data and privileged knowledge could prove beneficial in improving the generalizability and accuracy of SER models. For example, incorporating facial features associated with aging into age estimation algorithms could enhance the performance of automated age estimation systems, particularly when dealing with unseen or challenging data samples. Overall, the continued research and development efforts in SER hold immense potential for advancing human-computer interaction and improving various aspects of daily life, from education and healthcare to entertainment and beyond. By leveraging the power of speech to infer emotional states, we can create more empathetic and intelligent systems that better understand and respond to human needs and preferences.

II. LITERATURE SURVEY

"Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends" by M. Alippi, G. D'Angelo, A. Vinciarelli (2011) - This paper provides a comprehensive overview of the evolution of SER research over the past two decades. It discusses benchmark datasets, evaluation metrics, and challenges in SER, such as variability in emotional expression and cross-cultural differences. The authors highlight ongoing trends in SER, including the adoption of deep learning techniques and the integration of multimodal data for improved emotion recognition accuracy. [1]

"A Review on Speech Emotion Recognition: Challenges, Recent Advances, and Future Directions" by H. D. Sharma, D. K. Yadav (2019) - This review paper discusses the challenges and recent advances in SER, focusing on both traditional machine learning and deep learning approaches. It explores feature extraction methods, including prosodic, spectral, and temporal features, and compares the performance of various machine learning algorithms for emotion classification. The paper also highlights future directions in SER research, such as enhancing model interpretability, addressing data scarcity issues, and incorporating contextual information for more robust emotion recognition. [2]

"Deep Learning for Emotion Recognition: A Survey" by Z. Zhao, L. Liu, J. Li, Y. Yang, M. Zhang (2019) - This survey paper provides an in-depth analysis of deep learning techniques for emotion recognition, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants. It reviews the application of deep learning models to various modalities, such as speech, facial expressions, and physiological signals, and discusses their advantages and limitations. The authors also examine challenges in deep learning-based emotion recognition, such as the need for large, annotated datasets and model generalization across different domains and languages. [3]

"Multimodal Emotion Recognition: A Survey of Affective Computing Approaches" by M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic (2017) - This survey paper focuses on multimodal emotion recognition, which combines information from multiple modalities, including speech, facial expressions, body gestures, and physiological signals. It discusses various approaches to multimodal fusion, such as early fusion, late fusion, and hybrid fusion, and evaluates their effectiveness for emotion recognition tasks. The authors highlight the importance of multimodal data fusion in improving emotion recognition accuracy and robustness, especially in real-world scenarios where multiple sources of information are available. [4]

"Towards Privacy-Preserving Speech Emotion Recognition: A Survey" by H. Chen, Y. Zhang, Z. Zhao, Q. Liu, Q. Zhang (2020) - This survey paper explores privacy-preserving techniques for speech emotion recognition, addressing concerns related to data privacy and security. [5]

III.METHODOLOGY

A. *Project Scope Definition*

1) Develop a Speech Emotion Recognition (SER) system capable of accurately detecting and classifying emotions from speech signals.
2) Investigate the feasibility of applying SER in specific domains such as education, healthcare, customer service, etc.
3) Address privacy concerns in SER by exploring privacy-preserving techniques while maintaining performance.

B. *Data Collection*

1) Gather relevant data sources, including speech datasets, multimodal datasets (e.g., combining speech with facial expressions or physiological signals), and benchmark datasets for evaluating SER algorithms.
2) Ensure data quality and diversity to capture a wide range of emotional expressions and demographic characteristics.

C. *Feature Extraction*

Implement feature extraction techniques to extract relevant features from speech signals and face features. Commonly used features include:

1) Mel-frequency cepstral coefficients (MFCCs)
2) Prosodic features (e.g., pitch, intensity, duration)
3) Spectral features (e.g., spectral centroid, bandwidth)
4) Temporal features (e.g., speech rate, pause duration)
5) Facial Landmarks: These are specific points on the face such as the corners of the eyes, the tip of the nose, and the corners of the mouth. The positions of these landmarks can be used to characterize facial expressions.
6) Facial Action Units (AUs): These are specific facial muscle movements associated with different emotions, as described by the Facial Action Coding System (FACS). For example, raising of the eyebrows might indicate surprise, while a smile might indicate happiness.

