IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

Speech Emotion Recognition using Machine Learning With Real-time Audio Analysis

Prof. Sharda Dabhekar[1], Shivam K. Yadao[2], Tanay R. Tiwari[3], Pranjal Zode[4], Shubham B. Vaidya[5]
Guide, Department of Computer Science Engineering[1]
Students, Department of Computer Science Engineering[2]
Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, Maharashtra, India
shivamyadao284@gmail.com, tanaytiwari21@gmail.com,
pranjalzode007@gmail.com, shubhamvaidya599@gmail.com

Abstract: *This research paper presents a Speech Emotion Recognition (SER) system utilizing a Multilayer Perceptron (MLP) classifier and real-time audio analysis. The system records audio samples, extracts relevant features, and employs machine learning techniques to predict emotions in spoken language. The study focuses on the development of an intuitive Graphical User Interface (GUI) using the Kivy framework, providing a user-friendly platform for real-time emotion analysis.*

In shortly In this project, we attempt to detect underlying emotions such as (sad, happy, neutral, angry, disgust, surprised, fearful and calm) in recorded speech by analysing the acoustic features of the audio data of recordings and Created an application to implement the same on user input.

Keywords: python, python libraries, emotion recognition, Ravdess dataset, kivy framework for python

I. INTRODUCTION

This research paper presents a speech emotion recognition (ser) system implemented in python, integrating audio processing, machine learning, and a graphical user interface (gui) using the kivy framework. The system is designed to analyze and recognize human emotions expressed through speech. The core components include audio feature extraction, model training using a multilayer perceptron (mlp) classifier from the scikit-learn library, and a real-time emotion prediction functionality. The gui, developed with kivy, facilitates user interaction by providing a platform for voice input and displaying the predicted emotion.

The code utilizes various libraries such as pyaudio, wave, numpy, scikit-learn, soundfile, glob, librosa, and kivy. Audio features like mel-frequency cepstral coefficients (mfcc), chroma, and mel spectrograms are extracted to represent the emotional content of speech. The mlp classifier is trained on a dataset comprising various emotional states, allowing the model to generalize and predict emotions accurately.

The real-time emotion prediction feature is integrated into a kivy application, offering an intuitive interface for users to input their voice and receive instant feedback on the predicted emotion. This research contributes to the field of affective computing, offering a practical application for recognizing emotions in spoken language and providing a foundation for further research and development in human-computer interaction and emotional artificial intelligence.

II. LITERATURE REVIEW

Below is a literature review discussing key concepts and methodologies related to the code.

Speech emotion recognition (ser):

- Overview: ser involves the identification of emotional states from speech signals. It finds applications in human-computer interaction, affective computing, and mental health diagnostics.
- Feature extraction: the code utilizes three types of features - mel-frequency cepstral coefficients (mfccs), chroma features, and mel spectrograms. These features are commonly used in ser as they capture relevant information about the spectral and temporal characteristics of speech signals.
- Machine learning models for ser:multilayer perceptron (mlp): the code employs the mlpclassifier from scikit-learn for emotion classification. mlps, being a type of artificial neural network, have been successful in

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

capturing complex relationships in speech features for emotion recognition

- Training data: the model is trained on the ravdess dataset, which is a widely used dataset for ser research. it contains acted speech from actors expressing eight different emotions.
- Audio recording and processing: pyaudio and wave module: the pyaudio library is used for real-time audio recording, and the wave module is employed for handling audio file i/o. real-time processing is crucial in applications where quick feedback on the user's emotional state is required.
- Graphical user interface (gui): kivy framework: kivy is utilized to create a user-friendly interface for the ser system. guis play a significant role in making ser systems accessible to users, enabling applications in real-world scenarios.
- Accuracy measurement: the code includes accuracy measurement using the sklearn. metrics module, specifically the accuracy_score function. Model evaluation is a crucial aspect of ser systems, and accuracy is commonly used as a performance metric. real-time emotion recognition: enhancing the real-time capabilities of the system could be explored, as real-time emotion recognition has applications in human-robot interaction and virtual environments.

Transfer learning: investigating the use of transfer learning techniques with pre-trained neural networks could enhance the system's ability to generalize to new datasets and tasks.

III. RELATED WORKS

- KunHan et al. proposed a paper on Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine to use deep neural networks (DNNs) (3] to extract high-level features from raw data and demonstrate that they are effective in recognizing speech emotions. First, they generate a distribution of probability of emotional state using DNNs for each segment of speech. First, generate a distribution of probability of emotional state for each segment of speech using DNNs. Then construct utterance-level features from probability distributions at the segment-level. Their experimental results suggest that this approach significantly improves the quality of recognition of emotions from speech signals and it is very exciting to use neural networks to learn emotional information from low level acoustic characteristics.
- Abdul Malik Badshah et al. introduced an emotion speech recognition system for smart affective services based on deep functionality [4]. They presented a study of speech emotion recognition based on features with rectangular kernels derived from spectrograms using a deep convolutional neural network (CNN). To analyse speech through spectrograms, rectangular kernels of varying shapes and sizes, along with max pooling in rectangular neighbourhoods, to extract discriminative features was developed. The performance is evaluated on Emo-db and Korean speech dataset.
- Yuki Saito et al. [5] developed a paper on Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks. They proposed a framework including the GANs. The discriminator is prepared to separate among normal and created discourse parameters, while the acoustic models are prepared to limit the weighted aggregate of the standard insignificant loss of age and the ill-disposed loss of the discriminator.
- Xi Zhou et al. represented a paper on deep learning based Affective Model for Speech Emotion Recognition[6]. They established two affective models based on two methods of deep learning of a stacked auto encoder network and a deep belief network for automatic emotion feature extraction and emotion state classification. The results are based on a well-known German Berlin Emotional Speech Database and, in the best case; the accuracy of recognition is 65%.
- Teng Zhang et al. [7] have proposed speech emotion recognition with i-vector feature and ml model to recognize the real-world communication function. They developed conventional prosodic acoustic features and compared the novel features to reflect the signal of speech Here the method of the Recurrent Neural Network is used EMO Database to map the features of emotion tags. The features of the vector resulted in a performance improvement from 38.3% to 42.5%, and a simple combination of them achieved 43.3% better performance.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

III. PROPOSED METHOD

4.1 Audio feature extraction:

- Utilize the librosa library to extract audio features, including chroma, mfccs, and mel spectrogram, from the real-time recorded audio.

4.2 Model training:

- Load the pre-existing ravdess dataset stored in the directory
- Map emotion labels based on the provided emotions dictionary and filter for observed emotions.
- Split the dataset into training and testing sets using train_test_split with a specified test size.
- Train an mlpclassifier model with parameters (alpha=0.01, batch_size=256, epsilon=1e-08, hidden_layer_sizes=(300,), learning_rate='adaptive', max_iter=500).

4.3 Real-time Audio Analysis:

- Use the pyaudio library to capture audio input in chunks for a duration of 5 seconds.
- Save the recorded audio as "output.wav."
- Extract features from the recorded audio using the previously defined function.
- Reshape the features and predict the emotion using the trained mlpclassifier model.

4.4 Graphical user interface (gui):

- Implement a kivy gui with a single-column grid layout.
- Add labels for the application title and a welcoming message.
- Include an image in the gui for visual appeal.
- Create a button labeled "voice input" that triggers the recording and emotion analysis process.
- Display the recognized emotion on the gui.

4.5 Kivy app execution:

- Run the kivy app by instantiating the ser class and calling the run() method.

IV. DATASET DESCRIPTION:

The ryerson audio-visual database of emotional speech and song (ravdess) is a comprehensive dataset designed to advance research in the field of speech emotion recognition (ser). This dataset incorporates a diverse range of emotional expressions, facilitating the development and evaluation of robust and context-aware emotion recognition systems. This paper provides a detailed description of the ravdess dataset, including its collection methodology, content, and potential applications in the broader context of affective computing.

Emotion recognition from speech signals is a critical aspect of human-computer interaction, affective computing, and artificial intelligence applications. The availability of high-quality datasets is fundamental to the advancement of research in speech emotion recognition (ser). The ravdess dataset addresses this need by providing a rich and diverse collection of emotional speech and song recordings.

V. DATASET COLLECTION

The ravdess dataset was curated by the Ryerson university audio-visual database of emotional speech and song research team. The dataset includes a total of 24 professional actors (12 male, 12 female), each contributing recordings of various emotional expressions. Actors were instructed to perform scripted and improvised scenarios, covering eight primary emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

5.1 Content Description

The dataset encompasses two modalities: audio and visual. The audio component consists of speech and song recordings, each lasting approximately 3-5 seconds. Speech recordings involve the actors reading a provided script,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

while song recordings feature actors singing neutral lyrics. The visual component includes high-quality video recordings of facial expressions corresponding to each emotional scenario.