D. *Model Development*

Develop SER models using appropriate machine learning or deep learning algorithms. This may involve:

1) Experimenting with different algorithms such as Haar-Cascade, Multi-layer perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), etc.

2) Fine-tuning hyperparameters and model architectures to optimize performance.
3) Implementing multimodal fusion techniques if integrating data from multiple sources.

E. Evaluation

1) Evaluate the performance of SER models using appropriate evaluation metrics. Common metrics include accuracy, precision, recall, F1-score, and confusion matrix analysis.
2) Validate the models using cross-validation or holdout validation on separate test datasets.
3) Compare the performance of different models and techniques to identify the most effective approaches.

IV. IMPLEMENTATION

A. Tools And Technology Used

Machine Learning: In the realm of machine learning, supervised models undergo training using labelled datasets, a process that empowers them to refine their accuracy incrementally. Take, for example, algorithms tasked with recognizing objects such as dogs within images; they learn from annotated examples provided by human annotators. The reliance on supervised learning is prevalent across various applications and industries, as machine learning offers unparalleled precision when compared to manual intervention. At its core, machine learning empowers users to input extensive datasets into computer algorithms, which subsequently sift through the data, extracting insights, and providing recommendations or making decisions based solely on the input data. This capability not only streamlines processes but also enables businesses and organizations to leverage the wealth of information at their disposal to drive informed actions and strategies.

B. Mathematical Model

Let S be the Whole system S= {I, P, O}
I - Input
P - Procedure
O - Output Input (I)
I= Input as Live camera and Audio
Procedure (P),
P = I
Output (O) = Detect the Speech and Face Emotion.

C. Algorithm Used

1) *Harcascade Algorithm* - Haar Cascade is a feature-based object detection algorithm designed to detect objects in images. A cascade feature is trained on a set of positive and negative images for detection. This algorithm does not require extensive computation and can operate in real-time.
2) *MLP Algorithm* - MLP networks are used in a supervised learning format. The common learning algorithm in MLP networks is also called the backpropagation algorithm. A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of output data from a set of input data.
3) *CNN Algorithm* - Convolutional Layer: A convolutional layer is the fundamental component of CNN architecture that performs feature extraction, usually consisting of a combination of linear and non-linear operations, i.e., convolution operation and activation functions.
4) Non-linear activation function: The result of a linear operation, such as convolution, is passed through a non-linear activation function. The most common nonlinear activation function in use today is Rectified Linear Unit (ReLU).

- Pooling layer: The pooling layer provides a general down sampling operation that reduces the dimensionality of the in-plane feature maps, providing translation invariance for small shifts and distortions, and reducing the number of subsequent parameters to be trained.
- Fully connected layer: The output feature map of the final convolutional or pooling layer is usually flattened, converting it to a one-dimensional (1D) numeric array (or vector) and concatenated into one or more fully connected layers. It is a dense layer where each input is associated with each output through learnable weights. The features extracted from the convolutional layer and the down sampled pooling layer are then mapped via a subset of fully connected layers to the final network output, such as the probability for each class in the classification task. The last fully connected layer typically has a number of output nodes equal to the number of classes.

- Activation function of the last layer: The activation function applied to the last fully connected layer is usually different from other layers. The activation function applied to multi-class classification problems is a softmax function that normalizes the output actual values of the last fully connected layer by the probability of the target class, with each value in the range 0 to 1 and the sum of all values being 1.

V. RESULTS

Fig. 1 Home Page

Fig. 2 Registration Page

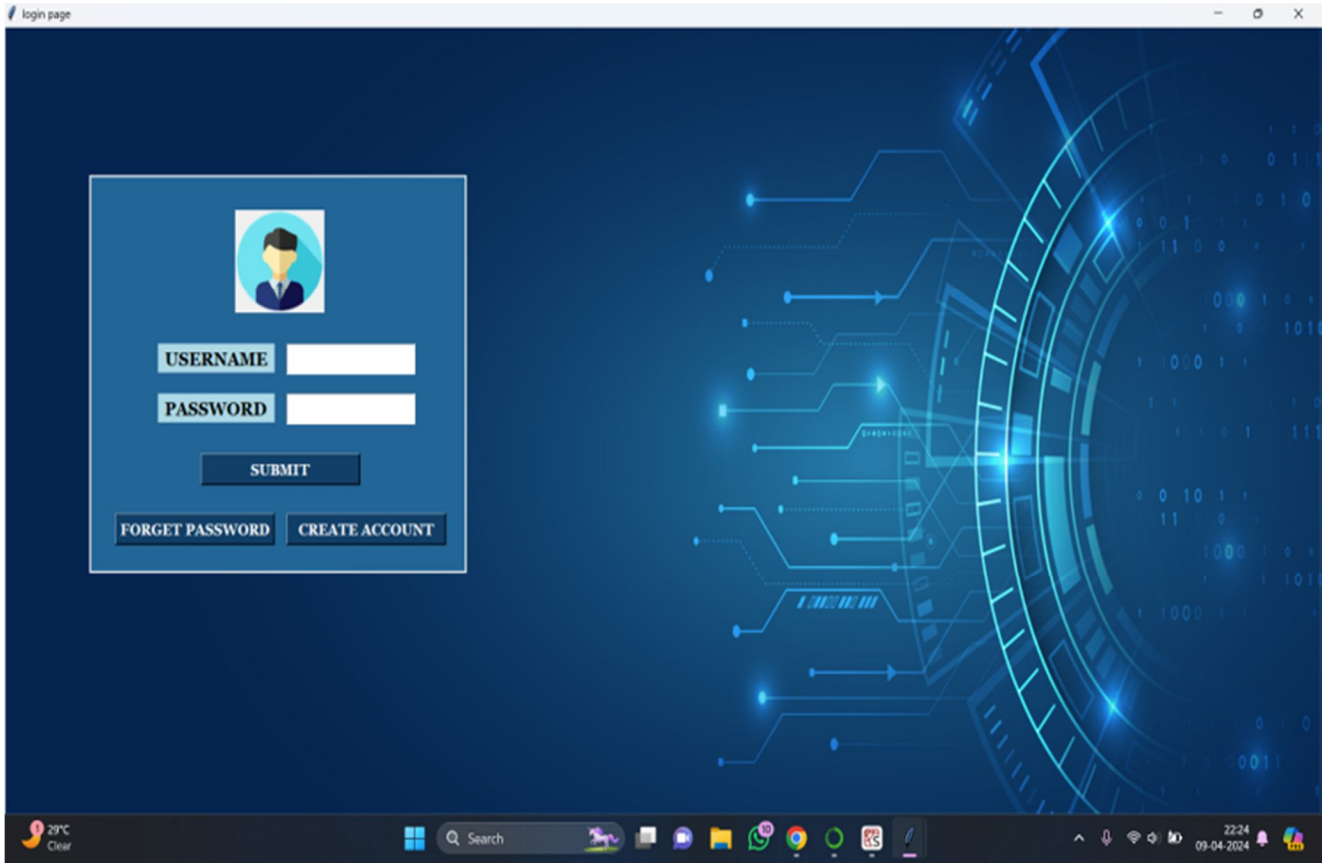


Fig. 3 Login Page

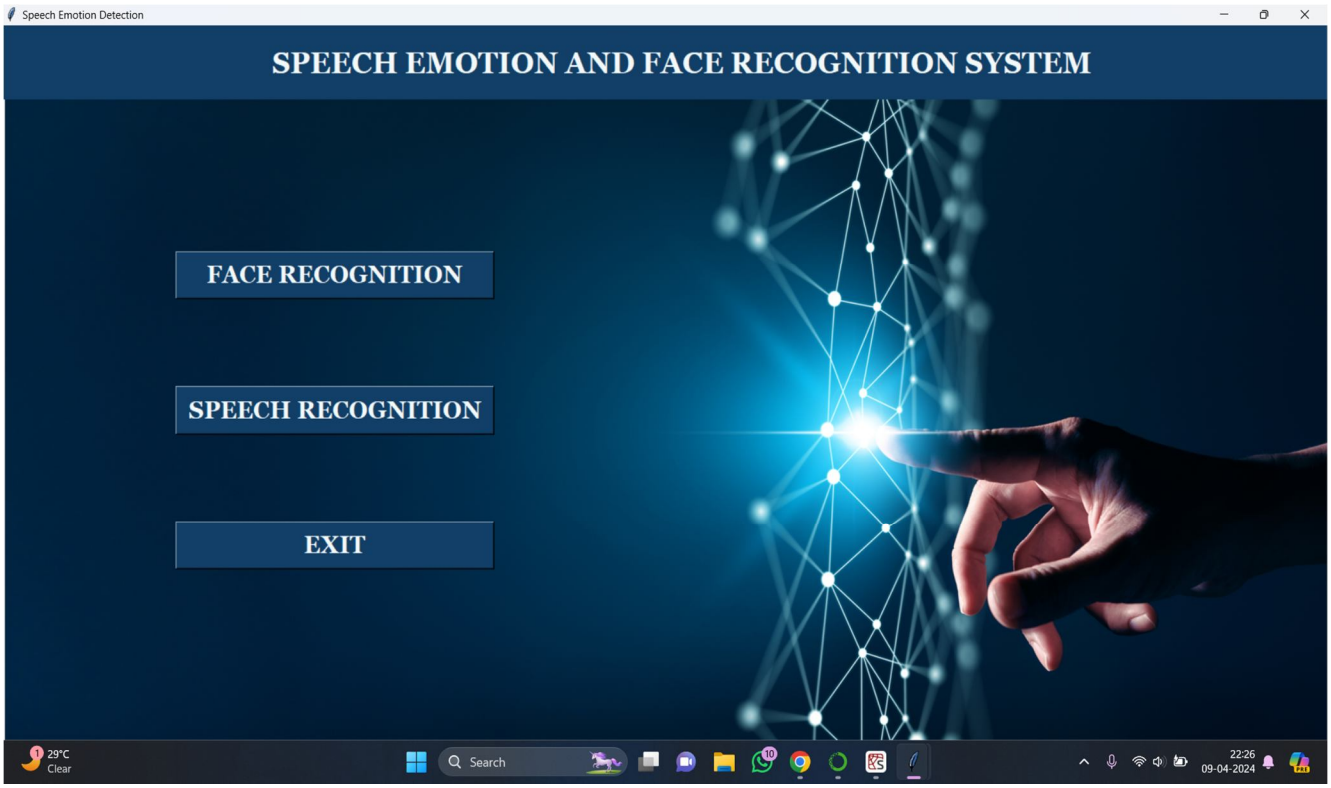


Fig. 4 GUI Main

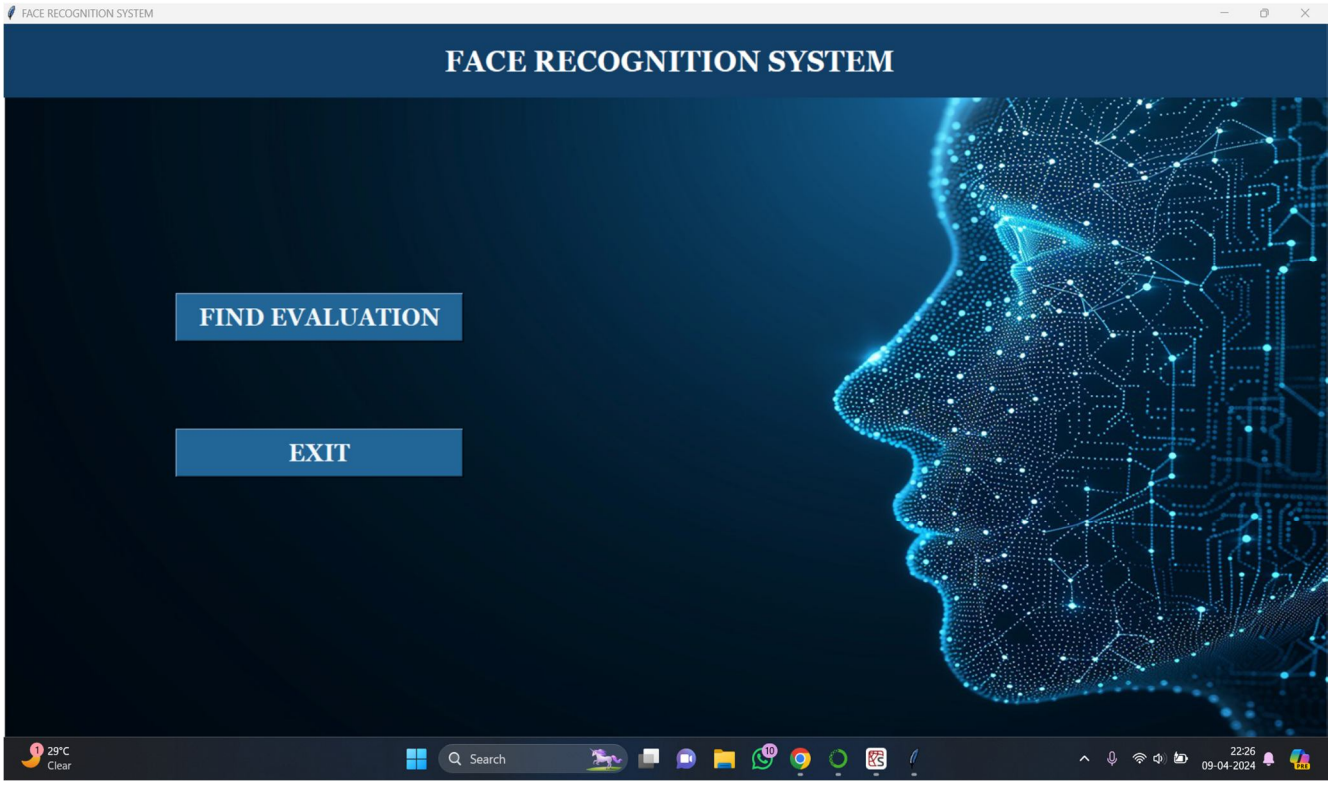


Fig. 5 Face Evaluation Page

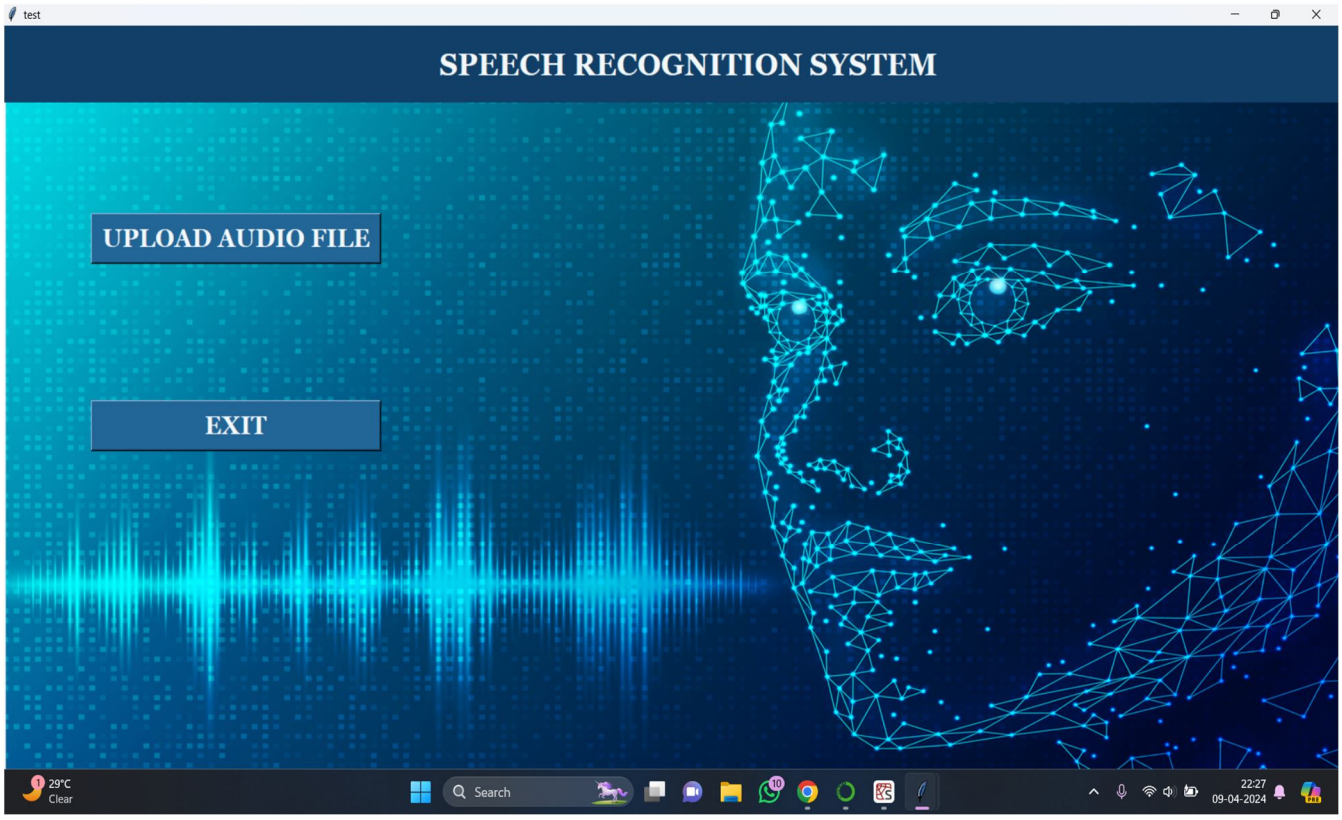