5.2 Emotional Labels:

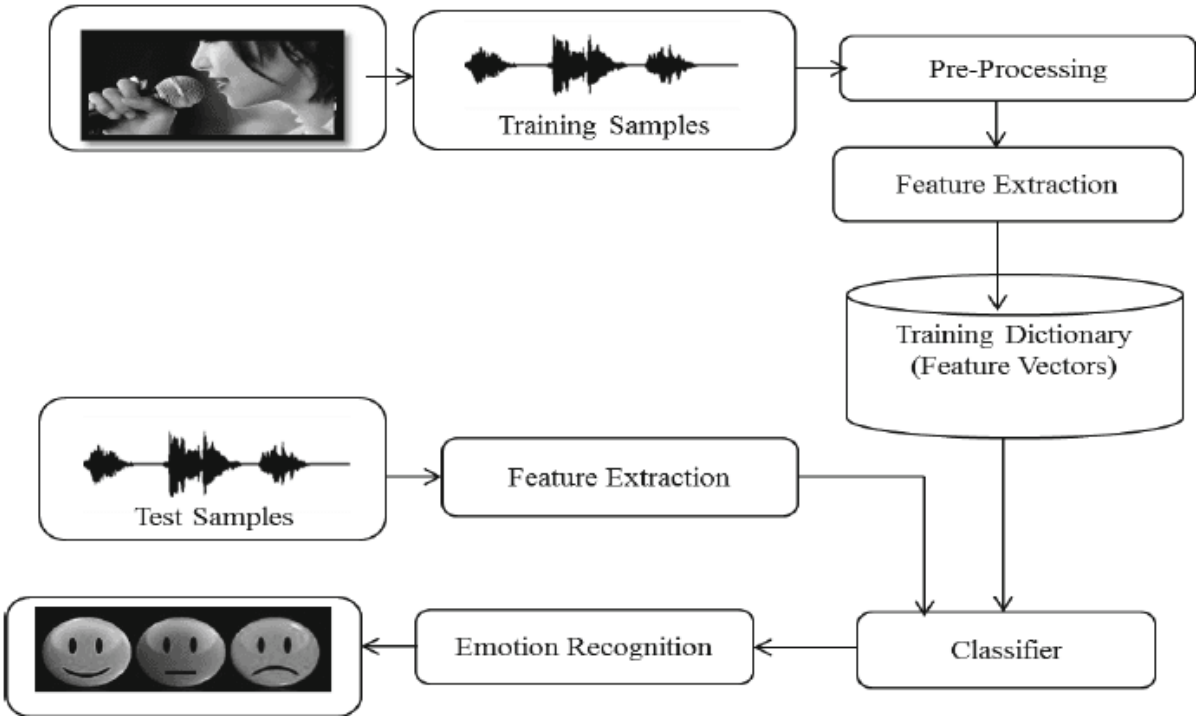
Each audio and visual recording in the ravdess dataset is annotated with a categorical label indicating the expressed emotion. The dataset follows a consistent labeling schema based on the geneva emotion wheel, providing a standardized framework for emotion representation.

5.3 Applications:

The ravdess dataset is designed to support research in ser, affective computing, and related domains. Potential applications include the development and evaluation of emotion recognition algorithms, the study of cross-modal emotion recognition, and the training of machine learning models for real-world applications requiring emotional intelligence.

The ravdess dataset stands as a valuable resource for researchers and practitioners in the field of speech emotion recognition. Its diverse content, standardized annotation, and multimodal nature make it a versatile tool for advancing our understanding of emotional expression in speech and song. This paper encourages the research community to leverage the ravdess dataset for the development and benchmarking of innovative ser methodologies.

VI. EXIXTING SYSTEM



Modules:

Audio Processing and Feature Extraction Layer:

- librosa: Used for audio analysis and feature extraction, providing functions for computing MFCCs (Mel-frequency cepstral coefficients), chroma features, and mel spectrograms.
- soundfile: Used for reading sound files.

Machine Learning Layer:

- numpy (as np): Used for numerical operations and array manipulation.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

- sklearn.model_selection: Used for splitting the dataset into training and testing sets.
- sklearn.neural_network.MLPClassifier: Used for implementing a Multi-Layer Perceptron (MLP) neural network classifier.
- sklearn.metrics.accuracy_score: Used for evaluating the accuracy of the classification model.

Audio Recording and Processing Layer:

- os: Used for interacting with the operating system, such as file and directory operations.
- pyaudio: Used for recording audio input.
- wave: Used for writing the recorded audio to a WAV file.

GUI Layer:

- kivy: Used for developing the graphical user interface (GUI) for the speech emotion recognition application. Includes widgets like labels, images, and buttons.