Fig 6 Speech Evaluation Page

Fig. 7 Speech Emotion Recognition

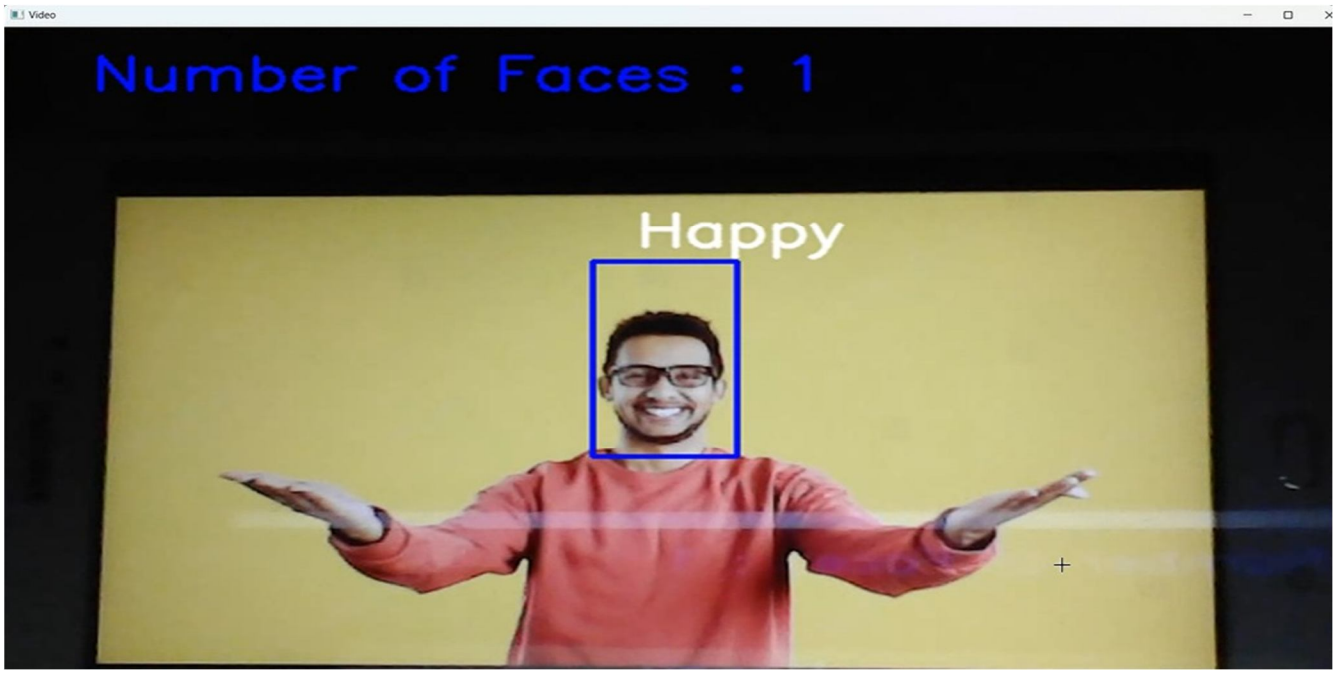


Fig. 8 Face Emotion Recognition

VI. CONCLUSION

Speech emotion recognition (SER) using convolutional neural network (CNN) models stands as a significant advancement in human-machine interaction. Speech, being a primary mode of communication, holds immense importance in understanding human emotions. SER technology finds extensive application in various fields, ranging from online communication platforms to virtual assistants, where machines are capable of discerning and responding to human emotions effectively. This technology enhances user experience by enabling more personalized and empathetic interactions, contributing to the seamless integration of machines into human-centric environments.

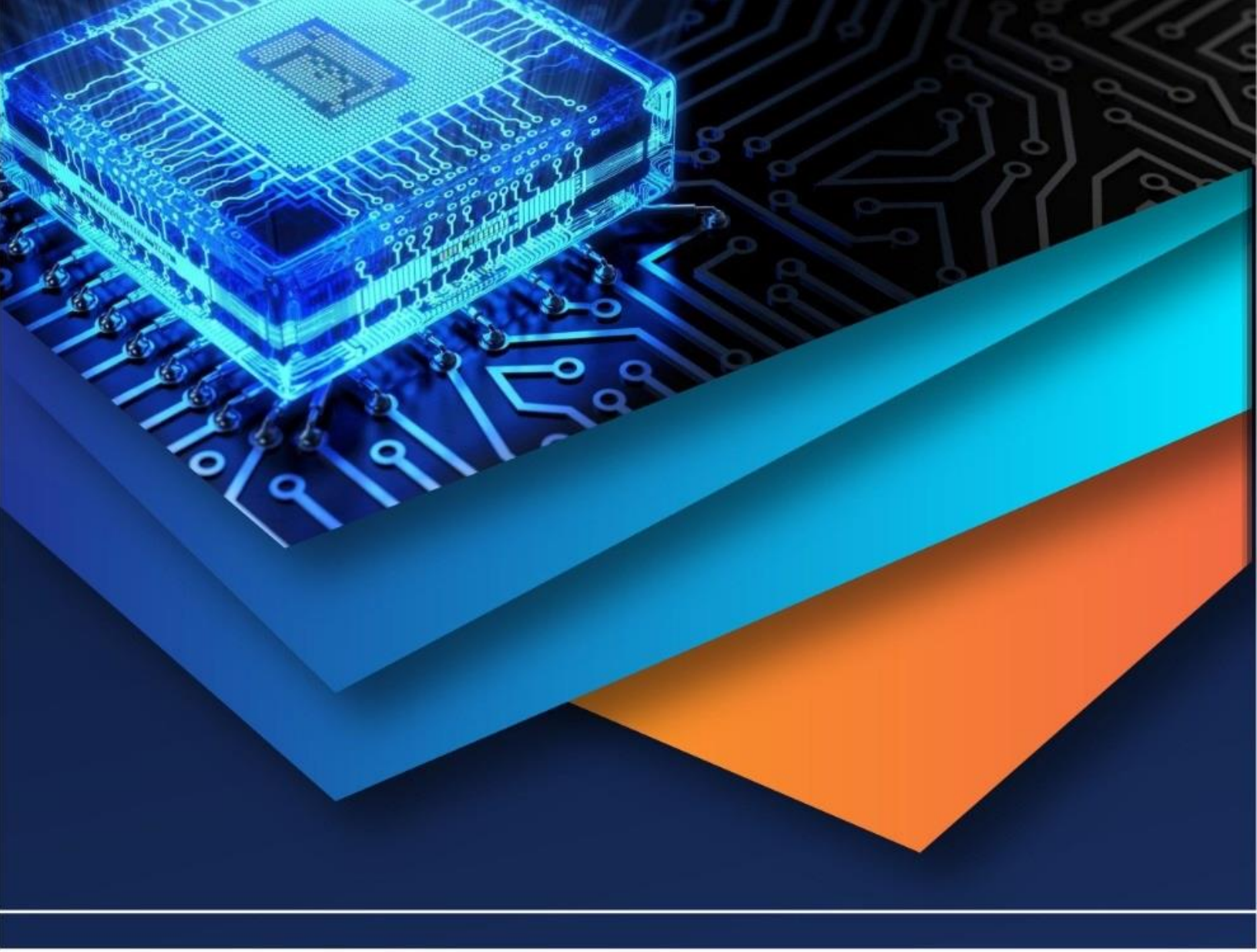
The Dataset RAVDESS provides crucial resources for training and evaluating SER models. These datasets offer diverse recordings of human speech, annotated with corresponding emotional labels, covering a wide spectrum of emotions. Such datasets serve as the cornerstone for the development of accurate and robust SER systems, facilitating the exploration of novel methodologies and approaches.