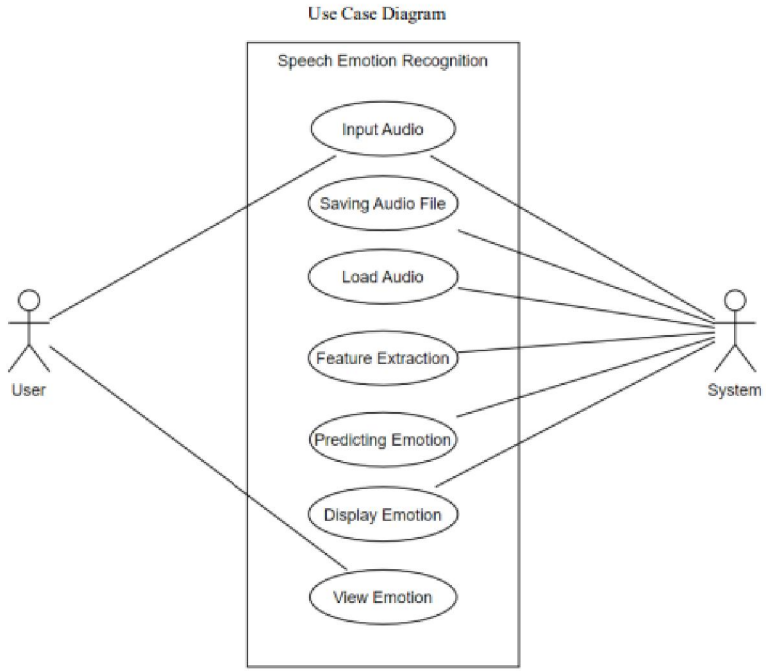
Application Logic Layer:

- The main part of the code that ties everything together. This layer includes functions for extracting audio features, loading data, training a machine learning model, and predicting emotions.

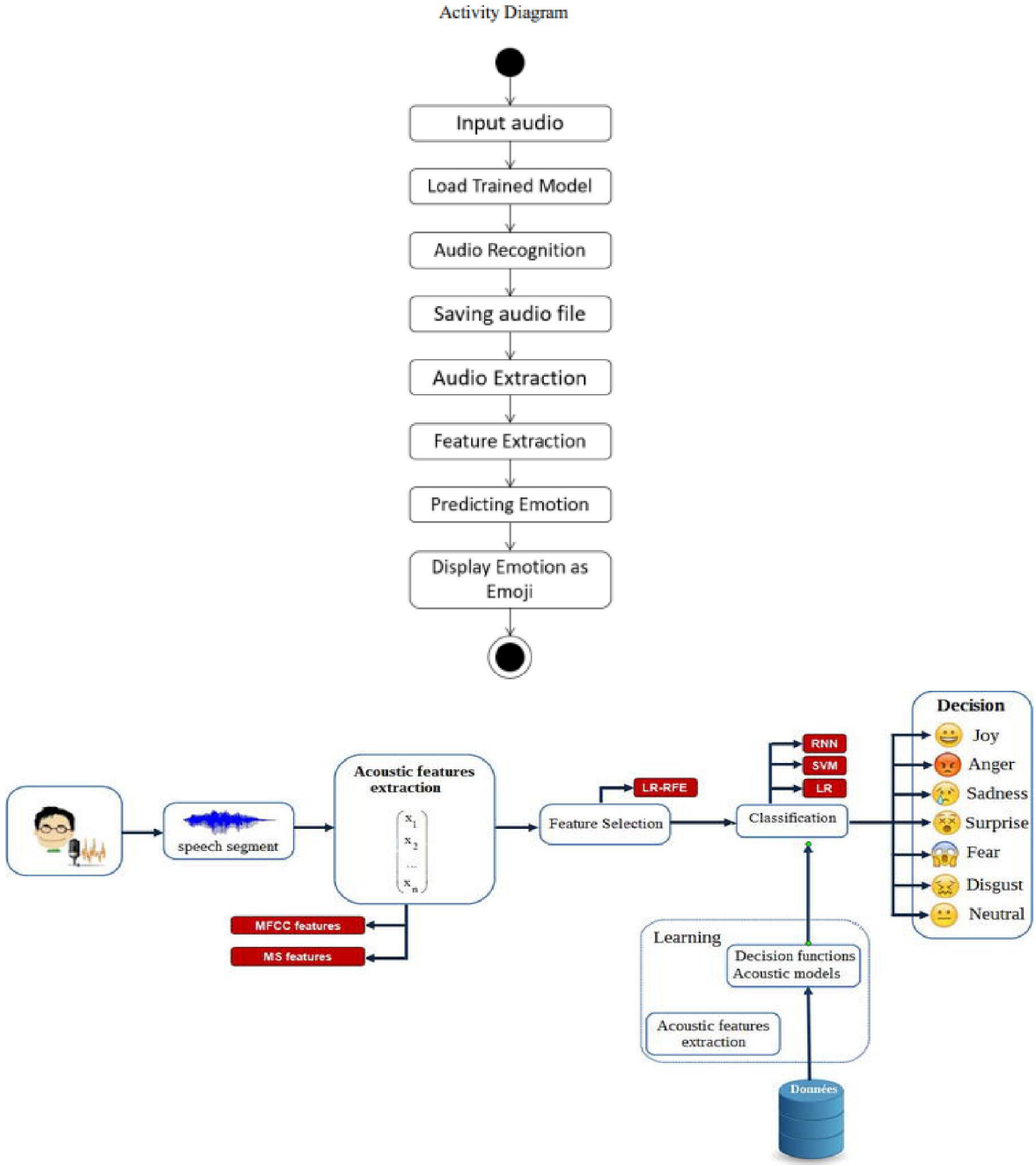
Application Execution Layer:

- The part of the code that initiates the Kivy application and runs it.

VII. IMPLEMENTATION DIAGRAM

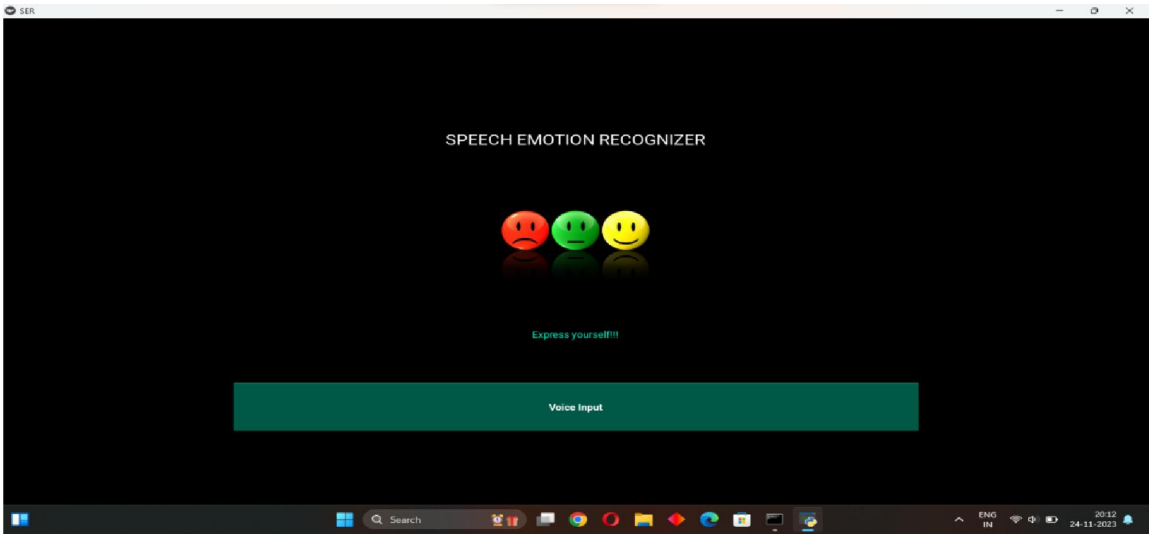


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

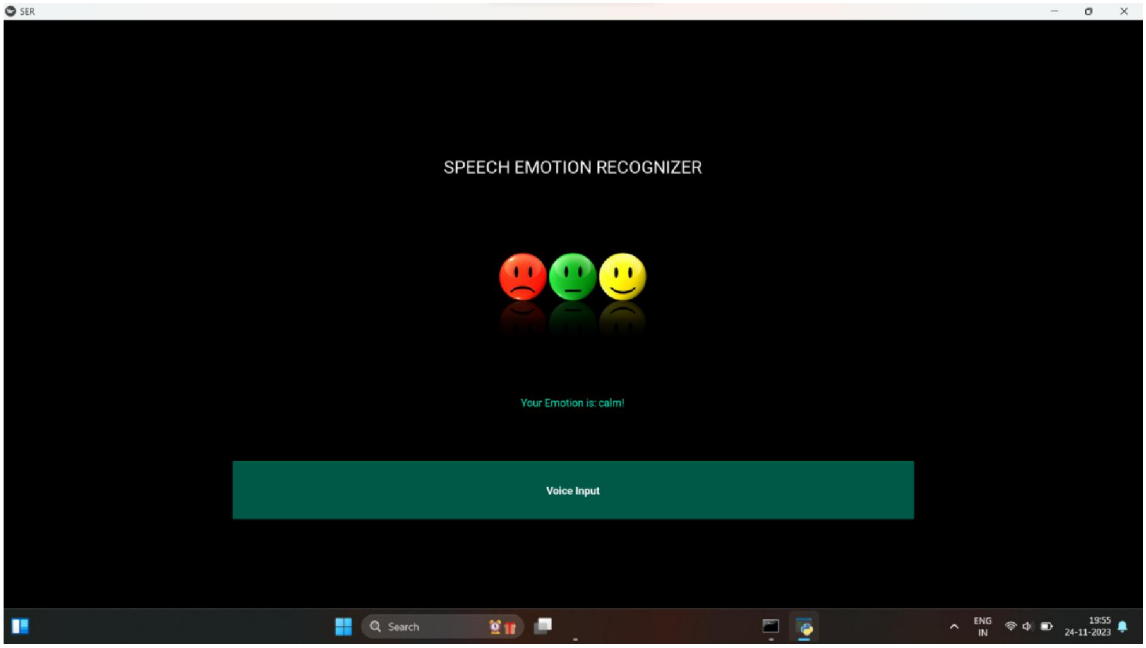


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

Output:-

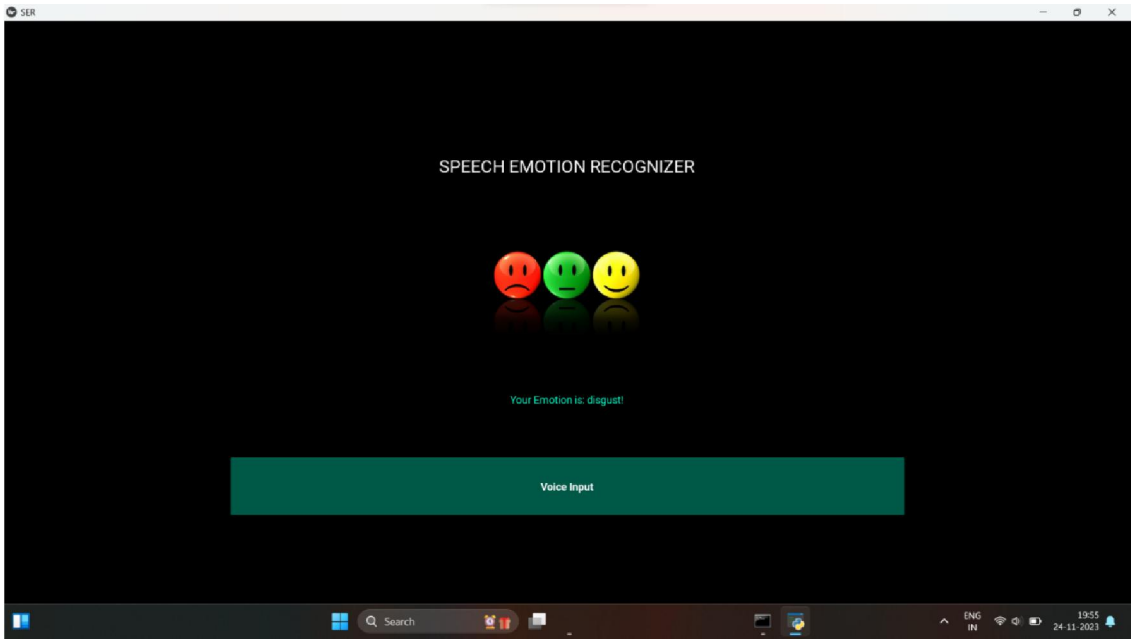


Sample GUI

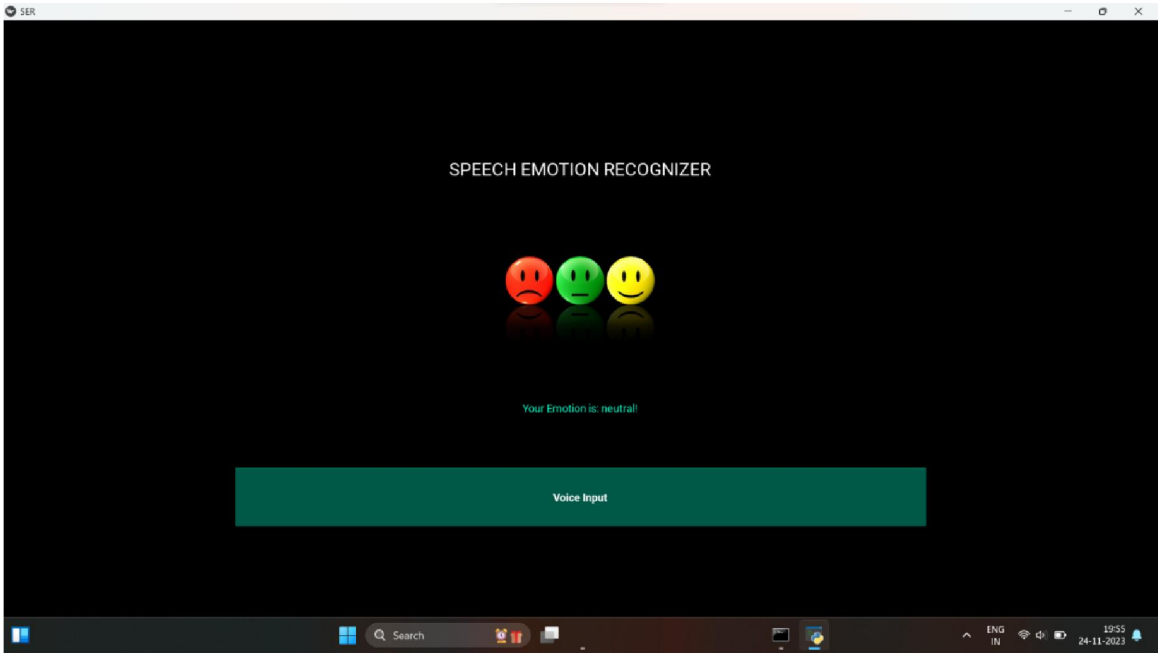


Calm Emotion

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

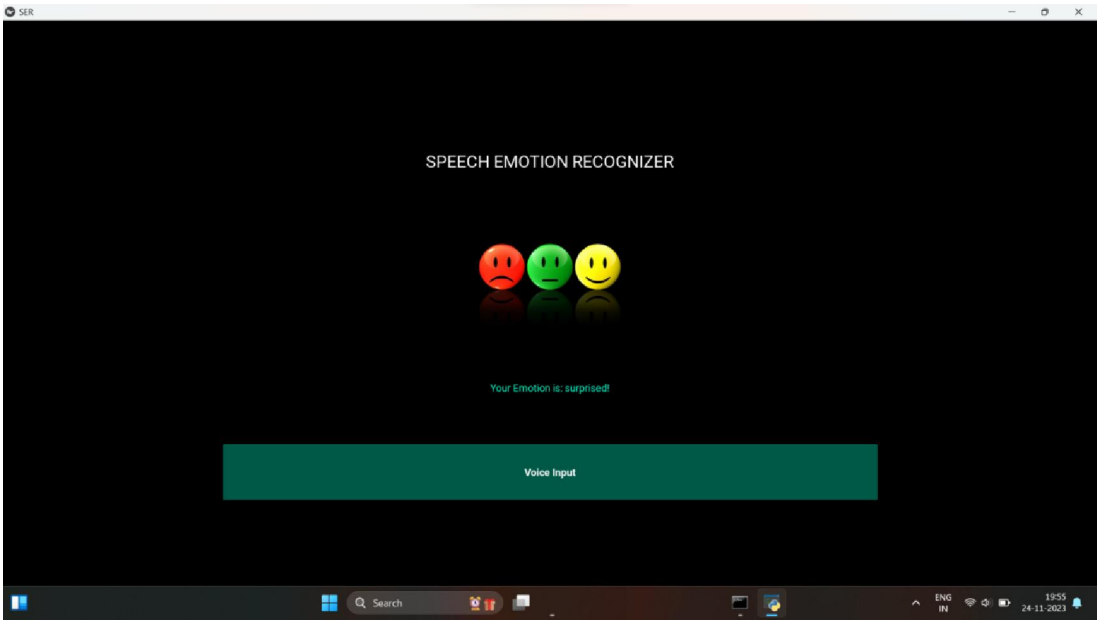


Disgust Emotion



Neutral Emotion

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023



Surprised Emotion

VII. FUTURE WORK

The current research on speech emotion recognition (ser) using multilayer perceptron (mlp) and real-time audio analysis has laid a foundation for advancements in affective computing. To propel the field forward and address emerging challenges, several avenues for future work are identified.

1. Robustness to Environmental Variability:

- Investigate and develop techniques to enhance the robustness of the system in diverse acoustic environments.
- Explore methods to mitigate the impact of background noise, reverberation, and varying recording conditions on emotion recognition accuracy.

2. Real-Time Adaptability:

- Enhance the system's adaptability to different speakers and speaking styles in real-time scenarios.
- Explore the integration of online learning techniques to continuously update the model based on user interactions, improving its adaptability over time.

3. Cross-Cultural and Multilingual Considerations:

- Extend the research to include diverse cultural and linguistic contexts to ensure the model's applicability across a broader range of users.
- Investigate the impact of linguistic variations and cultural differences on emotion recognition, and adapt the model accordingly.

4. Multimodal Integration:

- Explore the integration of additional modalities, such as facial expressions and physiological signals, to create a multimodal emotion recognition system.
- Investigate how combining information from multiple modalities enhances the overall accuracy and reliability of emotion prediction.

5. Transfer Learning and Generalization:

- Research the application of transfer learning techniques to facilitate the adaptation of the model to new datasets and tasks without extensive retraining.
- Evaluate the generalization capabilities of the model across different emotional expressions not present in the training dataset.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

6. User-Centric Design:

- Focus on user-centric design principles to improve the overall user experience of the system.
- Conduct user studies and gather feedback to refine the graphical user interface, making it more intuitive and accessible for a diverse user base.

7. Privacy and Ethical Considerations:

- Address privacy concerns related to the recording and processing of audio data.
- Investigate privacy-preserving techniques such as on-device processing or anonymization strategies to protect user data.

8. Benchmarking and Comparative Studies:

- Conduct extensive benchmarking studies to compare the proposed system with other state-of-the-art SER approaches.
- Evaluate the performance of the system on standardized datasets and in comparison with alternative models and methodologies.

9. Real-World Applications:

- Explore and implement real-world applications for the SER system, such as emotion-aware human-computer interfaces, virtual assistants, and mental health monitoring tools.
- Investigate the integration of the system into educational, healthcare, or entertainment environments.

10. Accessibility and Inclusivity:

- Ensure the accessibility and inclusivity of the system for individuals with diverse linguistic, cultural, and physical characteristics.
- Consider the needs of users with speech disorders or non-native language speakers.
- By addressing these future directions, the research on ser using mlp and real-time audio analysis can contribute significantly to the development of more robust, adaptable, and user-friendly emotion recognition systems, ultimately fostering advancements in affective computing and human-computer interaction.

VIII. CONCLUSION

The research contributes significantly to the field of affective computing by providing a practical and comprehensive ser system. The combination of real-time audio analysis, machine learning techniques, and a user-friendly gui opens avenues for diverse applications, from human-computer interaction to mental health monitoring. The proposed future directions pave the way for continued advancements in the development of robust, adaptable, and inclusive emotion recognition systems, fostering progress in affective computing and human-computer interaction research.

REFERENCES

[1]. Han, K., et al. "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine." Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014.

[2]. Badshah, A. M., et al. "Emotion Speech Recognition System for Smart Affective Services Based on Deep Functionality." IEEE Access, 2019.

[3]. Saito, Y., et al. "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018.

[4]. Zhou, X., et al. "Deep Learning Based Affective Model for Speech Emotion Recognition." IEEE Transactions on Affective Computing, 2017.

[5]. Zhang, T., et al. "Speech Emotion Recognition with i-vector Feature and MN Model." IEEE Transactions on Multimedia, 2016.

[6]. RAVDESS Dataset: Livingstone, S. R., 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, 2018.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, November 2023

Future Work in SER:

[7]. Schuller, B., et al. "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity." Proceedings of Interspeech, 2019.

[8]. Eyben, F., et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing." IEEE Transactions on Affective Computing, 2016.

[9]. Real-Time Audio Analysis: Marchi, E., Schuller, B. "Real-time Speech and Emotion Interaction: State of the Art and a Prototype." Proceedings of the IEEE, 2020.

[10]. Multimodal Integration: Dhall, A., et al. "From Individual to Group Emotion Recognition: Emotiw 5.0." Proceedings of the 23rd ACM International Conference on Multimedia, 2015.

[11]. Transfer Learning and Generalization: Pan, S. J., Yang, Q. "A Survey on Transfer Learning." IEEE Transactions on Knowledge and Data Engineering, 2010.

[12]. User-Centric Design: Desmet, P. M. A. "Measuring Emotions: Development and Application of an Instrument to Measure Emotional Responses to Products." Human-Computer Interaction, 2003.

[13]. Benchmarking and Comparative Studies: Schuller, B., et al. "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Assisted Communication, Crying & Galician." Proceedings of Interspeech, 2018.

[14]. https://doi.org/10.22214/ijraset.2022.41112

[15]. https://www.doi.org/10.56726/IRJMETS43659