In SER, CNN models play a pivotal role in extracting meaningful features from speech signals. Preprocessing techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and chroma features enable the capture of distinctive characteristics indicative of emotional content in speech. Integration with frameworks like TensorFlow further streamlines the development and deployment of CNN-based SER systems, enhancing their efficiency and scalability.

Despite the progress achieved, further enhancements are necessary to improve the accuracy and applicability of SER models. Expanding the diversity and size of voice datasets, refining precision in emotion labelling, and accommodating variations in accents and speaking styles are essential steps towards achieving greater accuracy in emotion recognition. Continued research and development efforts in these areas will advance the capabilities of SER technology, fostering deeper understanding and seamless integration of emotions in human-machine interaction scenarios.

REFERENCES

[1] Saikat Basu, Jaybrata Chakraborty, Arnab Bag and Md. Aftabuddin.” A Review on Emotion Recognition using Speech” (IEEE 2017)
[2] Ashwin V. Gatty, G. S. Shivakumar, Kiran Shetty.” Speech Emotion Recognition using Machine Learning” (IJRESM 2021)
[3] Chaitanya Singlaa, Sukhdev Singhb, Monika Pathakc” AUTOMATIC AUDIO BASED EMOTION RECOGNITION SYSTEM: SCOPE AND CHALLENGES”
[4] Ajay Gupta1, Siddhesh Morye2, Mukul Sitap3, Supriya Chaudhary4.” Speech based Emotion Recognition using Machine Learning” (IRJET 2021)
[5] Sudha Tushara S. and Y. Zhang, “Analyzing Tweets to Discover Twitter Users’ Mental Health Status by a Word-Frequency Method”, IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), Visakhapatnam, India, 2019
[6] Sudha Tushara S. and Y. Zhang, “Finding a Depressive Twitter User by Analysing Time Series Tweets”, 2020 IEEE India Council International Subsections’ Conference (INDISCON) Oct 3-4, 2020, 978-1-7281-8734-1/20/©2020 IEEE (Accepted)
[7] Sudha Tushara S. and Y. Zhang, “Finding a Depressive Twitter User by Analyzing Depress and Antidepressant Tweets”, 2020 IEEE India Council International Subsections’ Conference (INDISCON) Oct 3-4, 2020, 978-1-7281-8734-1/20/©2020 IEEE (Accepted)
[8] Kelly, Y., Zilanawala, A., Booker, C., Sacker, A. “Social media use and adolescent mental health: Findings from the UK Millennium Cohort Study”. EClinical Medicine. 2019.
[9] Twenge, J. M., Campbell, W. K. “Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study”. Preventative Medicine Reports, 12, 271-283. 2018.
[10] Mogg K, Bradley BP, Williams R, Mathews A. Subliminal processing of emotional information in anxiety and depression. J Abnorm Psychol. 1993; 102:304–11
[11] https://www.omnicoreagency.com/twitter-statistics/ (accessed on March 25, 2019)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY